Data Scientist: A Systematic Review of the Literature

Marcos Antonio Espinoza Mina[1,2(✉)]
and Doris Del Pilar Gallegos Barzola[3]

[1] Universidad Ecotec, Samborondón, Ecuador
mespinoza@ecotec.edu.ec
[2] Universidad Agraria del Ecuador, Guayaquil, Ecuador
mespinoza@uagraria.edu.ec
[3] MADO S.A., Guayaquil, Ecuador
doris@ecuaportales.com

Abstract. The commercial activities of services and production have accumulated plenty of data throughout the years, hence today's necessity of a professional agent to interpret data, generates information in order to produce valuable results and conclusions. The scope of the current article is to present a systematic review of the literature which main goal was to spot the work and career profile of the so called Data Scientist; realizing that, as a new work field, there are not concretely defined profiles, although knowledge areas are indeed defined, as well as characteristics that are needed to be counted, apart from some technologies that can serve as supporting means for the labor these new technicians do in the IT (Information Technology) area.

Keywords: Data scientist · Work profile · Career profile

1 Introduction

The Data Scientist turns into the type of professional that when facing huge data bases - which in major part have no specific goal - applies their knowledge in numerical areas such as: programming, math and statistics, in order to compile, extract and process the contained relevant information; to then explore and analyze multiple source data, too great in occasions, known as big data, with too different formats at times.

Moreover, the Data Scientist must have a strong business vision to enable them to extract and transmit recommendations to the representatives of a company. On top of that, this professional should provide useful insights and decision-making support; the professionals must be capable of understanding the business issues and framing the appropriate analytical solutions. The necessary business knowledge professionals ranges from general familiarity with the areas of accounting, finance, management, marketing, logistics, and operation management, to the domain knowledge required in specific [1].

Exceptional communication skills are required by a Data Scientist due to the necessity of explaining the process involved in the data analysis in terms of working precision, problem solving and experience on analysis and data cross- verification like

M. Botto-Tobar et al. (Eds.): CITT 2018, CCIS 895, pp. 476–487, 2019.
https://doi.org/10.1007/978-3-030-05532-5_35

SQL, among others. Their work is minimally supervised, for which it is compulsory to have planning and organizing skills, collaborative approach to share ideas and find solutions.

It is known, that the data set may be originated from any type of device and electronic technology media such as: cell phones, social media, medical data, web sites, among others. Data that, significantly affects the current research processes in many fields like: neuro-marketing, biologic science, health, social science and counting.

Activities related to: big data, data mining, machine learning and business intelligence have been studied by many researchers, however, few have been the works carried out that consolidate technical know-how and expertise on professionals who handle and use these tools. Therefore, not much is found on their professional profile and hardly anything on their working profile; or systematic reviews on related works of such profiles. This article seeks to know the career or work profile of a Data Scientist. Jaramillo [2], it indicates the differences between work and career profile. The career profile is delimited by a set of skills and knowledge that determine the professional practice and satisfy the work environment demands. For this reason, their tuition must be a fusion of a set of knowledge with the development of capacities and skills related to the area so that they provide a solid base, both theoretic and practical which allow their application to distinctive domains. The work profile corresponds to both, the charge-function-responsibility as well as the attitude, abilities and skill components required for the professional performance. Consequently, the work profile is made in parts: academic formation, working experience, and charge competences, since such profiles are made by business needs, and these respectively are fundamental in the eligibility of the professional individual; starting from the provable skills to the achieved goals.

2 Limitations of the Current Literature

Regardless of having identified some performance areas for the Data Scientist with certain profiles for such professional, the validity of each conclusion reached by the authors of this field's literature are difficult to evaluate, for being too many the fields in which it is currently performing. For instance, specific [3], unfortunately, data scientist with the analytical and software engineering skills to analyze these large data sets have been hard to come by; only recently have software companies started to develop competencies in software-oriented data analytics.

Additionally, [4] implies as suggested in the literature, skills requirements re-search should be conducted periodically to help information systems educators in redesigning curricula or courses that would better prepare future Data Scientists.

The goal of this study was to have a systematic review of the updated literature in which aspects of the Data Scientist profile are examined, for which there was an attempt to borrow from other works developed in this subject, but until the date when the information survey was done of the present study, other systematic reviews were not found that specifically examine the abilities and competences related to the professional and working profile of the Data Scientist.

3 Systematic Review

3.1 Generality

This section describes the followed process to make the systematic literature review using the reliable methodology of Kitchenham [5]. The basic goal of this reviewing system is to compile and evaluate all the investigations available related to an intended inquiry, wherefore achieving impartial, auditable and repeatable results. The systematic review presented by Kitchenham is aligned to the PICO process (Problem, Intervention, Comparison and Outcome), this framework was developed to facilitate the formulation of clinical queries [6]. Hence, it was configured to get a more efficiently scope, that allows getting to know the Data Scientist profile.

The execution of the systematic review allows identifying, evaluating and analyzing all the signifying studies concerning a research enquiry, theme area or interest phenomena. Some of the advantages provided by a systematic review are: the summary evidences about a specific technology and the existing gap identification in the current investigation which, likewise, allows establishing future investigations. The systematic reviews follow a defined research strategy which analysis unity makes part of the original preliminary studies, from which are meant to answer an investigation query clearly formulated through a systematic and explicit process. The systematic reviews synthesize the results of preliminary investigations through strategies that limit the bias and random error.

3.2 Definition of the Research Question

The characteristics and requirements that a Data Scientist must comply with, and are expected to be known, will be obtained through the formulated query: Is there a defined work and career profile for a Data Scientist?

As the key words used were: data scientist, work profile, career profile and technology.

The presented investigation query reported the following results:

- Which are the knowledge areas a Data Scientist must have access to?
- Which work and career characteristics must a data scientist count with?
- Which technological tools support a Data Scientist job?

3.3 Sources Selection

In this section, the query sources in which queries of the studies regarding the theme were made will be identified. The following pointers were considered:

- Digital versions of articles, magazines and documented conferences were queried about Data Scientist using the established keywords.
- The chosen bibliographic sources had to count with a search engine that allows executing advanced search queries.
- The studies had to be written in English.

With the selection criteria exposed, a study query was made in the digital research libraries: EBSCO host (EBS), IEEE Xplore (IEEEX) and ERIC. The approach was made based on the vast access and coverage in areas like technology, social and human sciences besides accessibility to complete revised works makes it an easy task.

The digital sources present more relevant investigation works, however it was added as bibliographic references to the web sites of companies and organizations that offer software in general to help in a Data Scientist job.

It is important to highlight that there were also made Science Direct and Springer data base searches, but such did not generate acceptable results previously defined in the search strategies; the few related articles in Science Direct required payment to gain total access to their content, and in Springer the articles were not found in English, and the book chapters were excluded.

3.4 Search Strategies

- A series of terms and key words were chosen to reply the query: is there a defined work and career profile for a Data Scientist? And therefore, obtaining the expected results.
- The searching strategy was based on the words "data scientist profile" and "data scientist".
- The structured research chain was: "data scientist" OR "data scientist profile".
- The chains were applied to: title and summary. When the summary matched with the research object, the article was obtained and reviewed in its totality.
- The research language used for the publications was English.
- Publications temporality, since 2014. It allows getting to know the current needs of the labor market that presents a more targeted training and entrepreneurial vision, avoiding outdated information.

3.5 Inclusion and Exclusion Criteria

Defined inclusion criteria:

- Articles published since 2014. It was made this way, since higher education standards are in increasing demand; organizations and markets change quickly in short periods of time, the charge pro-files keep evolving; for which it is needed people capable of acquiring tuition and continuously learning. For this reason, any published article before this year could be considered obsolete.
- Conferences, magazines and international workshops articles.
- When an article is duplicated in the digital libraries only one is selected.

Exclusion criteria:

- Articles which content are not related to a Data Scientist or their profile.
- Works content in slides or books.
- Published works outside the specified range.
- Grey literature.

3.6 Extraction of Information and Review of Works

This step allowed identifying relevant documents related to the objectives of a systematic review and to the reach of the investigation query. The main difficulty to achieve this objective was that the terms used in the query led to excessively wide results. For instance, the word "data" is widely used in many types of publications, and therefore a big quantity of files appeared in the first results obtained using the defined research chain.

For the extraction and reviewing processes of the investigation, were chosen those works that their quality was assumed to be granted by the evaluation made by the same bibliographic sources where they were obtained, because the platforms generate their results ranked by relevance. For the digest, only the works related to the query are taken into consideration.

Most documents were obtained through the completion of query based on the research chain in "any field" or "complete text". In this first phase, a big quantity of investigation documents were obtained, however such results were not the most pertinent because they consisted on comments, letters and duplicated works that not only were limited to query the Data Scientist profiles, but they expanded in contents far from the objective, see Table 1.

The database keywords used were: "data scientist" and "profile data scientist"; together with search strings: "data scientist" in fields "title" and "summary" and "profile of data scientist" in field "Full Text".

The second phase, based in the query chain in fields selected by "title" and by "summary", allowed eliminating some useless results, but that was not enough to satisfy the investigation, see Table 1. Finally, to obtain the definite list of preliminary studies, a chain query based review was made with fields selected by "title" and "summary", as well as selecting only the academic articles, see Table 2.

Table 1. Chain and additional options to queries by source.

Phase	Inclusion criteria	Found articles		
		EBS	IEEEX	ERIC
Based on query chain in any field		1543	722	16
Based on query chain in selected fields	● Data scientist in options "title" and "summary" and "profile of data scientist" in the option "any field"	197	31	16
Based on query chain in selected fields	● All the query terms with full text ● Additional filter within academic publications	13	31	13

4 Results of the Review

Selected studies and mainly reviewed are displayed in Table 2, with data that allows a quick information comparison among each other. They are presented in alphabetic order by title and it does not determine the importance level regarding the objectives of this work.

Table 2. Summary of articles with the knowledge area detail, characteristics and tools that support a data scientist.

Work	Approximation of knowledge areas for a DS	Characteristics of a DS	Supporting technological tools of a DS	DS requiring area
[7]	Scientific, investigative, IT	Leaders, communicative, Thinking skills	Data base	Academic
[8]	Basic math, statistics, communication, IT, data base	Researchers, communicative	Data base	Journalism, IT
[9]	IT, research		Data base IT systems	IT
[10]	Educative, scientific, economic and IT	Researchers, Data interpretation, Communicative	Data base, teaching method	Academic
[11]	Statistics, math, data analysis, scientific knowledge	Statistic, programmer, info graphic designer and narrator	Data base	IT
[12]	Statistics, data mining and engineering skills	Domain experience, knowhow, skills and attributes	Data base	Labor, IT and Academic
[13]	Biology, IT	Developer, researcher	Software, Data base	Medicine
[14]	Computer science, matrixes, calculus		Bi- dimensional matrix algorithm, storage technique	IT
[15	Scientific, IT, data base		GIDNA and CDM	IT and Data base
[16]	Biology, IT		Data base	Medicine
[17]	Math, probability, statistics, predictive analysis, uncertainty model, Information sciences			Medicine
[18]	IT, Scientific		Overflow, cloud applications, data crowd sourcing	IT
[19]	Educative, IT, data base	To know the business, science and technologic skills, communication skills	Software Hadoopecosystem	Academic
[20]	Educative, investigation, IT and data base	Administrator, quantitative researcher		Academic
[21]	Programming, statistics	Leader, decision-making	R Framework, java,.net	IT

(*continued*)

Table 2. (*continued*)

Work	Approximation of knowledge areas for a DS	Characteristics of a DS	Supporting technological tools of a DS	DS requiring area
[22]	Statistics, IT	Information analysis	PINQ (McSherry, SIGMOD 2009)	Statistics, math
[23]	Statistics, IT	Information analysis	Data base, applications, technologic tools	Statistics and IT
[24]	IT, Data base	Sharing knowledge	Data base	IT
[25]	IT research	Data consultant, researcher	Data base, Data library	Academic
[26]	Educative, IT and Data base	Formative	Data base	Academic
[27]	Data engineering, statistics, scientist, communicator	Domain expert, team captain	Data base	Academic
[28]	Statistics, business world, IT	Data segmentation, problem forecasting, development	Software TRANE, web site KAGGLE	IT

In the selected articles it was found that the Data Scientist must possess knowledge in many fields weighting more on the IT area, which would allow them to perform better with all the stored information in the data base, so they can use it orderly with quick access. Then, exact sciences knowledge is highlighted as of math and statistics, which are needed for formulation of hypothesis, using a suitable methodology and knowledge systematization. Additionally, a Data Scientist must know about communication and engineering in general. The necessity of this knowledge can be illustrated after reading Treadwell, Ross, Lee and Lowenstein article [8], where it is indicated that in the journalistic newsroom their personnel needs the skills in math and statistics as well as software tools inclusion, programming language like Python, Ruby, PHP or Perl to discover new information.

Zhai, Jocz, and Tan [7] point out that teachers with a positive attitude -from moderate to strong- towards implementation of scientific research teachings, create a pedagogic focus that deepens the learning of scientific concepts, which improves the self-regulation and develops thought skills of higher order that helps forging better leaders; a most required characteristic of a Data Scientist. Communicative and researcher are other of their very important ones.

Many technologies, among the algorithms and presented models, have been developed within the chosen articles that seek to support the data scientist work. Big data bases are studied as well as diverse programs for in-formation extraction; those are the already created software packages or new algorithms, developed by professionals in each area, as it is the case of Trane which was proposed by Schreck and Veeramachaneni [28]. That can be applied to any set of variables in time. Users can connect their data and wait for the software to synthesize forecasting problems. It also interacts

with general usage tools like Hadoop, which is a framework open source for the saving and processing big quantities of any type of data fast. Additional help models are presented such as G-DINA (Generalized DINA Model Frame-work) and CDM (Cognitive Diagnosis Modeling). It is additionally evidenced in the review that definitively data bases play an important role because they possess large saved information and therefore this generates the function of a Data Scientist.

It was found that many of the studies are oriented to give support to the academic area, because it is an enormous interest for educational institutions accurately determining the career profile of a Data Scientist, making even its projection and necessity that of fields different from IT, such is the case of nursery, journalism and agronomy. For instance, as Gold et al. indicate [10] students, such as those of agronomy can examine data to explain why the snow depth quickly diminishes in the early summer.

The performance of the Data Scientist is support by a variety of software and applications which make their job trustful and effective. Below, there is a review of five of them, being the last three, free of charge:

- DataRobot, is a platform that allows the production of any model with only a few code lines, making the validation of the model for every modeled technique, scientifically selecting the hyper parameters and options that optimize the performance of their models; this technology is prized [29].
- SubjectiveSystems, it permits to turn the data in the maximum value of the results, helping in everything, from optimizing the way of data acquisition, to designing and implementing panels, integrations and applications based on scalable data. To gain access to this technology it must be paid [30].
- GraphLab-Create 2.1 It allows the developers and Data Scientist apply automatic learning to build products from last generation data. This is a cost free tool, ideal for students [31].
- IPython, is a tool for parallel and interactive IT that is widely used in scientific IT. It is a cost free software [32].
- KNIME Analytics Platform, an open solution that leads to innovation based on data, helping discovering the hidden potential in data, quickly to implement, easy to escalate and intuitive learning. It has more than 2000 modules, hundreds of ready to execute examples, a wide range of integrated tools and the widest variety of available advanced algorithms [33].

5 Discussion

At the beginning, setting for search results was just for studies and no article with this type of systematic review and scope was found. But, after determining there was a shortage in works that make systematic review and orient the establishment on the profile of the Data Scientist, a mapping study was made being the following discussion content about its results.

It was defined that the main objective was to identify if the work and career profile defined for a Data Scientist exists, and after checking the preliminary articles it was found that eight of them, show how many educative institutions and higher education

centers are involved in the process to define a study plan adequate for this profession, because even without a third level degree anyone can do activities related to this work. Some academic institutions are opting for identifying and evaluating the work profile requested by private and state companies to define their study plan. Educating and training Data Scientist specialists requires a new model, which reflects by design the whole lifecycle of data, and is aligned by construction with the target research and industry domains context and technologies [12].

It is pointed out in the majority of the articles that the IT, math and statistics areas are the academic bases which the Data Scientist must count with; for this the business area must be added because they must understand about administration, projects, budgets, among other fields, which turn them into multidisciplinary professionals, a role not easily assumed by any human being. Data Scientist, bridges the gap between the programming and implementation of data science, the theory of data science, and the business implications of data. They can take a business problem and translate it to a data question, create predictive models to answer the question, and tell a story about the findings [34].

The characteristics and skills that a Data Scientist must possess are varied, predominating leadership, research and communication, which are the key for the development of their activity, being commercially or industrially. They must have an open and creative mind, being organized since they need to interact with different type of people to comply with assigned activities, dealing some days with operatives and other with directive and management to make them know the information found after analyzing great data bases, which generates great value and competitive advantages.

In the review it is intended to find out about the tools that give support to the Data Scientist and several new algorithms were found, specifically created for a necessity. In other cases, exhausting analysis is made of software tools results already developed and of general use. The most remarkable aspect in the investigation reviewed is the studies about the diverse data base proposed and the way in which the information is extracted. Many Data Scientists compete even on Kaggle, which is a platform for the forecasting modeled and analysis competences, to find better model to present in a logical manner this big quantity of information imbedded in them.

The Forbes magazine [35] points out that Data Scientists lead the group of the best Jobs in the US. This is due to companies in need of someone capable of analyzing dirty dataset, and applies simple statistic tools to ungrounded patterns. Many more technology companies are compiling copious quantity of data, but few managers or executives are capacitated in the IT code needed to compile all that in a report. This provides an advantage to Data Scientist in a world that appeals more to data for decision making.

6 Conclusions and Further Research

The objective of this paper is to make a systematic review of the mapping of the Data Scientist profile and it was made following the Kitchenham [5] model. After running the literature review, the first conclusion points out the existence of a tendency of few articles published where the work profile and the career profile of a Data Scientist is

established. Universities and education centers are seeking to make an adequate study plan, since (DS) it is anew profession, an academic program, adapted and updated to the new social circumstances, has not been established yet, in such a way that tuition level does not deteriorate, many of the reviewed articles have enabled clear the air terms in this matter.

In the evaluated articles it is rendered clearly that a Data Scientist must have solid knowledge in IT, math, and statistics. However, when it comes to business and administration knowhow, it is not defined how far the knowledge should be since, private industry and companies represent the higher demand for these type of professionals, it is expected being literate in finance, investment, production or human recourses.

The personal traits of a Data Scientist are very well defined, such as leadership, communication savvy and research, in order to achieve their goals in the data analysis and being able to explicitly communicate results; they must have many skills for accurate decision making and to forecast any type of problems.

Currently there is a lot of technology available to the Data Scientist, even algorithms are still being developed to ease up their job, which can be well applied if such an individual complies with all the appointed requirements; it is the only way they have to enclose the identification processes, capture, processing, analysis and data visualization.

Although the comparative analysis shows that there are studies in which knowledge areas of the Data Scientist are made known, as well as their abilities. No studies were found that emphasize about their work or career profile. This way it is shown that achieving a profile for this profession is still to be realized, to avoid premature obsolesce, with which it is expected to define the academic guidelines required for a groundbreaking profession, and as seen by many considered a new staple profession.

In the future it is expected to be able to execute an investigation that allows realizing an adequate division of the professional field of the Data Scientist, determining their participation levels in the labor market in order to define adequate profiles. Additionally, it is intended to carry on an investigation and a systematic review of the literature that includes more digital libraries.

References

1. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
2. Jaramillo, O.: Pertinencia del perfil de los profesionales de la información con las demandas del mercado laboral. Revista Interamericana de Bibliotecología. **38** (2015). https://doi.org/10.17533/udea.rib.v38n2a03
3. Kim, M., Zimmermann, T., DeLine, R., Begel, A.: The emerging role of data scientists on software development teams, pp. 96–107. ACM Press (2016). https://doi.org/10.1145/2884781.2884783. http://dl.acm.org/citation.cfm?doid=2884781.2884783
4. Ecleo, J.J., Galido, A.: Surveying LinkedIn profiles of data scientists: the case of the Philippines. Procedia Comput. Sci. **124**, 53–60 (2017). https://doi.org/10.1016/j.procs.2017.12.129

5. Kitchenham, B.: Procedures for performing systematic reviews. **33** (2004)
6. Huang, X., Lin, J.: Evaluation of PICO as a knowledge representation for clinical questions: In: Proceeding of the Annual Symposium oh the American Medical Informatics Association. AMIA Press (2006). http://users.umiacs.umd.edu/~jimmylin/publications/Huang_etal_AMIA2006.pdf
7. Zhai, J., Jocz, J.A., Tan, A.-L.: 'Am I Like a Scientist?': primary children's images of doing science in school. Int. J. Sci. Educ. **36**, 553–576 (2014). https://doi.org/10.1080/09500693.2013.791958
8. Treadwell, G., Ross, T., Lee, A., Lowenstein, J.K.: A numbers game: two case studies in teaching data journalism. Journal. Mass Commun. Educ. **71**, 297–308 (2016). https://doi.org/10.1177/1077695816665215
9. Younge, A.J.: Architectural principles and experimentation of distributed high performance virtual clusters. **24** (2017)
10. Gold, A.U., et al.: Arctic climate connections curriculum: a model for bringing authentic data into the classroom. J. Geosci. Educ. **63**, 185–197 (2015). https://doi.org/10.5408/14-030.1
11. Fuller, M.: BIG DATA: new science, new challenges, new dialogical opportunities: Zygon. Zygon® **50**, 569–582 (2015). https://doi.org/10.1111/zygo.12187
12. Manieri, A., et al.: Data science professional uncovered: how the EDISON project will contribute to a widely accepted profile for Data Scientists (2015)
13. Seo, D., Lee, M.-H., Yu, S.: Development of network analysis and visualization system for KEGG pathways. Symmetry **7**, 1275–1288 (2015). https://doi.org/10.3390/sym7031275
14. Shaikh, M.A.H., Omar, M.T., Azharul Hasan, K.M.: Efficient index computation for array based structured data. In: Efficient Index Computation for Array Based Structured Data, pp. 101–105. IEEE (2015). http://ieeexplore.ieee.org/document/7391930/. Accessed 18 May 2018
15. Rupp, A.A., van Rijn, P.W.: GDINA and CDM packages in R. Meas.: Interdiscipl. Res. Perspect. **16**, 71–77 (2018). https://doi.org/10.1080/15366367.2018.1437243
16. Webb, S.J., et al.: Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. J. Autism Dev. Disord. **45**, 425–443 (2015). https://doi.org/10.1007/s10803-013-1916-6
17. Brennan, P.F., Bakken, S.: Nursing needs big data and big data needs nursing: nursing needs big data. J. Nurs. Scholarsh. **47**, 477–484 (2015). https://doi.org/10.1111/jnu.12159
18. Tudoran, R., Costan, A., Antoniu, G.: OverFlow: multi-site aware big data management for scientific workflows on clouds. IEEE Trans. Cloud Comput. **4**, 76–89 (2016). https://doi.org/10.1109/TCC.2015.2440254
19. Asamoah, D.A., Sharda, R., Hassan Zadeh, A., Kalgotra, P.: Preparing a data scientist: a pedagogic experience in designing a big data analytics course: preparing a data scientist. Decis. Sci. J. Innov. Educ. **15**, 161–190 (2017). https://doi.org/10.1111/dsji.12125
20. Bowers, A.J.: Quantitative research methods training in education leadership and administration preparation programs as disciplined inquiry for building school improvement capacity. J. Res. Leadersh. Educ. **12**, 72–96 (2017). https://doi.org/10.1177/1942775116659462
21. Malviya, A., Udhani, A., Soni, S.: R-tool: data analytic framework for big data. In: R-Tool: Data Analytic Framework for Big Data, pp. 1–5. IEEE (2016). http://ieeexplore.ieee.org/document/7570960/. Accessed 18 May 2018
22. Ebadi, H., Antignac, T., Sands, D.: Sampling and partitioning for differential privacy. In: Sampling and Partitioning for Differential Privacy, pp. 664–673. IEEE (2016). http://ieeexplore.ieee.org/document/7906954/. Accessed 18 May 2018

23. Rojas, J.A.R., Beth Kery, M., Rosenthal, S., Dey, A.: Sampling techniques to improve big data exploration. Sampling Techniques to Improve Big Data Exploration, pp. 26–35. IEEE (2017). http://ieeexplore.ieee.org/document/8231848/. Accessed 18 May 2018
24. Gehl, R.W.: Sharing, knowledge management and big data: a partial genealogy of the data scientist (2015)
25. Kim, S., Choi, M.-S.: Study on data center and data librarian role for reuse of research data. In: Study on Data Center and Data Librarian Role for Reuse of Research Data, pp. 303–308. IEEE (2016). http://ieeexplore.ieee.org/document/7440517/. Accessed 18 May 2018
26. Eybers, S., Hattingh, M.: Teaching data science to post graduate students: a preliminary study using a « F-L-I-P » class room approach (2016)
27. Baškarada, S., Koronios, A.: Unicorn data scientist: the rarest of breeds. Program **51**, 65–74 (2017). https://doi.org/10.1108/PROG-07-2016-0053
28. Schreck, B., Veeramachaneni, K.: What would a data scientist ask? Automatically formulating and solving predictive problems. In: What Would a Data Scientist Ask? Automatically Formulating and Solving Predictive Problems, pp. 440–451. IEEE (2016). http://ieeexplore.ieee.org/document/7796930/. Accessed 19 May 2018
29. Data robot: Beneficios para los científicos de datos. https://www.datarobot.com/data-scientists/. Accessed 19 May 2018
30. SubjectivesSystems: Convertimos DATA en VENTAJA. https://www.subjectivesystems.com/. Accessed 19 May 2018
31. Turi create intelligence: GraphLab-Create. https://pypi.org/project/GraphLab-Create/. Accessed 19 May 2018
32. Ipython: Ipython interactive computing. http://ipython.org/index.html. Accessed 19 May 2018
33. KNIME: KNIME Analytics Platform. https://www.knime.com/knime-analytics-platform. Accessed 19 May 2018
34. Saltz, J.S., Grady, N.W.: The ambiguity of data science team roles and the need for a data science workforce framework, pp. 2355–2361. IEEE (2017). http://ieeexplore.ieee.org/document/8258190/. Accessed 19 May 2018
35. Forbes: Report: Why « Data Scientist » is the Best Job to Pursue in 2016. https://www.forbes.com/sites/gregoryferenstein/2016/01/20/report-why-data-scientist-is-the-best-job-to-pursue-in-2016/#13caba13a526