# EDISON Data Science Framework: A Foundation for Data Science Competence Management and Curricula Development

Yuri Demchenko, Adam Belloum, Wouter Los, Cees de Laat
University of Amsterdam, The Netherlands
{y.demchenko, A.S.Z.Belloum, W.Los, C.T.A.M.deLaat@uva.nl}@uva.nl

*Abstract*—The education and training of Data Scientists currently lacks a commonly accepted, harmonized instructional model that reflects by design the whole lifecycle of data handling in modern, data driven research and the digital economy. This paper presents the EDISON Data Science Framework (EDSF) that is intended to create 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 (DSP profiles). The MC-DS is built based on CF-DS and DS-BoK, where Learning Outcomes are defined based on CF-DS competences and Learning Units are mapped to Knowledge Units in DS-BoK. In its own turn, Learning Units are defined based on the ACM Classification of Computer Science (CCS2012) and reflect typical courses naming used by universities in their current programmes. The paper provides example how the proposed EDSF can be used for designing effective Data Science curricula and reports the experience of implementing EDSF by the Champion Universities that cooperate with the EDISON project.

***Keywords—Data Science, Data Scientist Professional, Big Data, EDISON Data Science Framework (EDSF), Data Science Competences Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), Data Science Professional profiles.***

## I. INTRODUCTION

Data Science is an emerging field of science, which requires a multi-disciplinary approach and has a strong link to Big Data and data driven technologies that created transformational effect to all research and industry domains. Their sustainable development requires re-thinking and re-design of both traditional educational models and existing courses.

This 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 [1, 2]. The paper provides information about the proposed EDISON Data Science Framework (EDSF) and its components that include the Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional profiles (DSP). The EDSF is intended to provide a basis for building effective Data Science curricula and enable the whole Data Science supply-demand-community ecosystem.

## II. DATA SCIENCE PROFESSIONAL DEFINITION AND SKILLS

There is no well established definition of the Data Scientist due to a diverse number of competences and skills expected from these specialists. We will take as a basis the definition provided in the NIST SP1500-1 document [5]: "*A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the **big data lifecycle**.*" The document defines the following groups of skills required from the Data Scientists: domain experience, statistics and data mining, and engineering skills [5].

Other definitions [6, 7] admit such desirable features as ability to solve variety of business problems, optimize performance and suggest new services for the organisation employing Data Scientist. Many practitioners admit a need for a successful Data Scientist to develop a special mindset, to be statistically minded, understand raw data and "appreciate data as a first class product" [8].

The qualified Data Scientist should be capable of working in different roles in different projects and organisations such as Data Engineer, Data Analyst or Data 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. EDISON DATA SCIENCE FRAMEWORK COMPONENTS

The EDISON vision for building the Data Science profession will be enabled through the creation of a comprehensive framework for Data Science education and training that includes such components as Data Science

Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS).

Figure 1 below illustrates the main components of the EDISON Data Science Framework (EDSF) and their inter-relations that provides conceptual basis for the development of the Data Science profession (references to published discussion documents are provided):

- CF-DS – Data Science Competence Framework [9]
- DS-BoK – Data Science Body of Knowledge [10]
- MC-DS – Data Science Model Curriculum [11]
- DSP - Data Science Professional profiles and occupations taxonomy [12]
- Data Science Taxonomy and Scientific Disciplines Classification (including Vocabulary)

The proposed framework provides a basis for other components of the Data Science professional ecosystem:

- EDISON Online Education Environment (EOEE)
- Education and Training Marketplace and Directory
- Data Science Community Portal (CP) that also includes tools for individual competences benchmarking and personalized educational path building
- Certification Framework for core Data Science competences and professional profiles

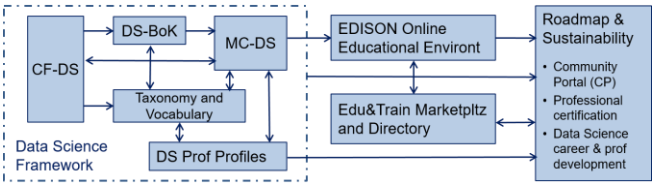


Figure 1 EDISON Data Science Framework components.

The CF-DS includes common competences required for successful work of Data Scientists in different work environments in industry and in research and through the whole career path.

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge
- *Data Management, Curation, Preservation (new)*
- *Scientific or Research Methods (new)*

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support required Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs defined based on the Classification Computer Science (CCS2012) [13], components taken from other BoKs and proposed new KAs to incorporate new technologies used in Data Science and their recent developments.

The MC-DS is built based on CF-DS and DS-BoK where Learning Outcomes are defined based on CF-DS competences 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.

## IV. Example of New Data Science courses development

MC-DS can be used for design a new Data Science curricula for target group of students or learners. In practice when designing a new programme it is necessary to decide on the set of courses with a specific number of credits. The standard in Europe is to use European Credit Transfer System, which defines bachelor programs to have 180 points and Master programs 120 points. This gives usually 30 points per semester. At American institutions credit hours systems are used and they are not fully uniform between institutions.

Required proficiency in each competence group for each professional profile is illustrated in Figure 2. It creates a basis for division of points between LOs and related LUs. In addition, each Learning Outcome can be achieved on three different knowledge or mastery levels (familiarity, usage, assessment). Typically, Bachelor programs focus on two lower levels and Master programs on two higher levels.

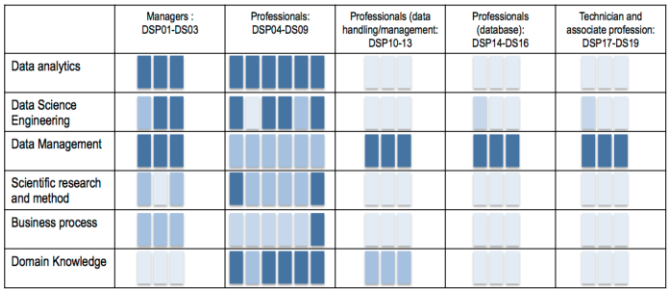


Figure 2. Proficiency/mastery level needed by different Data Science Profile for each of Data Science competence groups
Legend: (1) Bars represent individual DSP profiles [12], 2) mastery levels: familiarity –light blue; usage - blue; assessment – dark blue.

## V. Conclusion and Further Developments

The presented EDSF includes components 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 of the data related jobs. The proposed EDSF has been widely discussed at numerous workshops and community forums. It is already used by few institutions associated with the EDISON project.

## REFERENCES


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

[2] 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

[3] 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/

[4] 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

[5] 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

[6] What is a data scientist? 14 definitions of a data scientist! [online] http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

[7] Cortnie Abercrombie, What CEOs want from CDOs and how to deliver on it [online] http://www.slideshare.net/IBMBDA/what-ceos-want-from-cdos-and-how-to-deliver-on-it

[8] LinkedIn's Daniel Tunkelang On "What Is a Data Scientist?" [online] http://www.forbes.com/sites/danwoods/2011/10/24/linkedins-daniel-tunkelang-on-what-is-a-data-scientist/

[9] Data Science Competence Framework [online] http://edison-project.eu/data-science-competence-framework-cf-ds

[10] Data Science Body of Knowledge [online] http://edison-project.eu/data-science-body-knowledge-ds-bok

[11] Data Science Model Curriculum [online] http://edison-project.eu/data-science-model-curriculum-mc-ds

[12] Data Science Professional Profiles [online] http://edison-project.eu/data-science-professional-profiles

[13] The 2012 ACM Computing Classification System [online] http://www.acm.org/about/class/class/2012