The Human Side of Big Data

Understanding the skills of the data scientist in education and industry

Patrick Mikalef, Michail N. Giannakos, Ilias O. Pappas, John Krogstie,
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
e-mail: patrick.mikalef@ntnu.no

***Abstract*—It is widely recognized by public and private organizations, that the biggest challenge faced in light of the data revolution is finding people with the required set of skills to transform data into actionable insight. The growing interest on the role of the data scientist and the relating data analytics skills has seen an increasing amount of research on the importance of data analytics skills in the contemporary working environment. Yet, there is still limited understanding on the importance of data analytic skills, and even more, there is limited research on the discrepancies between the skills that are needed in the market and what graduates possess. To this end, this research uses a mixed-methods approach combining quantitative survey data from 113 IT executives, and qualitative interview data from 27 big data project managers to explore the significance, discrepancies, and aspects of data analytic skills. Our results show that data analytic skills significantly contribute firm performance, particularly for firms that are data-oriented. In addition, we find that the need for skills greatly exceeds those that graduates possess. Lastly, our analysis suggests that the data skills of the data scientist span multiple subject areas which are further discussed.**

***Keywords*— Data Analytics; Data Scientist; Graduate; Empirical; Big Data**

I. Introduction

The past few years have seen an unprecedented explosion in the interest of organizations around the areas of big data and analytics. Today, the topic is one of the most frequently discussed in research and practice, since many organizations are struggling to leverage the data they own or have under their control in order to improve their efficiency and effectiveness of operations [1]. In loose terms, big data refers to data sets that are too large and complex to processes using traditional storage and analysis technologies [2]. Researchers and practitioners differentiate big data from conventional forms of data warehousing on the basis of the ongoing expansion of four V's, volume, variety, velocity, and veracity of data [3]. Data from a recent survey conducted by the MIT Sloan Management Review and the IBM institute for Business Value on approximately 3000 executives, managers, and analysts working in over 100 countries found that the best performing organizations use big analytics five time more often compared to the lower performing ones [4]. Yet, despite the great potential of new technologies, analytics tools and applications, the biggest problem practitioners face in leveraging these technologies is finding employees with the required skills [4]. A number of similar industry reports indicate the shortage of skills relating to big data analytics, relating this shortage with a struggle to achieve profitability [5].

In a highly featured article, Davenport and Patil identify the mismatch of skills offered in today's curricula compared to industry expectations, following the dramatic increase in interest on big data [6]. According to the authors, the job of the data scientist, is forecasted to be the most sought after professional in industry in the upcoming years. The data scientist is a high-ranking professional with the training and curiosity to make discoveries in the world of big data [6]. According to a broad definition by van der Aalst, the data scientist is someone who understands how to fish out answers to important business questions from today's tsunami of unstructured information [7]. The data scientist, however, is more than just a statistician, but involves competences associated with behavioral and social sciences, industrial engineering, computer science, and visualization and representation [7]. According to a report by MIT Sloan Management Review (Ransbotham, Kiron, & Prentice, 2015), more than 40% of organizations are struggling to find and retain data analytics talent. Similar reports from the European Commission (EC) forecast a skill gap of 500.000 ICT professionals by 2020, with a particular emphasis on data analytics skills [8], while for the same period the International Data Cooperation predicts a need of more than 190.000 people with data analytics competences in the USA alone [9]. Recognizing this growing need for data analytics skills, the EC has urged the attention of all relevant stakeholder through the Digital Skills Agenda for Europe action to try to minimize the gap [10]. A report from McKinsey states that an additional 1.5 million managers and analysts will be needed with a sharp understanding of how big data can be applied [11]. The authors of the report urge the attention of relevant stakeholders on the importance of setting up recruitment and retention programs, while making substantial investments in the education and training of key data personnel.

Despite the surging interest from industry executives in finding employees with these sets of skills, current literature has only scarcely focused on the competences that are lacking in industry. Even more, there is a lack of understanding concerning the different roles of the data scientist and the skills that must be promoted in academic curricula to fulfil the gap that exists in today's market [12, 13]. The objective of this paper is threefold. The first, is to understand the importance of big data skills in relation to business objectives. To do so we draw on a sample retrieved from industry executives of 113 organizations. The weighted importance of a data-driven strategy is taken into account so as to be able to isolate the relative value of data skills in contemporary firms. The second is to empirically examine if there is a discrepancy in the skills that are required in today's organizations, compared to those available from graduates, and to provide a clearer understanding on the core areas that are of importance. We employ a mixed-methods approach using primary survey data from industry executives, and a complementary 27 interviews from Information Technology (IT) and digital innovation managers. Finally, the third objective, is to gain a deeper understanding based on the 27 interviews of the nature of these skill-gaps in order to better prepare future curricula and direct research attempts.

The rest of the paper is structured as follows. In the next section, the research questions are outlined and contrasted to the existing state of literature. In the third section, the methodology is described to understand that gap of big data-related skills in industry. Section four describes the quantitative and qualitative results of the empirical study, concluding in section five with a discussion on how these outcomes can inform research, practice, and education.

II. Background

The literature provides some insights into the range of competencies and skills that are associated with the role of the data scientist. Mikalef and colleagues delve into the subject of the skill-gap that exists, and the difficulty of managers in findings and recruiting employees that have the necessary competencies for the jobs of the future [12]. According to the authors, the data scientist's skills encompass technical and managerial skills. Specifically, technical skills according to the managers interviewed in the study, were the most major challenge faced by companies, since in the market there is a gap in skills that is noted as remarkable. The gap is also project to increase due to the rising demand from organizations for employees with such knowledge. In the skill-set of these employees are competencies of data architecture (i.e. designing blueprints of strategies for data acquisition), big data engineering (i.e. handling the storing, cleansing and coding of data), and big data analytics (i.e. applying the right mathematical and statistical techniques to make predictive and analytical models and visualize results). In addition to these technical skill-sets, what is becoming increasingly more apparent in the literature is the managerial thinking that such employees must have [14]. One of the core characteristics of a data scientist is his ability to understand the function of the domain areas he is working on, and not be solely restricted to technical knowledge. This is a critical component since asking the right questions, and understanding what sources of data can provide a valuable business solution is fundamental [15].

A definition by Hal Varian [16], the chief economist at Google, states that the data scientist must have the ability to take data and to be able to understand it, to process it, extract value from it, visualize it, and communicate it. Based on this process-oriented view, the specific data analytics skills can be distilled in terms of technologies, techniques, problem formulation, and communication. In a recent study of the skills relating to data analytics, Van der Aalst [7] shows that it is more than just analytics/statistics; it also involves behavioral/social science (e.g. for ethics and understanding human behavior), industrial engineering (e.g. to understand the value of data and to know about new business models and organizational strategy), and domain specific knowledge (e.g. to understand and communicate with relevant stakeholders). The general consensus in the academic community is that data analytics is indeed a new discipline, just like computer science emerged from mathematics when computers became abundantly available in the 1980's. The growing interest on the role of the data scientist and the relating data analytics skills has seen an increasing amount of research on the importance of data analytics skills in the contemporary working environment [17].

Recent empirical studies on the role of the data scientist argues that the required knowledge base exceeds that of the Computer Scientist or the one of the Statistician, but is in the umbrella of the Information Systems domain [18]. On identifying the core skills of data scientists, Power [19] looks at what data analytics skills are central for data-driven organizations, and highlights the need to develop research and academic activities that prepare these professionals. Following a scientific literature review approach, Costa and Santos [18] survey the knowledge base of the data scientist as well as the corresponding skills, comparing their findings to two of the most commonly recognized competences/skills frameworks in the field of Information and Communication Technology (ICT), the European e-Competence (e-CF) framework and the Skills framework for the Information Age (SFIA). Through their analysis they conclude that the data scientist profile is seen as a multi-disciplinary profile, combining expertise from different areas. Despite such research approaches that attempt to understand the profile of the data scientist, there are very limited studies employing empirically driven methods to identify the knowledge base and skillset of the data scientist, particularly in relation to current and future needs in industry and public administration. Even more, the is lack of understanding of the most effective ways of implementing educational curricula to fulfil these skill requirements, and limited knowledge on the fit that actual graduates have in terms of jobs requirements [20].

This problem is particularly critical for industry executives since from early studies the lack of human skills and knowledge relating to big data and analytics is one of the main barriers in achieving business value [21]. In a recent literature review, Mikalef, et al. [3] note that data skills are perhaps the most sought after resource in companies that are deploying big data solutions, since the skills encompassed in the data scientist profile enabled firms to ask the right questions, and transform data into actionable insight. From an analysis of past findings, they conclude that software, infrastructure, and even data itself is insufficient to provide any value if there is a lack of personal with the appropriate skills and knowledge to put them into action. Similar findings are noted in numerous studies. For instance, Gupta and George [22] state that currently, with the lack of technical skills in industry, big data skills are a potential source of a competitive advantage. Nevertheless, since these skills can be easily codified and taught they will eventually get dispersed among individuals working in the same (or different) organizations, thereby making this resource ordinary across firms over time. The authors do note however, that what will constitute a valuable resource over time are managerial skills relating to big data and analytics. In the words of the authors: *"dispersed among individuals working in the same (or different)* *organizations, thereby making this resource ordinary across firms over time"*. Doing so requires that managers have a solid ability to understand the current business environment and ae capable to predict the future needs of other business units, customers and partners. Managerial skills are also an invaluable asset in promoting a sense of organizational culture revolved around data-driven decision making [23, 24].

III. Research Questions

The breadth of knowledge required by the data scientist has ignited a discussion on how academic curricula should be changed to accommodate the rising needs for skilled and knowledgeable professionals [25]. Due to the increasing popularity of big data analytics in industry, several research papers have voices the opinions that understanding the exact skills and the gap that currently exists in practice is of outmost importance in developing future data-science programs [25]. While some of the required skills are already taught in computer science, engineering, and other technological-related degrees, there are a number of competencies that are not included and highly sought after [26]. In addition, there is an ongoing discussion about the value of data skills, and to what extend they can lead to performance gains by companies. This is an important question since it dictates the weight that should be put on the large skill-gap that currently exists in the market. Understanding the importance of data skills in data-oriented firms is the starting point of our study. Therefore, the main objectives of this paper can be summarized as follows:

RQ1. What is the importance of data analytics in contemporary organizations?

RQ2. What are the skills that are lacking in today's graduates in relation to data scientist profiles?

RQ3. What are the main challenges for firms in terms of finding employees with the necessary data analytics skills?

Prior IT-business value research suggests that the skills of the IT personnel is a critical resource in gaining any measurable business value [27]. Human skills are long regarded as better predictors of performance variance than strategy and economic factors [28]. The notion that firms should merge technology with human skills traces its roots back to the sociotechnical framework. This idea elevates the role of human skills and states that maximized technological performance requires simultaneous management and nurturing of an organizations human skills and knowledge [21]. This idea is particularly relevant in the context of big data, since skills are not only placed in the operation of technical resources, such as software and infrastructure, but more importantly in the generation of insight that drives organizational decisions [29]. From the above argument we can expect the following:

H1. Firms that possess human capital with better data analytic skills will realized increased performance gains.

While the value of data analytics skills is increasingly important in contemporary firms and organizations, it increasingly critical for those that heavily rely on data-driven insight to perform their operations [14]. It is generally agreed that firms that operate in highly uncertain environments rely more on data to enact their strategies and operations. In addition, high complexity forces companies to opt to big data analytics in order to be able to make sense of the vast amount of information and make appropriate decisions that allow competitive survival. As such, a class of firms is noted as having a more data-driven strategic orientation compared to others that do not rely so much on big data analytics to make sense of their competitive landscape [30]. We therefore expect that the level of data-driven strategic orientation will amplify the value that human data skills have on realizing performance gains. Therefore, we hypothesize the following:

H2. A firm's data-driven strategic orientation will positively moderate the impact of human data analytic skills have on performance gains.

IV. Methodology

A. Data Collection

To answer the research questions posited in this study, we draw on two separate samples. The first comprised of 113 industry executives that help managerial roles in IT-business operations (e.g. CIO, CTO and Digital Innovation Managers). These individuals were recruited from a contact list of industries that maintain strong ties with the Norwegian University of Science and Technology (NTNU), and frequently recruit graduates of the university. They were administered with a custom-built online survey and asked to evaluate the significance of data analytics for their companies, as well as the importance of a series of skills, and correspondingly, the level to which they are satisfied with the skills and knowledge of graduates. In addition, respondents were asked to assess the degree to which their firm was performing better than their competitors in a number of areas.

Respondents were contacted via email, which included a link to the electronic survey. A sample of 500 respondents was initially composed from Norwegian firms, with key employees being knowledgeable about their firm's level of investments in areas relating to big data analytics. After the initial contact, two reminder emails were sent out with a two-week internal between them. The final number of responses was 121, of which 113 were usable for further analysis. This sample yields a valid response rate of 22.6% which is similar to respective studies using a key informant. The collection of data was done over a period of approximately 4 months, between January and April 2016. Average completion time for the survey was approximately 17 minutes.

For the qualitative aspect of our study, we contacted a separate list of 50 representatives that held positions similar to those of the quantitative sample. Key respondents were contacted by phone and informed about the objective of our study and assured that any information they provided would remain strictly anonymous. From these, 27 agreed to partake in the interview, 18 of which were performed in person, while the other 9 through a Skype call. The interviews took place between October 2016 and April 2017. Respondents were contacted from multiple countries in Europe, including Norway, the Netherlands, Greece and Italy. The format of the interview was in a semi-structured way, starting with some information on the background of the company they worked for, main responsibilities and experience, and industry-related questions.

The next section involved questions on big data and analytics investments within the firms, including an extensive series of questions about the human capital, deficiencies in terms of skilled personnel, recruitment methods, and training programs and tutorials in order to get staff at a satisfactory level of competence. The final part of the interview involved question about performance of the company and strategic directions. On average, interviews lasted 65 minutes. All responses were transcribed by one of the authors, and sent back to respondents to verify their accuracy. In addition, sensate information was removed so that the interviews could not be traced back to the interviewee.

B. Sample Descritpion

For the quantitative sample, measures were taken to examine there was any non-response bias. We compared responding with non-responding firms in terms of industry, firm size, and revenues to make sure that there were no significant differences. Through independent t-tests performed with the use of the software package IBM SPSS 24.0 no significant difference was observed. In addition, we compared early responses (those received during the first 2 weeks of the study), with late responses (those received during the last two weeks of the study) in terms of performance indicators, size, and industry, and again found no statistically significant differences. These results confirmed that there were no issues of late-response bias. Sample demographics of respondents are presented in Table 1 below.

Factor		Sample	Percentage
Industry			
-	Oil & Gas	17	15.0%
-	Basic Materials	4	3.5%
-	Industrials	3	2.7%
-	Consumer Goods	8	7.1%
-	Health Care	3	2.7%
-	Consumer Services	3	2.7%
-	Telecommunications	11	9.7%
-	Utilities	5	4.4%
-	Financials	5	4.4%
-	Technology	49	43.4%
-	Education	2	1.8%
-	Transportation	2	1.8%
-	Other	1	0.9%
Size-class (in number of employees)			
-	1 - 9	11	9.7%
-	10 - 49	29	25.6%
-	50 - 249	30	26.5%
-	250 +	41	36.3%
Sector			
-	Private sector	94	83.2%
-	Public sector	8	7.1%
-	Non-profit organization (NPO)	1	0.9%
-	Non-government organization (NGO)	0	0.0%
Job Experience (in years)			
-	0 - 10	39	34.5%
-	11 - 20	42	37.1%
-	21 - 30	19	16.8%
-	31 +	3	2.7%

Table 1 Sample demographics of quantitative study

From the collected sample, most firms belonged to the technology sector (43.4%), followed by the oil & gas (15.0%), telecommunications (9.7%), and consumer goods (7.1%). The majority of companies, 36.3%, were large (250+ employees), 26.5% were medium-sized (50-249 employees), 25.6% small (10-49 employees), and 9.7% micro (1-9 employees). Private companies dominated the sample with 83.2%, and public sector organizations consisting a small proportion of the sample (7.1%). The majority of respondents were highly experienced professionals, with the largest proportion of responses coming from employees with 11-20 years of experience (37.1%), and a slightly smaller percentage having between 0-10 years in the firm they operated in (34.5%). Most respondents held the position of Chief Information Officer (CIO) (57.2%), followed by IT managers (31.2%), and business executives (11.5%).

With regards to the sample of firms that formed our qualitative study, the demographics are visible in Table 2. To extract the necessary demographic information, we asked respondents to give us an answer to each respective question, and then cross-referenced their response to publicly available data. Similar to the quantitative sample, the qualitative sample was predominantly populated by companies belonging to the technology industry (25.9%), followed by oil & gas (18.5%), basic materials, consumer goods, and telecommunications, which shared an equal proportion of the sample (11.1%). Regarding size, most companies were in the large size-class (55.5%), followed by medium-sized (25.9%), small (11.1%), and micro (7.4%). The whole sample constituted of private firms, and respondents were once again highly experienced with the largest group being those that had between 11-20 years of experience (51.8%), followed by those with 21-30 years of experience (44.4%).

Factor		Sample	Percentage
Industry			
-	Oil & Gas	5	18.5%
-	Basic Materials	3	11.1%
-	Industrials	1	3.7%
-	Consumer Goods	3	11.1%
-	Health Care	0	0.0%
-	Consumer Services	2	7.4%
-	Telecommunications	3	11.1%
-	Utilities	1	3.7%
-	Financials	1	3.7%
-	Technology	7	25.9%
-	Education	0	0.0%
-	Transportation	0	0.0%
-	Other	1	3.7%
Size-class (in number of employees)			
-	1 - 9	2	7.4%
-	10 - 49	3	11.1%
-	50 - 249	7	25.9%
-	250 +	15	55.5%
Sector			
-	Private sector	27	100.0%
-	Public sector	0	0.0%

- Non-profit organization (NPO)	0	0.0%
- Non-government organization (NGO)	0	0.0%
Job Experience (in years)		
- 0 - 10	1	3.7%
- 11 - 20	14	51.8%
- 21 - 30	12	44.4%
- 31 +	0	0.0%

Table 2 Sample demographics of qualitative study

C. *Measures*

The survey of the data consisted of three main parts, a) questions about the demographics of their organization, b) an assessment of the importance of data analytics in company decision making, and c) paired questions on the importance of certain skills for their company and an estimation of the fulfillment of these requirements from graduates they hire. In developing the set of skills that are evaluated by respondents, past empirical studies were utilized as well as industry reports and commentaries on the emerging requirement for well-trained data scientists [4, 6, 14]. Specifically, for the data analytics skills, past empirical studies were utilized and adapted measures were created to capture the breadth of related skills [12, 21, 22]. Respondents were asked to evaluate for each of the presented skills the extent to which a) these skills were important to their firm, and b) the level to which their current personnel were proficient with them All questions were assessed in terms of a 7-point likert scale measure. To capture the level of data-driven strategic orientation, adapted measures from past empirical studies were used [25, 31]. Respondents were asked to evaluate on a 7-point likert scale how much they agreed or disagreed with a number of sentences relating to the level at which their firm places a strong emphasis on data when making decisions. Finally, to determine the degree to which performance had been improved, we asked respondents to evaluate the level to which they believed their firm was in a better position than their competitors in several areas, such as market share, return on investments (ROI), increased customer satisfaction, and rapid confirmation of customer orders. The measures used to operationalize competitive performance were adopted from prior empirical literature. [32, 33].

D. *Coding*

The empirical analysis was performed through an iterative process of reading, coding, and interpreting the transcribed interviews and observation notes of the 27 case studies [34]. The first of the coding process involved the identification and isolation of a large number of concepts. This was done on the basis of the theoretical underpinnings that were discussed in the previous sections. Since the interviews included several questions spanning multiple areas within he big data analytics domain, those that were specifically revolving around human skills and knowledge were used for the purposes of this study. We employed an open coding scheme which allowed us to quantify the characteristics of each concept. This also enabled us to cluster primary data in a tabular structure, and through an iterative process we identified the relative concepts and notions that were related to each. Two of the co-authors completed the independent coding of the transcripts in accordance with defined notions of the previous sections. Each coder read the transcripts independently to identify statements that revolved around skills of big data analytics, as well as other statement of competences of human capital in general. This process was repeated until inter-rater reliability of the two coders was larger than 90 percent [35].

V. Results

A. *Measurement Model*

All constructs were assessed in terms of reliability, convergent validity, and discriminant validity. We examined reliability at both the construct and item level. At the construct level, composite reliability (CR) and Cronbach's alpha (α) indicators were assessed, with all values being above the threshold of 0.70, suggesting acceptable construct reliability [36]. Indicator reliability was assessed through construct-to-item loadings, with those below the threshold of 0.70 being omitted. All remaining items had loadings above 0.72. We confirmed convergent validity by examining if the average variance extracted (AVE) for each construct was above the threshold value of 0.50. All values exceeded this lower threshold, thus, confirming that convergent validity was not an issue. For discriminant validity, we examined each constructs AVE square root was greater than its higher correlation with any other construct (Fornell-Larcker criterion). The outcomes of these tests suggest that our constructs are reliable and valid to further analyze, and that items support their respective latent variables.

	1.	2.	3.
1. Data Skills	**0.78**		
2. Strategic Data Orientation	0.39	**0.76**	
3. Competitive performance	0.29	0.15	**0.88**
Mean	4.18	3.91	4.88
Standard deviation	1.57	1.52	1.34
Composite reliability	0.924	0.948	0.893
Cronbach's alpha	0.906	0.935	0.858
AVE	0.606	0.785	0.602

Table 3 Assessment of reliability, convergent and discriminant validity of constructs

B. *Structural Model*

We analyzed the data using partial least squares structural equation modeling (PLS-SEM), and specifically the SmartPLS software package [37]. The structural model from the PLS analysis is summarized in Figure 1, in which the explained variance of endogenous variables (R^2) and the standardized path coefficients (β) are presented. The significance of estimates (t-statistics) are obtained by running a bootstrap analysis with 5000 resamples. As illustrated in Figure 1 both hypotheses were confirmed. Specifically, we found that the data skills of employees had a positive and significant direct association to a firm's competitive performance (H1, β=0.383, t=8.715, p < 0.001). In addition, the strategic importance of data for the firm strengthened the aforementioned relationship. Our results show that strategic data orientation had a positive and significant moderating effect on the relationship between data analytic skills and competitive performance (H2, β=0.151, t=2.005, $p < 0.05$). Outcomes suggest that data skills, along with the moderating influence of strategic data orientation are good predictors of a firm's competitive performance, since the two variables account for 19.3% of variance in the dependent construct.

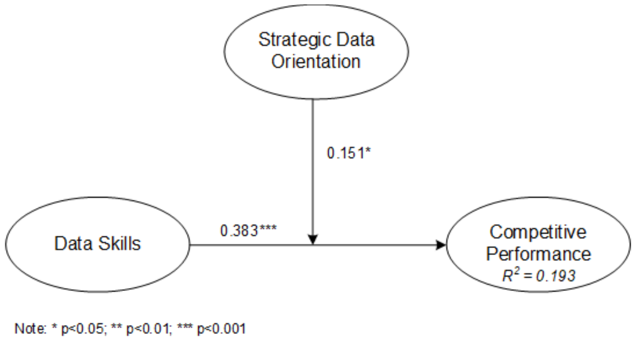
Strategic Data Orientation
0.151*
Data Skills
0.383***
Competitive Performance
$R^2 = 0.193$
Note: * p<0.05; ** p<0.01; *** p<0.001

Figure 1 Results

C. Paired Samples t-Test

To examine if there are any discrepancies regarding the required skills that companies place value to, and the level of competencies that graduate have, a paired-sample t-tests analysis was performed. Prior to comparing the values for each of the skills, we looked at normality of distribution data through the Kolmogorov-Smirnov test, which indicated no substantial skewness in our data. In addition, we used Grubb's test for outliers which indicated that there was no single outlier in our univariate data. To test for differences between data skills in firms, we made use of the statistical package IBM SPSS 24. In Table 4, the first column described the different skills that were assessed, the second column includes the level to which skills are important to the company, while the last column represents the extent to which graduates in the market have these skills.

Skills	Ind.	Edu.	Sig.
Big data analytics	4.08	3.07	***
Data and knowledge visualization	4.66	3.55	***
Statistical techniques	4.26	4.07	n.s.
Transforming raw data into business intelligence	4.37	3.15	***
Structuring and analyzing content (web-based, sensor-based) in a meaningful way	4.44	3.41	***
Research methods and empirical validation	4.15	4.25	n.s.
Working with high volume unstructured data	3.96	3.26	***
Machine learning	3.54	3.62	n.s.

Table 4 t-test between industry importance to data skills and level of skills by graduate hires

The outcomes of the paired samples t-test reveals that there is a significant difference in the need of skills within industry, and those that graduates actually have. Specifically, in five out of eight different categories, highly significant differences were observed. This result confirms the anecdotal evidence which claimed data analytics skills are currently lacking in industry, and that higher-education graduates are not adequately trained to fill in jobs relating to data science. While statistical techniques are lacking in graduates, their difference in our sample didn't warrant them as adequately significant. On the other side, research methods and empirical validation, as well as machine learning skills seemed to be on a greater level of competence obtained by graduate students, than that required by industry. Nevertheless, there was no statistical significance, so we can therefore conclude that they are adequately available in the market. Some of the most striking differences in terms of skill gap were observed for data and knowledge visualization, and for structuring and analyzing content in a meaningful way. This appears to be a major deficiency in graduate programs, who have a more traditional approach in fulfilling their curricula. The same is also valid for the case of big data analytics skills. Contemporary tools for managing and analyzing big data are something that needs to be added to education in the higher level [38].

D. Interview Outcomes

From the 27 conducted interviews, all respondents noted that data analytic skills were one of the most important resources in realizing business value from their big data investments. In addition, there was the unanimous opinion that they were also the most difficult to acquire, relative to other resources. One of the aspects we first identified from the respondents was the realization that data analytics skills were very broad in nature. This is reflected in the quote from one of the respondents.

"You must have the technical and somewhat analytical and program experience. In addition to that you should fit into the industry and be a selling variant."

The reply from the respondent also draws on the issue of having domain knowledge. This is something that is often overlooked since technical skills are more easily detectable as lacking over an array of different industries. In addition to this quote that signifies the relevance of domain-specific knowledge, many interviewees also stressed the value of having soft skills. For instance, one respondent noted the following:

"You must have people with analytical skills but you must also have the capabilities to communicate. It is very important to be able to discuss with people from different departments, understand the problems they are facing, and explain to them what your plan and solution is."

Perhaps the importance of, so called, soft skills is the most under-represented in academic literature. It is commonly accepted that data scientists are required to communicate and collaborate with employees of a technical and business background, but the necessary cooperation and communication skills are seldom talked about. Past literature has recognized that employees in inter-disciplinary teams should have training on skills needed to effectively work with other personnel, yet, in the domain of big data there is very little knowledge about what types of soft skills are important [39]. This has been referred to in past research of information systems as shared knowledge, the understanding and appreciation among IS and line managers of the technologies and processes that affect their mutual performance [40]. Several of the respondents explained this issue in more detail, with one of them making the following statement:

"Being a data scientist in our company is tricky. You have to be able to convince management about the importance of it (big data analytics). Only then will they allow you to have the resources you need. In addition, they are often skeptical of results when they hear words like data-pool, regression, and other statistical terms. It confuses them and reduces their trust in the outcome. I think that if they understood the value of it they would be more likely to invest more time and money in analytics."

This outcome is in line with what Bassellier and Benbasat [39] noted as a necessary condition for business-it alignment. In a recent review article, Chan and Reich [41] note the same thing, that have broad communication skills is essential in data scientists are to create linkages with other organizational units and have a broader perspective of strategy and business objectives. These communication skills and an overall understanding of company objectives and strategy stem from the fact that most data analytics projects are initiated as experimental projects within the IT group. A small amount of resources is usually allocated by higher executives to experiment with data, since it is a trend to do so. The projects that are more successful are the ones which managers have been convinced about the importance and value of data analytics, and have invested more resources in this area. In addition, top management support for creating data-driven culture is seen as imperative in this direction. As such, having professionals that are in place of communicating with top manager executives is fundamental. Several quotes state the fact that projects start as IT-based and eventually end up informing strategic decisions. Specifically, one respondent stated the following:

"It started out as a technology project where we tried Hadoop technologies and gathered some competency about how it should be built. We are now in the phase of finding out how we are to use this type of platform to create business opportunities."

These results support the suggestion that big data analytics education should also include modules on more business-oriented subjects so that data scientists can more easily understand the needs of business, and better communicate relevant results. Data scientists have to participate in social interactions and deal with group dynamics involving stakeholders from a diverse background. They are increasingly being involved in inter-disciplinary teams, therefore the business competence should also include the ability to interact with and manage others. Big data professionals are expected to be able to put away their specialized vocabulary to communicate effectively with their partners [17, 42, 43]. Apart from the required soft-skills, respondents also commented on the difficulty in attracting suitable candidates. Specifically, one respondent noted the following:

"...first of all, lack of competencies – this technology is very new, in the marked there are not people at all with experience. In the country operate, at the maximum, we have guys with three years of experience and these people already belong to our company, because the same guys started three years ago. There is a gap in the skill that is quite remarkable at the moment, and the gap is increasing because there are a lot of new demand for new skills. So the first of all is to have a good set of skills."

This quote clearly denotes the difficulty I attracting skillful data scientists. Many companies resolve in training individuals with a basic set of skills on job operations. A quote from a respondent in a financial firm clearly shows this.

"We are more or less hiring 2-3 people from the college at the time, and providing a lot of training not by sending this guys to training classes but by providing real use cases and a tutor in order to be sure they are working not on the theory but on the practical application."

This constitutes a problem for companies since they are forced to utilize their newly-hired employees in non-productive posts until they are sufficiently trained. While some competencies are indeed developed while working, a large part of the essential skills of the data scientist can be developed in higher education. Thus, we develop here a list, based on the findings of our interviews, on the core areas that a data scientist must be knowledgeable about. The goal is to use these main areas to promote education and skill acquisition for professionals engaged in the domain.

- Data management and challenges
- Security, anonymity, privacy and ethics of data
- Research thinking, hypothesis formulation and statistical analysis methodologies
- Data analytics tools
- Data flow management
- Visualization and presentation of results
- Programming and technical skills
- Artificial intelligence and machine learning
- Interpersonal and social skills
- Domain knowledge
- Business and strategy competences
- Distributed systems

While these are just categories of skills that emerged in our study, the goal is to expand them into specific skills and develop modules that are relevant to the context of big data.

VI. Discussion

In this paper we have delved into the issues of data analytic skills, and examined the significance the discrepancy that exists in contemporary organizations in terms of employing data scientists. The goal of the paper is threefold, a) to understand how important data analytics skills are in today's firms in relation to achieving competitive success, b) to identify if and at what competences there are deficiencies, and c) to develop a deeper understanding of the necessary skills of the data scientist as well as the process of educating employees in firms that utilize big data. To examine these research questions, we build on a mixed-methods approach, utilizing a survey-based dataset of 113 higher level executives in firms, and interviews with 27 IT managers. The use of a mixed-methods approach is appropriate for the context of examination since it allows us to examine a sufficient sample of respondents, and extract meaningful information.

From a research perspective, this study extends the existing understanding on the importance of data analytics skills in contemporary organizations. While past literature anecdotally stressed the significance of data skills, no attempts had been made to empirically investigate the weight of data analytics skills in attaining a state of competitive performance [14, 44]. While past studies on emerging technologies placed considerable attention on the necessary skills of IT professionals, in the era of big data, the competencies associated to the data scientist have been largely neglected. A disproportionate amount of attention has been placed on technologies and infrastructure supporting big data and analytics, and very scarce research has focused on the human side. Past literature on IT skills in the business environment have been very concrete on the types of skills that are needed in inter-disciplinary roles, such as the study of Bassellier and Benbasat [39]. In addition, it is frequently mentioned that the most valuable resource when it comes to information technology is the human one, since the skills and knowledge, as well as how they are weaved in the organizational fabric are the most difficult to replicate from competitors [45]. Our findings confirm these previous results in the IS context, and extend this understanding to the area of big data and analytics. Furthermore, while there is considerable discussion on the role of the data scientist and the large lack of the skills that encompass this role, this study has sought to empirically examine such claims. First, from our quantitative study it is clear that in most data-related skills, there is a significant difference between what is currently needed in industry and what is possessed by higher-education graduates. This is also confirmed from the sample of 27 interviews we conducted, in which managers consistently ranked data analytics skills as the scarcest resource. Finally, the qualitative analysis provided us with a deeper understanding of the multitude of skills that are required in today's data analytics jobs. While previous literature relied mostly on bibliographical references to support this claim, our study is one of the first to empirically examine the diverse set of skills of the data scientist. The insight generated from the qualitative study also helps understand how the skills of the data scientist influence the development of big data projects in firms and also helps relate their success.

From an educational point of view, the findings of the study raise several critical issues to policy makers and higher education curriculum developers. It is apparent that there is currently a large gap in the skills that are needed and the ones taught in academic curricula. As such, one of the implications of this research is that is important for policy makers and developers of academic programs to pay close attention to the needs of industry. One of the fundamental issues with big data is that it moves in very fast cycles, making existing skills obsolete very fast. It is therefore critical to keep up with new developments and needs, and also forge close ties with industry executives. It is quite common for companies nowadays to develop joined initiatives on education, particularly in fast-pasted industries. While it is not easy to formulate tailor-made programs based on each industry's needs, there are several common elements that underlie all domains. Specifically, one of the deficiencies that can easily be incorporated into data analytics programs is that of teaching soft skills, such as communication, collaboration, and inter-disciplinary team management. In addition, including domain specific subjects for industries that utilize big data analytics largely, like financials, telecommunications, media, and retail, for instance, could augment the skill-set of graduates and promote employability.

Despite the contributions of the paper, this study does not come without certain limitations. First, we rely on key respondents within firms to form an understanding of skill requirements. Specifically, our target respondents include higher level managers. Future studies should seek to examine the role of recently hired graduates and ask for their opinions on skills that they are deficient in or competences that are critical but they didn't acquire in their study programs. Second, while we look at skills and competences through a holistic lens, it is important to understand that the needs of each industry are unique, and there are possibly many different sub-profiles of data scientist within industry. An alternative means of analysis could highlight these sub-groups and provide greater clarity. Such methods could include qualitative comparative analysis (QCA) and cluster analysis [46]. Finally, a longitudinal study would be suitable for uncovering the learning process that takes place within industry. In the challenge of facing a large skill gap, firms and organizations have become very competent in training their staff for the required data analytics skills. Therefore, it is important to examine how they manage to do so, and what are the critical success factors when learning on the job. The process as well as choice of methods that are used to support education on data analytics skills could help inform future educational practices in higher academic institutions and drive teaching efficiency.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Sklodowska-Curie grant agreements No 751550 and No 704110

References

[1] M. Kowalczyk and P. Buxmann, "Big Data and information processing in organizational decision processes," *Business & Information Systems Engineering*, vol. 6, no. 5, pp. 267-278, 2014.

[2] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills," *Business & Information Systems Engineering*, vol. 6, no. 5, pp. 289-300, 2014.

[3] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," *Information Systems and e-Business Management*, pp. 1-32, 2017.

[4] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," *MIT sloan management review*, vol. 52, no. 2, p. 21, 2011.

[5] S. Ransbotham and D. Kiron, "Analytics as a Source of Business Innovation," *MIT Sloan Management Review*, February 2017.

[6] T. H. Davenport and D. Patil, "Data scientist: The Sexiest Job of the 21st Century," *Harvard business review*, vol. 90, no. 5, pp. 70-76, 2012.

[7] W. M. Van der Aalst, "Data scientist: The engineer of the future," in *Enterprise Interoperability VI*: Springer, 2014, pp. 13-26.

[8] European Commission. (2017, 31/10/2017). *Digital Skills and Jobs Coalition*. Available: https://ec.europa.eu/digital-single-market/en/digital-skills-jobs-coalition

[9] International Data Cooperation. (2017, 31/10/2017). *Market Analysis Perspective: Worldwide Developer Demographics, Communities, and Skills*. Available: https://www.idc.com/getdoc.jsp?containerId=US42054117

[10] M. Curtarelli, V. Gualtieri, M. S. Jannati, and V. Donlevy, "ICT for work: Digital skills in the workplace," European UnionMay 2017 2017, Available: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=44434.

[11] B. Brown, M. Chui, and J. Manyika, "Are you ready for the era of ‘big data’," *McKinsey Quarterly*, vol. 4, no. 1, pp. 24-35, 2011.

[12] P. Mikalef, V. A. Framnes, F. Danielsen, J. Krogstie, and D. H. Olsen, "Big Data Analytics Capability: Antecedents and Business Value," in *Pacific Asia Conference on Information Systems*, Langkawi, Malaysia, 2017.

[13] Y. Demchenko, E. Gruengard, and S. Klous, "Instructional model for building effective Big Data curricula for online and campus education," in *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on*, 2014, pp. 935-941: IEEE.

[14] A. McAfee, E. Brynjolfsson, and T. H. Davenport, "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60-68, 2012.

[15] B. Wixom *et al.*, "The current state of business intelligence in academia: The arrival of big data," *CAIS*, vol. 34, pp. 1-13, 2014.

[16] H. R. Varian, "Big data: New tricks for econometrics," *The Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3-27, 2014.

[17] A. De Mauro, M. Greco, M. Grimaldi, and G. Nobili, "Beyond Data Scientists: a Review of Big Data Skills and Job Families," *Proceedings of IFKAD*, pp. 1844-1857, 2016.

[18] C. Costa and M. Y. Santos, "The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age," *International Journal of Information Management*, vol. 37, no. 6, pp. 726-734, 2017.

[19] D. J. Power, "Data science: supporting decision-making," *Journal of Decision systems*, vol. 25, no. 4, pp. 345-356, 2016.

[20] M. Savin-Baden, "Education and Big Data," *Encyclopedia of Educational Philosophy and Theory*, pp. 1-7, 2017.

[21] S. F. Wamba, A. Gunasekaran, S. Akter, S. J.-f. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," *Journal of Business Research*, vol. 70, pp. 356-365, 2017.

[22] M. Gupta and J. F. George, "Toward the development of a big data analytics capability," *Information & Management*, vol. 53, no. 8, pp. 1049-1064, 2016.

[23] P. P. Tallon, R. V. Ramirez, and J. E. Short, "The information artifact in IT governance: toward a theory of information governance," *Journal of Management Information Systems*, vol. 30, no. 3, pp. 141-178, 2013.

[24] P. Mikalef, I. O. Pappas, M. N. Giannakos, J. Krogstie, and G. Lekakos, "Big Data and Strategy: A research Framework," in *MCIS*, 2016, p. 50.

[25] 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, 2013.

[26] R. Eynon, "The rise of Big Data: what does it mean for education, technology, and media research?," ed: Taylor & Francis, 2013.

[27] A. Bharadwaj, "A resource-based perspective on information technology capability and firm performance: an empirical investigation," *MIS quarterly*, vol. 24, no. 1, pp. 169-196, 2000.

[28] T. C. Powell and A. Dent-Micallef, "Information technology as competitive advantage: The role of human, business, and technology resources," *Strategic management journal*, pp. 375-405, 1997.

[29] T. Schoenherr and C. Speier-Pero, "Data science, predictive analytics, and big data in supply chain management: Current state and future potential," *Journal of Business Logistics*, vol. 36, no. 1, pp. 120-132, 2015.

[30] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Research*, vol. 2, no. 2, pp. 59-64, 2015.

[31] 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.

[32] Y. E. Spanos and S. Lioukas, "An examination into the causal logic of rent generation: contrasting Porter's competitive strategy framework and the resource-based perspective," *Strategic management journal*, vol. 22, no. 10, pp. 907-934, 2001.

[33] P. Mikalef and A. Pateli, "Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM

and fsQCA," *Journal of Business Research*, vol. 70, pp. 1-16, 2017.

[34] M. D. Myers and M. Newman, "The qualitative interview in IS research: Examining the craft," *Information and organization*, vol. 17, no. 1, pp. 2-26, 2007.

[35] M.-C. Boudreau, D. Gefen, and D. W. Straub, "Validation in information systems research: a state-of-the-art assessment," *MIS quarterly*, pp. 1-16, 2001.

[36] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of marketing research*, pp. 39-50, 1981.

[37] C. M. Ringle, S. Wende, and J.-M. Becker, "SmartPLS 3," *Boenningstedt: SmartPLS GmbH, http://www. smartpls. com*, 2015.

[38] B. Gupta, M. Goul, and B. Dinter, "Business Intelligence and Big Data in Higher Education: Status of a Multi-Year Model Curriculum Development Effort for Business School Undergraduates, MS Graduates, and MBAs," *CAIS*, vol. 36, p. 23, 2015.

[39] G. Bassellier and I. Benbasat, "Business competence of information technology professionals: Conceptual development and influence on IT-business partnerships," *MIS quarterly*, pp. 673-694, 2004.

[40] K. M. Nelson and J. G. Cooprider, "The contribution of shared knowledge to IS group performance," *MIS quarterly*, pp. 409-432, 1996.

[41] Y. E. Chan and B. H. Reich, "IT alignment: what have we learned?," *Journal of Information technology*, vol. 22, no. 4, pp. 297-315, 2007.

[42] B. H. Reich and I. Benbasat, "Factors that influence the social dimension of alignment between business and information technology objectives," *MIS quarterly*, pp. 81-113, 2000.

[43] D. Schuff, "Data science for all: a university-wide course in data literacy," in *Analytics and Data Science*: Springer, 2018, pp. 281-297.

[44] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, pp. 72-80, 2014.

[45] L. Fink and S. Neumann, "Gaining agility through IT personnel capabilities: The mediating role of IT infrastructure capabilities," *Journal of the Association for Information Systems*, vol. 8, no. 8, p. 440, 2007.

[46] P. Mikalef, A. Pateli, R. S. Batenburg, and R. v. d. Wetering, "Purchasing alignment under multiple contingencies: a configuration theory approach," *Industrial Management & Data Systems*, vol. 115, no. 4, pp. 625-645, 2015.

Appendix A. Questionnaire Items

Construct	**Question**	**Loading**
Data Analytics Skills	*How would you rate the importance of the following skills for your company (1 – Not important at all, 7 – Highly important)*	
[DAS_1]	Big data analytics	0.815
[DAS_2]	Data and knowledge visualization	0.836
[DAS_3]	Statistical techniques	0.724
[DAS_4]	Transforming raw data into business intelligence	0.853
[DAS_5]	Structuring and analyzing content (web-based, sensor-based) in a meaningful way	0.768
[DAS_6]	Research methods and empirical validation	0.709
[DAS_7]	Working with high volume unstructured data	0.867
[DAS_8]	Machine learning	0.710
Strategic Data Orientation	*Please indicate the extent to which you agree or disagree with the following statements (1 – strongly disagree, 7 – strongly agree)*	
[SDO_1]	Data-driven decisions are a core aspect of how our firm operates	0.853
[SDO_2]	Most decisions are made based on data analytics and not so much on managerial insight	0.858
[SDO_3]	We thoroughly analyze various types of data in order to improve and develop products and/or services	0.914
[SDO_4]	Operations are constantly monitored and improved upon by modeling streams of data	0.869
[SDO_5]	There is a strong emphasis on data and analytics to guide business strategy within our firm	0.932
Competitive Performance	*Compared with your key competitors, please indicate how much you agree or disagree with the following statements regarding the degree to which you perform better than them in the following areas (1 – strongly disagree, 7 – strongly agree)*	
[CP_1]	Return on investment (ROI	0.730
[CP_2]	Profits as percentage of sales	0.746
[CP_3]	Decreasing product or service delivery cycle time	0.745
[CP_4]	Rapid response to market demand	0.765
[CP_5]	Rapid confirmation of customer orders	0.724
[CP_6]	Increasing customer satisfaction	0.857