How to train a data scientist for the Global South?

Eeva NYGREN, Martti SUTINEN, Tomi WESTERLUND, Erkki SUTINEN
University of Turku Department of Future Technologies, Turku FI-20014, Finland
eeva.nygren@utu.fi, martti.sutinen@gmail.com
tomi.westerlund@utu.fi, erkki.sutinen@utu.fi

Abstract: The extensive use of mobile devices and modern technologies has led to the expansion of amount and quality of data in the Global South. With the help of professional data scientists, developing countries could benefit from big data effectively and leapfrog from traditional methods of data generation to modern data handling methods without going through the same process as in wealthier regions. The sky-high demand for data scientists is due to the enormous business value that they can bring to the table, deriving deep learning from the big data. In this paper, we give one solution how to train a data scientist for the Global South where data scientists face partly different challenges as their colleagues in the Global North. Based on more than ten years of collaboration in ICT education with various African partners, we developed the general framework for an innovative BSc degree programme in computer science with specialisation in data science including two essential threads - *learning-by-being* and *learning-by-doing.* Only by combining learning to both doing and being, we can train data scientists that are able to change the target context in a profound way. A built-in research program on the data science training increase Global South – North collaboration, which in turn combined with open collaboration and flexible way of overcoming hierarchical boundaries most probably leads to the innovative joint ventures.

Keywords: Data science, global south, education.

1. Introduction

Today's society is completely dependent on Information and Communication Technology (ICT). The amount of data in our world has been exploding and new technologies are emerging to organise and make sense of "big data" [1] [2] [3]. The analysis of big data and techniques mastering and getting use of Big data are comparable to "Industrial Revolution" in terms of technological innovations, structural change, and the sources of economic growth [4]. According to Ksehtri *et al.* [4] "a growing number of interesting and creative ideas and techniques involving big data and the cloud are being developed and deployed throughout the developing world and in different contexts" and "the uses of these technologies by businesses, governments, non-government organizations, and consumers are rapidly increasingly in the developing world" (pp.19-20). Data is seen even as a strategic asset driving or determining the future of science, technology, the economy, and possibly everything in our world today and tomorrow [2]. Data scientists not only help the business world but also solve problems related to human global challenges, like forced migration, disease, poverty and economic stagnation, and ecological and environmental crises [5]. Smart applications of data science, and challenge-aware data scientists in particular, can empower the latent human capital.

Mobile penetration is growing fast especially in developing countries, where the number of mobile-broadband subscriptions grows annually faster than anywhere else

reaching a penetration rate of close to 41% [6]. Owing to the fast growth of mobile technology usage as well as technologies like satellites, biometric tools, machine-to-machine communications, cheap sensors, social media sites and cloud-based storage the amount of data is increasing enormously [3]. However, available data is underutilized: there are plenty of possibilities that were non-existent before [5]. The demand for competent data scientists in the Global South is an emerging challenge that reflects the global need for data scientist, estimated to be two million worldwide [3]. Most of these professionals are expected to have a pragmatically oriented BSc degree in data science. Data scientists are valuable team members, but a rare find in the talent market. The scarcity is due to the skill set being very difficult to acquire, requiring (1) advanced math expertise, (2) solid technical skills, and (3) strong business acumen. Furthermore, if data scientists are supposed to give solutions to global challenges and change the world, they need multidimensional and comprehensive skills in addition to the more traditional skill set mentioned above.

To answer the challenge to educate new breed of data scientists for the Global South, we have developed an innovative BSc degree programme in computer science with specialisation in data science. The programme not only utilise the learning-by-doing model but also a learning-by-being model, in which students learn being skills referring to their assets in context awareness, sensibility and presence, and, thus, do not reduce to so called soft skills.

In this paper, we first described an overview of the problem domain and proposed solution. In the following section, we provide background context on the skills required and some use cases, and in Section 3 we give one solution how to train data scientists for this specific context. In the last section, we discuss how this solution benefit different parties involved.

2. Data science as a profession

In this section we discuss what data science is and what kind of skills a data scientist should have to be able to face and overcome the challenges in the Global South.

2.1 What is data science and how to benefit from it?

Data science is all about leveraging the available knowledge in data, giving new tools for business decision making and optimizing business processes, even creating new products and services. This is achieved by feeding historical and real-time data to algorithms which produce insights, predictions and recommendations on optimal actions. The algorithms are borrowed from research in multiple fields including information technology, engineering, statistics, artificial intelligence, computational linguistics, etc. [2]

Of course, data is already being used in controlling business processes [3]. However, recent advancements in processing and storage technology have made it possible to feasibly analyze data with more versatile methods. For example, to identify customers who are at risk of taking their business elsewhere, i.e. churning, there might be a simple manmade rule tracking their order volumes. This intuition driven approach can now be augmented or replaced with a data driven classifier, which learns patterns leading to churn events from a large history of data.

How to start benefiting from data science? One option is to first try products with existing data driven features. However, these may lack appropriate industry templates or underperform in the context of your business. The other option is to invest in data science professionals and produce tailored solutions. The demand for data scientists will be much greater than supply[1]. Gartner estimates that in 2018 more than half of large organizations

[1] https://www.datanami.com/2016/03/25/tracking-data-science-talent-gap/

will compete using advanced analytics.[2] Predictive analytics deployments have seen returns on investment (ROI) of one- to ten-fold.[3]

2.2 Example use cases

Below are some examples of how data science accelerates business in different industries.

Manufacturing / Predictive purchasing of raw materials: Imagine you are sourcing raw materials from South Africa. It will take a given amount of days before the materials reach your plants. Every now and then, the inventories run low before your employee makes the new order and a production line must be halted. This costs as unrealized sales or unsatisfied customers. But what if new orders were placed automatically taking into account the raw material demand and delivery times? Using sensor or camera data to track raw material volumes or simply existing production and raw materials order books, one could simulate inventory development and automate raw materials orders just-in-time.

Agriculture / Estimating soil condition based on historical weather and sensors: In order to increase crop yield, the soil condition should be kept good. Satellite images, rainfall history, and irrigation system data, as well as scheduled measurements, could be used to continually estimate soil condition. This enables focus on correct areas of the fields.

Tourism / Mining the social web for tourism trends: It is difficult to stay up-to-date on the latest discussion on tourism trends. There are just so many online channels where people share experiences and opinions. Text analysis can be utilised to structure and summarise this discussion into topics, sentiments, and influencers. Topics of interest serve as a basis for planning new tourist activities.

Marketing / Segmenting customers based on interests and spending patterns: Understanding customer segments is important. How many naturally occurring segments are there and what kind? What is the current ratio of high-value visitors and what activities are they interested in? Data can be used to explore clusters of similar customers and their traits. This will be helpful in guiding marketing and tourism infrastructure investment decisions.

Education / learning analytics: Data science has not only potential to help improve the quality of teaching, learning, and education management [7], but also potential to create new models to learning [8]. Learning analytics i.e. analytics and data mining applied to education bring new models that are also are suitable for many low-performing education systems characterized by poor-quality teaching and a lack of educational resources. Increased interest in mobile learning is resulting in more mobile learning initiatives and learning happens both formally and nonformally with help of mobile devices.

2.3 Data science best practises

Doing data science is not just about one person but a team backed by the whole organization. The most important factor for succesfully applying data science is senior commitment. It is important that the company management believes that data driven enhancements bring value and translate to earnings and satisfied investors.

Secondly, it is a cultural change and a matter of gaining trust through positive experiences. Many employees do not currently use advanced analytics but rely on reports that they are used to. Supporting new ways of doing things and requiring the data science team to communicate with the rest of the organization effectively is essential success factors.

[2] https://www.gartner.com/newsroom/id/3192717
[3] http://www.ibmbigdatahub.com/blog/analytics-roi-strategy-execution

Lastly, the data science team should be versatile and ever evolving, a center for innovation, fun, and results. Data science is a mixture of business domain knowledge, analytical and methodology capabilities, and technological skills for implementation. It also requires a creative visualization, design and communication orientation.

Data science projects last between one and six months and are very natural to manage in an agile manner. There is an existing industry framework, the Cross Industry Standard Process for Data Mining, to use as basis for project planning [9].

When viewed with a larger scale, McKinsey Global Institute [3] points out that "policy makers need to recognize the potential of harnessing big data to unleash the next wave of growth in their economies" (p. 23). Issues they should take into account are: providing the institutional framework to allow companies easily create value out of data, ensuring that data do not live in isolation bringing a large variety of different data sources together, protecting the privacy of citizens and providing data security, tackling the shortage of talent through education and immigration policy, organizing infrastructure such as communication networks, accelerating research in selected areas including advanced analytics; and creating an intellectual property framework that encourages innovation [3, 5].

2.4 Data scientistis perspective to the Global South

The skills and competencies of IT professionals differ in the Global South and the Global North [10] . Not only skills and competencies, but the amount of available data differs hugely. Chandy *et al*. [10] points out that "the information vacuum that still exists in many developing countries makes the potential for impact from big data much greater in these contexts". They also note that historically the situation of data context in many Global South countries has been the opposite of big data in terms of limited volume, limited variety, and limited velocity; even government statistics have been sparse and outdated, sharing of the data restricted, and formats of data incompatible or data is not even being digitized. [5].

Recently the amount and quality of data in the Global South has increased vastly as a result of the use of mobile devices and modern technologies by businesses, governments, non-government organizations, and consumers [4]. It is possible for developing countries to leapfrog from "small data" (traditional methods of data generation) to big data without going through the same process as in wealthier regions that have many more legacy data sources [5]. For example, when compared to the developed countries, the emerging economies in Asia (e.g., Indonesia and India) and Latin America (e.g., Argentina, Mexico) are experiencing the highest growth rates for cloud services [4].

For data Scientist the amount of data is only one factor in his/her profession. Data scientists are storytellers. Their task is to connect the dots between data fragments – create a story and visualise it. Storytelling connects us back to the Global South where storytelling has a long tradition. It all begins with understanding the audience. The listeners have their own backgrounds and interests. Data scientist has to create a new story of the data under review. Data scientist has to use data in a meaningful way and create a plot with characters, each overcoming or falling for a challenge. Data scientist has to be able to structure the objectives and needs of the audience into questions that can be answered with data. Data scientist becomes both creative with and critical of data. He/she begins to think in terms of information rather than details first. Data scientist learns to explore data, transforming it and visualizing it to show different aspects of the underlying phenomena. Instead of forcing the real world fit to his single model, he/she wants to understand many of the world's complexities through several lenses, developed with iterative experiments.

Data scientist must become familiar with common ways to approach data-driven problem solving, e.g., descriptive, diagnostic, predictive, or prescriptive techniques. Be it

clustering, classification, regression, linear programming, manifold learning, neural networks, or topic modeling, Data scientist understands what questions can be answered with given techniques and information. Data scientist is wary of common mistakes; missing or wrong data, data with exceptions or outliers, overfitting to learn noise instead of the signal. Inspiration comes from many fields of science; statistics, systems analysis, engineering, computer science, linguistics, behavioral science. The next-generation Data scientists should be multidisciplinary experts, or able to collaborate with domain-specific specialists and subject matter experts to achieve broader impact [2]. Data scientist develops an abstract mind, turning his characters' challenges into questions and techniques to answer them with data.

One more little thing. Implementation. Mostly Data scientist works using information technology: Data collection, storing, processing, and analysis require multiple tools. When Data scientist makes choices, Data scientist has to be mindful of the restrictions he/she scientist is placing on results. Not all phenomena generate data that is inherently relational or comes in batches. Data scientist has to be ready to get his hands dirty and program his/hers way to results. Some 90% of the time will go to getting and cleaning data to be able to answer the questions with various analysis.

3. Solution

As a solution to the question "How to train data scientists for the Global South?", we have designed a new 30-month intense and face-to-face BSc degree program[4]. The general framework for data science programme is developed iteratively using design science [11] as a methodology. We have extensive experience of the uses of information and communication technologies for development (ICT4D), computer science and ICT education, creative problem solving, and contextual design. Specific examples are designing and implementing contextualized Computer Science undergraduate program in Tanzania [10] and 2-years STIFIMO consulting for science, technology and innovation – program in Mosambique. Iterative development of data science program and its general framework (Fig.1) included co-operation with researchers, teachers and business sector from the Global South. We also visited Banglades, Ghana, Mosambique, Namibia, Nigeria and South Africa especially for the development of this program during years 2016-2017.

We identified two essential threads - *learning-by-being* and *learning-by-doing* - which are balanced and built-in for all courses. Learning-by-being provides future data scientists with *being skills*, whereas learning-by-doing induces *doing skills*. With *doing skills* we mean traditional cognitive skills, and with being skills we mean deeper layer of non-cognitive or soft skills including things like presence, listening, awareness, patience, vulnerability and authenticity [12]. Considering contextualization, both intercontextuality and context sensitivity are important. While context sensitivity emphasizes that learning (mostly by doing) takes place outside of the learner and the learner needs to take the external into account by being sensitive, inter-contextuality stresses that learning (mostly by being) takes place within a community of which the learner is a member of.

[4] www.databachelor.com

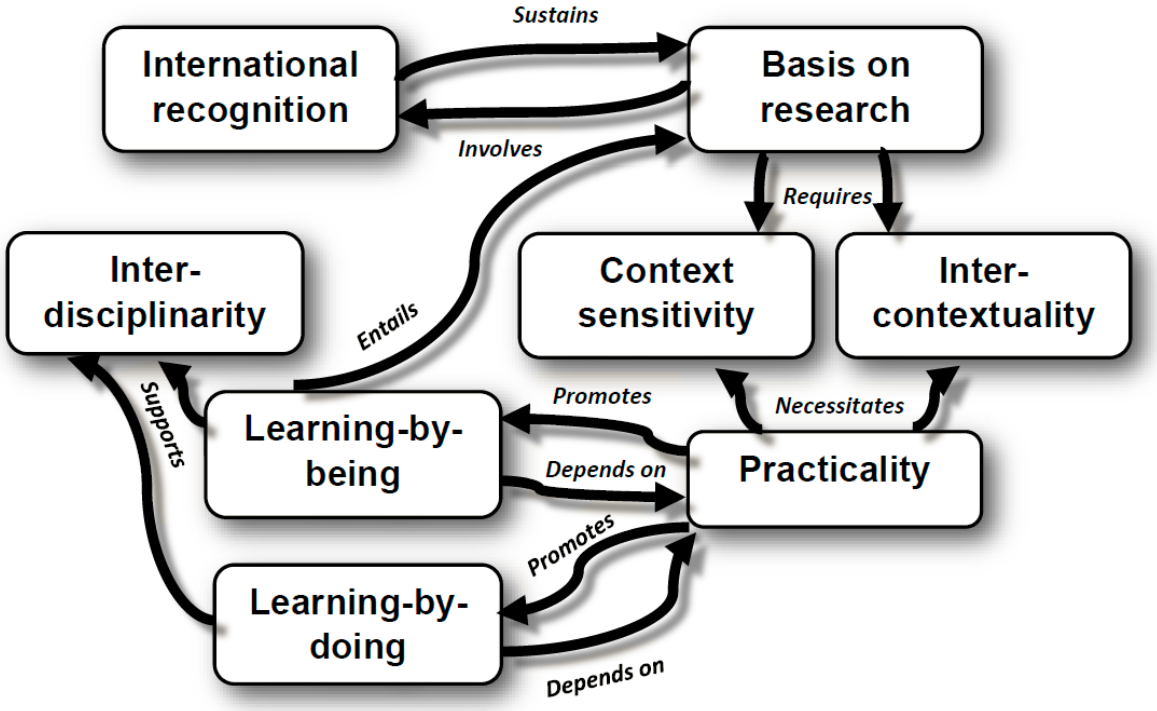

Figure 1: The general framework for data science programme

Table 1 illustrates the complementary approaches of learning-by-doing and learning-by-being. A plain learning-by-doing approach easily leads to training professionals who can solve a given problem without questioning its relevance (that would require a learning-by-being aspect). Data scientists trained as doers are good at incremental innovation: they create solutions that change aspects of the context, without a profound impact. In the context of the Global South, the outcomes of incremental innovation are sometimes called white elephants, or innovations without value. Those trained as plain be-ers, may understand the urgency of a radical change in the context, but due to lacking doing skills, cannot reach the goal. Thus, only by combining learning to both doing and being, we can train data scientists that are able to change the target context in a profound way.

Table 1. Learning-by-doing and Learning-by-being

	Learning-by-doing	Learning-by-being
Mode of learning	Pragmatic orientation to solve problems in the context	Interpretive orientation to receive from and understand the context
Direction	From the learner towards the outside environment	From the outside environment towards the learner
Leads to	Changes in the context	Changes in the learner's mindset
Orientation	Observation	Presence
Basis of curriculum	A predefined set (but not necessarily a sequence) of technical or soft skills that the learner is supposed to internalize by doing and external activities	A given community and context that the learner is getting increasingly aware of during the learning process.

3.1 Learning-by-being

Being skills are acquired by learning by being element that means using *community pedagogy*. The University of Turku, Finland, co-operates with a specific boarding school, Folk high school, and trains data scientists in an intensive, safe, and creative full-board learning environment. Community pedagogy has its roots in the early 20th century Danish

folk high school movement, which originated from the educational ideas of N. F. S. Grundtvig [13], who emphasized the "living word" and the importance of local people's experiences. Grundvig wanted to give education for the local people to enable them to take responsibility of the coming democratic society, and believed that dialogue based “living word” gives youth inspiration and kindles them to action - it is something more than the “dead word” i.e. academic knowledge of the Latin schools [13].

A community that our model requires sets out a microecosystem where learning takes place. The microecosystem serves as a living laboratory where future data scientists explore how data science can transform its immediate context in a safe environment that allows learning from mistakes, whether made in doing or being. While creating a culture where innovation thrives, the students are involved in full-time studies in intensive groups of 25 persons, including lectures, exercises, projects, and, furthermore, of personalized study and effective career counselling by a personal tutor. Like the students, also their teachers will live for lengthy periods in the same premises.

Stories have been successfully utilized in education from pre-primary to primary, secondary, and tertiary education and beyond to adult education (in-service, non-formal and informal). While a data scientist learns to become a storyteller of the data outside their personality, they will – at the same time, as another thread – compose a story of their personal learning odyssey that takes place in the community.

3.2 Learning-by-doing

Doing skills are more traditional cognitive skills that are obtained by learning-by-doing element. The studies include many projects and are open to collaboration with students’ employers and other partners from private, public and third sectors especially in the global South. The cognitive skills are demanding, because the program trains a new generation of highly competent specialists to solve real-world problems using complex and unstructured data. After the studies, the graduates will be able to identify computationally solvable problems rooted in human situations, design and create mobile applications that implement to solutions, analyze the data collected by the heavy use of mobile applications, market and sell the solutions, and operate in intercultural markets.

The data science BSc degree program offers an open and unique platform for extensive co-operation between companies, universities, and other public sector organizations. Experimental learning methods utilizing cooperative work and learning-by-doing will enable interdisciplinary real-life originated projects. Researchers and experts from different fields are being invited to join the projects and guide the students. Professors from diverse countries, cultures and disciplines enrich the research. Collaboration with international industry offers highly topical themes for projects.

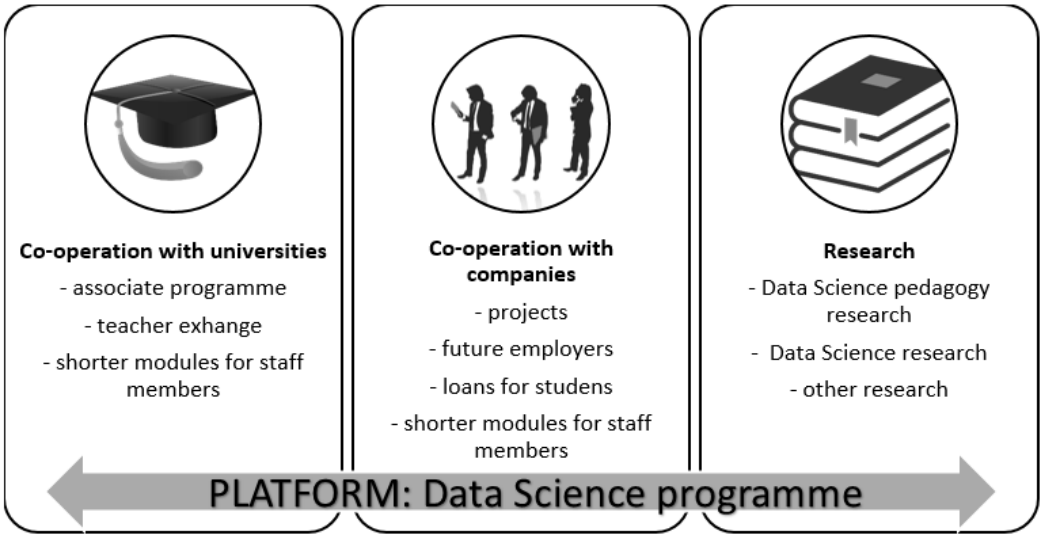

Figure 1: Data Science programme offers an open and unique platform for extensive co-operation

3.3 Curriculum

Data science bachelor programme's major subject is computer science. The curriculum includes 30 ECTS basic studies and 58 ECTS subject studies as well as Innovation & Business Creation (25 ECTS), Technologies and humanities (25 ECTS), Language and communication studies (17 ECTS) and Mathematics for data scientists (25 ECTS). The learning process proceeds in a cycle during which the student learns to answer different questions through the data (Figure 2).

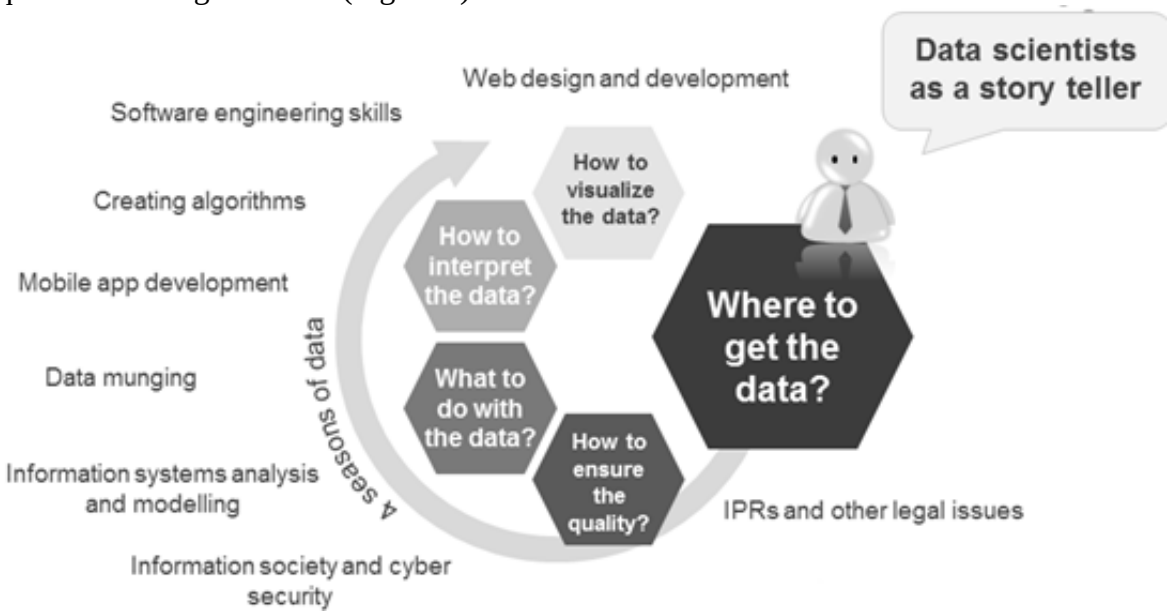

Figure 2: Contents of the data science programme

3.4 Prospects of Graduates' of the programme

The prospects of graduates from the programme are excellent and versatile. The degree gives their eligibility to work in industry or to pursue another degree at the University of Turku or any other prestigious university. Especially the African private, public and third sectors will highly value their education, and thus make them as a prominent employer for the graduates. To further improve the employability, pursuing another degree will be very

profitable for the graduates. Even continuing to postgraduate research training and getting a doctoral degree is possible for the most talented graduates especially in Finland, where PhD degrees are still free from tuition fees.

4. Discussion and Consideration for training data scientists

Increased amount of big data brings opportunities to accelerate business in different fields like manufacturing, agriculture and education. There is a high demand for data scientists due to the enormous business value and solutions to global challenges that they can bring. In this paper, we give one solution how to train a data scientist for the Global South where data scientists are working in specific environment. Based on more than ten years of collaboration in ICT education with various African partners, we have derived a two-thread approach to train data scientists for the Global South. These two-threads, which form the core for our new data science BSc programme, are learning-by-doing and learning-by-being. These elements induce both traditional cognitive skills as well as deeper soft skills. Other important factors of the programme are context sensitivity and intercontextuality, which both helps in creating data scientists competent to change the target context by creating stories from different data sets from different contexts.

We believe, that a built-in research program on the data science training and open collaborative leads to better quality of data scientists. The annual number of students in this 30 months' training programme is quite small, so it is not enough to fulfil the great need of workforce emerging in the Global South. However, we see that this new kind of education is a very important trendsetter in the data science field. We will develop new methodology for data science education; students do not only learn technical skills, but also thinking and problem solving skills. This relatively small group works as an innovative demo environment where teaching methods can be experimented and curriculum for larger mass education improved. The programme contains real projects from the Global South context enabling learning from them co-operatively with partners from companies, third sector and universities, and developing together data science education especially for the Global South needs. Simultaneously, students learn how to apply technical skills to practice. Companies can send their staff members to participate in the short courses as students or as teachers, and remote connections enable participation even from the other side of the world. We also co-operate with the Global South universities; their teachers, researchers and students are able to participate in the different courses for shorter or longer periods. Teachers and researchers give their own expertise of local context, and students from various countries guarantee multidimensional viewpoint to course content. It is also possible to develop joint degrees with the Global South universities. Altogether, Global South – North collaboration increases and the new way to overcome hierarchical boundaries hopefully leads to the joint ventures. With the help of experimental information obtained from the students and co-operators, we are able to continue developing the innovative framework as well as give recommendations how to combine learning-by-doing and learning-by-being also outside data science education.

References

[1] D. Lazer, R. Kennedy, G. King and A. Vespignani, "The parable of Google Flu: traps in big data analysis," *Science, 343(6176),* pp. 1203-1205, 2014.

[2] L. Cao, "Data science: a comprehensive overview," *ACM Computing Surveys (CSUR),* vol. 50, no. 3, p. 43, 2017.

[3] McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity.," https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf, 2011.

[4] N. Kshetri, T. Fredriksson and D. C. R. Torres, Big Data and Cloud Computing for Development:

Lessons from Key Industries and Economies in the Global South, Taylor & Francis, 2017.

[5] R. Chandy, H. Magda and P. Mukherji, "Big Data for Good: Insights from Emerging Markets," *Journal of Product Innovation Management,* vol. 34, no. 5, p. 703–713, 2017.

[6] ITU [International Telecommunications Union], "ICT facts and figures 2016," ITU, https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf, 2016.

[7] J. Traxler and S. Vosloo, "Introduction: The prospects for mobile learning," *Prospects,* vol. 44, no. 1, pp. 13-28., 2014.

[8] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics.," in *Learning analytics*, New York, Springer, 2014, pp. 61-75.

[9] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models.," *The Knowledge Engineering Review,* vol. 21, no. 1, pp. 1-24, 2006.

[10] M. Tedre, F. D. Ngumbuke, N. Bangu and E. Sutinen, "Implementing a contextualized IT curriculum: Ambitions and ambiguities," in *In Proceedings of the 8th International Conference on Computing Education Research. ACM.*, 2008.

[11] A. Hevner and S. Chatterjee, "Design science research in information systems," in *Design research in information system2*, Springer US, 2010, pp. 9-22.

[12] J. Ellard, "Doing Skills + Being Skills = Career Success," HuffPost Contributor Platform, https://www.huffingtonpost.com/entry/doing-skills-being-skills-careersuccess_us_597a1726e4b09982b73762b7, 2017.

[13] J. Kulich, "The Danish folk high school: can it be transplanted? The success and failure of the Danish folk high school at home and abroad," *International Review of Education*, vol. 10, no. 4, pp. 417-430, 1964.