Data Science Rex: How data science is Evolving (Or Facing Extinction) Across the Academic Landscape

Jennifer Lewis Priestley, Ph.D. Kennesaw State University

ABSTRACT

The discipline of data science has seen an unprecedented evolution from primordial darkness to becoming the academic equivalent of an apex predator on university campuses across the country. But survival of the discipline is not guaranteed. This session will explore the genetic makeup of programs that are likely to survive, the genetics of those that are likely to become extinct and explore the role the business community plays in that evolutionary process.

PAPER

Data science is a nascent discipline that is evolving out of primordial academic darkness to assume some evolutionary commonalities in university curricula across the country. The academic community should be applauded for pivoting in meaningful and unprecedented ways to respond to the demands of the private and public sectors for deep analytical talent[1]. Since 2007, the number of masters-level programs in "data science" or "analytics" has exploded from 0 to over 100[2]. In the last few years, there has been a parallel emergence of Ph.D. programs in data science, which is an important double-sided solution to the talent gap – not only are Ph.D.s uniquely qualified to directly engage in private sector research and problem solving, but with a "terminal degree" these individuals are also qualified to teach the next generation of data scientists.

A scan of about 100 "data science" and "advanced analytics" programs across the country is represented in Figure 1. As this constellation of data science-related programs illustrates, the proliferation of programs is not only geographically diverse, but is also evolving from completely different spheres of influence (the nodes represent the colleges where "data science" or "advanced analytics" programs are housed in universities across USA)[3].

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[2] http://analytics.ncsu.edu/?page_id=4184

[3] Program information from college websites collected September, 2016

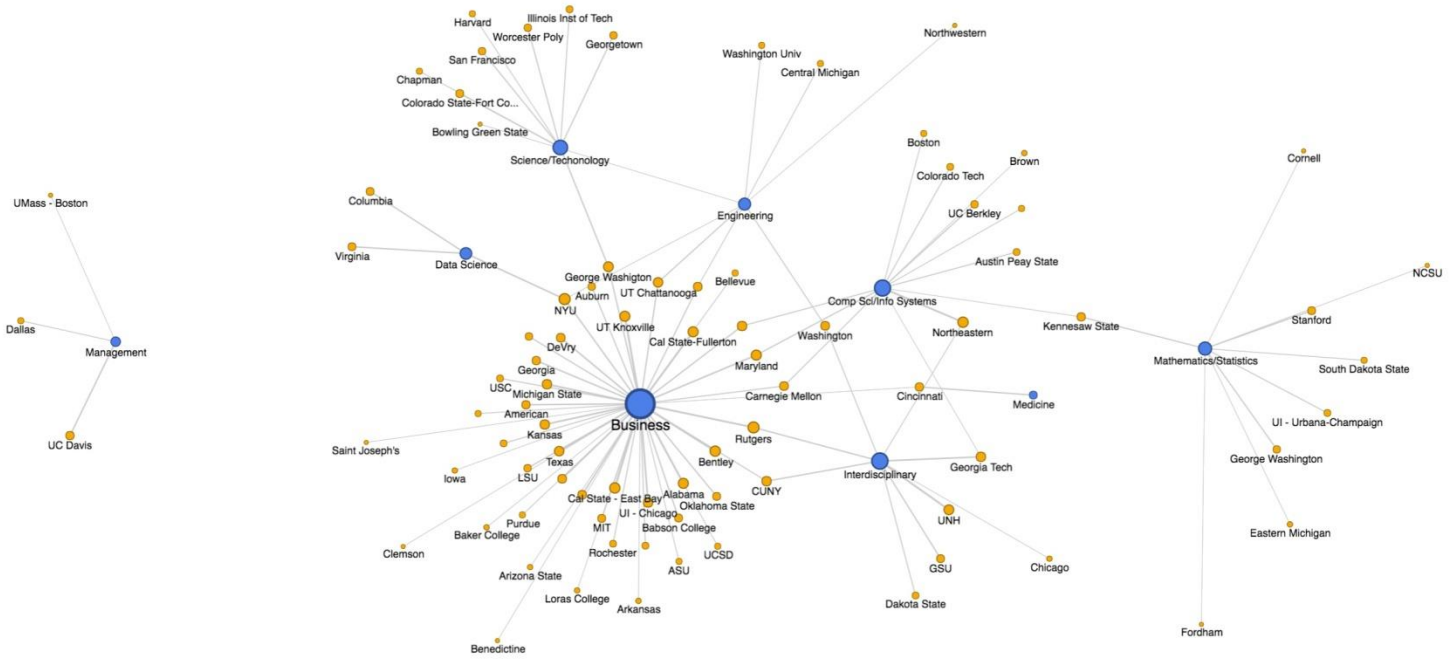

Figure 1: US Graduate Programs in Analytics and Data Science by Academic College

Select any of these programs at random and you will likely see commonalities of applied statistics, basic computer science and applied project work. However, the characteristics that are dominant will differ depending on their orientation and importantly on where they were conceived within the university.

Generally speaking, programs that evolved out of colleges of business are more likely to have a stronger orientation towards applied problem solving, case studies, and utilize more "point and click" software to translate data into information. They are also likely to have a stronger emphasis on the "softer", more latent skills associated with data science, including contextual interpretation, communication and visualization. These graduates are more closely aligned with "downstream" aspects of data science.

Alternatively, programs that have evolved out of a college of science (including computer science, information technology and mathematics), are often more oriented with the computational mathematics of working with big data structures, algorithmic design, scripting languages, machine learning, and statistics – on the "science of data" and applying the rigors of the scientific method to the translation of data into information for problem solving. They are less likely to emphasize domain expertise. These graduates are more closely aligned with "upstream" aspects of data science.

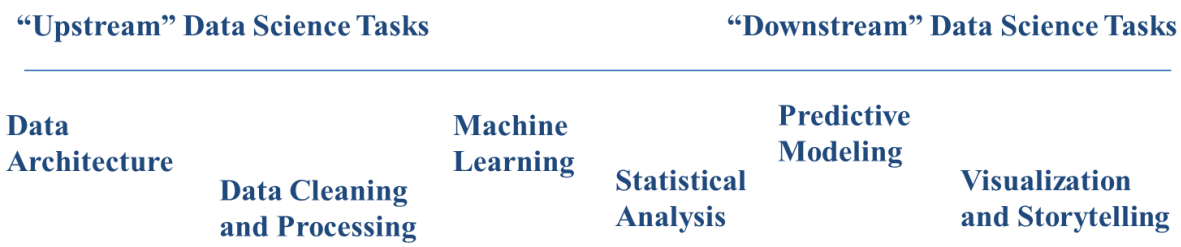

Figure 2: Upstream and Downstream Data Science Tasks

A scan of the emerging Ph.D. programs – the "youngest" evolutionary iteration of data science programs – illustrates that almost all are firmly grounded in mathematics and computer science even if they are housed in business schools. Because of the length of these programs (generally four or more years compared to two or fewer years for most masters programs), students have the luxury of time to develop deep technical skills combined with domain expertise and research experience to become the next generation of Chief Data Officers or assistant professors of data science.

For example, consider the constellation of the computing applications in Figure 3 taken from the programs above[4]:

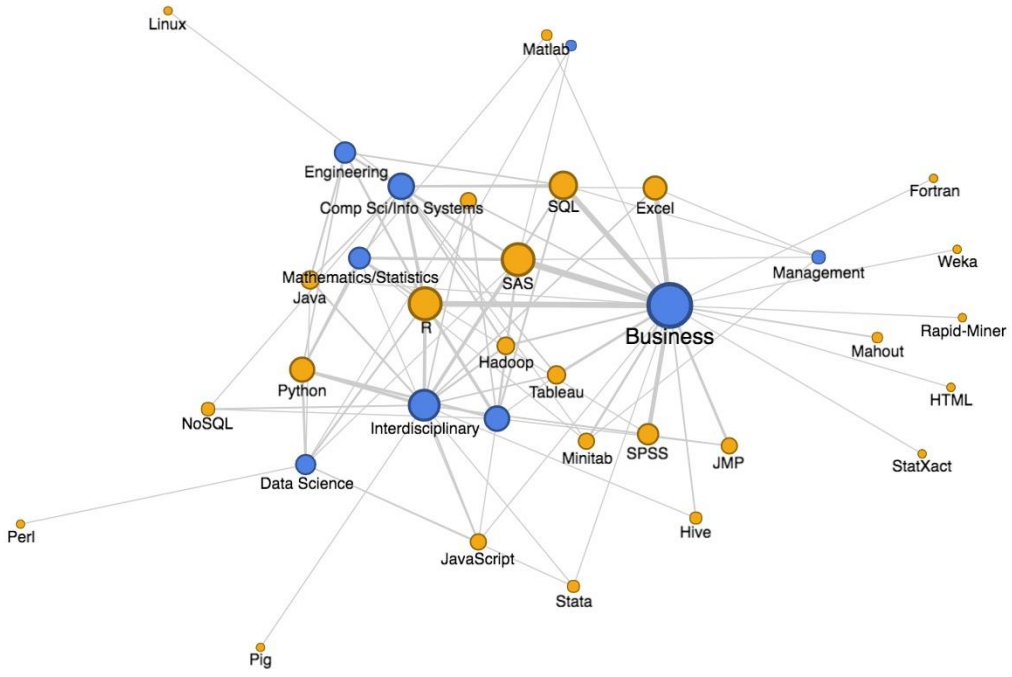

Figure 3: Programming Packages/Languages Used by Graduate Programs in Analytics and Data Science by Academic College

[4] Program information from college websites collected September, 2016

The wider paths represent the more frequently used language/software from the programs above. While all programs generally emphasize SAS, R and SQL, programs housed in business schools are more likely to emphasize "point and click" software (e.g., JMP, SPSS, Excel, Minitab) while programs housed in science colleges are more likely to use scripting languages (e.g., Java, Python). It is important at this point to emphasize that neither approach is "wrong". Both orientations are evolving organically within their own context.

This evolution of programs is analogous to the characteristics of Darwin's finches which have developed unique beaks to most efficiently thrive in the demands of the ecosystem from which they have evolved.

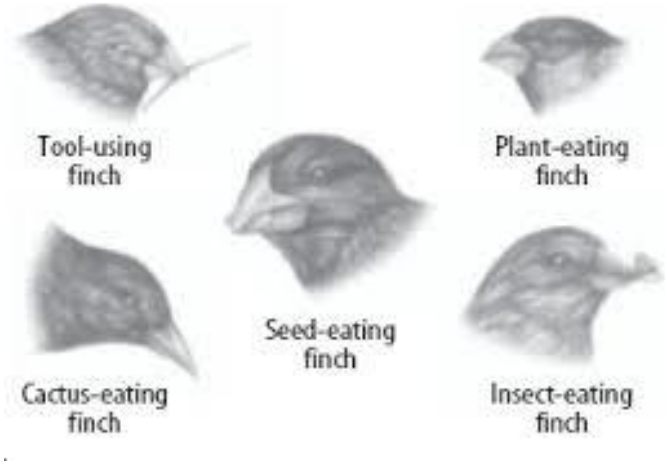

Figure 4: Darwin's Finches

As we witness these first stages of our discipline's evolution, there are several lessons from natural selection that we can learn from.

Lesson 1: ***In the process of natural selection, individuals in a population who are able to adapt to a particular set of (and changing) environmental conditions have an advantage over those who are less able to adapt***.

In data science, maybe more so than any other discipline, our environmental conditions are changing at lightning speed – data continues to be generated in differing ways, through differing channels in ever greater volumes. Many of the skills and concepts that were being taught in the classroom just a few years ago have almost no recognized value in the marketplace.

When asked what made him the greatest hockey player of all time, Wayne Gretzky responded "*I skate to where the puck is going to be, not to where it is*".

So, how do we ensure that graduates of data science programs can adapt? And "skate to where the puck is going to be?" One way is through helping students develop critical, non-linear thinking skills. This can be supported by incorporating unstructured problem solving, computational mathematical theory, graph theory (supporting big data relationships), algorithmic design (supporting computer science) and matrix algebra (supporting statistics). They need to see real data, from real companies – which is challenging, messy and complex. They need to understand that there is more than one right answer, and

sometimes there is no answer. This has to come from unstructured, incomplete data where the question is not even well understood – not just established case studies with pre-processed data. This is where university/private sector partnerships are critical - these are the contexts where students will draw from their complete toolset of applied and theoretical tools…and adapt.

Lesson 2: ***Popular interpretations of "survival of the fittest" typically ignore the importance of both reproduction and cooperation. To survive but not pass on one's genes to the next generation is to be biologically unfit. And many organisms are the "fittest" because they cooperate with other organisms, rather than competing with them.***

This concept of "cooperation" in the context of successful data science education takes two forms.

The first is recognition of the fact that the discipline of data science integrates concepts from mathematics, statistics, computer science, business, and social sciences. In 2012, Josh Wills[5] tweeted that the data scientist is the "*Person who is better at statistics than any software engineer and better at software engineering than any statistician*". The "Priestley Corollary" to this quote is that the data scientist is also the "*Person who is better at explaining the business implications of the results than any scientist and better at the science than any business person*" (can someone Tweet that out please?).

The data science programs which are most likely to "survive as one of the fittest" are those programs which directly or indirectly integrate an interdisciplinary orientation. This is admittedly difficult for most universities – no one does bureaucracy and "ivory tower" better than universities.

A second dimension to this concept of cooperation is that data science, more than any other discipline, is inherently interdisciplinary. This is evidenced by the explosive demand for the talent. We are seeing the employment equivalent of a "run on the bank" – all verticals are chasing the same talent at the same time. At our own university, we will have students in our MS in Applied Statistics and data science program receive job offers from the health care sector, the retail sector and the financial services sector simultaneously. Why? Because the core required skills are the same – the ability to extract, transport, load, clean, transform, analyze, translate and tell the story.

Students from multiple disciplines who engage in team-based project work – comprised of diverse teams with students from sociology, marketing, physics, mathematics, nursing, finance – have a very different experience than students who are engaged in team-based projects will students exclusively from their own discipline. There are important "latent" skills that students develop by working in teams with people who think differently but are tasked with the same problem. These latent skills will also contribute to their ability to adapt to changing environments.

Lesson 3: ***Individual organisms don't evolve. Populations evolve.***

Populations evolve. Programs evolve. Just like we have a responsibility to help our students adapt to a changing environment, our programs need to have built in flexibility to respond to changes in the market. This should not be confused with "teaching to the job" (or again, skating to where the puck is). Directors who oversee programs in data science need to be prepared to make curricular changes the changing environmental conditions. This is best done by integrating with a strong advisory board of executives and

[5] https://twitter.com/josh_wills/status/198093512149958656

practitioners who can provide regular and ongoing feedback regarding the performance of the program graduates.

Program evolution is the primary reason why the cries for "standardization" of data science curricula and required industry "certifications" for data scientists similar to those developed for accountants or actuaries are pre-mature. Standardizations at this stage would only suppress the valuable and innovative work that is evolving organically across the academic landscape: artificially imposing "standards" for the beaks of finches would have likely lead to the pre-mature extinction of some species. And who would set those "standards"? Operations Research? Computer Science? Marketing? Mathematics? The "invisible hand" that sets the standards would artificially manipulate the future discipline – likely creating the extinction of excellent programs that are in their infancy. The seed-eating finches would have died out if their beaks had been contrived to take the shape of the insect-eating finches.

However, this also raises an important issue. A business analyst has important skills to contribute to the talent gap. As does the data engineer. As does the graduate who was engaged in the science of data. But, they are not interchangeable. We do our discipline and the market a disservice when we fail to make the distinction of the business analyst and the data engineer. They both play a role in the data science ecosystem.

In the context of academia, data science could look to economics as a potential discipline role model. Why? One reason is related to organizational placement. A scan of economics programs across the country will reveal programs housed in business schools, schools of arts and science, and independent "schools" of economics. Economics programs exist at the undergraduate, masters and Ph.D. levels in all of these academic locations. A second reason is that these programs may emphasize microeconomics or macroeconomics - some have a strong mathematical orientation and require a series of calculus courses and some emphasize public policy and domain expertise. Finally, economics programs place graduates both equally as well in academia as well as in the private sector with titles like "Chief Economist". It is also worth noting that the now well-established discipline of economics has never instituted a "standardization" process – with no detriment to its evolutionary path.

As data scientists, we are living in our first evolutionary epoch – few people can claim that they were present and contributed to the development of a new discipline. As the evolutionary ancestors and thought leaders of a nascent discipline, we have responsibility to our intellectual descendants. Will data science evolve and flourish with rich and meaningful iterations like Darwin's finches? Or are we destined to simply have our moment as the Data Science Rex?