Architecting Data Science Education

Vadim Ermolayev[1[0000-0002-5159-254X]], Mari Carmen Suárez-Figueroa[2],
and Oleksii Molchanovskyi[3]

[1] Department of Computer Science, Zaporizhzhia National University, Ukraine
vadim@ermolayev.com
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
mcsuarez@fi.upm.es
[3] Ukrainian Catholic University, Lviv, Ukraine
olexiim@ucu.edu.ua

Abstract. Data scientists are currently among the most demanded professionals in many spheres, including industries, governments, public sector, among others. This is due to several good reasons. Probably an important one of those reasons is the growing demand to find proper ways to face the challenges of establishing data-driven economies and societies. As academics and educationalists, but also Data Science professionals, we look at how to bring up this kind of specialists such that to meet the current shortages but also mid-term demands. In this position paper we deliberate about how to architect thematically, didactically, and organizationally a university program under the thematic umbrella of Data Science. We focus on the selection of learning units or disciplines to be covered in order to produce the M.Sci. and Ph.D. graduates who will be ready to face the future challenges in the mid-term perspective. We outline our recommendation on using learning tools and materials. We also concisely present the approach for stimulating competitive and cooperative atmosphere in the class that stimulates intensive collective and individual learning. We recommend to reinforce an academic program by involving industrial partners intensively in the process. We ground our deliberation on our experience in implementing relevant M.Sci. and Ph.D. programs in Data Science and Semantic Technologies.

Keywords: Data Science education, topical scope, program structure, learning tools, didactics, collaboration with industries.

1 Introduction

The boost in the abundance, complexity, and variety of data in all spheres of human activity is a phenomenon that leaves a rare information professional negligent these days. Industries are entering into data driven economy, which demands having and using data as a primary asset. On the other hand, the shift to more intensive use of data results in the increase of data generation and storage at unprecedented scales in terms of volumes and rates. A few topical examples are as follows (c.f. [1]):

> “Exponential growth of data volumes is accelerated by the dramatic increase of social networking applications that allow non-specialist users create a huge amount of content easily and freely. Equipped with rapidly evolving mobile devices, a user is becoming a nomadic gateway boosting the generation of additional real-time sensor data. The emerging Internet of Things makes everything a data or content, adding billions of additional artificial and autonomic sources of data to the overall picture. Smart spaces, where people, devices, and their infrastructure are all loosely connected, also generate data of unprecedented volumes and with velocities rarely observed before.”

Hence, data generation is a phenomenon that fuels itself and so far we do not observe any signs of saturation for this process. Straightforwardly, the societal demand for the professionals capable of efficient and effective processing of these data also increases at unprecedented rate. These gave rise to Data Science as a discipline and community. As denoted by Hoehndorf and Queralt-Rosinach [2]:

> “Data Science has as its subject matter the extraction of knowledge from data. While data has been analyzed and knowledge extracted for millennia, the rise of “Big” data has led to the emergence of Data Science as its own discipline that studies how to translate data through analytical algorithms typically taken from statistics, machine learning or data mining, and turning it into knowledge. Data Science also encompasses the study of principles and methods to store, process and communicate with data throughout its life cycle, and starts just after data has been acquired”.

A data scientist is currently one of the most requested and highly paid jobs. The reason for it is the lack of such professionals in industries, but also in academia.

It is widely acknowledged that Big Data, which is the area of our interest, begins when the traditional methods for processing data do not work due to the excess in volume, variety, velocity, or complexity. The phenomenon of Big Data causes also a conceptual divide in the Data Science community in broad. Enthusiasts propagate that, faced with real big data, a scientific approach “… hypothesize, model, test – is becoming obsolete. … Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns …” (c.f. [3]). Pessimists however point out that Big Data is often not healthy, as it provides “… destabilizing amounts of knowledge and information that lack the regulating force of philosophy” (c.f. [4]).

Academia has to respond to this challenge by providing professionals capable of dealing with this phenomenon following a balanced path that equally accounts to the highlights and lowlights of both optimistic and pessimistic approaches. A pursuit to such a balanced path gives us a hint about what is the shortage in the required skills for a Data Scientist.

This is exactly the way we follow to architect and deploy the related programs, at M.Sci and Ph.D. levels. There are some results on this way which we want to share in this paper. To help a reader better understand our approach we structure this presentation along the three important facets that form, so to say, the environmental grid. These are thematic, didactic, and organizational.

The remainder of the paper is structured as follows. Section 2 focuses on the related work and therefore discusses the most prominent relevant academic programs to date. Section 3 deals with our approach to form the topical scope of the academic programs at M.Sci. and Ph.D. levels. It is also about choosing the teaching and learning tools and also didactics that help make, in our opinion, teaching and learning Data Science more efficient and effective. Section 4 is about our approach to propose a proper organizational environment for our students that enables seamless bi-directional interaction with the data generating and processing stakeholders in industries. Here we also report about our experience in having different kinds of cooperation, between universities and also with industrial partners, that help us achieve a surplus in providing quality education to our students. Finally, in Section 5 we draw some conclusions and outline our plans for future work.

2 The Most Prominent Related Work

There are an increasing number of institutions from around the World now offering M.Sci. courses and also Ph.D. programs in Data Science. Perhaps one of the most prevalent international efforts in Europe is EIT Digital Master Program in Data Science[1]. This Master offers to the students the possibility of studying data science, innovation, and entrepreneurship at leading European universities. In this program, students will learn about scalable data collection techniques, data analysis methods, and a suite of tools and technologies that address data capture, processing, storage, transfer, analysis, and visualization, and related concepts (e.g., data access, pricing data, and data privacy). This two-year Master promotes the geographical mobility by means of studying in universities in two different European cities.

One more relevant European initiative was the European Data Science Academy (EDSA)[2] an H2020 EU project that has been effective in 2015 - 2018. The objective of the EDSA project was to deliver the learning tools that are crucially needed to close the skill gap in Data Science in the EU. The project spinned off an Online Institute**,** based on the project foreground, to leverage the outcomes of the EDSA project. The Institute will continue to be operated by the EDSA project partners beyond the lifetime of the project: Open University (UK), University of Southampton (UK), Institut Josef Stefan (Slovenia), Fraunhofer Institut (Germany), KTH Royal Institute of Technology (Sweden), ideXlab (France), Persontyle Limited (UK), Technische Universitaet Eindhoven (TU/e) (the Netherlands), Open Data Institute LBG (UK).

At National scale in Europe, the Italian Data Science PhD program[3] needs to be mentioned. Data Science PhD is a joint initiative by Scuola Normale Superiore, University of Pisa, Sant'Anna School of Advanced Studies, IMT School for Advanced

[1] https://www.tue.nl/universiteit/faculteiten/wiskunde-en-informatica/studeren/graduate-programs/masteropleidingen/eit-data-science/

[2] http://edsa-project.eu/

[3] http://datasciencephd.eu/

Studies Lucca, and National Research Council. This program develops a mix of knowledge and skills on the methods and technologies for: the management of large, heterogeneous, and complex data; data sensing; data analysis and mining; data visualization and storytelling; understanding the ethical issues and social impact of Data Science. Ph.D. students admitted to the program have an opportunity to develop data science projects in different domains.

While Data Science is a relatively new term, the academic programs (M.Sci. and Ph.D. levels) in the neighboring areas, such as Statistics, Business Analytics, Artificial Intelligence, and Machine Learning, existed for a while. The appearance of the new term, Data Science, allowed to introduce "an umbrella" for many programs. That led to their (programs) broader comparison and analysis. A lot of web resources appeared for that purpose. For instance, "23 Great Schools with Master's Programs in Data Science"[4] that lists M. Sci. degree programs in Data Science in the U.S. universities. One of the most completed (and continuously updated) list of the academic programs in the field is provided by the Data Science Community list[5]. The list counts almost 600 colleges.

The structure and the curricula of the majority of the academic programs in Data Science are relatively the same worldwide - all admitting its interdisciplinary nature and synergetic character (e.g. [5]). There are courses focused on ramping-up the students regarding the necessary mathematical background (theory of probability, statistics, time series, linear algebra, etc.). Another group of courses teaches data and software engineering (mostly databases, database management, and related software infrastructures). The core group of courses, including machine learning, data mining, deep learning with various modifications provide the competencies in relevant enabling technologies for data scientists. Finally, there is a business or application oriented group of courses. This group teaches how the technologies could be effectively applied in specific business tasks and for solving specific business problems. Application domains are seen quite broadly and span across marketing analytics, natural language processing, computer vision, bioinformatics, etc. In the recent years Data Science curricula started to integrate the courses that take into account ethical problems in the context of data processing, analytics and interpretation. Some programs see the landscape for ethical issues even broader and include Data Science and Artificial Intelligence applications.

In terms of building Data Science curricula, the Data Science Model Curriculum [6], created in the context of the EDISON project[6], is of particular relevance. This Model Curriculum was built as a part of the EDISON Data Science Framework (EDSF) that provided a foundation for the Data Science profession definition. The EDSF includes the following core components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional Profiles definition (DS-PP).

[4] http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/

[5] http://datascience.community/colleges

[6] http://edison-project.eu/

3 Topical Scope and Didactics

In this section we present our positions by putting together: a structural and topical organization of a Data Science program, mainly for its M.Sci. level; our views on appropriate tools and media for teaching and learning; and some tips on didactics that helps make the learning process more effective and efficient.

3.1 Topical Scope

For architecting Data Science Education the approach elaborated in the EDISON project looks like very relevant. The EDISON approach to defining the Data Science Model Curriculum followed a competence-based education model and has been summarized in the following **steps** (c.f.[6]):

1. For each enumerated competence from CF-DS, the Learning Outcome is defined according to knowledge or mastery level (Familiarity, Usage, Assessment for current MC-DS version)
2. A DS-BoK includes Knowledge Area Groups (KAG) from the available BoK elements and also those defined based on the 2012 ACM Computing Classification System[7] (CCS 2012)
3. Each Knowledge Area Group (KAG) is mapped to existing academic subject classification groups that are based on ACM CCS 2012 complemented with the domain or technology specific classifications to be defined by subject experts.
4. For each KAG or Knowledge Unit, related Learning Units are specified according to academic subject classification or current university practices
5. For each Learning Unit, its category as core/mandatory (Tier 1 or Tier 2), elective, or prerequisite is assigned
6. For Core and Elective Learning Units, the list of Learning Outcomes is defined

In addition, as suggested by the Government of the UK in their Guidance document[8], a data scientist has to have, among others, the following **skills**:

- Good knowledge about applied mathematics, statistics and scientific practices
- Good knowledge about data engineering and manipulation

To best account for an interdisciplinary and synergetic character of Data Science and also its focus on practice, [7] suggests the following **guiding principles** to forming the curricula:

- Organize the course around a set of diverse case studies
- Integrate computing into every aspect of the course
- Teach abstraction, but minimize reliance on mathematical notation
- Structure course activities to realistically mimic a data scientist's experience

[7] https://www.acm.org/publications/class-2012

[8] https://www.gov.uk/government/publications/data-scientist-skills-they-need/data-scientist-skills-they-need

- Demonstrate the importance of critical thinking / skepticism through example

Taking into account the aforementioned guidance and desired skills, we suggest that a Data Science program needs to **cover**:

- Foundations, such as mathematical apparatus (e.g. statistics, algebra), machine learning, and artificial intelligence,
- Technologies, such as text mining, semantic web, open data, data storage and processing,
- And also specific domain applications, such as in healthcare, humanities, public governance, social studies, finance, management, media, journalism, etc.

Therefore, we propose, following the EDISON framework [6], that a Data Science program is structured as follows and the parts further comprise the following units/disciplines:

- **Core subjects/courses**. The examples of the courses can be: Elementary statistics; Computational thinking; Advanced algorithms and data structures; Web technologies; Digital entrepreneurship; Qualitative research methods; Interdisciplinary thinking; Data-centric decision making
- **Subjects/courses for the DS Analytics Itinerary**. This set includes subjects for using appropriate statistical and data analytics techniques on available data to deliver insights and discover information, providing recommendations, and supporting decision-making. The examples of the courses can be: Data mining; Supervised and unsupervised machine learning; Statistical modelling; Predictive analytics.
- **Subjects/courses for the DS Engineering Itinerary**. This set includes disciplines subjects for using engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management. The examples of the courses can be: Software and infrastructure engineering; Manipulating and analyzing complex, high-volume, high-dimensional data, structured and unstructured data; Cloud based data storage and data management. In this part it is also important to inform student About: Symbolic Artificial Intelligence; Semantic technologies and knowledge graphs; Open Data and respective initiatives; Cognitive Computing - an initiative by IBM Watson Research Center; Automated Machine Learning[9]; Computer Vision; Natural Language Processing; Reinforcement Learning; Network Analysis.
- **Subjects/courses for specific applications**. This set includes different scenarios for disciplines as management, healthcare, social sciences and humanities, media and communication, and astronomy, among others.

3.2 Tools and Learning Materials

A number of MOOCs offer data science courses to large, global audiences, offering the opportunity for data science training to occur online and remote from the host

[9] http://www.ml4aad.org/automl/

organization. For example, the University of Warwick offers a MOOC in Big Data: Measuring and Predicting Human Behaviour, while in the US, Stanford offers a dedicated MOOC on Machine Learning. The University of Southampton[10] has held a highly successful Web Science MOOC[11] hosted on FutureLearn and a MOOC on the use and implementation of open data and innovation in organizations. UPM offers also MOOCs related to semantic technologies and ontologies in MiriadaX platform, which is focused on Spanish.

A good approach for learning resources is to create them as reconfigurable course components that can be adapted and customized for different learning contexts. Such learning materials should be created by means of a participatory approach (using for example the SlideWiki platform). It is also important that such resources include also real world data sets, including publicly available data[12] and well as data assets and benchmarks used in different scientific disciplines.

3.3 Tips on Didactics

Active competitive learning model is suggested for the courses in order to make the learning process more effective and efficient. Among didactical patterns we already used or plan to use in different related courses are:

- The use of peer review for course assignments
- Organizing final competitions to rate the results of course works or practical assignments
- Real-life project work

A peer review model assumes that the students in a class take an active part in evaluating the results of the other students. This model has been practiced in the classes on Advanced Algorithms and Data Structures, and Linear Algebra, two of the core courses in the Computer Science with Data Science (CSDS) program at Ukrainian Catholic University (UCU) for evaluating coursework reports. The approach was based on the former practices reported in [8] and resembled a conference peer review process. The students played the role of a “Program Committee Member”. They got their assignments from the Chair played by their instructor. The chair also did the reviews. In order to make student reviews more unbiased and structured, a detailed review form has been developed, which included all the required aspects to be evaluated. The individual grades given by the students to the others work have been compared to the average grades, also counting for the instructor’s grades. The students with smaller deviations from the averages received more additional points for their review work. Overall, a student had a chance to receive up to 80 points for his or her

[10] The University of Southampton, UK, launched the first data science related MOOC on the platform on web science, which attracted around 15,000 students worldwide.

[11] https://www.futurelearn.com/courses/web-science

[12] E.g. http://open-data.europa.eu/ and http://publicdata.eu/

work and up to 20 points for the reviews. EasyChair conference management system[13] was used to manage this peer review process.

A competition of the software developed by students as a practical component of a course is planned to be introduced in the course on Automated Term Extraction (ATE) and Ontology Learning from Texts. The course will be given in Spring 2018 at the Ukrainian Catholic University. The course will be organized in the form of several short tutorials, each dealing with a particular aspect in the ATE pipeline. Each tutorial comprises a hands-on component taking circa 50 percent of teaching time. After each lecture, except the introductory one, the students are offered to:

- Use the instrumental software and the document collection(s) / dataset(s) provided by the tutor
- Refine the software in some advised way, e.g. by introducing a more sophisticated metric or an improvement in an algorithm
- Perform a cross-evaluation experiment to compare the initial revision of the software and their refined revision

These practical tasks are organized in a way to finally assemble a simple instrumental tool suite that helps performing a basic ATE workflow.

The final slot of the course is organized as a cross-evaluation contest for the solutions by students. They are offered to apply their tool suites to the same document collection and measure the quality of ATE results using objective metrics. The ranked list of the solutions is built based on the comparison of these results. So, the students are rated according to their achievements in the cross-evaluation.

Another important and highly useful approach to didactics includes integration of the real-life project-based syllabus for various classes. One of the relevant examples was implemented as a part of the course Introduction to Data Science (CSDS program at UCU). During the project work (which was the only one practical part of the course) students had to participate in the Queen's International Innovation Challenge[14] organized by Smith School of Business from Queens University, Canada. As the participants of the contest, the student teams worked on the challenges provided by various Canadian companies (in 2016) and United Nations (in 2017). Several topmost teams have been qualified for the final in Toronto, Canada, where they presented their projects to a jury. The whole aim was to not only show a technical solution but provide its business background and value. The latter is crucial for the data scientists.

4 Partnerships and Cooperation

Several activities related to data science training should be performed as part of M.Sci. programs. These activities could be included in short-term certified programs (e.g. summer and winter) schools and exchange programs. These training activities should be also aligned with institutional practices in human resources and profes-

[13] http://easychair.org/

[14] https://smith.queensu.ca/centres/scotiabank/competition/index.php

sional development, as well as with current company creation courses and activities; as for example ActuaUPM at UPM, which has created more than 100 technology companies in the past 10 years. Universidad Politécnica de Madrid (UPM) has been among the first IT labs in Europe to establish long-term collaborations via research projects, researcher exchange programs, and training in life sciences, healthcare, management, or Earth and Space science. UPM.is active in organizing summer schools and also involves students, at M.Sci. and Ph.D. levels, in academic and research exchange programs. Several good examples are:

- Organized summer schools:

UPM has been involved in the ISSGC summer school series (International Summer School on Grid Computing), until its end in 2009, in the SSSW summer school series (Summer School on Ontology Engineering and Semantic Web), and in the Marie Curie ITNs SCALUS and BigStorage.

- Exchange programs / projects:

SemData[15] was the project coordinated by UPM and funded under the International Research Staff Exchange Scheme (IRSES) of the EU Marie Curie Actions. It was focused on facilitating exchanges of the research group members, including Ph.D. students, among the participating institutions. SemData brought together research leaders and young researchers across the globe from the relevant communities: Linked Data, Semantic Web, and Database Systems. Research cooperation between the members of project partner organization continues beyond the lifespan of SemData, for example between the Ontology Engineering Group (OEG, UPM) and Intelligent Systems Research Group (ISRG, Zaporizhzhia National University, ZNU).

UPM also was partner in the PlanetData Network of Excellence[16], in which one of the main highlights was the founding of the ESWC Summer School. This school was created to provide an opportunity for Master’s and Ph.D. students in the area of the Semantic Web to learn the key topics in the field from the leading researchers in the area. This summer school was designed using experiences from previous summer schools run within the OntoWeb, KnowledgeWeb and S-Cube networks of excellence.

UPM also participated in the UNIVERSAL (Universal Exchange for Pan-European Higher Education) project, which offered a solid basis for curriculum alignment.

Ukrainian Catholic University (UCU) also runs their summer school in Data Science[17] in Summer semesters. The school aims to provide a broad and rapid introduction to the field of Data Science. The audience mainly consists of the senior-year bachelor, master, and Ph.D. students, and young professionals. The curricula include introductory courses (e.g., Statistics, Machine Learning), domain-specific courses (in healthcare, finance, marketing analysis, urban analytics, among others), and project work. The latter is based on the topics provided by the third-party organizations and

[15] http://www.semdata-project.eu/
[16] https://www.planet-data.eu/
[17] http://cs.ucu.edu.ua/en/summerschool/

commercial companies. The activities framing out short-to-long-term stays in different organizations should include industrial partners in addition to the involved academic institutions in different countries around the world. Long-term collaborations via research projects, researchers exchange programs, training in several disciplines of Science, and involvements in public administration work are topical sorts of cooperative relationships that help framing, broadening and deepening the professional horizons of future data scientists. So, these need to be established, evolved, and refined for every Data Science program, especially at a Ph.D. level.

ISRG at ZNU is active in establishing and rationally exploiting cooperative partnerships with industrial entities regarding their M.Sci. and Ph.D. level students enrolled on the Data Science and Semantics program (the part of Computer Science program). In their cooperation with BWT Group[18], the M.Sci. students are enrolled for short professional internships in their second year. On these internships, the students are fully involved as junior software engineers and data scientists in the commercial projects performed by the company. The Ph.D. students in this cooperation may apply for a part-time work at BWT and hence combine the benefits of academic and industrial professional environments put together for their Ph.D. projects. For example, a Ph.D. student may borrow a use case for his research from his industrial work. From the other hand, a new approach or technique developed in a Ph.D. project may find its validation and swift transfer to industry in this cooperative setting. last but not least, Ph.D. students are paid for their part-time work at an industrial scale. So, they also earn enough money during their Ph.D. term and also learn what is industrial research consulting.

One more good example of the cooperation of ISRG with industry is their activity with the LOD project by Springer Nature[19]. In this case, Springer provides the use case for one of our Ph.D. projects (c.f. [9]) which develops an approach to detect terminological saturation in high-quality document collections. The use case by Springer is on journal papers in Knowledge Management. The company provides the documents and looks forward to evaluate the outcome of this research in their industrial setting. One more partner in this research project is UPM.

The CSDS program at UCU which was designed tightly with the local IT industry includes the same internship approach as well. During the last semester of the program, students visit the companies for the internship work. The diploma thesis could be grounded on the projects that students made during that visits.

Yet one more line of cooperation includes the invitation of lecturers for various MSc and Ph.D. courses. To our experience, that could be done at least in two scenarios: (i) when calling for external academic expertise may bring a super-additive effect on the quality of teaching and learning; and (ii) to attract industrial professionals for more specific and state-of-the-art educational content. One of the interesting examples for the first scenario is a course on Computer Vision at UCU. For that relatively large course (5 ECTS), the university didn't possess a top-qualified instructor. Thus the idea to split the course into several interlinked modules and invite different instructors for

[18] https://www.groupbwt.com/

[19] http://www.springer.de/

teaching these units has been implemented. This approach also allows decreasing the overall load on one particular professor and makes the whole course more flexible and diversified in terms of been based on several approaches, opinions and experiences.

The involvement of industrial leaders was piloted at UCU as a part of the season certified programs (Machine Learning Winter School[20] and Machine Learning Summer Workshops[21]). This allows introducing most relevant knowledge and practical techniques to the students and making them familiar with the real-life industry cases and projects.

5 Recommendations

In fact, a Data Science program, like any other academic program, becomes successful if a proper balance between efficiency and effectiveness is achieved. Efficiency is traditionally regarded as a function of spent resource per achieved result, which has to be minimized. Effectiveness is related to impacts on the students, and also on the society in broad. Contrarily to efficiency, effectiveness needs to be maximized. Notably, improving efficiency – i.e. decreasing resource and effort spent – brings a risk of reducing effectiveness. Luckily for Data Science education and due to a high demand for data science professionals, efficiency can be improved by not reducing but sharing and balancing resource and effort with interested partners, such as industries.

In this section we summarize our positions presented in Sections 3-5 in the form of recommendations for architecting an efficient and effective academic Data Science program.

Recommendation 1. Make your curriculum competence-based, modular, flexible, and adaptable to feedbacks. In our case, as explained in Section 3, these are achieved by following the EDISON framework for competencies, forming the topical scope, and using reconfigurable teaching units for flexibility and reacting to feedbacks.

Recommendation 2. Reduce efforts and increase impacts by using appropriate teaching and learning tools. For achieving that, we suggested using a MOOC approach and infrastructures like SlideWiki for up-scaling impacts through the re-use of teaching materials. This approach also facilitates to making course units reconfigurable and more flexible.

Recommendation 3. Improve effectiveness by incorporating proper motivations in didactics. To implement this recommendation, we use peer-review and competitions in the teaching process. Competitiveness is balanced with teamwork. Hence, as mentioned above, the motivation of students and the quality of teaching and learning are improved. Yet one more way to increase these impacts is the introduction of the real-life project-based syllabus for various courses.

Recommendation 4. Balance resources and increase impacts by intensively involving industrial partners and international programs in the process. In our experience, this is achieved by actively involving different stakeholders as academic program partners, such as companies, international associations, public funding bodies.

[20] http://cs.ucu.edu.ua/en/winterschool/winter-school-2017/

[21] http://cs.ucu.edu.ua/en/mlworkshop/machine-learning-summer-workshops-2017/

As presented in Section 4, this works very well and proves to be effective both at M.Sci and Ph.D. levels.

It looks like the recommendations we gave may work well not only for Data Science Education, but also broader – for other educational domains. We did not check it though, so far. The results of our checks show that, for Data Science, these architectural tips are (i) modular – so can be exploited individually to improve programs; and (ii) super-additive – so the more of these are used, the better the balance between the efficiency and effectiveness of a program becomes.

6 Conclusions and Future Work

In this position paper we presented our views on how an effective and efficient academic Data Science program may be architectured. These views are of course biased as we take active parts in developing such programs at our universities (UCU, UPM, and ZNU). We also believe that we had the right to share these views as the programs we contribute to are deemed successful. Despite the bias and some indicators of individual success, we think that the paper presented some considerations and choices that are broadly applicable, because these are parts of the best practices in Europe and overseas. We referred to these practices, and also some of their resources, in our concise review of the related work in Section 2.

Section 3, focusing on the proposed topical scope and outlining some tips on tools, learning modes, and didactics, is deemed as our contribution in this paper. based on the relevant results, we outlined how a Data Science program should be structured and look topically. Further we presented our views on the effective use of learning tools and materials, like MOOC. Yet further, we shared our experience in using an active and competitive learning model in the courses. These didactics use a peer review approach, students' software competitions, and real-life project work.

In Section 4, we scoped out our views and shared relevant experience on how a productive environment for "breeding" data scientists may be organized via several sorts of partnerships, including cooperation with industries. By pointing out to our experience, we articulated that a collaborative environment involving industrial parties proves to be useful and effective for making a Data Science program successful.

Finally, we summarize our positions on architecting a successful academic Data Science program in the form of four recommendations in Section 5.

Regarding the future work, our plan is to analyze and review our proposal with respect to well-known teaching practices and novel teaching methods. In addition, we will update the proposed program organization taking into account the developed educational content of the courses within Data Science programs in EU (as for example the one proposed by the EIT Digital School). We also think about ways to include the relevant ethical and legal issues in the curricula in a harmonized way. Ethical issues could be taught based on the existing frameworks like the Guidance document by the Government of the UK on ethical issues[22]. The discussion of legal issues in

[22] https://www.gov.uk/government/publications/data-science-ethical-framework

(Big) data analytics is on the hype today. So the lessons could be learnt and taught to students, e.g. based on the case of Cambridge Analytica[23].

Acknowledgements

This work has been done in cooperation between the authors which has been kicked off in the EU FP7 SemData project funded by Marie Curie International Research Staff Exchange Scheme, grant agreement No PIRSES-GA-2013-612551. The authors would also like to thank the anonymous reviewers for their valuable suggestions on improving the paper.

References

1. Ermolayev, V., Akerkar, R., Terziyan, V., Cochez, M.: Toward evolving knowledge ecosystems for Big Data understanding. In: Akerkar, R. (ed.) Big Data Computing, pp. 3--56, Taylor & Francis, ISBN 978-1-46-657837-1 (2013)
2. Hoehndorf, R., Queralt-Rosinach, N.: Data Science and symbolic AI: synergies, challenges and opportunities. Data Science, vol. 1, no. 1-2, pp. 27-38 (2017)
3. Anderson, C.: The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16:07 (June 23). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (2008)
4. Berry, D.: The computational turn: thinking about the digital humanities. Culture Machine 12 (11 July). http://www.culturemachine.net/index.php/cm/article/view/440/470
5. De Veaux, R. et al.: Curriculum guidelines for undergraduate programs in Data Science. Annu. Rev. Stat. Appl. 2017.4:15-30, (2016) DOI: 10.1146/annurev-statistics-060116-053930
6. Demchenko, Yu., Belloum, A., Wiktorski, T.: EDISON Data Science framework: Part 3. Data Science model curriculum (MC-DS). Release 2. EDISON project deliverable, 03 July 2017, http://edison-project.eu/sites/edison-project.eu/files/attached_files/node-447/edison-mc-ds-release2-v03.pdf
7. Hicks, S.C., Irizarry, R.A.: A guide to teaching Data Science. arXiv:1612.07140, 15 May 2017 (v2)
8. Ermolayev, V., Keberle, N., Borue, S.: Coursework peer reviews increase students' motivation and quality of learning. In: Ermolayev, V., et al. (Eds.) ICT in Education, Research, and Industrial Applications. Revised Selected Papers of ICTERI 2012, CCIS Vol. 347, pp. 177–194, Springer-Verlag, Berlin-Heidelberg (2013)
9. Kosa, V., Chugunenko, A., Yuschenko, E., Badenes, C., Ermolayev, V., Birukou, A.: Semantic saturation in retrospective text document collections. In: Mallet, F., Zholtkevych, G. (eds.) Proc. ICTERI 2017 PhD Symposium, CEUR-WS, vol. 1851, pp. 1--8, Kyiv, Ukraine, May 16-17 (2017) online

[23] https://amp.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election?CMP=share_btn_tw&__twitter_impression=true