The Data Scientist: Specialist or Generalist

Written by Richard Boire at Boire Analytics

As a practitioner with over 30 years of experience in the field, the discipline of data science has evolved dramatically. As I have discussed at length in previous articles and in my book, the approach and process in conducting data science exercises has not changed. Yet technology has transformed the discipline in that many tasks are now being automated thereby allowing the data scientist to focus more of their intellectual energy as problem solvers. At the same time, data access and processing have increased to the point where analysis can now be conducted on data which was previously inaccessible. These new capabilities, though, have seen the emergence of deep learning techniques, the real essence of AI, as being the cause of transformational business changes across virtually all business sectors.

Certainly much discussion and content has been written about these disruptions caused by Big Data and AI. Yet, much of this disruption discussion seems to be remiss in addressing how the roles and responsibilities within data science have evolved over the years. But in order to truly understand the current nature of data science, an historical perspective does provide a reference point in how the discipline has evolved.

The early date science practitioners such as direct marketing organizations and credit card issuers, focussed their efforts on hiring individuals with mathematical skills related to statistics. At the same time, it was a required expectation that the newly hired data scientist be able to code in some programming language. In fact, when I was hired at American Express, the organization was cognizant of my statistical skills and more importantly my understanding of how they could be applied in a given business situation. Yet, within my first week, I was made aware that these skills, although vital to the data science role, were insufficient if I could not code or program. So, I focussed the next 6 months in learning how to program in SAS as other programming languages such as R and Python were simply unknown at that time in the emerging field of data science. I still remember my first SAS program in conducting an analytics exercise for one of the American Express product groups. It was at that point that I had my epiphany where I realized my passion for data science. I could “work the data” without the utilization of any external I/T resources. Programming was indeed the core function of the data scientist at that time and I loved it.

As my programming skills improved, knowledge and understanding of data and how to maximize its value became the inevitable outcome given my newfound technical skills. This data knowledge, given the technical constraints at that time, also required statistical knowledge as sampling of data was the norm in virtually all analytics exercises. Yet, the real core of my work was to build predictive models which would have again leveraged my statistical knowledge.

In those early days, I could extract my own data, create the analytical file, and use the right statistics in developing a predictive model. But these technical skills were always directed towards the intention of solving the business problem. I also had to ensure that I could communicate my results to the business unit. Otherwise, my work would not be utilized. The practitioner in many ways was a generalist in the early days of data science. The early days also saw large volumes of data but in those days, batch processing of the data, particularly in applying a given solution, was the norm as the technology was incapable of providing real-time data processing. But in the development of a solution with large volumes of data, sampling of these large volumes was the approach in order to have quicker turnaround time when conducting a variety of different statistical analyses.

Big Data and AI have now introduced more specialization into the data science roles. Let’s explore this more closely. Big Data has exponentially increased the volumes of data that now can be processed for

analytics. Not only do we have to deal with these vast volumes of data, but the data often arrives in varying formats and is arriving at increased velocity. ETL processes that were once used in a batch environment are no longer applicable. Increased programming capabilities in being able to program in map reduce are fundamental requirements that are needed in this new ETL paradigm. Suppliers have emerged to provide capabilities in this area and to operationalize many of these tasks. But technical skills are still required in order to function in this more automated environment. For example, besides how the data is being processed, one needs to understand the movement of data and where it is being stored. If we think about the data science process where most of the work is done in the creation of the analytical file, this ETL process, which is a component of this stage, now comprises more tasks and technical knowledge than what was required in the past. More computer-science and engineering type knowledge are the requisite skills for these type of activities.

But once the data requirements are established within the ETL process, the traditional data science tasks of the data audit and how to integrate all this data with meaningful variables still remain the major proportion of work in creating the analytical file. The subsequent remaining stages of conducting the analytics alongside the final stage of measurement and implementation all require the typical data science skills and expertise. However, the challenge , here, is that the data scientist needs to implement and measure the solution within a more operationalized and automated environment, thereby once again requiring data engineering and the more complex computer science type skills.

AI or deep learning is not a new technique but Big Data can now provide the extremely large volumes of data which are a requirement for successful AI solutions. Sampling, which has traditionally been the data science foundation when building solutions has now become less relevant where one can model all the data. But within AI, there is a deep technical need to understand how different parameters can yield different solutions. Much of the mathematics tends to be more engineering-based as many of the equations are based on calculus concepts such as derivatives and integrals. This is due to the fact that much of the math is based on rate of change within a given solution and trying to find an optimal solution where the rate of change converges to some type of minimum. In traditional statistical models, the practitioner is often analyzing the variance between a predicted solution and observed values given an assumed distribution. For example, in multiple regression, the distribution is linear while for logistic regression, it is log linear or a sigmoid type curve.

In building data science solutions, one will explore a variety of techniques which will include both AI and non AI solutions. But as newer techniques and technologies are now required, we are seeing the need for these more specialized technical type skills. Yet, even with these newer technical skills, the growing demand for data science as a core business discipline has resulted in a greater need for business analysts. The primary function of these roles is to be able to work with the solution or data in order to communicate what will be meaningful to the business. The growing emergence of data visualization and BI tools is industry's attempt to empower as many people as possible within the area of data science and analytics. In some cases, these tools have enough sophistication where the user can build an analytical file without knowing how to code or program. But the user still needs to have a deep understanding of data and how to "work" it into an analytical file. These type of roles are becoming more and more specialized as practitioners now focus their efforts on learning certain tools whether it be the hard core programming skills of R, Python or SAS, the visualization tools such as Tableau, Qlik, or the non-programming data manipulation tools such as Alteryx.

But what does this all mean for data science as a discipline? Like in many professions such as medicine, there will be both generalists and specialists. Historically, one learned data science as a generalist since the practitioner had to do all the tasks within a given project. Now, the learning route involves working in a specialized role within a given project. Ideally, the data scientist should gain exposure across the all these areas of specializations, thereby increasing his or her generalist knowledge.

The need for generalists will grow in a world of increasing automation. Think ahead to 10 or even 5 years from now as newer and more automated tools emerge on the scene to allow us to solve even more problems. The generalist will be the one who can identify problems, determine how to best organize the data, and utilize the right analytics in solving the problem. The generalists will then manage the specialists to ensure the proper execution of a given project. Generalists, who have been able to gain experience across these different data science specializations, will see the demand for their skills continue to grow. In the future world of business leadership, the generalist or citizen data scientist will become a permanent fixture on a company's Board of Directors.