EDISON Data Science Framework (EDSF): Addressing Demand for Data Science and Analytics Competences for the Data Driven Digital Economy

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl
Steve Brewer
University of Southampton
S.Brewer@soton.ac.uk

Cuadrado Gallego Juan José
University of Alcala
jjcg@uah.es
Tomasz Wiktorski
University of Stavanger, Norway
tomasz.wiktorski@uis.no

***Abstract*—Emerging data driven economy including industry, research and business, requires new types of specialists that are capable to support all stages of the data lifecycle from data production and input to data processing and actionable results delivery, visualisation and reporting, which can be jointly defined as the Data Science professions family. Data Science is becoming a new recognised field of science that leverages the Data Analytics methods with the power of the Big Data technologies and Cloud Computing that both provide a basis for effective use of the data driven research and economy models. Data Science research and education requires a multi-disciplinary approach and data driven/centric paradigm shift. Besides core professional competences and knowledge in Data Science increasing digitalisation of Science and Industry also requires new type of workplace and professional skills that rise importance of critical thinking, problem solving and creativity required to work in highly automated and dynamic environment. The education and training of the data related professions must reflect all multi-disciplinary knowledge and competences that are required from the Data Science and handling practitioners in modern, data driven research and the digital economy. In modern conditions with the fast technology change and strong skills demand, the Data Science education and training should be customizable and delivered in multiple form, also providing sufficient lab facilities for practical training. This paper discusses aspects of building customizable and interoperable Data Science curricula for different types of learners and target application domains. The proposed approach is based on using the EDISON Data Science Framework (EDSF) initially developed in the EU funded Project EDISON and currently being maintained by the EDISON Community Initiative.**

Keywords—Data Science, Data Scientist Professional, Big Data, EDISON Data Science Framework (EDSF), Data Science Competences Framework, Data Science Body of Knowledge, Data Science Model Curriculum, FAIR principles in Open Education, Curriculum Design.

I. Introduction

Emerging data economy, as a part of more general The Fourth Industrial Revolution (referred to as Industry 4.0) is powered by convergence of previously disconnected fields such as Cloud Computing, Big Data, Data Science and Analytics (DSA), Artificial Intelligence (AI), robotics, mobile technologies, 3D printing, nanotechnology and biotechnologies, that all are based on automation and digitalisation of organisational, industrial and business processes. The Industry 4.0 will be characterized by fast development, high level of technologies convergence and increased role of knowledge, skills and human factors to enable continuous and sustainable science and technology development. Such type of economy requires new type of data driven and Data Science and Analytics enabled competences and workplace skills.

Sustainable development of the modern data driven economy requires re-thinking and re-design of both traditional educational models and existing courses reflecting multi-disciplinary nature of Data Science and its application domains. However, at present time most of the existing university curricula and training programs cover a limited set of competences and knowledge areas of what is required for multiple Data Science and general data management professional profiles and organisational roles required by research and industry. In conditions of continuous technology development and shortened technology change cycle, Data Science education requires effective combination of theoretical, practical and workplace skills. Importance of effective use of existing data analytics and data management platforms and tools and corresponding hands on experience is growing and their elements need be generically incorporated into modern curriculum design.

The EDISON Data Science Framework (EDSF) [1, 2-5], which is the products of the EDSION Project, provides a basis for building effective educational environment combining educational or training components and practical hands on experience with virtual and data labs. The future educational model and approach should also solve different aspects of the future professionals that includes both theoretical knowledge and practical skills that must be supported by corresponding education infrastructure and educational labs environment.

The paper refers to the previous authors works that researched new approaches to building effective curricula in Cloud Computing, Big Data and Data Science [6, 7. 8, 9] and provides examples of curricula that are important to enable digital transformation of organisations.

The paper is organized as follows. Section II provides reference to recent studies on the skills demand for Industry 4.0 and data economy and describes the challenges in professional education and training of the data scientists and data workers. Section III describes in details the proposed EDSF and its components. Sections IV describes the example curricula developed by the EDISON project partners, and University of Amsterdam in particular, specifically targeted to facilitate the digital transformation of organisations. Section V provides summary and suggestions for future work, including vision for adopting FAIR principles in Open Education.

II. Demand for Data Science and Data Skills

Growing demand for Data Science and Analytics enabled and general data driven professions is confirmed by multiple European and global market studies. Demand for data related professions will grove even more with the emerging Industry 4.0 [10] that will bring tremendous changes both to business models and labour market with the strong change in the skills set needed to strive in the new economy landscape.

The key Industry 4.0 elements that both empower new data economy and will be facilitated by the new business and consumer models:

- Cyber-physical systems
- Internet of things
- Internet of services
- Smart factory
- Mobile technologies
- Cloud computing
- Big data

The World Economic Forum (WEF) published report “The Future of Jobs” (2016) [11] that is focused on the employment, skills and workforce strategy for the future economy. The report summarised vision of the leading high-tech companies on future skills demand. The following 10 top skills are identified as critical for 2020 (reflecting shift from currently required skills in direction of independent critical thinking, creativity and cognitive flexibility) [12, 13]:

1. Complex problem solving
2. Critical thinking
3. Creativity
4. People management
5. Coordinating with others
6. Emotional Intelligence
7. Judgement and decision making
8. Services orientation
9. Cognitive flexibility

The IDG report 2017 [14] provided deep analysis of the European data market and growing demand for data workers, the value of the data market, the number of data user enterprises, the number of data companies and their revenues, and the overall value of the impact of the data economy on EU GDP. The EU data market is estimated as EUR 60 Bln with growth to EUR 106 Bln in 2020. With the total number of data workers to grow 6.1 mln (2016) 10.4 million in 2020 the data worker skill gap is estimated as 769,000 or 9.8% (2020). Addressing this demand and gap is becoming critical for European economy and challenge for universities. The report stresses that not satisfied demand in data workers with lead to under-performing economy, industry, research and loss of competitiveness.

III. EDISON Data Science Framework (EDSF)

Designing future effective Data Science educational environment will require developing and widely accepted a general framework for Data Science education, curriculum design and competences management that can be based on the proposed EDISON Data Science Framework (EDSF) that is a core product of the EDISON Project. EDSF provides a basis for the definition of the Data Science profession and other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification and career transferability.

Figure 1 below illustrates the main EDSF components and their inter-relations:

- CF-DS – Data Science Competence Framework [2]
- DS-BoK – Data Science Body of Knowledge [3]
- MC-DS – Data Science Model Curriculum [4]
- DSPP - Data Science Professional profiles and occupations taxonomy [5]
- Data Science Taxonomy and Scientific Disciplines Classification.

The proposed framework provides the basis for the definition and design of other components of the Data Science professional environment such as

- Data Science Education Environment (DSEE) intended to be cloud based, customizable and aligned with the new workplace practices and skills
- Education and Training Directory connected to Marketplace and Virtual Data Labs
- Data Science Community Portal (CP) that provides information and community support services. It also provides gateway to DSEE, Marketplace and Virtual Data Labs. CP is intended to include tools for individual competences benchmarking and personalized educational path building

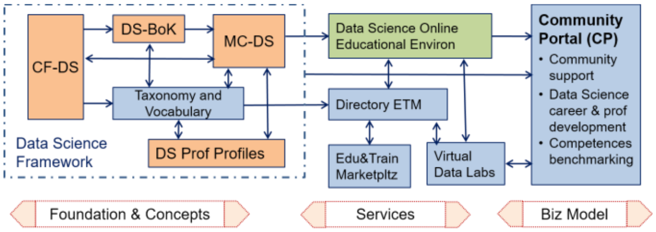

Figure 1. EDISON Data Science Framework components and Data Science Educational environment.

A. *Data Science Competence Framework (CF-DS)*

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path. The CF-DS is defined using the same approach as e-CFv3.0 [15] (competences defined as abilities supported by knowledge and skills with applied proficiency levels) but have competence structured according to the major identified functional groups (as explained below).

The CF-DS is structured along four dimensions (similarly to European e-Competence Framework e-CFv3.0 [15]) that include (1) competence groups, (2) individual competences definition, (3) proficiency levels, and (4) corresponding knowledge and skills. In this context, each individual competence includes a set of required knowledge topics and a set of skills type A and skills type B. Such CF-DS structure allows for competence based curriculum design where competences can be defined based on the professional profile (see DSPP [5] for mapping between professional profiles and competences) or target leaners group when designing a full curriculum, or based on competence benchmarking for tailored training to address identified competences and knowledge gaps.

The following core CF-DS competence and skills groups have been identified (refer to CF-DS specification [2] for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science competences must be supported by knowledge that are defined primarily by education and training and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills:

- Skills Type A which are related to the professional experience and major competences, and
- Skills Type B that are related to wide range of practical computational skills including using programming languages, development environment and cloud based platforms (refer to CF-DS [2] for full definition of the identified knowledge and skills groups).

B. Workplace skills

Workplace skills, also referred to as “soft” skills or professional attitude skills, are becoming increasing important in modern data driven and future Industry 4.0 economy.

The CF-DS defined two groups of skills that are demanded by employers and required for Data Scientist to efficiently work in modern data driven agile companies:

- Data Science Professional and Attitude skills (Thinking and acting like Data Scientist) that defines a special mindset that be developed by a practicing Data Scientist along their career progression
- 21st Century skills that comprise a set of workplace skills that include critical thinking, communication, collaboration, organizational awareness, ethics, and others.

University should pay attention to developing such skills and include them into curricula or extra-curricula activity. Refer to CF-DS for detailed skills definition.

C. Data Science Professional Profiles (DSPP)

Provides effective for organisational skills management and capacity building including

The proposed Data Science professional profiles definition is based on the analysis of the research and industry demand in data related professions. The identified professional profiles are classified using ESCO taxonomy [16], and necessary extensions are proposed to support the following hierarchy of the data handling related occupations:

- Managers: Chief Data Officer (CDO), Data Science (group/department) manager, Data Science infrastructure manager, Research Infrastructure manager
- Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.
- Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators
- Professional (data handling/management): Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists
- Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators
- Support and clerical workers: Support and data entry workers.

The individual profiles are defined in accordance with the CWA 16458 (2012): European ICT Professional Profiles [17] standard (and its revision 2018)

The DSPP document defines also mastery levels and corresponding learning outcome for the targeted education or training. The following mastery levels are defined (using workplace terminology that can be easy mapped to mastery levels defined in MC-DS):

A - Awareness
1) Understand Terminology
2) Understand Principles
3) Apply principles
4) Understand Methods
U - Use/Application
5) Apply basics
6) Supervised use
7) Unsupervised Use
P - Professional/Expert
8) Development of applications using wide range of technologies
9) Supervise project development, team of professionals,
where borderline mastery levels 4 and 7 actually belong to both higher level and lower level groups.

D. Data Science Body of Knowledge and Model Curriculum

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [18], incorporates best practices in defining domain specific BoK’s and provides reference to existing related BoK’s. It also includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS [4] is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS competences (specifically skills type A), and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. Practical curriculum should be supported by corresponding educational environment for hands on labs and educational projects development.

The formal DS-BoK and MC-DS definition will create a basis for Data Science educational and training programmes compatibility and consequently Data Science related competences and skills transferability.

IV. Example Curricula to Facilitate Digital Transformation

This section provides example curricula "Data Science Accelerator" developed and taught by the EDISON project university partners specifically oriented on supporting digital transformation of organisations to adopt Agile Data Driven Enterprise model (ADDE).

A. *Data Management and Data Governance*

Establishing effective Data Management and Data Governance (DM&DG) in organisation is considered as the first step in digital transformation. Best practices in DM&DG are well defined by the DAMA (Data Management Association) and published as Data Management Body of Knowledge (DMBOK) [22] and corresponding guidelines.

The DM&DG course uses DMBOK as general framework covering majority of topics and extending them with the Data Science and Big Data Analytics platforms. The following are the main topics included in the course:

- Introduction. Big Data Infrastructure and Data Management and Governance.
- Data Management concepts. Data management frameworks: DAMA Data Management framework, the Amsterdam Information Model. Extensions for Big Data and Data Science.
- Enterprise Data Architecture. Data Lifecycle Management and Service Delivery Model. Data management and data governance activities and roles. Data Science Professional profiles family. Skills management and capacity building.
- Data Architecture, Data Modelling and Design. Data types and data models. Data modeling. Metadata. SQL and NoSQL databases overview. Distributed systems: CAP theorem, ACID and BASE properties.
- Enterprise Big Data infrastructure and integration with enterprise IT infrastructure. Data Warehouses. Distributed file systems and data storage. Cloud based data storage services: data object storage, data blob storage, Data Lakes (services by AWS, Azure, GCP).
- Big Data storage and platforms. Big Data Security and Compliance. Data security and data protection. Security of outsourced data storage. Cloud security and compliance standards and cloud provider services assessment.

B. *Data Science and Analytics Foundation (DSAF)*

The goal of this course is to introduce the students into the whole spectrum on Data Science and Analytics technologies and at the same time provide strong statistical background for future mastering core data analytics methods and Machine Learning technologies. This defines the main stress in DSAF course on statistical methods, probability theory, hypothesis testing, data preparations, methods of qualitative and quantitative analytics. The primary analytics for this course is recommended to have a low programming threshold to enable fast learning. The RapidMiner visual data analytics environment (https://rapidminer.com/) was identified as preferable choice against R or python tools to introduce the trainees into the key data analytics methods and enable active experimentation.

The following topics are included in DSAF curriculum:

- Introduction and course overview: Data Science and Big Data technologies, Data Science competence and skills, Research Methods in Data Science, Machine Learning and Data Mining overview.
- Statistical methods and Probability theory
- Data description and Statistical Data Analysis
- Data preparation: data entry, data cleaning, data pre-processing
- Qualitative and Quantitative data analysis
- Classification: methods and algorithms
- Cluster analysis basics and algorithms
- Performance of data analytics algorithms and tools

C. *Professional Issues in Data Science*

The goal of this course is to equip the students and practitioners with the knowledge and skills for further focused study of more specific Data Science and Analytics areas and courses.

- Data Science Competences and Skills management and capacity building, EDISON Data Science Framework.
- Data Science professional skills ("Act and think as Data Scientist") and 21st Century Skills.
- Data Science and Analytics methods and technologies overview.
- Research Methods; Business processes management.
- Data Management in research, industry and personal: standards and best practices.
- FAIR (Findable, Accessible, Interoperable, and Re-usable) principles in Open Data and enterprise data management.
- Ethical and legal principles and regulations. Privacy enabling technologies.

It is also beneficial to supply this course with the guided/tutored groups and/or individual training on essential professional skills such as complex problem solving, critical thinking, creativity, etc. defined as critical for Industry 4.0 workforce.

D. *Cloud based DSEE and Virtual Data Labs*

The educational Data Science labs and project development environment can benefit from using clouds and available data analytics and data handling applications and services that can be made available on demand for specific time periods when the education of training take place. Using cloud resources to build effective and up-to-date professional Data Science education environment is inevitable with current fast technology development and required computational performance that can be requested on-demand.

Major Cloud Service Providers (CSP) provide wide range of data analytics and business analytics services and platforms that can be equally used by big, small and medium companies and individuals on the pay-per-use basis. In addition to a possibility to use same resources for education and training purposes, the major CSPs provide also designated education and self-training resources that are in many cases supported also educational grants for students and teachers.

Important component of Data Science education is educational datasets that often need to be provided with their specific applications. While many educational datasets are available from mentioned above cloud platforms, from community run Kaggle (https://www.kaggle.com/) and UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php), use of cloud based VDLabs allows to instantiate the whole experimental setup or environment together with used data sets in case of specific domain focused education or training.

V. Conclusion and Further Developments

EDSF provides a common semantic basis for interoperability of all forms of the Data Science curriculum definition and education or training delivery, as well as knowledge assessment based on fully enumerated definition of EDSF components and individual units. Besides defining academic components of the effective and consistent curriculum, EDSF provides also advice on the required Data Science Education Environment to facilitate fast practical knowledge and skills acquisition by students and learners.

Business Higher Education Forum (BHEF) has published two important reports in cooperation with PriceWaterhouseCoopers (PwC), IBM and Burning Glass Technologies (BGT) [19, 20] that studied Data Science and Analytics (DSA) job market in US and identified a number of actions to be addressed by business, higher education, government and professional organisations to address increased demand and growing gap in demand and supply of skilled DSA workforce capable to effectively work in modern data driven economy.

The authors' experience of developing a pilot project for re-/up-skilling employees of one of the Dutch governmental organisations confirmed trend that organisations, in a way to become data driven and agile, will intend to make the existing organisational roles DSA enabled and require corresponding DSA training in a customizable and flexible form.

An effective professional education needs to provide a foundation for future continuous professional self-development and mastering new emerging technologies, that can provide a basis for the life-long learning model adoption. Wide use of available online resources and platforms for so demanded Data Science and other digital and data skills will facilitate adoption of FAIR principles in the future Open Education to become Findable, Accessible, Interoperable, and Re-usable that were initially proposed for Open Data [21]. The universities can contribute to building FAIR life-long educational space that can serve both organisational and individual needs of students and learners, including support for widely apprised citizen scientists.

The EDSF and the proposed in this paper its further integration with the Data Science Education Environment will facilitate education and training for highly demanded Data Science and Analytics competences and skills.

Acknowledgment

The research leading to these results has received funding from the Horizon2020 projects FAIRsFAIR (grant number 831558), MATES (grant number 591889) and EDISON (grant n. 675419).

References

[1] EDISON Data Science Framework (EDSF). Available at http://edison-project.eu/edison/edison-data-science-framework-edsf
[2] Data Science Competence Framework. Available at http://edison-project.eu/data-science-competence-framework-cf-ds
[3] Data Science Body of Knowledge. Available at http://edison-project.eu/data-science-body-knowledge-ds-bok
[4] Data Science Model Curriculum. Available at http://edison-project.eu/data-science-model-curriculum-mc-ds
[5] Data Science Professional Profiles. Available at http://edison-project.eu/data-science-professional-profiles
[6] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.
[7] Manieri, Andrea 2015, et al, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, Proc. The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada
[8] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Steve Brewer, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc.The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg.
[9] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.
[10] The Fourth Industrial Revolution: what it means, how to respond. [online] https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond
[11] The Future of Jobs, World Economic Forum Report, 18 January 2016 [online] http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf
[12] The 10 skills you need to thrive in the Fourth Industrial Revolution [online] https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/
[13] Are you ready for Industry 4.0. [online] http://www.delivered.dhl.com/en/articles/2017/02/skills-for-industry-4-0.html
[14] Final results of the European Data Market study measuring the size and trends of the EU data economy, EC-IDC, March 2017

[online] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

[15] e-CF3.0, 2016 European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. Available at http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[16] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at https://ec.europa.eu/esco/portal/#modal-one

[17] European ICT Professional Profiles CWA 16458 (2012) (Updated by e-CF3.0) [online] http://relaunch.ecompetences.eu/wp-content/uploads/2013/12/EU_ICT_Professional_Profiles_CWA_updated_by_e_CF_3.0.pdf

[18] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012

[19] PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017) http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent

[20] Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017) https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF

[21] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] https://www.nature.com/articles/sdata201618

[22] DAMA Data Management Body of Knowledge (DMBOK2), DAMA International, 2017