SciDataCon 2016

Conference paper

EDISON Data Science Framework for defining the Data Science Profession

Yuri Demchenko, Adam Belloum, Wouter Los, University of Amsterdam, Netherlands
Steve Brewer, University of Southampton, UK
Andrea Manieri, Engineering Ingegneria Informatica S.p.A., Italy

Abstract

The effective use of Data Science technologies requires new competences and skills and demands for new professions that should support all stages of the research data lifecycle from data production and input to data processing, storing, and obtained scientific results publishing and dissemination. This paper introduces the EDISON Data Science Framework (EDSF) that include conceptual, instructional and policy components required to establish sustainable graduation and training of the future Data Science professionals.

Introduction

Modern research requires new types of specialists that are capable to support all stages of the research data lifecycle from data production and input to data processing, storing, and scientific results publishing and dissemination, which can jointly defined as the Data Science professions family. The future Data Scientists must posses knowledge (and obtain competencies and skills) in data mining and analytics, information visualisation and communication, as well as in statistics, engineering and computer science, and acquire experiences in the specific research or industry domain of their future work and specialisation. Although the Data Scientist is a key occupation in the data related professions family, other occupation are focused on other stages of the data lifecycle and supporting infrastructure.

The paper describes the main components of the proposed EDISON Data Science Framework (EDSF) that is used as a basis for defining the Data Science Professions family. More extended information is provided about the Data Science Competence Framework (CF-DS) and the Data Science Body of Knowledge (DS-BoK) which are essential for defining consistent and customizable Data Science curricula.

EDISON Data Science Framework

Figure 1 below illustrates the main components of the EDISON Data Science Framework (EDSF) that provides conceptual basis for the development of the Data Science profession (including reference to available documents):

- CF-DS – Data Science Competence Framework (CF-DS, 2016)
- DS-BoK – Data Science Body of Knowledge (DS-BoK, 2016)
- MC-DS – Data Science Model Curriculum (MC-DS, 2016)
- Data Science Taxonomy and Scientific Disciplines Classification
- Data Science occupations taxonomy and professional profiles (DSP, 2016)

The proposed framework provides basis for other components of the Data Science professional ecosystem such as

- EOEE - EDISON Online Education Environment
- Education and Training Marketplace and Directory
- Data Science professional profiles definition and certification
- Community Portal (CP)

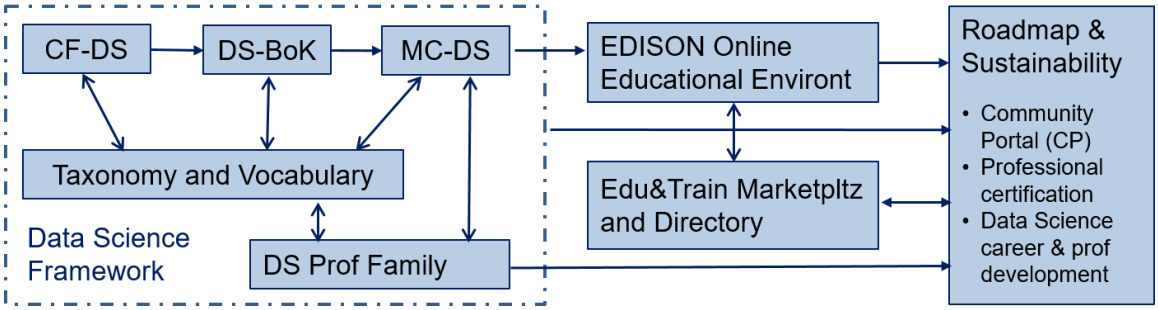

Figure 1: EDISON Data Science framework components.

Data Science Competence Framework and Body of Knowledge

The Data Science Competences Framework (CF-DS) is a cornerstone of the EDISON Data Science Framework and used for defining such components as Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS). The CF-DS is defined in compliance with the European e-Competence Framework (e-CF3.0) and provides suggestions for e-CF3.0 extension with the Data Science related competences and skills.

Figure 2 illustrates the main CF-DS competence groups and their inter-relation:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge
- Data Management, Curation, Preservation
- Scientific or Research Methods (for research professions) and Business Process Management (for business related professions)

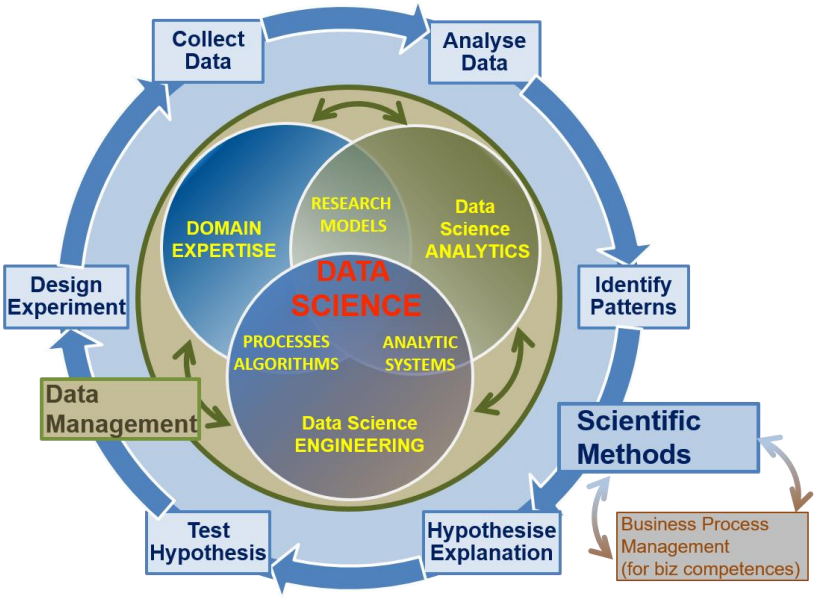

Figure 2: The Data Science competence groups inter-relations.

The identified competence areas provide a better basis for defining education and training program for Data Science related jobs, re-skilling and professional certification. Knowledge of the scientific research methods and techniques makes the Data Scientist profession different from all previous professions. It is recommended that both Data Management (or specifically Research Data Management) and Research Methods are included into all Data Science curricula. The CF-DS provides a basis for the definition of the Data Science Body of Knowledge (DS-BoK), the knowledge needed by the professionals to perform all data related processes of their profession. The BoK typically defines the content of a curriculum and is linked to CF-DS via learning outcomes that can be defined for the specific groups of trainees.

Data Science Body of Knowledge and Model Curriculum

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

Universities can use DS-BoK as reference to define knowledge areas that they need to cover in their programs depending on their primary demand groups in research or industry. The domain specific knowledge can be acquired as a part of the academic education or as a post-graduate professional training at the graduate's work place. It is also commonly recognized that a "fresh" Data Scientist would require 2-3 year to become proficient in his/her profession.

The initially proposed the Data Science Model curriculum provides two basic components for building customisable Data Science curricula: (1) definition of the learning outcomes (LO) based on the CF-DS competences, including their differentiation for different proficiency levels, e.g. using Bloom's Taxonomy, (2) definition of the Learning Units (LU) that map to the LOs for target professional groups, which need to be defined in accordance with the existing academic disciplines classification such as Classification Computer Science (CCS, 2012).

Conclusion and Further Developments

The presented EDSF 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 of the data related jobs. 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 for public comments DS-BoK and MC-Ds documents will require further development and validation by experts and communities of practice to define specific knowledge areas. It will be done by involving experts in the related knowledge areas, 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 MC-DS and collecting feedback from practitioners. All EDISON project products are provided openly under Creative Common license.

Acknowledgements

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

Competing Interests

The authors declare that they have no competing interests.

References [alphabetically by surname, then year]

Andrea Manieri, et al, 2015, 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

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

CF-DS, 2016, 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

Demchenko, Y., E.Gruengard, S.Klous, 2014, 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

DS-BoK, 2016, 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

DSP, 2016, Data Science Professional profiles definition (CF-DS). EDISON draft v0.1, 11 July 2016 [online] http://www.edison-project.eu/data-science-professional-profiles-dsp

eCFv3.0, 2014, 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-

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

ESCO, 2016, ESCO (European Skills, Competences, Qualifications and Occupations) framework [online] https://ec.europa.eu/esco/portal/#modal-one

MC-DS, 2016, Data Science Model Curriculum (MC-DS), EDISON Draft v0.1, 11 June 2016 [online] http://www.edison-project.eu/data-science-model-curriculum-mc-ds

NIST, 2015, 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