The Role of Data Scientists in Modern Enterprises - Experience from Data Science Education

Thoralf Mildenberger
Zurich University of Applied Sciences, Winterthur, Switzerland

Martin Braschler
Zurich University of Applied Sciences, Winterthur, Switzerland

Andreas Ruckstuhl
Zurich University of Applied Sciences, Winterthur, Switzerland

Robert Vorburger
Zurich University of Applied Sciences, Wädenswil, Switzerland

Kurt Stockinger
Zurich University of Applied Sciences, Winterthur, Switzerland

ABSTRACT

"Data Scientist" has often been considered as the sexiest job of the 21st century. As a consequence, the spectrum of data science education programs has increased significantly in recent years, and there is a high demand for data scientists at many companies. However, what training is required to become a data scientist? What is the role of data scientists in current enterprises? Is the training well-aligned to the practical needs of a job?

In this article, we will address these questions by evaluating a survey of people who were trained in a continuing education program in data science in Switzerland. Our study sheds lights on the practical aspects of the data science education and how this newly-gained knowledge can successfully be applied in an enterprise. One of the highlights from the point of view of the database community is the important role of SQL in data science.

Keywords

Data Science, Data Science education, Data Science methods

1. INTRODUCTION

Data Science is an important new discipline with many applications in various business domains [3, 2]. Over the past few years, it has often been claimed that the profession of data scientist is especially attractive [1]. As the wider public has become aware of the potential of data for value-creation, there has been a concurrent increase in attention towards the field and its practitioners. Such attention should ideally translate into earnings potential, which explains why many new education programs branded as data science have recently been proposed [8]. The underlying curricula of these programs diverge widely in terms of the amount of technical vs. managerial skills covered.

The question that we want to tackle in this paper is how these curricula align with the real needs of the industry. Which of the technology trends discussed in academia, such as big data, machine learning, statistical analysis, natural language processing, artificial intelligence etc., are really taken up in practice? In this article, we focus on the data science education in Switzerland, with its strong university system.

Prior research on the skill set needed to become a data scientist or data analyst is comprehensively reviewed in [4]. Other empirical surveys [6, 7] were conducted in Norway and, in contrast to our work, focus on the managerial side, while the focus of our study is on data science graduates. Hence, our paper fills an important open gap in understanding the impact of data science education with respect to industry.

To answer the above-mentioned questions of how curricula align with industry needs, we conducted a survey among former and current participants of the continuing education program in data science[1] at Zurich University of Applied Sciences (ZHAW). Notably, this program is one of the oldest in continental Europe, being established in 2013. This gives us a uniquely broad spectrum of alumni to base our analysis on.

Currently, the overall program is structured into the following 5 different Certificates of Advanced

[1]https://www.zhaw.ch/de/engineering/weiterbildung/detail/kurs/mas-data-science/

Studies (CAS). For each of the course, we list the main topics and skills that are taught: (1)*Data Analysis*: R, descriptive statistics, introduction to probability and inferential statistics, linear regression, time series, clustering and classification. (2)*Information Engineering*: Python, databases and data warehousing, information retrieval and big data. (3) *Statistical Modelling*: Advanced topics in R, generalized linear models, time-to-event data, network analysis, graphical models and causality. (4) *Machine Intelligence*: Machine learning, deep learning, text analytics and advanced big data. (5) *Smart Service Engineering (Data Product Design)*: Smart service & data product design, data-specific business model design, praxis workshop, data protection and data security.

Each CAS runs for one semester, with one day of teaching per week, and is worth 12 ECTS points (European Credit Transfer System). For the successful completion of each module, participants are awarded a *Certificate of Advanced Studies* (CAS). A *Master of Advanced Studies* (MAS, 60 ECTS) is awarded after finishing four such modules along with an additional master's thesis.

2. EVALUATION METHOD

We conducted a survey[2] among participants of our continuing education program in data science. We sent out invitations for an online questionnaire to all former and current participants (over 1300 potential respondents). These ranged from participants who were, at the time of the study, currently undertaking their first course; to some who completed the full MAS Data Science Program with 60 ECTS several years ago. There were 159 respondents, of which 115 filled out the survey completely. The survey included multiple choice questions, which allowed respondents to give other answers apart from the listed ones; as well as some open questions, where participants were able to write a short text.

In Section 3, we collate and plot the responses to this survey and provide a descriptive analysis of these results. In some cases similar answers have been grouped together manually. The grouping were rather straightforward and uncontroversial, so no formal investigation into aspects of internal consistency was conducted.

3. RESULTS

This section discusses survey results.

[2]The survey was conducted using the open source survey tool LimeSurvey (www.limesurvey.org)

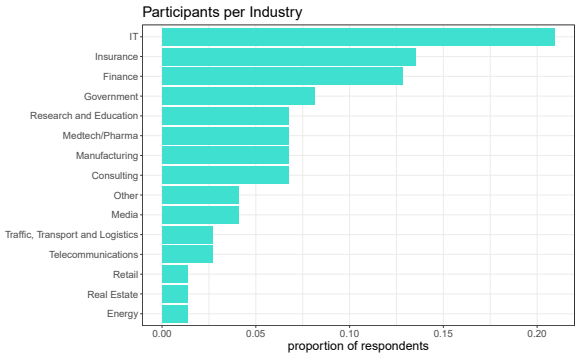

Figure 1: Industries the participants are employed in.

3.1 Demographics and Industries

Most participants are between 35 and 54 years old and the majority identifies as male. Nearly 80% hold a bachelor or master's as the highest academic degree and around 10% hold a PhD. Holding an academic degree is a formal requirement for enrolling, although this can be waived; as such, approximately 10% of respondents did not have a degree from an institution of higher learning.

Data Science is an academic field with a strongly interdisciplinary nature. Digitization affects nearly every line of business. This is reflected in Figure 1, which shows that our participants originate from a broad spectrum of industries. The most dominant sector is IT, followed by Finance and Insurance. In addition, Manufacturing, Medtech / Pharma, Consulting, Research and Education and Government each make up more than 5% of respondents. In a similar study from 2015 (see Chapter 4 in [9]), the consulting sector dominated, which is now in midfield in the current study. By contrast, the IT sector was rather weakly represented at that time.

3.2 Skills - Before and After

Participants from such widely different backgrounds start the program with very different preexisting skill sets. Whether or not they apply their newly-gained knowledge will also be highly dependent on the needs of their existing jobs or new positions. Figure 2 shows the skills participants had *before* enrolling and contrasts this with the skills used in their jobs *after* completing the program. The skills listed are the ones covered by the curriculum of our continuous education program.

Most participants were familiar with the mature concepts of databases and SQL beforehand, and the majority continued to use this skill set in their job after their education. The high demand for SQL

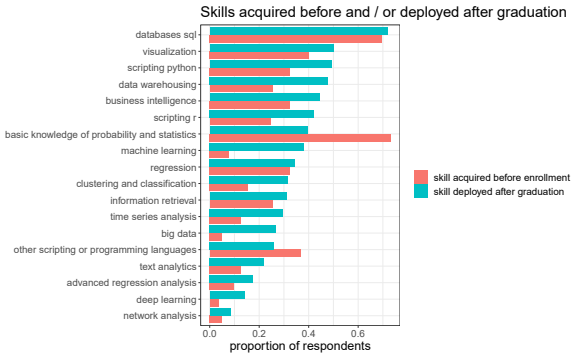

Figure 2: Data science skills acquired before enrolling in data science program (red) and skills used after graduating from the program.

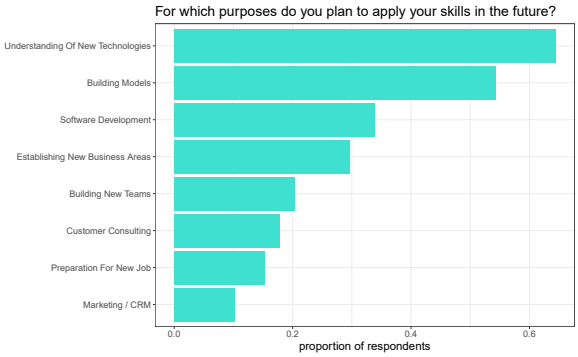

Figure 3: Participants were asked for which purposes they are likely to apply their newly acquired skills (multiple answers possible).

skills is especially encouraging for the database community as an important contributor for data science.

By contrast, machine learning and big data are skills widely used in practice, and are part of our curriculum; but they were unfamiliar to many participants before entering the program.

Our curriculum also includes some topics which are prominently associated with data science and widely discussed in academia; such as text analytics, network analysis and deep learning. Only a minority was familiar with these techniques before entering the program. However, somewhat surprisingly, these seem to be much less widely used in practice than expected. We consider this the first remarkable result of our analysis: *the hype level of topics in academia is not necessarily a good guide for building an industry-relevant data science curriculum.*

3.3 Effects on the Job

Figure 3 indicates the purposes to which partic-

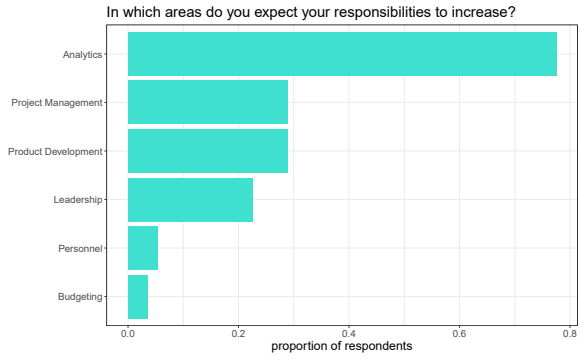

Figure 4: Participants were asked in which areas their responsibilities were likely to increase (multiple answers possible).

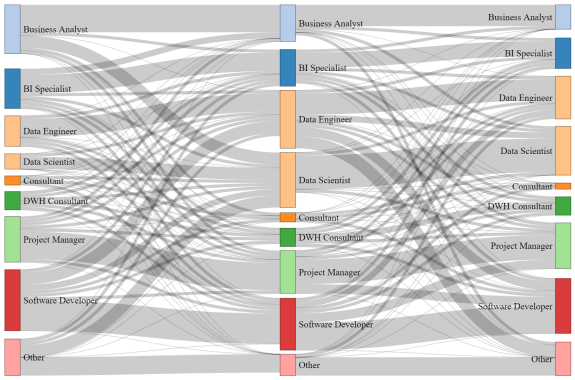

Figure 5: Change of job titles over three years. Left: Job title before education. Middle: Job title one year after education. Right: Job title three years after education.

ipants are planning to apply their newly acquired skills. The most prominent are the understanding of new technologies and building models. Even after formal education in data science, participants will inevitably meet new challenges on the job for which they did not receive formal training.

Figure 4 shows the areas in which participants believe their responsibilities will increase. The most prominent area is analytics, with nearly 80%, followed by project management and product development, with nearly 30% each. The large gap may, to some extent, be an artifact of the technical orientation of our program.

Figure 5 shows the change in job titles over a span of 3 years. The colored blocks represent the percentage distribution of job titles. The data representing three years after education has to be interpreted especially carefully, since many participants had graduated more recently or were still matriculated when

the survey was conducted. However, a clear shift to the job titles *Data Scientist* and *Data Engineer* is visible after one year. This switch is coming mostly from former *Business Analysts* and *Software Developers*. Participants that have responded with more specific, not predefined job titles, are covered under the *others* category.

Most of the participants who were *Data Scientists* after one year, remained *Data Scientists* after three years. However, more than half of the *Data Engineers* changed their job title again after three years to *Software Developers*, *Data Scientists*, or more individual job titles represented by *others*. Comparing these results with the 2016 study (see Chapter 4 in [9]), it is noticeable that certain job titles are hardly used today, such as *Data Miner*, but others are have become more widely accepted, such as *Data Engineers*. It may well be that other job titles will become established in this area and others will disappear (again).

Participants were also asked in open questions whether their employer benefited from these newly acquired skills, and in which way. While 10% stated there was no benefit (or no benefit yet), 39% gave concrete areas where their performance improved, and 46% named a deeper understanding of data science technologies and data analysis. (Missing from 100%: answers that could not be grouped under these three categories.)

4. DISCUSSION

Our survey showed some interesting and sometimes surprising relationships between the curriculum of our continuous education program and the job requirements in industry. It has to be noted, however, that the analysis was conducted over all participants combined. While we collected demographic data and information about the participants' industries, we did not include these parameters in the subsequent statistics. It would definitely be interesting to see if the discovered results would differ for some specific industries, or if big differences exist based on age or gender, but here we focus first on the big picture, as the relatively moderate number of participants in our cohort did not allow for detailed statistical analysis.

One particularly surprising result was the distribution of applied skills in industry *after* graduating from the program. As mentioned in Subsection 3.2, some of the skills in the curriculum prominently associated with data science, such as text analytics (also known as natural language processing) and deep learning, are, at least today, scarcely used afterwards on the job. An explanation for this might be the fact that it is often hard to directly include these skills in the existing work processes. There may be more potential for changing and improving work processes in the future. But as with any change, this requires substantial investment which is difficult to get from the management. In addition, it might simply be that the necessary training data for such methods are not available and the effort to generate it is too costly. Another reason could be that some industries are not ready yet for fully embarking on advanced data science technology yet - either because they lack Data Scientists or the industry sector still uses a fairly low amount of digitization.

A recent article [5] comes to similar conclusions, arguing that AI researchers should not be the first Data Scientists to be hired. The major reasons are that these young Data Scientists typically have no real-world experience of applying their skills, or companies do not have the right projects for them.

Another reason why some skills included in the program are only scarcely used might be given in Figure 3. It seems many participants were joining the program to learn data science methods with the goal of understanding new technologies better and being capable of seeing the potential of these technologies, rather than applying these methods immediately.

Combining Figure 3 and Figure 4 allows us to conclude that a majority of the participants have a strong interest in being capable of building models for analytical purposes in their job. The required skills for doing this are definitely covered by our curriculum. Especially *machine learning* with methods such as *clustering and classification* appear to be a strong driver here, as reflected by Figure 2.

5. CONCLUSIONS

In this work, we demonstrate the positive impact of our continuous education program in data science on its participants. This is reflected not only by the drift in job titles, but foremost by the newly-applied skills in their jobs, for example *machine learning* and *big data*. On the other hand, we discovered a discrepancy between academically prominent data science skills such as *deep learning* or *text analytics* and the application of these skills in industry.

Our analysis concludes that the role of data scientists is important in modern enterprises, and the demand for further education in the field, that we experience at ZHAW, remains high. In the interest of continuously improving the program, it is important to re-evaluate all parts of the curriculum.

REFERENCES

[1] S. Baškarada and A. Koronios. Unicorn data scientist: the rarest of breeds. *Program: electronic library and information systems, Vol. 51 No. 1*, 2017.

[2] M. Braschler, T. Stadelmann, and K. Stockinger. *Applied Data Science*. Springer, 2019.

[3] T. H. Davenport and D. Patil. Data scientist: The sexiest job of the 21st century. *Harvard business review*, 90(5):70–76, 2012.

[4] S. Klee, A. Janson, and J. M. Leimeister. How data analytics competencies can foster business value– a systematic review and way forward. *Information Systems Management*, 38(3):200–217, 2021.

[5] C. Kozyrkov. Why an ai researcher shouldn't be your first data science hire. 2022.

[6] P. Mikalef, M. N. Giannakos, I. O. Pappas, and J. Krogstie. The human side of big data: Understanding the skills of the data scientist in education and industry. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 503–512, 2018.

[7] P. Mikalef and J. Krogstie. Investigating the data science skill gap: An empirical analysis. In *2019 IEEE Global Engineering Education Conference (EDUCON)*, pages 1275–1284, 2019.

[8] D. Rzeznikiewicz. How to become a data scientist: A guide to the education, skills, and necessary experience. 2022.

[9] K. Stockinger, T. Stadelmann, and A. Ruckstuhl. Data Scientist als Beruf. In D. Fasel and A. Meier, editors, *Big Data: Grundlagen, Systeme und Nutzungspotenziale*, chapter 4, pages 59 – 81. Springer Wiesbaden, 2016. DOI:10.1007/978-3-658-11589-0.