Data science for the liberal arts

Kevin Lanning

2018-12-21

```
#devtools::install_github('yihui/tinytex')
#tinytex:::install_prebuilt()
#options(tinytex.verbose = TRUE)
#Sys.which("pdflatex")
```

This work-in-progress includes my notes for Introduction to Data Science at the Wilkes Honors College of Florida Atlantic University.

Chapter 1

Preface: an invitation

This work-in-progress includes my notes for Introduction to Data Science at the Wilkes Honors College of Florida Atlantic University.

what is this course about?

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.

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 include all similar goals (which we may again draw from) at Chicago, Georgia Tech, UC Santa Barbara, Princeton, UC Berkeley, at Berlin's Hertie School of Governance, and in Columbia's School of Journalism.

type C data science = data science for the liberal arts

Type C data science does not merely integrate 'domain expertise' with statistics and computing, it places content squarely at the center. We are appreciative of the compelling logic and power of statistics and the elegance of well-written code, but our perspective is that these are means to an end, as tools in the service of cultural concerns and social and scientific problems. Type C Data Science aims for work which is not merely cool, but responsible and meaningful.

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 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. In any event, **there will be computing throughout the course.** You will be expected to bring a laptop every day. (Please let Dr. Lanning know ASAP if you don't have access to this).

The place of data science in the WHC (and FAU) curriculum

At this writing, there is enthusiasm across units of FAU and its affiliated institutes, including Max Planck and FAU's Colleges of Science and Engineering as well as the WHC, for integrating data science into our curriculum. Within the WHC, a data science minor and a multi-track concentration are under development. Until these proposals have been formally approved, students interested in concentrating in Data Science are encouraged to pursue an individual concentration (see Dr. Lanning for details).

In addition, there are several integrated '4 + 1' pathways which will lead to a master's degree in the College of Engineering. These programs are also in progress; again, see Dr. Lanning for additional details.

should you enroll?

It's my intention that this course should serve every Wilkes Honors College student, regardless of concentration. The WHC was built and funded by the state of Florida to train tomorrow's leaders. That's you. The skills and insights that you will gain in this course will help you in graduate and professional schools, will help you in your careers, and will help you in your goal of making a better world. And it will help you train the next generation of data scientists, too.

How do I sign up?

If you are interested in taking the class, complete the Google form here. Note that enrollment in the class and lab will be limited to 30.

(PART) Introduction

Chapter 2

Data science for the liberal arts

Hochster, in Hicks and 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**. 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:"

Type C data science

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 fact that the world and workforce are undergoing massive **change**, which puts the classic liberal arts focus of "learning how to learn" at center stage. And it's shaped, not least, by the **creepiness** of living increasingly in a measured, observed world.

The incompleteness of the Venn diagram

The Venn diagram model is incomplete. In addition to *statistics, computing/hacking,* and *domain expertise,* a number of additional skills contribute to the success of the data scientist.

These include *collaboration,* which is arguably the most distinctive feature of contemporary scholarship in the natural and social sciences as well as in the private sector (Isaacson 2014).

Communication is central to data science because results are inconsequential unless they are recognized, understood, and built upon; facets of communication include oral presentations, written texts and, too, clear data visualizations.

Reproducibility is related to both communication and collaboration. There has been something of a crisis in recent years in the social and natural sciences as many results initially characterized as "statistically significant" have been found not to replicate. The reasons for this are multiple and presently contentious, but one path towards better science includes the public sharing of methods and data, ideally before experiments are undertaken. Reproducible methods are a key feature of contemporary data science.

Pragmatism refers to the relevance of work towards real-world goals. Ideally, these pragmatic concerns take into account *ethical concerns* as well.

The expertise dimension

Cutting across these eight facets (statistics, computing, domain expertise, collaboration, communication, reproducibility, pragmatism, and ethics), a second dimension can be articulated. No one of us can excel in all eight domains, rather, we might aim towards goals ranging from *literacy* (can understand) through *proficiency* (can get by) to *fluency* (can practice) to *leadership* (can create new solutions or methods).

That is, we can think of a *continuum* of knowledge, skills, interests, and goals, ranging from that which characterizes the data *consumer* to the data *citizen* to the data science *contributor.* A Type C data science includes this dimension as well.

Google and the liberal arts

Data science is at its core empirical, and all of this rhetoric would be meaningless if not grounded in real world findings. Although it was recently reported that soft skills rather than STEM training were the most important predictors of success among Google employees, it's difficult to know whether these results would generalize to a less select group. Nonetheless, there is a clear need for individuals with well-rounded training in the liberal arts in data science positions and, conversely, learning data science is arguably a key part of a contemporary liberal arts education.

Data sci and TMI

It has been said (by whom?) that the biggest difference between traditional stats and data science is that the former is typically concerned with making inferences from datasets that are too *small*, while the latter is concerned with extracting a signal from data that is or are too *big*.

The struggle to extract meaning from a sea of information - of finding needles in haystacks, of finding faint signals in a cacophany of overstimulation - is arguably the question of the age. It is a question we deal with as individuals on a moment-by-moment basis. It is a challenge I face as I wade through the many things that I could include in this class and these notes.

The *primacy of editing* or selection lies at the essence of human perception and the creation of art forms ranging from novels to film. And it is a key challenge for the data scientist faces as well.

A challenge

Imagine it is ten years from today. You are working in a cool job (yay). How, ideally, would 'data science' inform your professional contributions?

More proximally (closer to today) - what are your own goals for progress in data science, in terms of the model described above?

Chapter 3

Pretest and setup

You are - *where?* (an informal pretest)

Reflect on your own knowledge of data science, including the necessary-but-not-sufficient areas of computer programming and statistics.

You may know more than you think you do. Even if you haven't had formal programming, you may well have experience with spreadsheets such as Excel or Google Sheets.

What does '=SUM (A1:A15)' mean? In a spreadsheet, if '=SUM (A1:A15)' was in cell A16, and you copied this to cell B16, what would the result be?

Statistics is, of course, relevant in a myriad of ways. Consider, for example, the prospects for Susie, a young woman who is applying to two med schools. At School A, 25% of students are accepted, and at School B, 25% are accepted as well.

You are Susie. Are you going to get in to at least one of these programs? What is the probability? Does your estimate depend upon any assumptions?

Questions like these are important for us. If the combined probability is low, it *likely* (another probability concept) makes sense for Susie to apply to additional programs. If it is high, it may not. But problems like this are hard - our estimates of probability are frequently poorly calibrated, and combining probability estimates is challenging (see, e.g., Tversky and Kahneman (1974), and consider taking a course in *Behavioral Economics* or *Thinking and Decision Making* to learn more).

A thought experiment

Would it be possible to assign grades in a class based not just on what students know at the end of the term, but also on how much they have learned?

You could, in principle, do this using regression analysis. That is, you could predict final exam scores from pretest scores, and use the residuals - the extent to which students did better or worse than expected - as a contributor to final exam grades.

Interestingly, there would be an unusual incentive for students on this 'pretest' to do, seemingly perversely, as poorly as possible. How might you address this?

Setup

Install R and R studio on your own laptop. If you get stuck, reach out to others on Slack; if you don't get stuck, help your classmates.

Some basic tools: Slack, Markdown, Google Docs

On Markdown, see Freeman and Ross (2017), Chapter 3). Some Markdown editors include Typora, Atom, and Ghostwriter.

R markdown is a 'flavor' of - wait for it - markdown. For beginning with R markdown, consider Carmichael (2017) Getting started as well as the first chapter of Wickham and Grolemund (2016).

Slack is a messaging and collaboration platform. There is a simple markdown editor in Slack (for 'posts').

You are likely to be familiar with Google Docs as well. One interesting feature of Google Docs is that it provides some basic tools for *version control,* a critical skill in information management particularly (but not only) for collaborative work. You can learn more about how to see and revert to prior versions of a project in Google Docs here .

Version control can help you avoid the chaos and confusion of having a computer full of files that look like Cham's (2012) comic:

.

Never call anything 'final.doc'.

We'll be talking about the challenge of version control throughout this text - and I am hoping that my own habits in file management can improve as we move forward together.

Chapter 4

An introduction to R

R is a system for reproducible science, and reproducibility is essential (Baumer et al. 2014). R is a system for Representing data in cool, insight-facilitating ways. R is really popular, and really growing. Learning R will make you a more attractive candidate for many graduate programs as well as jobs in the private sector.

One does not 'learn R'

Unlike, say, learning to ride a bicycle, fry an egg, or drive a car with a manual transmission, learning R is not a discrete accomplishment that one can be said to have mastered and from which one then moves on. Rather, R is an evolving, open system of applications and tools which is so vast that there is always more that one can achieve, new lessons that one can learn.

What R stands for

Historically, R grew out of S which could stand for Statistics. But what does R stand for?

R does not stand for 'argh,' although you may proclaim this in frustration ('arggh, why can't I get this to work?) or, perhaps, in satisfaction ('arggh, matey, that be a clever way of doing this').

R might stand for *relatively high level.* Programming languages can be described along a continuum from high to low level, the former (like R) are more accessible to humans, the latter (like assembly language) more accessible to machines. Python, Java, and C++ are all more towards the middle of this continuum.

R stands, in part, for *resources.* There are many resources, including, for example -

- Online resources include the simple (and less simple) lessons of SwirlR, which offers the possibility of "learning R in R," as well as DataCamp, the Data Science Certificate Program at Johns Hopkins, and other MOOCs.

- Books include Peng (2015) - which includes not only videos of his lectures in the program at Hopkins, but also a brief list of still more resources - and Wickham and Grolemund (2016).
- You'll also learn (more directly) from people, including your classmates, as well as the broader community of people around the world. There are hundreds if not thousands of people, young and old, who are on the road with you. As I am, just a step or two (hopefully) ahead.

On R Studio and commercial vs. open software

R Studio is a commercial enterprise whose business model, judged from afar, is an important one in the world of technology and open science. Most of what R Studio offers is free (97% according to Garrett Grolemund in the video below). The commercial product they offer makes sense for a relative few, but it is sufficiently lucrative to fund the enterprise. The free product helps to drive the popularity of R studio; this widespread use, in turn, makes it increasingly essential for businesses to use.

This mixed free/premium model characterizes Slack as well, but while the ratio of free to paid users of Slack is on the order of 3:1, for R it is, I am guessing, an order of magnitude higher than this.

The openness of R is an important feature, not just because it saves you money, but because contributing to the world of R is an act of digital democracy, for in contributing to the world of R we open up knowledge to others who may lack our privileges. The corporate clients of R studio, in essence, support the rest of us.

A digression and thought experiment

All contemporary cars have computers. Many of these are quite sophisticated, allowing cars to drive themselves in increasingly autonomous ways. But the potential of autonomous cars will not be realized until they are connected on the 'net. Communication with other cars, stoplights, etc. will reduce travel times, increase fuel efficiency, and increase safety, too.

Until, that is, they are hacked.

Who should be the guardians of the net linking cars? Should this be closed (with Ford having it's own safeguards), or public?

A few characteristics of R

R includes the base plus thousands of **packages**. These packages are customized add-ons which simplify certain tasks, such as text analysis. But there are at least 40 different packages for this - where do you begin? The most recent answer, and where we will start, is the curated list of packages which jointly comprise the tidyverse (Wickham and Grolemund 2016).

Peng (2015) speculated that "it would be straightforward to build an R package for ordering pizza." Does one exist now?

R is an object-oriented language. At an atomic level, these include *characters, real numbers, integers, complex numbers, and logical.* These atoms are combined into vectors, which (unless they are lists) include objects of the same type (Peng 2015). Missing values may be characterized by NA (not available) or NaN (not a number, implying an undefined or impossible value). Attributes of R include such things as name, dimensions (for vectors and arrays), class (that's the atomic thing), length, etc. Vectors can be combined into a **data frame** (or, in a simplified form, a *tibble*), which greatly facilitates statistical analysis.

(The same can be said of data science: The field is changing, and much of the low-level knowledge that you will gain in this class will be obsolete in just a few years. This is another sense in

Finding help

To get a sense of some of the ways you can get help in R studio (and to see how a master uses the R Studio interface), consider the video at RStudio Essentials Webinar Series – Programming Part 1 – RStudio

For us, the key ideas in "looking for help" will include not just the tools on the R Studio IDE, but also (a) using google searches wisely, and (b) reaching out to your classmates on Slack. At the next level, the search for help should be built around reproducible errors. There is a package for this.

Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J Horton. 2014. "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics." *arXiv Preprint arXiv:1402.1894*.

Freeman, Michael, and Joel Ross. 2017. *Technical Foundations of Informatics, U Washington Info 201*. https://info201.github.io.

Hicks, Stephanie C, and Rafael A Irizarry. 2017. "A Guide to Teaching Data Science." *The American Statistician*, no. just-accepted. Taylor & Francis: 00–00.

Isaacson, Walter. 2014. *The Innovators: How a Group of Inventors, Hackers, Geniuses and Geeks Created the Digital Revolution*. Simon; Schuster.

Peng, Roger D. 2015. *R Programming for Data Science*. Lulu. com.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (4157). American association for the advancement of science: 1124–31.

Wickham, Hadley, and Garrett Grolemund. 2016. "R for Data Science." Sebastopol, CA: O'Reilly. http://r4ds. had. co. nz.