Data Science postgraduate education at University of Dar es Salaam in Tanzania: Current demands and opportunities

Betty Mbwilo [0000-0001-6603-2070], Honest Kimaro[0000-0001-6162-3277] and Godfrey Justo [0000-0003-2003-3320]

University of Dar es Salaam, Dar es Salaam, Tanzania
engbettie@gmail.com, honestck@gmail.com and njulumi@gmail.com

Abstract. Several studies indicate that there are no enough people in the market with data science skills and even those graduates in ICT from universities do not possess skills required by employers. Thus, researchers have suggested the urgency for universities to review their curricular as the world is heading towards data era. The aim of this research was to analyze the current skill-gaps needs from stakeholders and opportunities to establish data science postgraduate programme that reflects the current technological trends and market demands at the University of Dar es Salaam (UDSM). A questionnaire was administered to 85 identified organizations to solicit information on the needs for data scientists and existing skill gaps. A total of 61 filled questionnaires response were received out of the 85 that were administered to selected organizations indicating a turn out rate of over 70%. Overall the analyzed data articulated a compelling evidence for the local industry growing need for data scientist. The survey that was conducted was followed up by conduct of various workshops and meetings to solicit inputs from different experts and stakeholders on different versions of the developed curriculum. Finally, a new programme in MSc in data Science was approved and established from April 2018 at UDSM. Despite its late approval and without formal advertisement on the public media, the programme attracted a large number of applicants for 2018/19 academic year, compared to other several postgraduate programmes in ICT offered at UDSM.
.

Keywords: Data Science, Data Scientists, UDSM, Data Science Education, ICT, Skills-gap.

1 Introduction

Significant advances in ICT and underlying communication infrastructure within the last decade has prompted the new era of data [1]. Data and information have taken an unprecedented turn in recent years due to the new ways that it is being created and collected including the speed and diverse nature of its format, fostered by the ever expanding new digital telecommunication technologies that bring a close inter-twine of societies, including the internet and mobile devices. This has demanded change in methods of storing and analyzing such data in order to benefit from gaining insight

for evidence based decision making. Further, it has prompted need for change in computing device architecture and algorithms thereof for data analytics and processing [2]. The volume of new body of knowledge resulting from such activities put demands for more focused specialization, not just for researchers but also data professionals.

The potential for this relatively new and exciting data era has led to data science sub-discipline in fueling growth of the industrial and knowledge economy is immense and anchored within data-driven decision making. Data Science, an emerging field that combines techniques of computer science, Mathematics, statistics and numerical computational for scientific and business data analytics in order to foster better decision making. The universal nature for need of data driven decision making across sectors is now vivid. For example, the urgent need for countries to mobilize data to support monitoring of the Sustainable Development Goals (SDGs) targets and indicators to meet the 2030 global development targets have seen tremendous efforts for countries to build capacity and strengthen national data landscape to guide monitoring of global, regional and national development frameworks. Besides, data is a vital asset to any organization. It holds valuable insights into areas such as customer behavior, market intelligence and operational performance which can aid organizations improve its resource planning and better direct the investments [2]. As universities have been strategic partners to meet skilled human capital demands for numerous industrial disciplines, it is imperative and timely now for universities to assume and respond to the growing local and regional skill gaps within the data science sub-discipline.

In Tanzania, the need to improve the national capacity for data-driven decision making, policy planning and industrial investment is at high stake now than ever in an effort to sustain the national industrialization agenda. Globally and particularly in Tanzania, lack of adequate data science skilled individuals is apparent, while the need for this expertise is set to rise as the role of data in development globally increases. It is imperative to bridge the current national knowledge and skill gaps in data science, and bring the country to harness the existing and future generated massive data set to foster better policy, budgetary planning, decision making and direct for efficient and effective investments, vital for propelling forward the national industrialization agenda.

The aim of this research was to analyze the current skill-gaps needs from stakeholders and opportunities to establish data science postgraduate programme that reflects the current technological trends and market demands.

2 Relevant literature on Data Science

Aalst [3], Discussed how data science came into existence after availability of large amount of data. Competences that data scientist needs are also revealed as presented in Figure 1; data scientist should have a mix of quantitative and technical skills, should be creative, communication skills as well as apprehending end to end solution.

Motive behind data science discipline includes; realization of many organizations that their survival is solely dependent on capability to exploit available data intelligently.

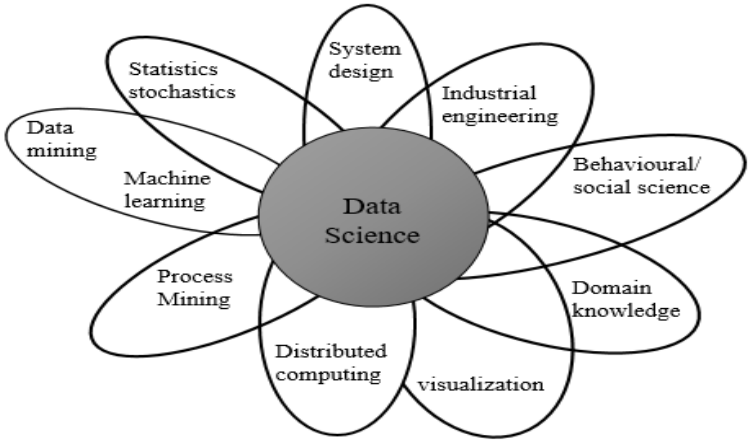

Fig. 1. Profile of a data scientist (source: Aalst [3])

Mikalef, Giannakos, Pappas, and Krogstie [4], highlighted the skill gap that exists in the market. The study was done by survey and interviews to ICT companies to see if they have enough skilled labor who are able to transform data into actionable insights, also to see if what is taught in training institutes meets their needs. It was found that there are no enough people in the market with data science skills and even those graduates from universities does not possess skills required by employers. The paper further suggested the urgency for universities to review their curricular as the world is heading towards data era.

Education for Data Intensive Science to Open New science frontiers [5] extracted from Data Science Competence Framework, Data Science Body of knowledge, Model Curriculum and Data Science Professional Profile; Data Science Competence Framework has been extended based on already existing data science framework and ICT competences and skills. The report defines five groups of competences of data science; data analytics, data science Engineering, domain knowledge, data management and research methods. These are basic skills which data scientists need to have to accomplish daily activities.

Moreover Song and Zhu [6], discusses the close relation between big data and data science in this era of data. Due to advances in ICT there had been dozens of data produced characterized by 5Vs; Volume, Veracity, Velocity, Variety and Value. The emerging trend in big data led to need for new methods to analyze and making insights out of it. Thus new body of knowledge, data science came into existence. The concept of big data, data science and data scientist are well discussed from a survey, hence advice on approach on how to teach data science in colleges.

However Holman, Stuart-fox and Hauser [7] acknowledged the fact that minority of women comprise Science, Technology, Engineering, Mathematics and Medicine (STEMM) workforce. The study was done by analyzing publishing outlets; PubMed and arXiV it was established that large percentage of authors were men. The authors'

recommends reforms in educational system, mentoring and academic publishing. The results are also related to Zhang [8] which explains well the gender gap in data science jobs in USA where on 26% are employed. The article went further by explaining reasons for the gap. The reasons for the gender gap included: a lack of STEM education for women early on in life, lack of mentorship for women in data science, and human resources rules and regulations not catching up to gender balance policies, to name a few.

The National Academies of Sciences, Engineering, and Medicine established the Committee on Envisioning the Data Science Discipline, as in [9]. The Committee study report provide candid discussion on the outlook for the data science education demand in various forms and potential for academic institutions to implementation data science training at different levels, including high schools, undergraduate and graduate levels. The committee reports provided an in-depth update on key issues pertaining to data science education including the knowledge and skills required for a data scientist; the success and challenges of existing data science programs; and the challenges of creating new such programs and strategies around leveraging synergies. We note, however, that the tremendous progress and the future outlook on data science education paints a picture of the developed world than the globally. The developing world trend is by large at infancy stage of data science education with notable initiative for strengthening the ecosystem pioneered by a few players including Data Science Africa [10] fostering a networking platform for data science researchers across Africa.

South African educational institutions stands by far as African leading institutions in provision of different forms of data science education including the African Institute for Mathematical Sciences (AIMS) South Africa [11] which offers two months intensive data science training and the University of Witwatersrand [12], offering a one year degree program in Big data Analytics. Tanzania has not been left behind, in this dynamics as there has been a few initiatives that aim to strengthen data use and data driven innovation although largely funded through international donor programs. As part of open government partnership (OGP) membership the country implemented a basic statistics public data portal, and received data capacity building support through the World Bank and other international donor agencies [13]. The basic statistics portal [14] was established to provide data in a machine readable format to be used and re-used by anybody. The data provided apply only to data and information produced or commissioned by government especially the prioritized sectors of Education, Water and Health. The most recent notable program for strengthening country data capacity is the Tanzania Data Lab project (dLab) [15] that started in April 2016 to June 2018 that sought to promote greater data use in the country through fostering data availability and data literacy. The dLab project was led by the UDSM through PEPFAR/MCC funding, and one key outcome from the project has been the motivation to expand data literacy through mainstreaming the data science education in higher learning institutions.

Tanzania Development vision of 2025 [16] emphasizes the need to have a well-educated and learned society and Tanzania to have competitive economy. Furthermore,

the ICT policy [17] of which was formulated in the context of Tanzania development vision 2025 recognize ICT being central to economic development of a nation.

3 Methodology

The objective of the study was to assess demand and opportunities for data science education in the Tanzanian context and in turn inform the curricular development need for a data science program at the University of Dar es Salaam. The dLab project presence was a natural catalyst to the hypothesis that demand for data science skilled professionals was, indeed, in the raise in the country, based on its internal need assessment analysis on stakeholders training need assessment in basic data management, data analysis and visualization course that were offered during the period of 2016-2018 as core project activities. The baseline data culminated with internal post training follow-up surveys consistently reflected the belief that there was a growing section of data community that were looking for cutting edge skills for working with growing data sets emerging from different communities including government, non-governmental organizations (NGOs), private organizations, researchers and innovators.

In order to sustain the ecosystem both dLab and the University partnered to carry this research for in depth assessment for demand and opportunities on data science education to lead into a concrete program at the university that pioneers the data science landscape in the country. A field study was designed to reach out to the key potential data stakeholders to solicit information through a questionnaire instrument. Through dLab's stakeholders mapping, it was established that most organizations including government agencies, private companies and NGOs were based in Dar es Salaam with operational branches in the regions, therefore for purpose of the study, the survey was conducted in Dar es Salaam.

The questionnaire was developed and administered to 85 purposively sampled organizations drawn from amongst dLab mapped data stakeholder organizations in which a consultation with individual recipient organization was conducted to identify suitable informant from within the organization. A total of 61 duly filled questionnaires response were received out of the 85 that were administered to selected organizations indicating a turn out rate of over 70%. The respondent organizations who participated in this study were specifically drawn from sectors such as the education and research sector, government agents, health sector, energy and mineral sector, agricultural sector, international agencies, financial sector and industrial sector. The informants' gender disaggregation constituted 53 male and 8 female; while participant age profile was 68% being youth with age ranging from 18 to 35 years. A graphic in Figure 2, provides the gender and age profiling for the study participants.

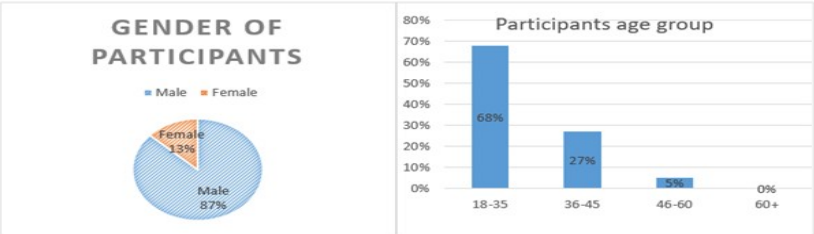

Fig. 1. Informant gender and age profile

To ensure that the informants clearly understood the information sought for by this study additional written and oral information was provided to each participant to help them understand the purpose of the study and why they were the best individuals to participate. It was made clear to each informant that filling of the questionnaires was voluntary and the information provided is categorically meant to inform the data skills needed to empower them and better perform the organizational tasks they are involved with informed decisions.

Filled questionnaires were compiled and coded in Microsoft Excel spreadsheet where analysis was done using the newly introduced Data Analysis add-ins in Microsoft Office version 2013. Different coding and analysis techniques were employed to analyze data depending on the type of information collected for each topic. Analysis was done independently for each topic covered in the questionnaire and necessary relationships between topics were established as is discussed in subsequent sections. Overall the analyzed data articulated a compelling evidence for the local industry growing need for the shortlisted skill sets relevant for data scientist.

4 Results

In this section the key results and findings of the study are presented which are summarized in four major subtopics; the informants' profession and level of education, the level of applicability or relevance of different skill areas related to data science education within the probed organizations, a review of informants opinion on demand for data science education and the relative preference of different available modes of learning. In terms of professional background most informants were from the fields of computing and ICT, Statistics, Finance, Engineering and Mathematics, while only a few were from other disciplines that had little or no relation to the data science field. Moreover, the informants' level of education ranged from advanced diploma to PhD, in which about 69% of respondents were yet to receive a master degree. Figures 3(a) and 3(b) provide a graphical view of the disaggregation of informants along the professional background and levels of education category.

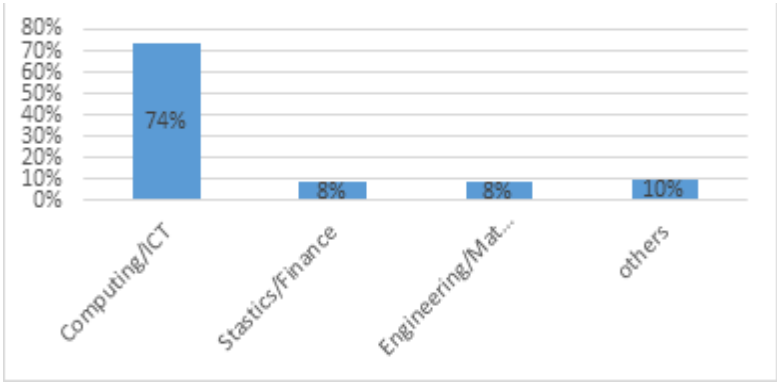

Fig. 1(a). Professional backgrounds

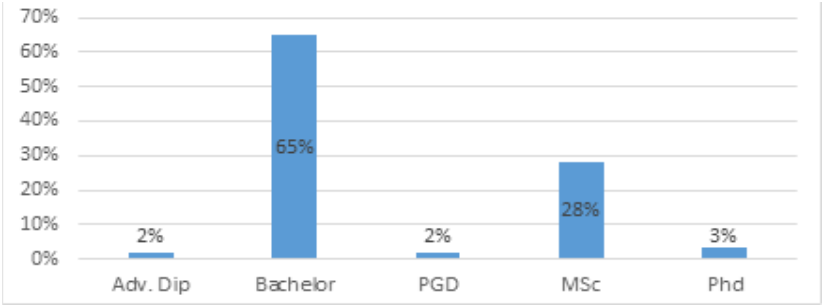

Fig. 3(b). Education Levels

In recent years many higher learning institutions have invested in providing choices for learners in terms of delivery modes, including on campus full-time or the part-time evening counterpart, on the other hand the technology mediated online learning or corresponding dual/blended mode of learning. It was imperative to find out learning mode preferences from respondents to help understand how optimal the investment should be directed along the various learning modes. When asked whether they preferred full-time, evening, online or dual mode of learning, most respondents indicated the part-time evening mode of learning as the most preferred as graphically depicted in Figure 4.

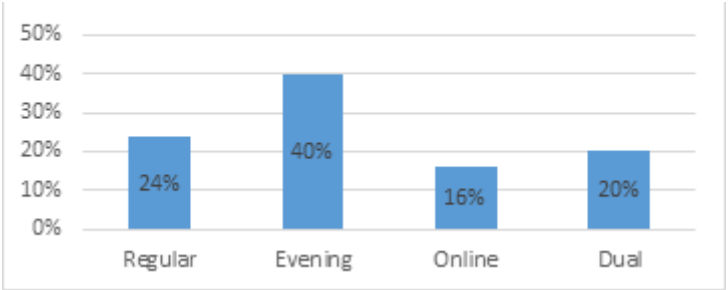

Fig. 1. Learning mode preference

The informants were subjected to a list of course topics relevant to the data science body of knowledge in order to rank the level of importance and applicability to their organization as shown in Table 1. In particular, from the given list of 29 data science related topics the respondents were to identify the relevance to their work related functions and their organizations' needs. They were asked to rank by using 6 levels of Likert scale ranging from highest relevance level 5 for "very important" to the lowest relevance level 0 for "not applicable". Besides, the encoded list the informants were given an option to add in the list missed out topics of relevance. Table 1 shows the aggregated response on level of importance for respective selected course topics in addition two additional missed out relevant topics were listed by respondents, namely, High Performance Computing and Digital signal processing topics.

S/n	Topics	Level of importance	S/n	Topic	Level of importance
1	Statistical models and regression	3.96	16	Data management	4.67
2	Distributed systems	4.09	17	Decision Support System	4.09
3	Data Analytics and Processing	4.60	18	Data led Information Systems	3.96
4	Data Mining and pattern recognition	4.32	19	Action led Information Systems	3.67
5	Social media Analytics	3.59	20	Optimization methods	3.58
6	Spatial databases and Applications	4.12	21	Forecasting and Predictive Analytics	3.98
7	Data Visualization	4.42	22	Strategic Business Analytics	4.07
8	Statistics for data analysis	4.34	23	Big data processing	4.42
9	Information Science	4.07	24	Artificial Intelligence	3.41
10	Data Driven Application-Emerging Applications	4.26	25	Data interpretation	4.48
11	Problem Driven learning	3.87	26	Knowledge management	4.10

12	Scalable Data Systems and Algorithms	3.84	27	Expert systems	3.83
13	Data integration and sharing	4.37	28	Data standards	4.15
14	Monitoring and evaluation	4.29	29	Data security and privacy	4.77
15	Data use	4.31			

Table 1. Topics and respective level of importance on Likert scale

In order to determine whether a data science program would be in demand, and a market for such program existed, respondent opinion level for demand and market was sought with respect to data science knowledge to solve day to day problems. Figure 5, show a graphic representation on market demand opinion for skilled labor force on the data landscape.

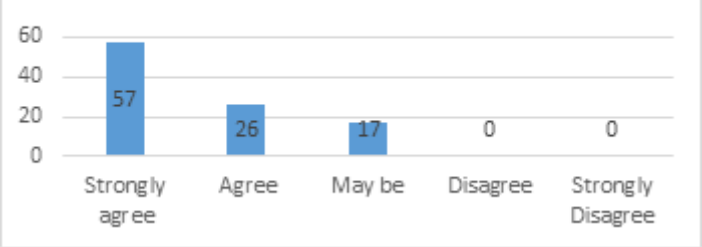

Fig. 5. Market demand

In addition to individual level of opinion, respondents were asked to further comment on how data science could be vital in respective organizations. A number of interesting approval comments emerged, as highlighted below by a few selected representative comments:

"The demand for the programme is vital and as well the programme will be used to solve lots of data demanding operations. We give it go ahead and look forward for its initiation".

"The programme is highly needed in the country. Apparently the richest companies are data companies, thus managing data should be priority."

"Strongly agree with the proposed programme as with data, professional and experts might produce a future analytic oriented results".

"Importance of big data analytical tools, need for exposure to big data analysis. The market is so fragile hence forecasting of customer behavior is essential. By integrating social media data usage in telecom services/ product the organization will be in position to come up with products needed by markets hence increase revenue".

"The outlined topics in this proposal are in line with current market demand and will help organizations to leverage on information for operational improvements and

increase their competitiveness. To be able to benefit large number of Tanzanians (currently employed) especially those in remote regions, I suggest the programme to offered in online mode as well as full time."

As a proof of concept the findings triggered a curriculum development process at the department of Computer Science and Engineering of the College of Information and communication Technologies for Masters in Data Science Program at the UDSM in partnership with dLab organization, the successor of the dLab project. The partnership aimed at creating a linkage between the industry and academia, making this one of the innovative program that harness such a linkage to foster both theory and practice. The curriculum has been approved by the University Senate under regular/full-time mode and the first cohort is ready for admission for the forthcoming academic year 2018/2019. Out of the admission process, it emerged out that this program had attracted a high number of applicants more than the second runner postgraduate programme at the host department which currently offers five postgraduate programmes as shown in Table 2.

Table 1. Summary of admission into masters programmes for 2018/2019

Programme	MSc Computer Science	MSc Data Science	MSc Health Informatics	MSc Information System Management	MSc Computer and IT Systems Engineering
Applicants	15	39	10	30	8

5 Discussion

The objective of the study was to assess demand and opportunities for data science education in the Tanzanian context and in turn inform the curricular development need for a data science program at the University of Dar es Salaam. There is compelling evidence of growing data science skill shortage in Tanzania as more data sets are becoming readily available and demand for harnessing such data is on the increase not just from the public sector but also researchers, innovators, NGOs and the private sector, it is also a concern for [3]. For, instance Government agencies such as ministries regulatory authorities need skilled labor on data science to draw insights from routine, administrative and survey data for development of the nation. Furthermore data science postgraduate education responds to [16, 17] since it prepares skilled human resource to manage the emerging data management and analytics demands in organizations. Data Science is vital to emerging data driven economy which will play a vital role in any industrialized and service driven business economy. Data Science postgraduate education will contribute towards the country's drive for industrial economy by increasing the number and quality of data scientists in organizations who will drive evidence-based decisions and policy planning in order to meet market demands.

The professional background and level of education of most data practitioners who are currently involved in organizational data related tasks confirmed that the existing graduate programs lack the necessary course topics that fosters knowledge and skill required for a competitive data scientists as in [4]. However, such professions and level of education can be leveraged as a prerequisite for entry into a data science educational program, in particular, through a postgraduate program in data science. Therefore, this bring new opportunity for educators to established postgraduate program focused in data science.

The gender disaggregation data confirmed that the field of data and ICT is skewed towards male as 87% of respondents who are also data practitioners in the shortlisted organizations were men. Indeed, this shows that the ICT related jobs in Tanzania are male dominated as discussed in hence, the same gender disparity would be expected for candidates of the anticipated data science program as discussed in [7, 8]. Higher learning institutions would need to put in place a deliberate strategy that seeks to overcome the gender disparity in the long term, possibly might attempt to reach out and engage female practitioners in the industry and schools to consider opting for ICT related programs. In this there is a lot that higher learning institutions can learn from a number of initiatives that the dLab had pioneered to promote women and young girls' involvement in data and data driven innovations. It is also noted that most data practitioners are young people below 35 years while 95% are under the age of 45. In such age group it is likely that a practitioner has acquired a relatively good experience and a career focus, hence also suitable to enroll for a postgraduate education.

There are delivery constraint for learning institutions to consider in terms of appropriate mode of learning for a cross-section of potential learners which indicate varying needs. It is clear that although most practitioners had showed willingness to take a stride for data science education, they are constrained with keeping their jobs, therefore, it is not surprising for the overwhelming majority being in preference of undertaking an evening mode of learning. That is, the evening mode of learning is the most preferred learning mode for postgraduate education as most candidates are on employment thus can find time for studies after work.

In spite of uncovering the opportunity for data science education, the issues of policy and decision making to leverage the uncovered opportunities should came into spotlight. On one hand it has been shown that the most preferred learning mode is part-time/evening mode, but as part of curriculum implementation process, it emerged that current curriculum approval policies in some higher learning institutions including the UDSM, did not support evening mode, which impact on the majority learners' constraint. On the other hand, the country being part of the OGP membership acted as a catalyst for increased data related activities leading to demand for data science skills across national sectors. It is likely that the recent national policy shift leading to withdrawal from OGP and the further tightening to access and use of public statistics through the new enacted statistical act may humper the dynamics for uptake of and anticipated outcome of strengthening the data science education.

6 Conclusion

As for the global trend, in Tanzania data scientists will rapidly be in demand following technological adoption in new and existing business organizations faced with global competitions, leading to appetite for growth through evidence based planning and decision making. Our analysis indicates that data science field attracts growing importance nationally as demand for skilled data scientists to cater for envisaged growing demands, fueled by country policy drive for industrial economy. It is essential for local performance monitoring and the global requirement for SDGs monitoring and reporting. Higher learning institutions need to actively play the role of mainstreaming the data science education in order to sustain the near term and future supply of data scientists.

References

1. J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," 2011.
2. T. Becker, E. Curry, A. Jentzsch, and W. Palmetshofer, *New horizons for a data-driven economy: Roadmaps and action plans for technology, businesses, policy, and society*. Switzerland: Springer International Publishing AG, 2016.
3. W. M. P. Van Der Aalst, "Data Scientist: The Engineer of the Future," in *Enterprise Interoperability VI: Interoperability for Agility, Resilience and Plasticity of Collaborations*, Enterprise., K. Mertins, Ed. Switzerland: Springer International Publishing Switzerland, 2014, pp. 13–26.
4. P. Mikalef, M. Giannakos, I. Pappas, and J. Krogstie, "The Human Side of Big Data Understanding the skills of the data scientist in education and industry," in *IEEE EDUCON 2018 Global Engineering Education Conference*, 2018, no. January, pp. 503–512.
5. I. Demchenko and A. Belloum, "EDISON: Discussion Document: Part 1. Data Science Competence Framework (CF-DS) release 2," 2017.
6. I.-Y. Song and Y. Zhu, "Big data and data science: what should we teach?," *Expert Syst.*, vol. 33, no. 4, pp. 364–373, 2016.
7. L. Holman, D. Stuart-fox, and C. E. Hauser, "The gender gap in science : How long until women are equally represented ?," *PLOS Biol.*, pp. 1–20, 2018.
8. V. Zhang, "Breaking Down The Gender Gap In Data Science," 2018. [Online]. Available: https://www.forbes.com/sites/womensmedia/2017/08/03/breaking-down-the-gender-gap-in-data-science/#d2aac904287e. [Accessed: 12-Oct-2018].
9. E. and M. National Academies of Sciences, Data Science for Undergraduates: Opportunities and Options. 2018.
10. Data Science Africa, "Data Science Africa," *Data Science Africa*, 2018. [Online]. Available: http://www.datascienceafrica.org/. [Accessed: 12-Oct-2018].
11. AIMS, "a Data Science Intensive Program," 2018. [Online]. Available: https://aims.ac.za/2018/07/02/data-science-intensive-program/. [Accessed: 11-Oct-2018].
12. Wits University, "Big Data Analytics," 2018. [Online]. Available: https://www.wits.ac.za/course-finder/postgraduate/science/big-data-analytics/. [Accessed: 12-Oct-2018].
13. OGP, Available: https://www.opengovpartnership.org/. [Accessed: 12-Oct-2018].
14. "Basic Statistics Tanzania," 2018. [Online]. Available: http://www.opendata.go.tz/en. [Accessed: 12-Oct-2018].

15. dLab, “Promoting innovation and data literacy through a premier cente of excellence .,” 2018. [Online]. Available: https://dlab.or.tz/. [Accessed: 12-Oct-2018]
16. MOF, “The Tanzania Development Vision 2025”, 1995[online]. Available: http://www.-mof.go.tz/mofdocs/overarch/Vision2025.pdf. [Accessed: 31-01-2019]
17. MoWTC, “National Information and Communications Technology Policy”. Available: https://tanzict.files.wordpress.com/2016/05/national-ict-policy-proofed-final-nic-review-2.pdf [accessed: 31-1-2019]