Teaching Data Science Annotated Bibliography

Table of Contents

I. Executive Summary

(A) Introduction

As data science skills become increasingly important, the teaching of data science becomes increasingly valuable. In the context of data science, I think of teaching not as the traditional process of talking at a whiteboard or grading worksheets, but instead as the process of creating environments and designing experiences that leads to measurable learning for students.

As a matter of definition, I am interested in developing in students the ability to extract value out of data (which I'll refer to as data science). The students I am interested in are anyone with the numeracy background to not need additional training in, say, algebra, but not necessarily background in either programming or any level of serious statistics.

(B) American Statistical Association Suggestions

The American Statistical Association (ASA) has recognized changes in training necessary with a changing landscape. In 2005, they first put out suggestions for a direction for the future of introductory statistics, and they updated these suggestions again in 2016 (ASA 2016). Their target is teaching introductory statistics at a college, but the wisdom contained also applies to teaching data science in any environment. Here are their suggestions (page 3):

1. Teach statistical thinking.

    a. Teach statistics as an investigative process of problem-solving and decision making.
    b. Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Most of these recommendations are simply generally good teaching. For example, there has been a great deal of recent educational research pointing out the significant measurable benefits of active learning (C. E. Wieman 2014; Koedinger et al. 2015). It's also worth pointing out that some scholars argue that these suggestions point in the right direction but are simply not enough (G. Cobb 2015).

In any case, absolutely essential to the ASA suggestions is critical thinking or decision making. Holmes, Wieman, and Bonn (2015) define critical thinking as "the ability to make decisions based on data, with its inherent uncertainties and variability, is a complex and vital skill in the modern world" (page 1). The ASA doesn't provide any explicit guidance as to how to build critical thinking in students, but Holmes, Wieman, and Bonn (2015) offer strikingly simple advice: "the key element for developing this ability is repeated practice in making decisions based on data, with feedback on those decisions" (page 1).

(C) Unsolved Problems in Teaching Data Science

The ASA suggestions and educational research offers somewhat obvious advice such as have students learn actively while analyzing real data with technology; when new ideas are introduced, focus on conceptual understanding as opposed to rote memorization; and measure student learning through assessments. What, then, are the difficult or unsolved parts of teaching data science? I argue that there are three: *first, building a social environment or culture that is conducive to learning; second, choosing technology and designing learning activities in which students are asked to make decisions; and third, measuring or assessing learning.*

1) Creating Culture

Building an environment in which students feel free to expose their own knowledge state and compassionately help each other learn is incredibly difficult. The idea of growth mindset was first put forth by Carol Dweck (Dweck and Sumoreads 2017). Jo Boaler discusses the growth mindset especially in the context of the math classroom and fittingly refers to it as the "mathematical mindset" (Boaler 2016). Their research points out the importance of students believing that they can improve through practice. More interestingly, perhaps, they offer suggestions as to how to foster growth mindsets such as praising effort as opposed to skill (Khan 2014).

I like the idea of a "data science mindset." Especially important to building such a mindset is getting students to think critically and make decisions as a group. For this to be effective, students need to be especially comfortable with each other socially. Humans can be shy creatures. Creating environments that overcome this tendency to get students to engage deeply with each other, share thinking, and challenge each other's ideas productively is incredibly difficult.

The closest thing to magic that I've seen for accomplishing this goal is improv. From my experience, students can enter a room shy and unsure and a few hours later be completely comfortable with each other by doing improv together. The book "The Second City Guide to Improv in the Classroom" offers full details on the benefits of improv in the classroom along with specific activity suggestions (McKnight and Scruggs 2008). These techniques could easily be adapted to a data science environment by, for example, having students act out what a function does as a group, having a classroom norm of yelling out "yay" enthusiastically when getting an error, or requiring students to start sentences with "yes, and…" when brainstorming how they might solve a data science problem. These ideas will strike some as silly, but that's actually the essential ingredient.

Teacher's often overlook purposefully building culture. In my opinion, nothing could be a greater mistake. Investing in classroom culture especially at the beginning pays great dividends. If a teacher doesn't explicitly build culture, the culture of the classroom is likely to become similar to the "real-world" culture outside of the classroom, which is particularly damaging to students from backgrounds that are often marginalized in analytical circles. In essence, building culture is a both an equity and effectiveness issue (P. Cobb and Hodge 2011).

2) Choosing Technology and Designing Learning Activities (in which students are asked to make decisions)

I'll dive more deeply into this in the "top 3" section.

3) Measuring/Assessing Learning

Data science is about decision making and decision making is hard to measure. Machine gradeable formats don't seem to do the trick and alternatives are time-intensive and subjective. In addition, desirable data science learning outcomes are often longer-term and have as much to do with student's relationship to data science as actually mastering skills. With that in mind, perhaps course surveys and gathering data on student confidence and sentiment is as valuable as traditional educational assessments (Wirth and Perkins 2005).

In summary, we have mountains to climb in the way of effective measurement and assessment of data science learning. Maurer and Lock (2016) offer an interesting starting place where they randomly assign Iowa State University students to a simulation-based or traditional inference curricula and compare students pre to post-test gains on questions that target specific learning objectives. A similar example is found in C. Wieman and Holmes (2015). A promising future direction is to consider program evaluation which takes a wider-view of measurement. Moore and Kaplan (2015) provide an example by evaluating the undergrad statistics major at the University of Georgia.

II. Top 3

1) Learning Statistics Using Motivational Videos, Real Data and Free Software – Harraway 2012

2) The fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses – Kim, Ismay, and Chunn 2018

3) Web Application Teaching Tools for Statistics Using R and Shiny – Potter et al. 2016

For my top 3, I did a deep dive on three papers that each suggest different technology in the classroom. The choice of learning technology comes down to managing the amount of energy a course has. If energy is spent on specifics of learning technology, then there will be less energy available for other topics. However, it's also important to recognize that some learning technologies are more valuable for students to learn than others. For example, using R compared to specific-use learning applets might come at a higher cost for the course, but also has greater potential value to the student after the course.

In *Learning Statistics Using Motivational Videos, Real Data and Free Software,* Harraway (2012) advocates the use of software GenStat for Teaching and Learning (GTL) which is a watered down version of the professional software GenStat 14. GTL is built to occupy a middle-ground between spreadsheets and advanced data analysis software by being entirely menu driven and focusing on ease of use of basic statistical needs such as transforming data, model fitting, hypothesis testing, and bootstrapping. The creators measure the effectiveness of the software by piloting it in a few classrooms and then observing and surveying results – for example, reporting that out of 41 students, only 6 students described GTL has hard to use. GTL also has learning activities and videos directly associated with its features.

Harraway (2012) admits that R is the largest competitor of GTL but continually emphasizes the needless complexities and frustration induced by requiring programming and syntax. I find this to be the most interesting question this paper raises. Taking professional software, removing complexity for beginners, and designing purposeful learning activities seems like a good idea, but I have a hard time buying the argument that being menu based is a required or even a desirable feature. My prior is that it's a good exercise for all students to interact with a syntax based program, and it can be made much less painful with good curriculum design and instruction. In any case, the fundamental point of having lighter versions of software designed specifically for learners is an interesting one.

In *The fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses,* Kim, Ismay, and Chunn (2018) doesn't describe adapting software to be simpler, but advocates for taming data in such a way that it requires students to focus on important obstacles. For example, at first providing data in a tidy format so students don't yet have to worry about reshaping. They put this advice into action by providing a well-documented R package with datasets from the data-driven journalism website FiveThirtyEight.

Their fundamental principle is that students need to be taught the “whole game” of data science as opposed to just isolated pieces one at a time, but doing so by tossing students too far in the deep end is counterproductive because students will often spin their wheels on low yield concepts or concepts they are not yet ready for. The solution, then, is to design learning experiences based on data that has been purposefully curated to both let students experience the whole game of data science and encounter the right obstacles at the right time (e.g. productive struggle). This strikes me as the exact right mindset data science teachers should hold. The extraordinarily difficult part is designing the datasets and the learning experiences with this level of purpose. The provided R package serves as a nice start.

The third paper with regard to technology in teaching data science is *Web Application Teaching Tools for Statistics Using R and Shiny* (Potter et al. 2016). They point out that applets can be powerful teaching tools especially for helping students build intuition. However, building applets requires technical expertise and a great deal of time. Shiny simplifies this process greatly. The paper comes with a companion site that illustrates a variety of impressive applets.

Applets seem somewhat unimportant relative to the main software used in the course, but I actually think teachers designing shiny apps for their classrooms could be incredibly valuable, especially if combined with thoughtful tame data exercises. Imagine a student tackles a tame data analysis but bumps into hairy statistical concepts along the way for which there is an applet specifically designed with that data analysis in mind to help the student build statistical intuition. This could be very powerful.

In summary, the biggest gap I see in the literature on technology use and activity design for data science is not enough focus on actually asking students to make decisions. Don’t ask students what the probability of someone arriving to work on time is, instead ask them what time they recommend the person leaves. Don’t ask students what the treatment effect is and if there are heterogeneous effects, ask them who they would give the medicine to with a fixed budget of a million dollars.

Then give feedback on both the process and the decision. For example:

> *Your group gave the medicine to young men instead of older women because you didn’t properly account for selection bias. 10 young men were cured but at the opportunity cost of 1000 older women’s lives. You should have controlled for zip code instead of state in your regression. Perhaps you overlooked this because you didn’t keep your geographic variables together when you were cleaning the data (see line 22 of your first script).*

Imagine how different that is than the usual feedback:

> *Your heterogeneous effect estimates seem to be off perhaps because you didn’t properly account for selection bias.*

My takeaway is that the essential piece to high quality data science activities is asking questions that focus on decision making and designing the activity such that students bump into important obstacles while obstacles that students are not ready to face are discretely removed by careful curation of data. The teacher also needs to deeply understand every wrinkle of the data so that they can quickly identify where students are and nudge them in the right direction. Real world data is preferred for sure, but if this requires using simulated data that feels real, I think that's okay too.

III. Bibliography

1) ASA. 2016. "Guidelines for Assessment and Instruction in Statistics Education (GAISE) in Statistics Education (GAISE) College Report College Report." http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.
    a) Somewhat stuffy report but interesting to see how official institutions think about things
2) Boaler, J. 2016. "Mathematical Mindset." San Francisco, CA: Jossey-Bass.
    a) Insight book on instilling mindsets that are productive for learning math, and I think it transfers pretty well to data science, too.
3) Cobb, George. 2015. "Mere Renovation Is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up." *The American Statistician* 69 (4). Taylor & Francis: 266–82.
    a) Insight and provocative commentary on movement by institutions
4) Cobb, Paul, and Lynn Liao Hodge. 2011. "Culture, Identity, and Equity in the Mathematics Classroom." In *A Journey in Mathematics Education Research: Insights from the Work of Paul Cobb*, edited by Anna Sfard, Koeno Gravemeijer, and Erna Yackel, 179–95. Dordrecht: Springer Netherlands.
    a) Overly academic language for my liking but does make interesting theoretical points with regard to culture and equity issues in the classroom
5) Dweck, Carol S., and Sumoreads. 2017. *Summary of Carol S. Dweck's Mindset: Key Takeaways & Analysis*. CreateSpace Independent Publishing Platform.
    a) Nice simple summary of Dweck's mindset work
6) Harraway, John A. 2012. "Learning Statistics Using Motivational Videos, Real Data and Free Software." *Technology Innovations in Statistics Education* 6 (1). https://escholarship.org/uc/item/1fn7k2x3.
    a) Suggests using light software for teaching statistics
7) Holmes, N. G., Carl E. Wieman, and D. A. Bonn. 2015. "Teaching Critical Thinking." *Proceedings of the National Academy of Sciences of the United States of America* 112 (36): 11199–204.
    a) Key point is get students practice making decisions
8) Khan, Sal. 2014. "The Learning Myth: Why I'll Never Tell My Son He's Smart." Khan Academy. 2014. https://www.khanacademy.org/talks-and-interviews/conversations-with-sal/a/the-learning-myth-why-ill-never-tell-my-son-hes-smart.
    a) Nice accessible example from Dweck's mindset work
9) Kim, Albert Y., Chester Ismay, and Jennifer Chunn. 2018. "The Fivethirtyeight R Package:` Tame Data' Principles for Introductory Statistics and Data Science Courses." *Technology Innovations in Statistics Education* 11 (1). https://escholarship.org/uc/item/0rx1231m.
    a) The most insightful piece written. All about removing and presenting the right obstacles at the right time.
10) Koedinger, Kenneth R., Jihee Kim, Julianna Zhuxin Jia, Elizabeth A. McLaughlin, and Norman L. Bier. 2015. "Learning Is Not a Spectator Sport: Doing Is Better Than Watching for Learning from a MOOC." In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, 111–20. L@S '15. New York, NY, USA: ACM.

    a) Best evidence that I know of that active learning is vital in MOOCs

11) Maurer, Karsten, and Dennis Lock. 2016. "Comparison of Learning Outcomes for Simulation-Based and Traditional Inference Curricula in a Designed Educational Experiment." *Technology Innovations in Statistics Education* 9 (1). https://escholarship.org/uc/item/0wm523b0.
    a) Good analysis but I'm a bit bored by this debate already

12) McKnight, Katherine S., and Mary Scruggs. 2008. *The Second City Guide to Improv in the Classroom: Using Improvisation to Teach Skills and Boost Learning*. John Wiley & Sons.
    a) Lovely book for strategies and thoughts behind getting people to feel comfortable together and creating a productive environment

13) Moore, Allison Amanda, and Jennifer J. Kaplan. 2015. "Program Assessment for an Undergraduate Statistics Major." *The American Statistician* 69 (4). Taylor & Francis: 417–24.
    a) Interesting to see measurement work at this broad of a level

14) Potter, Gail, Jimmy Wong, Irvin Alcaraz, Peter Chi, and Others. 2016. "Web Application Teaching Tools for Statistics Using R and Shiny." *Technology Innovations in Statistics Education* 9 (1). https://escholarship.org/uc/item/00d4q8cp.
    a) Nice summary and examples of how applets (especially Shiny) can be powerful

15) Stern, Dan. 2017. "Teach Yourself Data Science: The Learning Path I Used to Get an Analytics Job at Jet.com." 2017. https://medium.freecodecamp.org/a-path-for-you-to-learn-analytics-and-data-skills-bd48ccde7325.
    a) I liked reading a couple of these personal anecdotes but eventually they all feel the same

16) Straten, Savina van der. 2017. "Leaving My VC Job to Learn about Data Science and Machine Learning." 2017. https://towardsdatascience.com/leaving-my-vc-job-to-learn-about-data-science-and-machine-learning-4dbc15427fc5.
    a) What I expected

17) Wieman, Carl E. 2014. "Large-Scale Comparison of Science Teaching Methods Sends Clear Message." *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8319–20.
    a) I love this meta-study. I feel like it gets at the heart of our key educational inefficiency which is passive learning.

18) Wieman, Carl, and N. G. Holmes. 2015. "Measuring the Impact of Introductory Physics Labs on Learning." *arXiv [physics.ed-Ph]*. arXiv. http://arxiv.org/abs/1507.00264.
    a) Cool to see Wieman take his broad meta-study and dive deep into a single environment

19) Wirth, Karl R., and Dexter Perkins. 2005. "Knowledge Surveys: An Indispensable Course Design and Assessment Tool." *Innovations in the Scholarship of Teaching and Learning*. serc.carleton.edu. http://serc.carleton.edu/files/garnet/knowledge_surveys_indispensabl_1313423391.pdf.
    a) Kind of obvious or silly but also interesting to think about measurement by just directly asking