EDISON Project: Building Data Science Profession for European Research and Industry

Yuri Demchenko, Adam Belloum, Wouter Los, Spiros Koulouzis, Cees de Laat
University of Amsterdam, Science Park 904, 1098XH, Amsterdam; The Netherlands
{y.demchenko, A.S.Z.Belloum, W.Los, S.Koulouzis, C.T.A.M.deLaat}@uva.nl

Abstract— The digital revolution made available vast amounts of data both in industry and in the research landscape. The ability to manipulate and extract knowledge and value from this data represents a new profession called the Data Scientist: expected to be the most visible job in future years. This paper provides information about the EU funded project EDISON (which is coordinated and lead by University of Amsterdam team) and describes its major products such as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and other component required to establish sustainable graduation and training of the future Data Science professionals.

Keywords—Data Science, Data Scientist Professional, Big Data, Data Science Competences Framework (CF-DS), Data Science Body of Knowledge (DS-BoK)

I. INTRODUCTION

In recent years, Europe created advanced Research e-Infrastructures (eRI) supporting numerous European research communities. The complexity of eRIs is continuously growing and their normal operation requires more and more qualified engineers, specialists and researchers. The tasks of these experts expanded from equipment maintenance and operation in the past to solving complex tasks with data management and assisting researchers with new scientific and data analytics tools. With the growth of data driven research, the IT and data specialists are getting directly involved into the research process, and their ability to provide deep insight into generated and collected data becomes an important component of modern research. Subject domain and data domain specialists need to work together to benefit from available technologies and to obtain reliable scientific results for practical products and implementation.

The paper presents a research and coordination activity done in the framework of the EU funded EDISON project to establish the new profession of Data Scientist. This European cross-sectoral project with wide international collaboration is coordinated by the University of Amsterdam. The project established and built cooperative relations with many Big Data and Data Science related initiatives.

II. THE EMERGENCE OF THE DATA SCIENTIST PROFESSION

As described in the visionary book by Tony Hey and others "The Fourth Paradigm" [1] and confirmed in the report "Riding the wave: How Europe can gain from the rising tide of scientific data", computational (and statistical) methods and data mining on large sets of scientific and experimental data play a key role in discovering hidden and obscure relationships between processes and events that are necessary in order to make new scientific discoveries and support innovation in industry and the modern digital economy. Industry also recognises the benefits of Big Data technologies and the use of scientific methods in business/operational data analysis and in problem solving for managing enterprise operations, staying innovative and competitive, and being able to provide advanced customer-centric service delivery. These changes have increased the demand for new types of specialists with strong technical background and deep knowledge of the Data Intensive Technologies. These have been identified as the new profession of the Data Scientist.

The U.S. National Institute for Science and Technology (NIST) defined the following groups of skills required/expected from Data Scientists: domain experience, statistics and data mining, and engineering skills [3]. The qualified Data Scientist should be capable of working in different roles in different projects and organisations such as Data Engineer, Data Analyst or Architect, Data Steward, etc., and possess the necessary skills to effectively operate components of the complex data infrastructure and processing applications through all stages of the Data lifecycle till the delivery of expected scientific and business values to science and/or industry.

III. RE-ENGINEERING OF DATA SCIENCE EDUCATION

Educating and training Data Science specialists requires a new model, reflecting in its design the whole lifecycle of data in research and industry domains. Such a model must be built on a thorough analysis of the requirements of modern Data Science to define the Body of Knowledge (BoK) and Competences Profile (CP) of a Data Scientist.

Currently there is no widely accepted Data Science professional education profile, neither generic Big Data technologies training programs. Further, there is no common approach to effectively build professional level Data Science curricula. Universities both in Europe and in the USA (and beyond) do not offer sufficient possibilities for educating the large number of this new type of specialist. Universities, Industry and Research Infrastructures (RIs) must cooperate to establish a common Data Science competences profile and a common component-based curriculum for education and training to realise this profile. For achieving these goals, we foresee a options for re-engineering current education programs.

IV. EDISON DATA SCIENCE FRAMEWORK COMPONENTS

Figure 1 below illustrates the main components of the EDISON framework that provides conceptual basis for the development of the Data Science profession and includes components that to be implemented by the main stakeholder of the supply and demand side: universities, professional training organisations, standardisation bodies, accreditation and certification bodies, companies and organisations and their Human Resources department to successfully manage competences and career development.

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- Data Science Taxonomy and Scientific Disciplines Classification
- EOEE - EDISON Online Education Environment

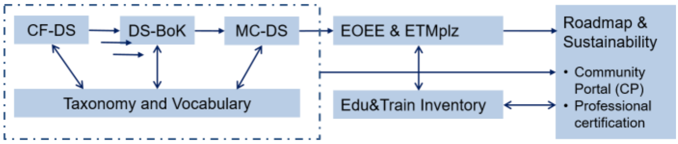

Figure 1. EDISON Data Science framework components

V. DATA SCIENCE COMPETENCE FRAMEWORK

The first pillar of the EDISON project [4.5] is to formally define the Data Science Professions in terms of defining the required Competences Framework and Skills Profile (CF-DS) and mapping these to the existing knowledge domains and academic disciplines currently taught at universities whilst also defining the new disciplines and training subjects/topics to be created for emerging data science needs. The CF-DS will include the common hard and soft skills (i.e., technical and collaborative skills) required to have Data Scientists engaged in a team and to act in the modern agile data-driven enterprise, as well as the subject-specific knowledge and skills allowing to work in different scientific and technical domains.

The EDISON CF-DS development follows the European e-Competences Framework (e-CF3.0) guiding principles [6, 7]. It was presented at the e-CF CEN workshop on 9 December 2015 as possible extension to future European ICT profiles standard. The CF-DS definition also intends to incorporate the ESCO taxonomy [8] by extending it with specific Data Science competences and skills.

The initial EDISON study on Data Science competences revealed that two new groups of competences should be included that have not been explicitly identified in previous studies and frameworks (see Figure 2 and [9]). The figure presents the following competences.

3 competence groups identified in the NIST document and confirmed by analysis of collected data:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge

2 new identified competence groups that are highly demanded and are specific to Data Science

- *Data Management, Curation, Preservation (new)*
- *Scientific or Research Methods (new)*

Data Management, curation and preservation are already included in the existing (research) data related professions such as data archivist, data manager, digital data librarian, data steward, and others. Data management is an important component of European Research Area policy. Knowledge of the scientific research methods and techniques makes the Data Scientist profession different from all previous professions.

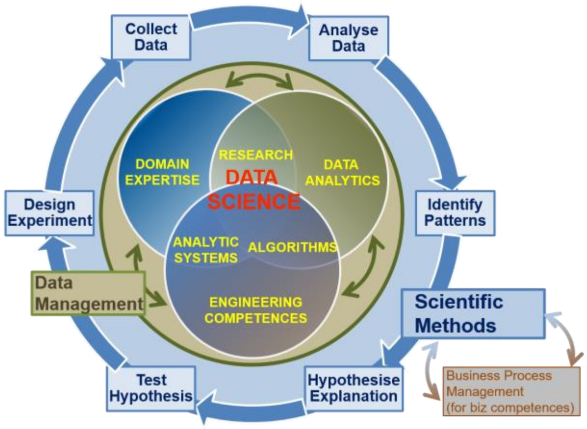

Figure 2. Data Science competences groups.

The identified demand for general competences and knowledge on Data Management and Research Methods needs to be implemented in the future Data Science education and training programs, as well as to be included into re-skilling training programmes. It is important to mention that knowledge of Research Methods does not mean that all Data Scientists must be talented scientists; however, they need to know general research methods such as formulating hypothesis, applying research methods, producing artefacts, and evaluating hypothesis (so called 4 steps model). Research Methods training are already included into master programs and graduate students.

The newly identified competence areas provide a better basis for defining education and training programmes for Data Science related jobs, re-skilling and professional certification.

VI. DATA SCIENCE BODY OF KNOWLEDGE/

The CF-DS provides a basis for the definition of the Data Science Body of Knowledge (DS-BoK), the knowledge needed by the practitioner to perform all the data related processes of his/her profession. The BoK typically defines the content of a curriculum and provides the basis for the definition of the expected educational results and the related knowledge assessment.

Following the CF-DS competence group definition, the DS-BoK should contain the following Knowledge Area groups (KAG):

- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering

- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Scientific or Research Methods group*
- KAG5-DSBP: Business process management group
- KAG6-DSDK: Data Science Domain Knowledge group includes domain specific knowledge

The subject domain related knowledge group (scientific or business) KAG6-DSDK is recognized as essential for practical work of Data Scientist what in fact means not professional work in a specific subject domain but understanding the domain related concepts, models and organisation and corresponding data analysis methods and models. These knowledge areas will be a subject for future development in tight cooperation with subject domain specialists.

The project will perform an analysis of existing academic and professional courses and curricula, including available books and other training resources to identify common conceptual elements and as well as identified gaps (by discipline and market sectors).

This work will be strongly grounded in education theory, for example including the application of Bloom's Taxonomy, Constructive Alignment and Problem-based Learning. It will be also based on the authors' experience in using advanced instructional methodologies for the education of new technologies [11].

Fine-tuning of both the model and the curriculum will benefit from pilot programs and champion institutions. And this will assist in wider adoption in the large community of universities in search of a well-constructed Data Science education program.

VII. Conclusion and Further Developments

The proposed DS-CF has been widely discussed at numerous workshops and community forums. It is already used by few institutions associated with the EDISON project. The published currently an initial version of DS-BoK will require further development and validation by experts and communities of practice to define specific knowledge areas by involving experts in the related knowledge areas, possibly also engaging with the specific professional communities such as IEEE, ACM, DAMA, IIBA, etc. The project will engage with the partner and champion universities into pilot implementation of DS-BoK and collecting feedback from practitioners.

It is anticipated that real life implementation and adoption of the EDISON Data Science framework will includes both approaches top-down and bottom-up that will allow universities and professional training institutions to benefit from EDISON recommendations and adopt them to available expertise, resources and demand of the Data Science competences and skills.

Acknowledgment

This paper presents the ongoing results of the EDISON project performed by the UvA team in close cooperation with the whole project team and individual partners.

The EDISON project is supported under H20202 Grant Agreement n. 675419 by the European Commission.

References

[1] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[2] Riding the wave: How Europe can gain from the rising tide of scientific data. *Final report of the High Level Expert Group on Scientific Data. October 2010*. [Online]. Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, Sept 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[4] EDISON Project: Building Data Science Profession [online] http://www.edison-project.eu/

[5] Andrea Manieri, 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

[6] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-

[7] 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

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

[9] Data Science Competence Framework (CF-DS). EDISON draft V0.6, 10 March 2016 [online] http://www.edison-project.eu/data-science-competence-framework-cf-ds

[10] Data Science Body of Knowledge (DS-BoK). EDISON draft V0.1, 20 March 2016 [online] http://www.edison-project.eu/data-science-body-knowledge-ds-bok

[11] Demchenko, Y., E.Gruengard, S.Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. In Proc. 6th IEEE Intern Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 Dec 2014, Singapore