Landscape of the Data Industry

Porter Bagley
Brigham Young University, chaselarsonmsr@gmail.com

Follow this and additional works at: https://scholarsarchive.byu.edu/marriottstudentreview

Part of the Accounting Commons, Business Administration, Management, and Operations Commons, Business and Corporate Communications Commons, and the Entrepreneurial and Small Business Operations Commons

Marriott Student Review is a student journal created and published as a project for the Writing for Business Communications course at Brigham Young University (BYU). The views expressed in Marriott Student Review are not necessarily endorsed by BYU or The Church of Jesus Christ of Latter-day Saints.

Recommended Citation

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Marriott Student Review by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
LANDSCAPE OF THE DATA INDUSTRY

BY PORTER BAGLEY
In an article published by Harvard Business Review in 2012, “data scientist” was named the “Sexiest Job of the 21st Century” (Davenport and Patil 2012). Seven years later, the field of data science is still as hot as ever. Data science is an extremely versatile set of tools and methods that allow us to draw powerful insights from massive amounts of data. While useful in virtually any application, it has especially important implications in the business, medical, and tech sectors. Many organizations have been clamoring to fill positions in their newly-formed data science teams. However, it is important that both those seeking to find work in the data industry as well as those hiring for data positions step back and look at data science as more than just the hottest job of the century. In this article, I would like to answer the following question: “What does the landscape of the data industry currently look like, and what implications does that have for the future?”

With all the hype surrounding the data industry, many of the associated terms have turned into buzzwords thereby losing much of their original meaning. So, to begin, I would like to clarify some of the essential terms within the industry:

**Artificial intelligence (AI):**
Using technology to make “smart” decisions that previously would’ve only been possible with a human mind involved (this is a very broad term).

**Machine learning (ML):**
A branch of AI wherein computers utilize statistical tools to “learn” from data.

**Data science:**
The use of advanced models, statistics, machine learning and other tools to extract insights from data — especially with the intent to make future predictions (RapidMiner 2019).

**Data analysis:**
Similar to data science, but less complex. Focused more on drawing insights from known data rather than making predictions.

**Data engineering:**
The process of collecting (mining), cleaning, processing, and generally preparing data for use in data analysis and data science. “Data wrangling” is a term used to sum up all of these functions as a whole.

It is important to note, as I will address later in this paper, that currently in the data industry these different methods aren’t always performed exclusively by those who have the matching job title. For example, data engineering isn’t done only by the data engineers.

### Popularity of Data Science

So why is data science so popular in the first place? Well, first off, data science has become more accessible in terms of practicality—the exponential growth in computing speed, power, and memory allows us to process and store more data than could’ve been imagined a few decades ago. But more importantly, data science can solve very impactful problems. With the sudden rise in popularity of data science, some may believe it is just a fad that will die out in a couple of years. However, the influence data science has in a myriad of different applications is far too large for us to ever rationalize abandoning data science. Here are a few cases where data science has made a major difference:

In combating Alzheimer’s, early diagnosis is crucial. Using AI, researchers devised an algorithm that could identify the signs of Alzheimer’s nearly a decade before the clinical symptoms would appear (Anathaswamy 2017, 103-10). The researchers designed this AI with two applications: “providing a reliable tool for clinical trials and a disease signature of neurodegenerative pathologies” (Amoroso et al. 2017, 1).

IBM, the technology giant, designed artificial intelligence that uses data science to predict with 95% accuracy which employees are considering leaving their jobs. Due to its high accuracy, this tool has saved IBM nearly $300 million in retention costs since its implementation (Rosenbaum 2019).

The term “natural language generation” (NLG) refers to the ability of computers/robots to generate text or speech automatically in order to communicate (think of Siri, Google, or Am-
In the past, much of the work of NLG had to be done through creating templates that AI would simply fill in. However, recent increases in data-driven techniques have allowed us to make massive progress in the NLG field — bringing us closer to having seamless verbal communication with computers and robots (Foster 2019).

The above examples are only a small sampling of the hundreds of thousands of recent applications of data science. If you begin to pay attention, you will see that many of the things you interact with on a daily basis have data science as an underlying backbone.

The Landscape of the Data Industry

Now, just how popular is data science? With companies quickly realizing all the incredibly impactful applications of data science, many are rushing to create and fill positions in their data science teams. There were over two million job listings within the data industry in 2015, and jobs are projected to grow 15% by 2020, with the specific roles of “data scientist” and “data engineer” projected to grow 39% (Markow et al. 2017, 3-16). The workforce is struggling to keep up with all of this growth; there is an incredible amount of unmet demand for workers with the needed skills, knowledge, and experience. Gartner made predictions that in 2014, 64% of large enterprises would intend to execute projects involving big data, but 85% of the Fortune 500 would fail to do so due to lack of human capital (Kandel 2014).

Representatives at higher education institutions aren’t blind to this growth. They are responding by racing to create curriculums which can prepare graduates to enter the industry. However, a study published in 2015 was only able to find 13 data-related undergraduate degree programs which were complete enough to review and analyze. They found that while most of the programs cover data mining and modeling/analytic techniques, there is a general lack in courses covering data capture, data preparation, data storage, and data security (Asheim et al. 2015, 110). Certainly, there have been additional institutions which have since designed programs in response to this growth, but there likely remains a shortage of programs that are teaching all of the required skills.

Still, there is great potential for success with universities that can create their programs appropriately. Brigham Young University (BYU), where I am a student, has a new major called Applied and Computational Mathematics (ACME). In a recent info session, one of the BYU math professors who designed the ACME curriculum said that BYU’s ACME program was among the most modernized and progressive data science programs in the nation, and that as such, BYU ACME graduates were in high demand. In 2018, the median salary for new ACME graduates was $102K (including signing bonuses), and job offers made to ACME graduates included companies such as Amazon, Goldman Sachs, and Microsoft (“Outcomes” 2019).

However, despite the rise in need and popularity of data jobs, it would seem that many companies don’t know much beyond the fact that they need to hire data scientists. Across almost every industry with job listings for data positions, there is a lot of ambiguity. The following graphic diagramming needed skills in different data roles can help illustrate the reason for some of the confusion:

In Figure 1 (Nair 2018), it is easy to see that many of the skills needed in these positions overlap and that professionals with the title “data scientist” should be able to do almost all of them proficiently. For this reason, a company may figure that a data scientist can and should perform all of these functions. However, many of the above skills are also buzzwords — leading to even more ambiguity. The result is that companies are putting out job descriptions that are far too vague or have a far-fetched wish list composed of dozens of different skills they hope their future data scientists possess (Verma et al. 2019, 243).

A Need for Balance

Unfortunately, even if they can find a candidate who somehow fulfills the requirements listed in their job posting, many companies find it unclear what it is they want the data scientist to do. In many cases, a data scientist might end up doing things that a simple business analyst would...
typically do — such as organizing tables in Excel — rather than utilizing the expensive and highly-demanded skills that the company hired them for — such as creating neural networks to make statistical predictions. It’s as if the companies just hired a data scientist because they heard it was a good idea to have one, but they don’t have any work cut out for the data scientist to do.

Much of this problem lies in companies who don’t already have an established data science team, or who have a team that is unbalanced. (To help better understand the hierarchy within a data science team, refer to the Figure 2 on page 38.)

Not only does this graphic help illustrate the different duties associated with certain job titles, it also provides reference for the proper proportion of those titles within an organization or team (Bolard 2018). The two parts of this pyramid that I would like to focus on are the “Data Engineer” and “Data Scientist Analyst” sections. As was defined at the beginning of this paper, data engineers are responsible for the flow and cleaning of data, while data scientists and analysts are responsible for analyzing, learning from, and optimizing the data for future predictions. It can be seen clearly that the “Data Engineer” section is closer to the foundation of the pyramid, and also that it is wider than the positions above it. This reflects the proper proportion of data engineers to data scientists.

Unfortunately, within the industry we often see a hierarchy that looks more like a tall rectangle than a pyramid — with organizations hiring too many data scientists and not enough data engineers. In a study conducted by Sean Kandel and published in the Harvard Business Review, he found that most of the average data scientist’s time was spent “turning data into a usable form rather than looking for insights.” He continues by saying that 50-80% of a data scientist’s time may be spent doing data cleansing and preparation tasks — two responsibilities that lie in the job description for data engineers. Because data scientists are spread thin in this way, both the quality of the data within an organization and the accuracy of insights drawn from the data suffer: “Poor data quality is the primary reason for 40% of all business initiatives failing to achieve their targeted benefits” (Kandel 2014).

In fact, the industry may currently be headed in exactly the wrong direction by seeking out more data scientists. I shared an interview with a data engineer for Instructure (interviewee name omitted) during which he answered some of my questions about data science. When I asked him what he be-
The Data Science Hierarchy of Needs

Learn/ Optimize
Aggregate/ Label
Explore/ Transform
Move/ Store
Collect

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

Research Scientist
Core Data Science Machine Learning Engineer

Data Scientist Analyst

Data Engineer

Software Developer

Figure 2: The hierarchy within data science teams.

Bagley: Landscape of the Data Industry
Published by BYU ScholarsArchive, 2020
A few companies have taken notice of this shift and reacted appropriately. Companies such as Google, IBM, and Apple have removed the requirement for a college degree from their job applications. The vice president of talent at IBM said that “instead of looking exclusively at candidates who went to college, IBM now looks at candidates who have hands-on experience via a coding boot camp or an industry-related vocational class” (Connley 2018). This is a good start, and changes like this will help bring more qualified workers into the market to fill the current demand.

Moving Forward

So, what can we learn from all of this, and how should we react moving forward? We can see that currently, despite the popularity of jobs within the data industry, the job title of “data scientist” is ill-defined, there is a shortage of universities offering adequate programs to meet the demand in the industry, and there is widespread imbalance in data teams with an especial need for more data engineers and fewer data scientists. As we as a community of workers and employers come to understand the landscape of the data industry more, there are several things that we can do to bring more efficiency and balance:

Companies need to define more clearly what they need the data scientists they hire to do. Upon closer look, they may find they need a few more data engineers to make it worthwhile to bring on a data scientist so that the data scientists can spend their time doing those high-level tasks that they do best.

Companies who are hiring should remove as many barriers to entry as possible, such as Ph.D.’s and lengthy qualification wish lists, and instead understand that the way people are acquiring the needed skills is changing. The more open-mindedly recruiters approach the hiring process, the better success they will have building the data teams they are envisioning.

More educational institutions should work to implement updated curriculums that will provide their students with the needed and highly-demanded skills to enter the ever-growing data industry. Proper weight should be placed on data engineering skills in comparison with data science skills within these curriculums to reflect the demand in the job market.

The world is changing at an excitingly rapid pace, and data science is a major part of that. The future of the data industry will see much more balance, as an increased number of people begin to implement the aforementioned actions. The better we refine our understanding of the data industry, and the sooner we take these actions to bring balance to it, the more incredible data science solutions we will see to important problems affecting each industry and the business community as a whole.
Notes


