Changing Paradigms: Faculty Moving to Data Science from Other Disciplines

Bernard Ricca, Bruce Blaine, Kathleen Donovan, & Anne Geraci (St. John Fisher College)

Introduction

Student demand for data science has increased in many schools. However, in many institutions hiring new data science faculty has proven difficult. Assisting existing faculty from a variety of traditional disciplines in making a change to data sciences approaches - particularly by moving away from more parametric approaches to more resampling-based approaches using R - can help meet the need for faculty. This qualitative study explores the transitions made by two faculty at a small liberal arts college from other disciplines (political science and computer science) to data science, both from the perspective of the faculty members making the transition and from the perspective of faculty who had already made the transition. Although both faculty members had some background in statistics, the transition to data science required professional development, mentoring, and a deep commitment on the part of the faculty and the program. From these experiences, insight into the transition process have been gained; these insights may be useful to programs at other institutions.

Context

St. John Fisher College is a small liberal arts school. It has had an undergraduate Statistics major for 13 years, and a Data Science minor for the last three. Statistics began with a single dedicated faculty member; it now has a faculty member dedicated to the program, two faculty members who are assigned a majority of the time to the program, a faculty member shared with another department, and occasionally other faculty teaching in the program. (These programs are also supported with a number of mathematics, computer science, and other disciplinary courses taught by faculty in those disciplines.)

STAT 160 – Introduction to Data Science – serves as the entry point for both programs. STAT 160 is taught using R, and includes basic inferential and predictive modeling (using resampling approaches; see Tintle, et al.) and data wrangling. This course has gone from a single section offered each year to six sections in the last academic year, with additional growth anticipated. The growth of this course has necessitated the need for additional faculty to teach the course.

All four faculty regularly teaching in the statistics & data science (SDS) programs come from varied backgrounds: The program was founded by a faculty member from psychology, and the three "new" faculty come from physics/science education, computer science, and political science. As there are still relatively few data science doctoral programs compared to the growing need for data science faculty, transitions from these fields to data science will likely be common in the near future.

Method and Analysis

This exploratory, retrospective study began with notes from the meetings of SDS faculty. These have been augmented with semi-structured interviews and informal discussions among the faculty. From the written records of these sources, a process of analytic induction (Hammersley, 2004) was used to identify the themes presented here. All four faculty members commented on and edited these themes and the supporting anecdotes presented here.

The Dynamics of Change: Emerging Themes

Faculty not forced, although encouraged, to make the change.

Five main themes emerged from this exploratory study: Faculty dispositions, strands in the shifts that faculty made, details about the professional development necessary to facilitate the change, the need for on-the-ground support for faculty, and possible pathways for the change.

Dispositions

Primary among the pre-requisites for making a change is openness to the change on the part of the faculty member. Pedagogy must match both the content being taught and the faculty member's personality and beliefs (Ricca, 2012). Clearly, coding and resampling are both part of the content of data science, but that is not sufficient for faculty to adopt these approaches easily. There are ways that programs can promote that openness (see *Programs: How to Help*) but a willingness on the part of faculty is a necessary condition for making the change. One faculty member making the change "knew people used R" in graduate school, but was out of graduate school for several years before shifting to a data science approach.

Shifts

Faculty coming into data science often must make three shifts in their teaching: coding, using empirical distributions, and the cumulative nature of data science.

The advent of computer technology brought about the development of statistical software (e.g., SPSS). The most common of these software packages, however, typically use a "point and click" approach. Shifting to a coding language such as R requires not only learning the language, but also adopting a different attitude toward the analysis of data: Instead of a cognitive focus on the software menus and submenus, the focus shifts to the data and what is done with those data. E.g., piping from a data frame, or the parameter *data = <data frame>* within a command explicitly bring the data into the command. In an environment like SPSS, the data are always present, but in the background. This shift also has a pedagogical dimension, as faculty must move to seeing students as producers of knowledge and not merely consumers. (Not all faculty view students as consumers, but those who do must shift.)

A second shift that must be made is from the normal distribution paradigm (Guastello, 2011) to working with empirical distributions and resampling. Resistance to this shift has its origin in the paradigm of faculty: If resampling is so important, then why isn't it done that way (in other programs and disciplines)? Faculty noted that having a textbook with resampling approaches would have been helpful; all of the faculty here made the transition to empirical distributions before we adopted Tintle et al, and in some cases, before that text existed.

The last shift that some faculty experienced, a pedagogical one, is that data science content is cumulative, as opposed to the "hub and spoke" structure of some disciplinary programs. (This is not to say that these disciplines have a hub and spoke structure, only that their teaching has that structure.) This shift highlights the need for ongoing assessment – Do students really understand the content? – before moving on. Failure to do so has consequences that aren't always evident in other disciplines.

Professional Development

All of the faculty made use of various available professional development opportunities, but their choices were varied and all found it helpful to have guidance when choosing from the possibilities. However, all of the faculty found that professional development tends to be very specialized, and hence, works better when one already has a sense of what to fill in rather than as an aid to the process of shifting. Further, relatively little of the professional development focuses on pedagogy.

Support on the Ground

All of the faculty noted that having a community of other data science instructors was very valuable (or in the case of the founding faculty member, was greatly missed). The value came in two related forms. One benefit of having a community of data science instructors was the availability of answers to the question "[A]re you struggling with this, or is it just me?" regarding various classroom difficulties. The other benefit was assistance in translating ideas learned in one context (e.g., an agricultural context using JMP) to another context; even when only one faculty member knew the source idea, the ability to discuss and critique ideas with another faculty member was beneficial. (This is sort of an Anna Karenina situation: two people, even when neither possesses the entire ability, can help each other develop the necessary knowledge because each person knows different things.)

Paths

The paths taken from one statistical paradigm to another, even among this small sample, were idiosyncratic. Incremental approaches, jumping in all at once, and so on, were all in evidence. The paths appear to be taken in accordance with each faculty member's personality; there is nothing obvious in the disciplinary background of the faculty that better explains the various paths. Furthermore, the time it takes to make the various shifts appears to be at least two semesters of active work, but there was clearly some prior work (in becoming open to the shift) and some post-transition work (to become comfortable in the new paradigm). The length of the process is not surprising, however, and is in line with what is generally known about the intensity and length of time necessary for effective faculty professional development, and the need for professional development to be connected to classroom teaching (e.g., Dori & Herscovitz, 2005). Regardless of the path, faculty shifted their teaching, but always in ways that fit with their established teaching styles.

Programs: How to Help

There are four formal actions that programs can take to help faculty make a transition from other disciplines to data science.

Although it may be obvious, the first thing that departments and programs can do is to provide financial support for professional development. Free professional development opportunities do exist, but some useful sites involve a cost. Providing these funds can sometimes be difficult because, as was true here, the faculty needing support may formally reside in a different department from the one that provides the support.

Second, departments must plan as far in advance as possible. Changing to a modern data science approach requires time, especially for faculty who do not have a coding background.

Third, faculty noted that opportunities for co-teaching and/or auditing courses were (or would have been) helpful. Programs should make such opportunities regularly available to faculty, and should incentivize those activities.

Lastly, programs can help promote openness on the part of non-data science faculty through outreach. In particular, it appears that helping outside faculty integrate either coding or resampling approaches into their own research projects may build some confidence in coding and resampling; many journals are heeding the call for reproducibility in research and for greater care with non-parametric distributions. For many faculty, these external pressures open them to coding and resampling, and in the process may lower resistance for making these transitions within the classroom.

What to Expect: Faculty

Faculty making the shift should be prepared for several resulting changes.

Chief among these changes, perhaps, is that students often think that a course involving coding and statistics, even when taught in an environment such as R, should be taught as two courses, one statistics and a second coding course.

As a consequence of this, faculty can expect an increase in the number of students wanting consultations during (and outside of!) office hours.

Lastly, faculty making the change from "hub and spoke" situations may find themselves asking the same question – Do students really understand the content? – in their other courses. Although this self-evaluation is inherently a good thing, it can be disconcerting.

References

Dori, Y., & Herscovitz, O. (2005). Case-based long-term professional development of science teachers. *International Journal of Science Education*, 27(12), 1413–1446. http://doi.org/10.1080/09500690500102946

Guastello, S. (2011). Frequency distributions and error functions. In Gregson, R., & Guastello, S. (eds) *Nonlinear dynamical systems analysis for the behavioral sciences using real data*, p. 33-72. New York, NY: CRC Press

Hammersley, M. (2004). Analytic induction, in Lewis-Beck, M., et al. (eds) *The Sage Encyclopedia of Social Science Research Methods*, Thousand Oaks CA: Sage.

Faculty PD

Ricca, B. (2012). Beyond Teaching Methods : A Complexity Approach. *Complicity: An International Journal of Complexity and Education*, 9(2), 31–51.

Tintle, N., Chance, B., et.al. (2016). Introduction to Statistical Investigations. Hoboken, NJ: John Wiley & Sons, Inc.