On Understanding Data Scientists

Paula Pereira
University of Minho
Braga, Portugal
a77672@alunos.uminho.pt

Jácome Cunha
University of Minho & NOVA LINCS
Braga, Portugal
jacome@di.uminho.pt

João Fernandes
CISUC, University of Coimbra
Coimbra, Portugal
jpf@dei.uc.pt

***Abstract*—Data is everywhere and in everything we do. While, on its own, data has little value, its analysis under the lenses of data science currently supports extremely valuable functions and systems. In recent years, the need to obtain knowledge from huge amounts of data lead companies to hire many professionals able to work creatively on data for data science position, regardless of their academic and professional background. Our analysis of the collected information allowed us to conclude that there is still a need to distinguish professionals concerning their past experiences, since they impact the way they perform data science tasks and the technologies they use. The results of this research are particularly useful for the scientific community and the industry in the design of solutions that are effectively useful to all data science workers.**

***Index Terms*—Data Science, Data Scientists, Interviews.**

I. Introduction

Every day huge amounts of data are created and manipulated to glean insights and extract value [1]–[4]. This data comes from diverse sources and is used for all kinds of purposes in various sectors, for example, life and physical sciences, medicine and healthcare, education, retail, government, communication and media [5], [6]. According to the DOMO's annual report on how much data is generated on the most popular platforms, in the year of 2019, 188.000.000 emails and 18.100.000 messages were sent every minute, representing an increase of more 5.000.000 messages compared to 2018 [1]. In the aviation area, the development of new technologies has led to the existence of new, more sophisticated aircraft, with Forbes magazine reporting that, for every flight performed, five to eight terabytes of data are generated regarding the crew, the passengers, the condition of the engines, etc., i.e. 30 times more data than those generated by the previous generation aircraft [7]. In fact, this growth has had an impact in all different sectors and led the International Data Corporation to estimate that, by 2020, the worldwide data will increase up to 40 zettabytes [8], [9].

The emerging of this amount of data gave origin the the term *big data*. When it comes to big data, the size is not the most important aspect to consider. When dealing with data we have to take into consideration three important characteristics: *volume*, *variety*, and *velocity* [10]–[13]. This has been widely accepted as the "three V's of big data". Based on this model, big data involves a very large volume of data that is being created, stored, and analyzed exponentially faster than at any period in the history of mankind. This data can be in a variety of formats, such as structured (attributes in a database) or unstructured data (images, video footage, audio, handwritten notes).

By itself, data, even if in massive amounts, has little value [14]. Indeed, it is the information that is extracted from data that has the potential to keep changing and improving our lives [14]. Because of this, the need for people capable of gathering, cleaning and using data to extract knowledge has led to the consolidation of a field called *data science*. Data science can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems [14], [15]. Generally, it involves the application of quantitative and qualitative methods to solve relevant problems and its ultimate goal is to improve decision making [14], [16]. Because of that, a data science worker must have a diverse range of skills, and a great domain knowledge, in order to be able to work creatively with their data.

In the last decade, the interest around the field of data science has been overwhelming. With data science being called the *sexiest job of the 21st century* [16], [17], many people with all types of backgrounds have been trying to secure their place in this field. This interest has been accompanied by significant growth in the number of job offers, and in recent years it has been found that the job offer in this area exceeds demand [18]. When analyzing some of the offers for data science positions recently published on the *LinkedIn* platform, such as data scientist, data analyst, data engineering, etc., we see that the vast majority ideally pursue people with a background in Engineering, Computer Science, Mathematics, Statistics, Physics, and other related fields. The description of the desired skills can be either very vague and with little information about the tasks to be performed or super detailed. Usually, these positions require: experience in the use of data batch and streaming tools (e.g. Spark, AWS, Hadoop); understanding and experience in the use of high-performance machine learning algorithms and deep learning techniques; experience with programming languages such as Python, R, or Java; knowledge of relational and non-relational databases; and experience in the use of statistical techniques to analyze data and provide ongoing reports.

However, the exponential growth of data science, as well as the rapid and urgent need for people able to manipulate data creatively to their advantage, may have had a negative impact in the definition of its boundaries. Given this scenario, it is currently unclear the position of a data science worker, the set of skills that most value their work and the tasks

for which they are more suitable [19]. For this reason, we intend to place data science professionals in the focus of our study to understand how their past experiences affect their work, what difficulties they feel and which technologies they consider most appropriate in the tasks they perform. For this, we decided that the best way to collect information about these professionals was through semi-structured interviews. In these interviews, we discussed topics related to their academic background, the jobs they have, the tasks they perform and the difficulties they experience in their daily lives.

In what follows, we first review some similar works on data science and data science workers (Section II); and we present the research method (Section III). Then, we review the information collected and point some interesting facts on how the participants feel about their work (Section IV). Finally, we discuss some limitations (Section V) and conclusions (Section VI).

II. Related Work

Along with the growing interest in big data and data science, some similar works were published by authors who were trying to understand data scientists and how they work.

Harris et al. [20] describe the results of a survey on data scientists, their experiences and how they viewed their own skills and careers. This survey was particularly interesting because it was design by data scientists. They used the survey results to identify a new, more precise vocabulary for talking about data science work, based on how data scientists describe themselves and their skills. The authors also showed that tools are critical to data scientists' effectiveness. They also managed to distinguish several sub-groups of professionals. However, the study does not have in consideration the academic background of the participants, nor their preferences regarding the tools and techniques they use in their daily lives. We believe this information is paramount to better understand the different kinds of data scientists.

Miryung et al. [21] conducted 16 interviews with data scientists from eight different product organizations within Microsoft to understand their responsibilities, considering their education and training backgrounds, their missions in software engineering contexts, and the type of problems on which they worked. The authors then used the information to characterize the roles of data scientists in a large software company and to explore various working styles of data scientists, having identified five different styles (insight providers, modeling specialists, platform builders, polymaths, and team leaders). However, the results only considered interviewees working at Microsoft which do not allow to generalize the conclusions. We will interview workers in different companies and in different markets.

More recently, Muller et al. [22] also conducted several interviews with 21 data science professionals. These interviews allowed the authors to focus on the way data science workers work with their data. The authors found that they are involved in various steps of the process and perform tasks like data collection, data cleaning, data integration, and engineering features. They also showed that, often, data is not ready for analysis, and must be designed to meet the requirements of an algorithm. This study involved only IBM workers which again may limit the generalization of conclusions.

Zhang et al. [23] focused on the collaboration of data workers during the several steps of a data science workflow. To do so, they conducted an online survey with 183 participants who work in various aspects of data science and learned that data science teams are extremely collaborative and work with a variety of stakeholders and tools during a data science project. Similarly to [22], the results of these studies may be difficult to generalize considering that all of the respondents worked at IBM.

III. Methodology

To understand who are the data science workers of today, as well as the difficulties experienced by them and the technologies they use the most, we decided to conduct semi-structured interviews, as they enable a researcher to collect the interviewees' perceptions, thoughts, and attitudes [24], [25]. Besides, this type of interviews are more like conversations, allowing the interviewer to change the topic of the conversation according to the interview progress [25], [26]. In this work, the non-probabilistic sampling methods *convenience* and *snowball* were used, meaning that responses were obtained from those people who were available and willing to take part, and people they believe would be willing to take part [27].

Although this project is in an initial phase, several interviews with people who are conducting data science-related tasks have been possible so far. There were a total of eight interviewees: two data analysts, one business intelligence manager, one big data architect and four data scientists. This group is composed of three women and five men who live and work in Portugal, except for one case which lives and works in the UK. The interviews were conducted in Portuguese and lasted about 30 minutes. The interview with participant P3 was not considered in our analyzes since the participant did not leave space for us to ask the questions initially prepared. Table I summarizes the information about the interviewees.

During the interviews, we used a script that addressed several aspects related to: (1) the academic and professional experience of the participants; (2) the type of data they work with and the way they acquire it; (3) the data cleaning and pre-processing; (4) the application of mining techniques to extract knowledge from the available data; (5) and the tools and programming languages they use the most.

The scheduling process for these interviews was carried out by e-mail, and initially, all participants were informed of the goals of the conversation. Besides, a consent form was sent to each participant to authorize the recording of the interviews. All interviews were conducted by videoconference.

IV. Results

A. Academic and professional background

As expected, the participants have quite different academic backgrounds, with the most striking cases being the case of

TABLE I
INTERVIEWEES INFORMATION.

ID	Sex	Age	Job Title	Education Level	Domain
P1	F	30	Data Analyst	Master, Marketing	Music Manag.
P2	M	36	Business Intelligence Manager	Master, Data Analysis and Decision Support Systems	Retail
P3	M	37	Big Data Architect	Bachelor, Math and Computer Science	Software Dev.
P4	M	34	Data Scientist	PhD, Electrical and Computer Engineering	Telecom.
P5	M	42	Data Scientist	PhD, Data Mining	Virtual Call Center
P6	F	26	Data Analyst	Master, Mathematics and Computation	Web Dev.
P7	F	32	Data Scientist	Master, Mathematics Engineering	Virtual Call Center
P8	M	32	Data Scientist	PhD, Evolutionary Biology	Telecom.

participant P1, who graduated in Hotel Management, and the case of participant P2, who graduated in Economics. However, these two participants ended up beginning their careers in data science-related areas and, after a few years of experience, in order to evolve professionally, they felt the need to obtain skills that were more suitable to this area. Therefore, participant P1 is currently taking a second master's degree in Business Intelligence and Knowledge Management, and participant P2 began a master's degree in Data Analysis and Decision Support Systems, two years after having worked in auditing information systems, as himself stated[1]:

> *"I had been working in auditing information systems for two years, and at that time I decided that data was 'the thing' and I went to get a master's degree in Data Analysis and Decision Support Systems."* – P2

Regarding the tasks that these professionals perform in their day-to-day lives, we realize that, in this group of people, only those who have training in computer science or engineering do tasks related to the creation of machine learning and deep learning models, as is the case of participants P4 and P5. The remaining participants dedicate themselves to more direct analysis, based on statistical parameters such as average, standard deviation, distributions, etc., which ends up fitting more into the profile of a mathematician or statistician. A similar idea was slightly mentioned in [20], [21], with the authors saying that data scientists from fields such as business and social sciences have strong numerical reasoning skills and are particularly good in using advanced statistical techniques in their analysis.

[1]The quotes represent our best translation of the Portuguese statements.

B. Data sources and data types

Regarding the data sources used by this group of professionals, and similarly to what is mentioned in [22], there is a great variety. In addition to the data generated internally by the various teams of the organizations where they work, the use of public data sources is also frequent. Also, the manipulated data is quite different, ranging from customer data, to operational data.

One aspect that several participants mentioned was the lack of metrics that would enable the quality of the data to be assessed beforehand. Participant P2 was the only participant who reported using any form of data quality metrics. In this case, the participant points that the existence of specific metrics is possible because there is a great knowledge about the data, and that if some drastic changes happens, they end up being easily detected.

C. Data cleaning process

In terms of the data pre-processing in order to increase data quality and improve the analysis of results, the majority of the participants agree that this remains one of the most time-consuming and laborious tasks, which goes along with the findings in [21], [22].

> *"In terms of time, I would say that I spend 80% of the time cleaning data and only 20% analyzing it."* – P1

The only exceptions were participant P5 and participant P7, since in such cases the participants work with audio files, and the pre-processing only involves converting all files to the same format.

> *"The pre-processing of our data is not difficult. The only thing we do is to convert all files to the same format."* – P5

In the remaining cases, participants claim that most of the anomalies that affect the quality of the data concern inputs that have been wrongly introduced by humans. Among the most common errors are duplicated records, missing values, inconsistencies in values (e.g. date formats) and outliers.

D. Data mining process

In terms of the techniques most used, participants use a wide variety according to the solutions that are intended to be built, ranging from machine learning techniques, such as clustering, prediction, and classification, to more complex deep learning models. These techniques are usually applied with two main and distinct goals: client segmentation and service customization; cost reduction and optimization of internal processes. In [21], were made some similar observations.

> *"We have projects with a strong emphasis on the customer and the user experience, but the main focus is on internal projects that aim to enhance the quality of our processes and minimize costs."* – P4

As far as data analysis concerns, all participants report that they do not follow any work methodology. Instead, everyone agrees that the best way to work with data is to adapt to the situations at hand. Even so, all participants agree that

the analysis of the data, be it statistical or through machine learning models, does not dispense a great knowledge of the data itself and the business in which the problem is inserted.

E. Tools and programming languages

Concerning programming languages, we find that in this group of participants there is a great tendency in the use of programming languages such as R and Python, as well as all of the state-of-the-art packages that the two languages provide both for data visualization and the data analysis. The choice of which language to use is made according to personal preferences and the type of tasks to be performed. In the case of participants P5 and P7, this choice was made as a team so that all elements of the same project used the same technologies.

Regarding the use of data analysis tools, most participants stated that they do not usually use them in their day-to-day lives since these tools end up limiting their analysis, which does not happen when they produce all the code they need in the data analysis. Even so, participant P1 and the participant P2 reported that a large part of their analysis is done using only MS Excell since this software allows for very fast results.

F. General difficulties

When we asked the participants what were the biggest difficulties they felt regarding their work, several situations were mentioned. For participant P1, she says that most of the difficulties she feels are related to the fact that she has no training in the field of data science, and that was the reason that led her to pursue a master's degree focused on data analysis. For participant P4, the biggest difficulty is the access to information with quality and relevant to the problems in which he works,a difficulty that is shared with some of the participants in the study referred to in [22].

> *"I believe that access to quality information and information relevant to our problems is the greatest challenge."* – P4

For the participant P6, the most challenging part of her job is that she is the only person on her team to perform these types of tasks. She also notes that she often finds it very difficult to convert business problems into data science issues. In fact, in [21] the lack of clear problems and questions is also mentioned, as the authors argue that the academic background of a data scientist may have a significant impact on the way they identify important questions.

> *"In my case, being alone is a big limitation, . . . and initially, it is very difficult to have the required business expertise to understand what are its needs."* – P6

Participant P8, on the other hand, says that his greatest challenge, after several years of experience, is the development of stable and scalable code, since he has no training in software engineering.

> *"On a personal level, I think my biggest challenge is to write a stable and scalable code because my training is not very oriented for software engineering."* – P8

In these conversations, some of the participants mentioned that there are also difficulties associated with professionals being hired for data science positions that, in reality, should be occupied by other type of professionals. This can be one of the causes that leads to a considerable overlapping amongst several data science roles as stated in [23].

> *"Companies look at the market and, because there is a demand for data scientists, they also want to hire one. However, looking at the job's requirements, their needs would be easily mitigated by other types of professionals."* – P7

Participant P2 pointed out that, in his opinion, it is important to clarify which are the different areas of data science, so that the professionals who wish to work in this field can position themselves correctly in this environment and understand what are the opportunities that meet their aspirations.

> *"In my opinion, there are two main areas: technological and application data science. The data scientist of the future must know how to put himself in the right area of data science to avoid regretting what (s)he is doing."* – P2

In general, all participants agree that it is a great advantage to have people with different backgrounds in data science teams because, although some are better suited to certain tasks than others, they all bring different perspectives on the data.

V. Limitations

As already mentioned, this is an early-stage project. In this sense, the number of interviews carried out is not yet enough to allow us to draw major conclusions about professionals in this field. Another aspect to be taken into account is that the interviews, although carried out with professionals from several different Portuguese companies, end up reflecting a part of the reality of what is like to be a data science worker in Portugal, which means that the worldwide panorama may be different.

VI. Conclusions

With this study, we intended to investigate whether the rapid evolution of data science, as well as the growing demand for people capable of taking advantage of information hidden in large amounts of data, has had an impact on professionals in this field.

Assuming that there is a heterogeneity of data science professionals, as a result of their distinct academic backgrounds and professional experiences, we decided to conduct several interviews with people who currently work in positions dedicated to the analysis of large amounts of data. With these interviews we tried to explore several aspects related to their academic background, the jobs they have, the tasks they perform and the difficulties they experience in their daily lives. Although the analysis of these conversations has allowed us to reach a few initial conclusions about these professionals,

we consider that the reduced number of interviews carried out does not allow us to generalize the conclusions. For this reason, we intend to continue conducting interviews with data science professionals and collecting their feedback. Besides, these interviews should support a large-scale study through the distribution of a public questionnaire with the same goals as the interviews conducted so far. This questionnaire should allow us to reach a much higher number of data science professionals and, more importantly, to collect information from people in very different professional and cultural contexts.

The main motivation for this research is the need to understand what are the difficulties that affect these professionals' work and what should be the focus of both the scientific community and the industry in the design of solutions that are effectively useful to them.

REFERENCES

[1] J. James, "Data Never Sleeps 7.0," 2019. [Online]. Available: https://www.domo.com/learn/data-never-sleeps-7

[2] A. Holst, "Data created worldwide 2010-2025 — Statista," 2019. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/

[3] D. Parkins, "Regulating the internet giants: The world's most valuable resource is no longer oil, but data," *Economist (United Kingdom)*, vol. 413, no. 9035, 2017. [Online]. Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[4] M. Kubina, M. Varmus, and I. Kubinova, "Use of Big Data for Competitive Advantage of Company," *Procedia Economics and Finance*, vol. 26, pp. 561–565, 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212567115009557

[5] J. Manyika, C. Michael, B. Brad, B. Jacques, D. Richard, R. Charles, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., 2011. [Online]. Available: www.mckinsey.com/mgi.

[6] D. Zhang, "Granularities and inconsistencies in big data analysis," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, no. 6, pp. 887–893, 2013.

[7] O. Wyman, "The Data Science Revolution That's Transforming Aviation," 2017. [Online]. Available: https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/

[8] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 10 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046413001007

[9] V. Turner, J. F. Gantz, D. Reinsel, and S. Minton, "The Digital Universe of Opportunities: Rich Data and Increasing Value of the Internet of Things," *IDC White Paper*, no. April, pp. 1–5, 2014. [Online]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[15] W. van der Aalst, *Process Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-662-49851-4

[10] C. Austin and F. Kusumoto, "The application of Big Data in medicine: current implications and future directions," *Journal of Interventional Cardiac Electrophysiology*, vol. 47, no. 1, pp. 51–59, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10840-016-0104-y

[11] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 4 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0268401214001066

[12] D. Laney, "Application Delivery Strategies from META Group," *META Delta*, vol. 949, no. February 2001, p. 4, 2001.

[13] J. Wieczorkowski and P. Polak, "Big data: Three-aspect approach," *Online Journal of Applied Knowledge Management*, vol. 2, no. 2, 2014.

[14] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, 3 2013.

[16] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/

[17] T. H. Davenport and D. J. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Business Review*, vol. 90, no. 10, p. 5, 2012.

[18] T. S. Perry, "Demand and Salaries for Data Scientists Continue to Climb," 2019. [Online]. Available: https://spectrum.ieee.org/view-from-the-valley/at-work/tech-careers/demand-and-salaries-for-data-scientists-continue-to-climb

[19] V. N. R. W. R. F. Chatfield, Akemi Takeoka; Shlemoon, "Data Scientists as Game Changers in Big Data Environments," in *Proceedings of the 25th Australasian Conference on Information Systems, 8th - 10th December, Auckland, New Zealand*, 2014, pp. 1–11.

[20] H. Harris, S. Murphy, and M. Vaisman, *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, 1st ed., M. Loukides, Ed. O'Reilly Media, Inc., 2013.

[21] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "The emerging role of data scientists on software development teams," *Proceedings - International Conference on Software Engineering*, vol. 14-22-May-, pp. 96–107, 2016.

[22] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. Vera Liao, C. Dugan, and T. Erickson, "How data science workers work with data," in *Conference on Human Factors in Computing Systems - Proceedings*. ACM, 2019. [Online]. Available: https://doi.org/10.1145/3290605.3300356

[23] A. X. Zhang, M. Muller, and D. Wang, "How do Data Science Workers Collaborate? Roles, Workflows, and Tools," 1 2020. [Online]. Available: http://arxiv.org/abs/2001.06684

[24] J. Rowley, "Conducting research interviews," *Management Research Review*, vol. 35, no. 3/4, pp. 260–271, 3 2012.

[25] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*, 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. [Online]. Available: http://link.springer.com/10.1007/978-3-642-29044-2

[26] B. L. Leech, "Asking questions: Techniques for semistructured interviews," *PS - Political Science and Politics*, vol. 35, no. 4, pp. 665–668, 2002.

[27] B. A. Kitchenham and S. L. Pfleeger, "Principles of Survey Research Part 5: Populations and Samples," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 5, pp. 17–20, 2002.