EDISON
building the data
science profession

EDISON Data Science Framework:
Part 1. Data Science Competence Framework (CF-DS)
Release 2

Project acronym: EDISON
Project full title: Education for Data Intensive Science to Open New science frontiers
Grant agreement no.: 675419

Due Date	
Actual Date	3 July 2017
Document Author/s	Yuri Demchenko, Adam Belloum, Tomasz Wiktorski
Version	Release 2, v0.8
Dissemination level	PU
Status	Working document, request for comments
Document approved by	

Document Version Control			
Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
0.1	2/11/2015	Initial draft	YD
0.4	30/11/2015	Added information about DS skills based on analysed information, reference to ACM Information Technology Competency model	YD, AB, TW, WL
0.5	30/12/2015	Update version for public comments	YD
0.7	4/07/2016	Updated after D2.2: Revised and added enumerated competences definitions. Data Science professional profiles span off into a separate document on the Data Science Professions family taxonomy	
Release 1	10/10/2016	Release 1 after ELG03 meeting discussion	
0.8	03/07/2017	Updated after multiple discussions and comments, DARE Project alignment, mapping CRISP-DM, ELG04 comments	YD, AM, ES
Release 2	10/10/2016	Release 2 documents published	

Document Editors: Yuri Demchenko		
Contributors:		
Author Initials	Name of Author	Institution
YD	Yuri Demchenko	University of Amsterdam
AB	Adam Belloum	University of Amsterdam
AM	Andrea Manieri	Engineering
TW	Tomasz Wiktorski	University of Stavanger
WL	Wouter Los	University of Amsterdam
ES	Erwin Spekschoor	Independent expert

Executive summary

The EDISON project is designed to create a foundation for establishing a new profession of Data Scientist for European research and industry. The EDISON vision for building the Data Science profession is enabled through the proposed comprehensive EDISON Data Science framework (EDSF) 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), and Data Science Professional Profiles (DSPP). The EDSF provides a conceptual basis for the Data Science Profession definition targeted education and training, professional certification, organizational and individual skills management and career transferability.

The definition of the Data Science Competence Framework (CF-DS) is a cornerstone component of the whole EDISON framework. CF-DS will provide a basis for the Data Science Body of Knowledge (DS-BoK) and Model Curriculum (MC-DC) definitions, and further for the Data Science Professional Profiles definition and certification. 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.

The proposed EDISON framework comprising of the mentioned above components will provide a guidance and a basis for universities to define their Data Science curricula and courses selection, on one hand, and for companies to better define a set of required competences and skills for their specific industry domain in their search for Data Science talents, on the other hand.

This document presents ongoing results of the Data Science Competence Framework definition based on the analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for Data Scientist profession in industry and research. The presented CF-DS Release 2 is significantly extended with the skills and knowledge subjects/units related to competences groups. The document also contains the Data Science professional (workplace) skills definition and provides reference to the general "soft" skills often referred to as 21st century skills.

- The presented CF-DS defines five groups of competences for Data Science that include the commonly recognised groups Data Analytics, Data Science Engineering, Domain Knowledge (as defined in the NIST definition of the Data Scientist) and extend them with the two additional groups *Data Management* and *Research Methods* (or Business Process management for business related occupations) that are recognised to be important for the successful work of Data Scientist.
- The document provides example of the individual competences mapping to identified skills and knowledge for the Data Science Analytics competence group.
- The identified competences, skills and knowledge subjects are provided as enumerated lists to allows easy use in applications and developing compatible APIs.
- The report suggests possible extensions to e-CF3.0 on the Data Science related competences.

It is intended that the proposed CF-DS can provide a basis for building interactive/web based tool for individual or organizational Data Science competences benchmarking, Data Science team building, and creating the customized Data Science education and training program.

The EDSF documents are available for public discussion at the project website at http://edison-project.eu/data-science-competence-framework-cf-ds

TABLE OF CONTENTS

1 Introduction

Data Science Competence Framework is a part of the EDISON Data Science Framework (EDSF) that comprise of the following documents: Data Science Competence Framework (CF-DS) [1], Data Science Body of Knowledge (DS-BoK) [2], Model Curriculum (MC-DC) [3], and Data Science Professional Profiles (DSPP) [4].

The CF-DS definition is a cornerstone component of the whole EDISON Data Science Framework. CF-DS provides a basis for the Data Science Body of Knowledge that defines a set of knowledge required from the Data Scientist or related Data Science enabled roles to support required competences and effectively operate in their organisational roles which are in its own turn defined based in functions, responsibilities and competences. Competences defined in CF-DS are used for defining learning outcomes when defining the Data Model Curriculum. It is intended that the CF-DS will be defined in compliance with the European e-Competence Framework (e-CF3.0) and will provide suggestions for e-CF3.0 extension with the Data Science related competences and skills.

This document presents the ongoing results of the Data Science Competence Framework definition based on overview and analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for Data Scientist profession in industry and research.

The presented CF-DS defines the five groups of competences for Data Science that include the commonly recognised groups Data Science Analytics, Data Science Engineering, Domain Knowledge (as defined in the NIST definition of Data Science) and extends them with the two additional competence groups *Data Management* and *Research Methods* (or Business Process management for business related occupations) that are recognised to be important for a successful work of Data Scientist. The Research Methods competences are essential for the Data Scientist to discover new relations and provide actionable insight into available data, and have ability to formulate good research questions, hypothesis and evaluate them based on collected data.

The proposed EDSF comprising of the mentioned above components will provide a guidance and a basis for universities to define their Data Science curricula and courses selection, on one hand, and for companies to better define a set of required competences and skills for their specific industry domain in their search for Data Science talents, on the other hand. The proposed CF-DS can be used for building interactive/web based tool or applications for knowledge and skills (self-) assessment, job vacancy design, assessment of the candidate's profile for a specific profile/role or job vacancy. All individual competences, knowledge subjects and skills are enumerated to allows easier design of API for applications that may use CF-DS.

The document has the following structure. Section 3 provides an overview of existing frameworks for ICT and Data Science competences and skills definition including NIST Special Publication 1500-1, e-CF3.0, ACM Computing Classification System (2012). Section 4 presents the full CF-DS definition that in current release 2 includes identified competence groups, identified skills, and knowledge that all together should enable the Data Scientist to effectively work with variety of Data Analytics methods and Big Data platforms to deliver insight and value to organisations. Section 4 also provides description of the Data Science professional (workplace) and general attitude skills demanded from the modern specialists/professional intended to work in modern agile data driven companies. Section 5 provides examples of the individual competences definition together with their linking to knowledge and skills. Section 6 provides suggestions for practical use of CF-DS in particular for other ESDF components definition and possible uses by organisations for competences assessment and skills management.

Appendices to this document contain important supplementary information: information about approach and data sets used for deriving the proposed CF-DS competences groups; overview of known studies, reports and publications related to Data Science competences and skills; concepts and models related to the Data Science competences definition, such as data lifecycle management models, scientific methods, and business process management lifecycle models.

2 EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework provides a basis for the definition of the Data Science profession and enabling the definition of the other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification.

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:

- CF-DS – Data Science Competence Framework [1]
- DS-BoK – Data Science Body of Knowledge [2]
- MC-DS – Data Science Model Curriculum [3]
- DSPP - Data Science Professional profiles and occupations taxonomy [4]
- Data Science Taxonomy and Scientific Disciplines Classification

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

- EDISON Online Education Environment (EOEE)
- Education and Training Directory and Marketplace
- 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

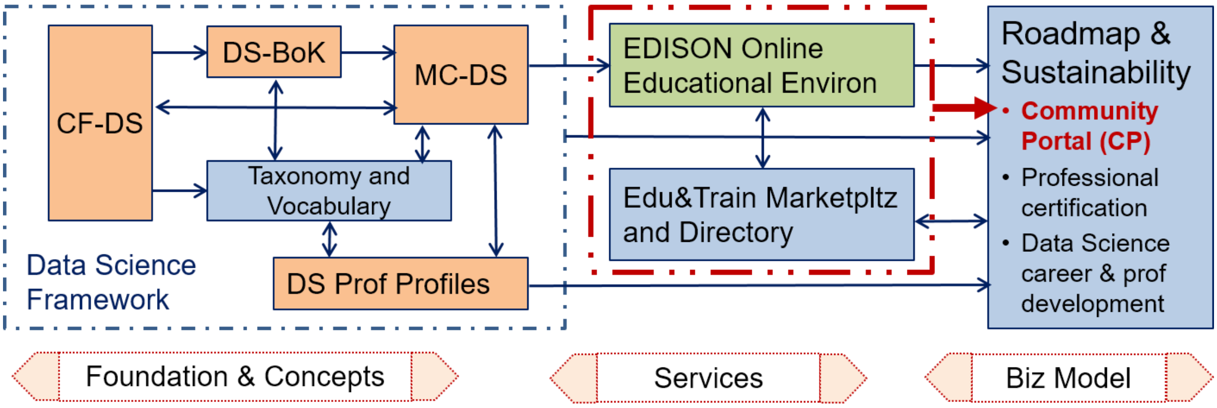


Figure 1 EDISON Data Science Framework components.

The CF-DS provides the overall basis for the whole framework, it first version has been published in November 2015 and was used as a foundation all following EDSF components developments. The CF-DS has been widely discussed at the numerous workshops, conferences and meetings, organised by the EDISON project and where the project partners contributed. The core CF-DS competences has been reviewed

The core CF-DS includes common competences required for successful work of Data Scientist in different work environments in industry and in research and through the whole career path. The future CF-DS development will include coverage of the domain specific competences and skills and will involve domain and subject matter experts.

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 follows the same approach to collect community feedback and contribution: Open Access CC-BY community discussion document is published on the project website. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs defined based on the Classification Computer Science (CCS2012), components taken from other BoKs and proposed new KA 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. The proposed Learning outcomes are enumerated to have direct mapping to the enumerated competences in CF-DS. The preliminary version of MC-DS has been discussed at the first EDISON Champions Conference in June 2016 and collected feedback is incorporated in current version of MC-DS.

The DSPP are defined as an extension to European Skills, Competences, Qualifications and Occupations (ESCO) using the ESCO top classification groups. DSPP definition provides an important instrument to define effective organisational structures and roles related to Data Science positions and can be also used for building individual career path and corresponding competences and skills transferability between organisations and sectors.

The Data Science Taxonomy and Scientific Disciplines Classification will serve to maintain consistency between four core components of EDSF: CF-DS, DS-BoK, MC-DS, and DSP profiles. To ensure consistency and linking between EDSF components, all individual elements of the framework are enumerated, in particular: competences, skills, and knowledge subjects in CF-DS, knowledge groups, areas and units in DS-BoK, learning units in MC-DS, and professional profiles in DSPP.

It is anticipated that successful acceptance of the proposed EDSF and its core components will require standardisation and interaction with the European and international standardisation bodies and professional organisations. This work is being done as a part of the ongoing EDSF dissemination and sustainability activity.

The EDISON Data Science professional ecosystem illustrated in Figure 1 uses core EDSF components to specify the potential services that can be offered for professional Data Science community and provide basis for the sustainable Data Science and related general data skills sustainability. In particular, CF-DS and DS-BoK can be used for individual competences and knowledge benchmarking and play instrumental role in constructing personalised learning paths and professional (up/re-) skilling programs based on MC-DS.

3 Existing frameworks for ICT and Data Science competences and skills definition

This section provides a brief overview of existing standard and commonly accepted frameworks that have been used for defining Data Science and general Computer Science and ICT competences, skills and subject domain classifications that can be, with some alignment, built upon and re-used for better acceptance from research and industrial communities... The information in this section is also complemented with the overview of other works and publications to define required Data Science competences and skills which are placed in Appendix A.

3.1 NIST definition of Data Science

NIST Big Data Working Group (NBD-WG) published their first release of Big Data Interoperability Framework (NBDIF) in September 2015 [5][1] consisting of 7 volumes. Volume 1. Definitions provides a number of definitions in particular Data Science, Data Scientist and Data Life Cycle, which we will use as a starting point for our analysis:

> **Data science** is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing. Data science can be understood as the activities happening in the processing layer of the system architecture, against data stored in the data layer, in order to extract knowledge from the raw data.
>
> Data science across the entire data life cycle incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering. Data science applications implement data transformation processes from the data life cycle in the context of Big Data Engineering.
>
> A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle.
> Data scientists and data science teams solve complex data problems by employing deep expertise in one or more of these disciplines, in the context of business strategy, and under the guidance of domain knowledge. Personal skills in communication, presentation, and inquisitiveness are also very important given the complexity of interactions within Big Data systems.
>
> The **data life cycle** is the set of processes in an application that transform raw data into actionable knowledge.

The term analytics refers to the discovery of meaningful patterns in data, and is one of the steps in the data life cycle of collection of raw data, preparation of information, analysis of patterns to synthesize knowledge, and action to produce value. Analytics is used to refer to the methods, their implementations in tools, and the results of the use of the tools as interpreted by the practitioner. The analytics process is the synthesis of knowledge from information.

The NBDIF Volume 1 also provides overview of other definitions of Big Data and Data Science from IDG, McKinsey, O'Reilly reports and popular blogs published by experts in a new technology.

Figure 2 from the BDIF publication provide graphical presentation of the multi-factor/multi-domain Data Science definition.

[1] Currently the NBDIF is undergoing revision that will also include the major definitions update

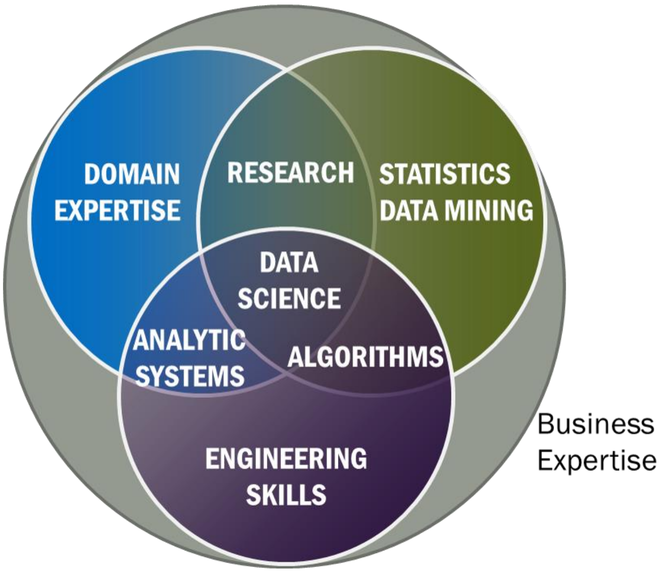


Figure 2. Data Science definition by NIST BD-WG [5].

3.2 European e-Competence Framework (e-CF)

The EDISON CF-DS development follows the European e-Competences Framework (e-CF) approach and guiding principles:

- CF-DS adopts a holistic e-CF definition: "Competence is a demonstrated ability to apply knowledge, skills and attributes for achieving desirable results" in organisational or role context.
- Competence is a durable concept and although technology, jobs, marketing terminology and promotional concepts within the ICT environment change rapidly, the e-CF remains durable requiring maintenance approximately every three years to maintain relevance.
- CF-DS should work as an enabler for multiple applications that can be used by different types of users from individual to organisational; it should support common understanding and not mandate specific implementation.
- A competence can be a part of a job definition but cannot be used to substitute similar named job definition; one single competence can be assigned to multiple job definitions.

The European e-Competence Framework (e-CF) [6,7, 8] was established as a tool to support mutual understanding and provide transparency of language through the articulation of competences required and deployed by ICT professionals (including both practitioners and managers).

The e-CF is structured from four dimensions:

Dimension 1:
5 e-Competence areas, derived from the ICT business processes PLAN – BUILD – RUN – ENABLE – MANAGE

Dimension 2:
A set of reference e-Competences for each area, with a generic description for each competence. 40 competences identified in total provide the European generic reference definitions of the e-CF 3.0.

Dimension 3:
Proficiency levels of each e-Competence provide European reference level specifications on e-Competence levels e-1 to e-5, which are related to the EQF levels 3 to 8.

Dimension 4:
Samples of knowledge and skills relate to e-Competences in dimension 2. They are provided to add value and context and are not intended to be exhaustive.

Whilst competence definitions are explicitly assigned to dimension 2 and 3 and knowledge and skills samples appear in dimension 4 of the framework, attitude is embedded in all three dimensions.

Dimension 1. Competence Area defined by ICT Business Process stages from organisational perspective:
A. Plan: Defines activities related to planning services or infrastructure, may include also elements of design and trends monitoring.
B. Build: Includes activities related to applications development, deployment, engineering, and monitoring
C. Run: Includes activities to run/operate applications or infrastructure, including user support, change support, and problems management
D. Enable: Includes numerous activities related to support production and business processes in organisations that include sales support, channels management, knowledge management, personnel development and education and training.
E. Manage: Includes activities related to ICT/projects and business processes management including management of risk, customer relations, and information security.

e-competences in Dimension 1 and 2 are presented from the organisational perspective as opposed to from an individual's perspective. Figure 3 illustrates the ICT process stages as they are defined in the e-CF3.0 document. Dimension 3 which defines e-competence levels related to the European Qualifications Framework (EQF), is a bridge between organisational and individual competences. Table 3.1 below contains competences defined for areas A-E. For more detailed definition of e-CF3.0 dimensions 1-3 and dimension 4 refer to the original document [6].

Table 3.1. e-CF3.0 competences defined for areas A-E

Dimension 1: 5 e-CF areas (A – E)	Dimension 2: 40 e-Competences identified	Dimension 1: 5 e-CF areas (A – E)	Dimension 2: 40 e-Competences identified
A. PLAN	A.1. IS and Business Strategy Alignment	D. ENABLE	D.1. Information Security Strategy Development
	A.2. Service Level Management		D.2. ICT Quality Strategy Development
	A.3. Business Plan Development		D.3. Education and Training Provision
	A.4. Product / Service Planning		D.4. Purchasing
	A.5. Architecture Design		D.5. Sales Proposal Development
	A.6. Application Design		D.6. Channel Management
	A.7. Technology Trend Monitoring		D.7. Sales Management
	A.8. Sustainable Development		D.8. Contract Management
	A.9. Innovating		D.9. Personnel Development
			D.10. Information and Knowledge Management
B. BUILD	B.1. Application Development		D.11. Needs Identification
	B.2. Component Integration		D.12. Digital Marketing
	B.3. Testing		
	B.4. Solution Deployment	E. MANAGE	E.1. Forecast Development
	B.5. Documentation Production		E.2. Project and Portfolio Management
	B.6. Systems Engineering		E.3. Risk Management
			E.4. Relationship Management
C. RUN	C.1. User Support		E.5. Process Improvement
	C.2. Change Support		E.6. ICT Quality Management
	C.3. Service Delivery		E.7. Business Change Management
	C.4. Problem Management		E.8. Information Security Management
			E.9. IS Governance

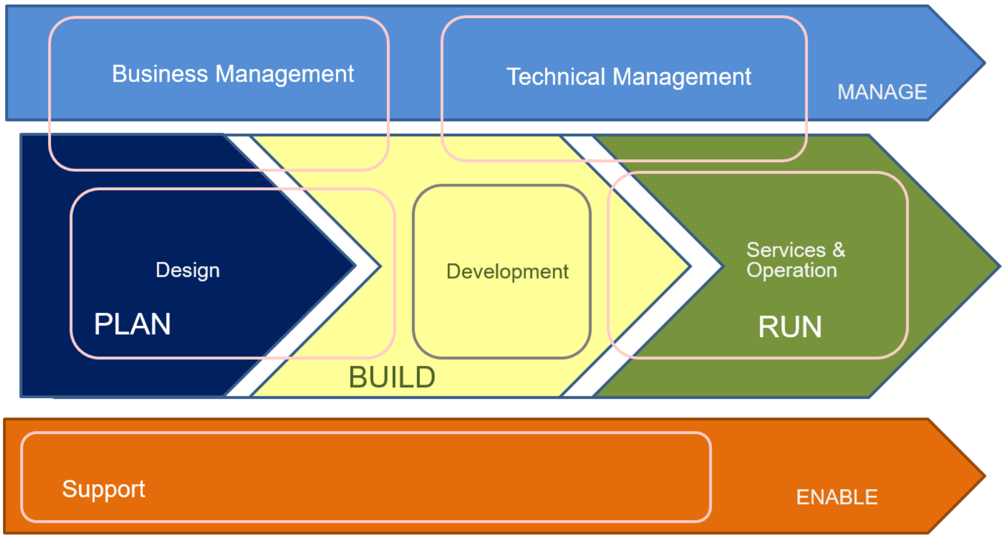


Figure 3. ICT process stages aligned with the organisational production workflow (as used in e-CF3.0)

Figure 4 illustrates the multi-purpose use of the European e-Competence Framework within ICT organisations. The e-CF has a multidimensional structure and is flexible in using for different purposes, it can be easy adopted for organisation specific model and roles. The e-CF3.0 is used for job-profiles definition in CWA 16458 (see [9] and EDSF DSPP document [4]) that are linked to the organisational processes what creates limitations for cross-organisational professional profiles and roles such as Data Scientist. However, combining competences from different competence areas and using them as building blocks can allow flexible job-profiles definition. This enables the derived job-profiles to be easily updated by changing set of competences related to profiles without the need to restructure the entire profile.

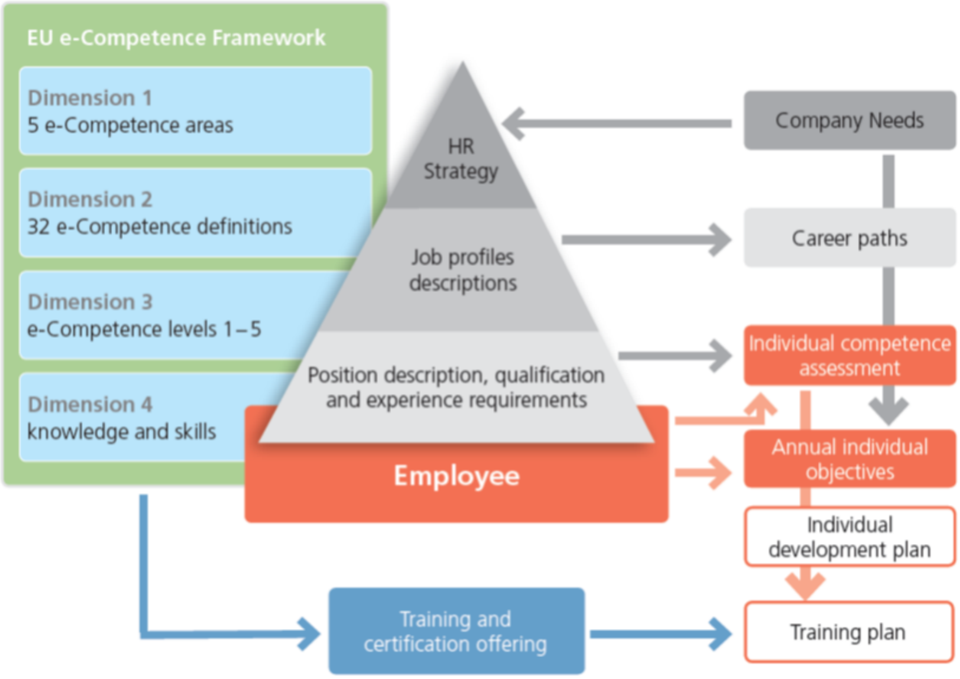

EU e-Competence Framework
Dimension 1
5 e-Competence areas
Dimension 2
32 e-Competence definitions
Dimension 3
e-Competence levels 1 – 5
Dimension 4
knowledge and skills
HR Strategy
Job profiles descriptions
Position description, qualification and experience requirements
Employee
Company Needs
Career paths
Individual competence assessment
Annual individual objectives
Individual development plan
Training and certification offering
Training plan

Figure 4. e-CF3.0 structure and use for definition of the job profile definition and training needs.

3.3 ACM Information Technology Competencies Model

The ACM Information Technology Competency Model (IT-CM) of Core Learning Outcomes and Assessment for Associate-Degree Curriculum (2014) has been developed by ACM Committee for Computing Education in Community Colleges (ACM CCECC) [11-12].

ACM currently categorizes the overarching discipline of computing into five defined sub- disciplines (ACM, 2005): computer science, computer engineering, software engineering, information systems and information technology. This report specifically focuses on information technology defined by the ACM CCECC as follows:

> *Information Technology involves the design, implementation and maintenance of technology solutions and support for users of such systems. Associated curricula focus on crafting hardware and software solutions as applied to networks, security, client- server and mobile computing, web applications, multimedia resources, communications systems, and the planning and management of the technology lifecycle (ACM CCECC, 2009).*

The document refers to the U.S. Department of Labour Information Technology Competency Model [19] that was one of sources that provided a foundation for the curricular guidance outlined in IT-CM report

Competencies are used to define the learning outcome. In formulating assessment rubrics, the ACM CCECC uses a structured template comprised of three tiers: “emerging”, “developed”, and “highly developed”, that can actually be mapped to the level of Bloom’s verbs from the lower order thinking skills (LOTS) to the higher order thinking skills (HOTS), including “analysing” and “evaluating.”

The ACM Competencies Model provides a basis for the Competency -based learning that is Instead of focusing on how much time students spend learning a particular topic or concept (Carnegie unit credit hour), the outcomes-based model assesses whether students have mastered the given competencies, namely the skills, abilities, and knowledge.

The document defines 50 learning outcomes (that also define the Body of Knowledge) that represent core or foundational competencies that a student in any IT-related program must demonstrate. Curricula for specific IT programs (e.g., networking, programming, digital media, and user support) will necessarily include additional coursework in one or more defined areas of study. The core IT learning outcomes are grouped into technical competency areas and workplace skills.

The ACM CCECC classification is supported by the web portal http://ccecc.acm.org/. The portal provides related information, linking and mapping between different classification systems, in particular:

- ACM Computing Classification System 2012
- U.S. Dept. of Labor IT 2012 Competency Model [13]
- Bloom's Revised Taxonomy [14]
- European e-Competence Framework 3.0 (Proficiency Levels 1 & 2) [7, 8]

4 EDISON Data Science Competence Framework (CF-DS)

This section describes the proposed Data Science Competence Framework (CF-DS) that serves as a foundation for the definition of other EDSF components. The presented CF-DS provides full and comprehensive view of the demanded Data Science competences, skills and knowledge comparing to the existing Data Science definitions that primarily cover the data analytics and software engineering competences while modern data driven enterprises and processes require advanced skills for heterogeneous data management and use of research methods to uncover full data value. In current version, the proposed Data Science Competence has evolved from the initially proposed as a result of the job market study supported by extensive desk research covering professional blogs, community discussions, and existing standards and best practices overview, to mature framework reviewed by expert groups and individual experts, and feedback received from multiple practical implementations by universities, professional training organisations, and different projects dealing with data related skills management.

4.1 Relation to and use of existing framework and studies

The following describes what existing frameworks and documents were used for defining the proposed set of Data Science competences and skills.

a) NIST NBDIF Data Science and Data Scientist definition [5]

It provided the general approach to the Data Science competences and skills definition, in particular, as having 3 groups: Data Analytics, Data Science Engineering, and Domain expertise, that may define possible specialisation of actual Data Science curricula or individual Data Scientists competences profile.

b) European e-Competence Framework (e-CFv3.0) [6, 7, 8]

e-CF3.0 provided a general framework for ICT competences definition and possible mapping to Data Science competences. However, it appeared that current e-CF3.0 doesn't contain competences that reflect specific Data Scientist role in organisation. Furthermore, e-CF3.0 is built around organisational workflow while anticipated Data Scientist's role is cross-organisational bridging different organisational roles and departments in providing data centric view or organisational processes.

c) European ICT profiles CWA 16458 (2012) [9]

European ICT profiles and its mapping to e-CF3.0 provided a good illustration how individual ICT profiles can be mapped to e-CF3.0 competences and areas. Similarly, the additional ICT profiles are proposed to reflect Data Scientist's role in the organisation.

d) European Skills, Competences, Qualifications and Occupations (ESCO) [10]

ESCO provides a good example of a standardised competences and skills taxonomy. The presented study will provide contribution to the definition of the Data Scientist as new profession or occupation with related competences, skills and qualifications definition. The CF-DS definition will re-use, extend and map the ESCO taxonomy to the identified Data Science competences and skills.

e) ACM Computing Classification System (ACM CCS2012) [11]

ACM Computing Classification System will be used as a basis to define the proposed Data Science Body of Knowledge, and extension to ACM CCS2012 will provided to cover the identified knowledge and required academic subjects. Necessary contacts will be done with the ACM CCS body and corresponding ACM curriculum defining committees.

f) O'Reilly Strata Survey (2013) [15]

It was a first extensive study on Data Scientist organisational roles, profiles and skills. Although skills are defined as very technically and technologically specific, the proposed definition of profiles is important for defining required competence groups, in particular identification of Data Science Creative and Data Science Researcher profiles indicates an important role of scientific approach and need for research method training in Data Scientist professional education. This group of competences is included in the proposed CF-DS.

g) EC Report on the Consultation Workshop (May 2012) "Skills and Human Resources for e-Infrastructures within Horizon 2020" [16].

This report provided important information about EC and European research community vision on the needs for Data Science skills for e-Infrastructure, in particular to support e-Infrastructure development, operation and scientific use. The identified nine skills gap areas provide additional motivation for specific competences and skills training for future Data Scientists who will work in e-Infrastructure that in particular include data management, curation and preservation.

In the course of defining, validating and refining the proposed CF-DS, the EDISON project interacted with following external projects and studies contributing to them and influencing their approach and alignment with EDSF:

- DARE project [2] their the EDISON contributed to the definition of recommended Data Science Analytics competences [25] and their alignment with the EDSF
- PwC and BHEF report "Investing in America's data science and analytics talent: The case for action" (April 2017) [23]
- Burning Glass Technology, IBM, and BHEF report "The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market" (April 2017) [24]

4.2 Identified Data Science Competence Groups

The results of the job market study and analysis for Data Science and Data Science enabled vacancies, conducted at the initial stage of the project, provided a basis and justification for defining the main competence groups that are commonly required by companies, including identification such skills as Data Management and Research methods that were not required formerly required for data analytics jobs.

The following CF-DS competence and skills groups have been identified:

Core Data Science competences/skills groups defining profile of the Data Science related professional profiles

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Additional common competence groups demanded by organisations

- Data Management and Governance (including data stewardship, curation, and preservation)
- *Research Methods for research related professions and Business Process Management for business related professions*

Data management, curation and preservation competences are already attributed to the existing (research) data related professions such as data stewards, data manager, data librarian, data archivist, and others. Data management is an important component of European Research Area and Open Data and Open Access policies. It is extensively addressed by the Research Data Alliance (RDA) and supported by numerous projects, initiatives and training programmes[3].

[2] DARE (Data Analytics Rising Employment) project is commissioned by Asia Pacific Economic Cooperation (APEC) council and is focused on defining the Recommended Data Science Analytics competences. The DARE project recommendation is to include the basic competences or literacy in the overall Data Science competences definition.

[3] Research Data Alliance Europe https://europe.rd-alliance.org/

Knowledge of the research methods and techniques is something that makes Data Scientist profession different from all previous professions. It should be also coupled with the basic project management competences and skills.

From the education and training point of view, the identified competences can be treated or linked to expected learning or training outcome. This aspect is discussed in detail in relation to the definition of the Data Science Body of Knowledge and Data Science Model Curriculum.

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

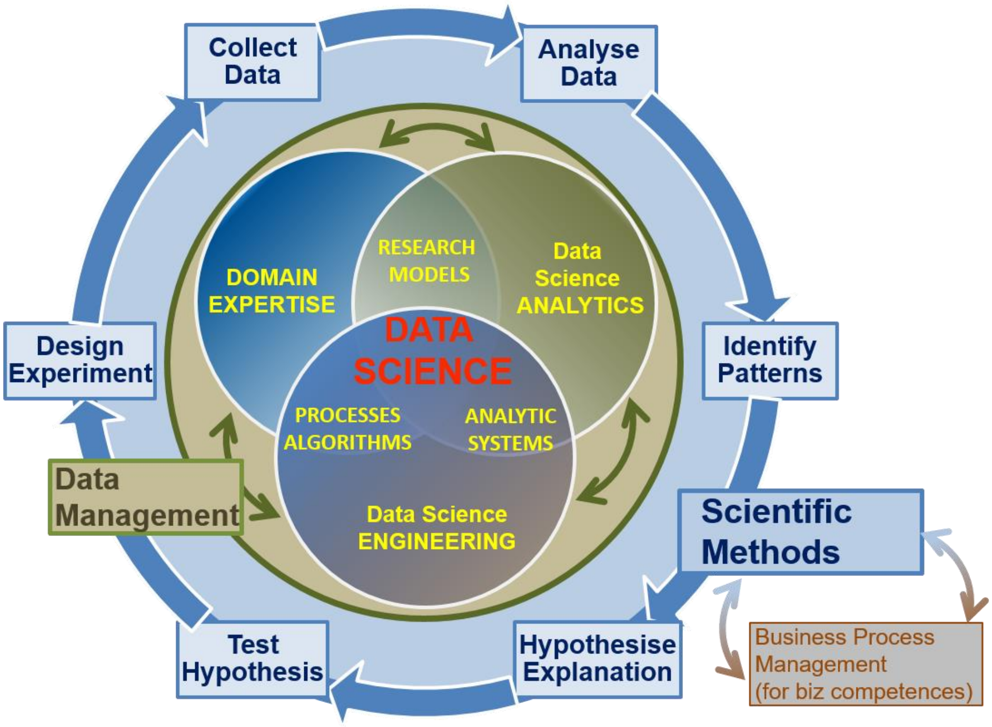
Table 4.1 provides the proposed Data Science competences definition for different groups supported by the data extracted for the collected information. The presented competences definition has been reviewed by a number of expert groups and individual experts as a part of the project EDISON engagement and network activities. The presented competences are required for different professional profiles, organisational roles and throughout the whole data lifecycle, but not necessary to be provided y a single role or individuum. The presented competences are enumerated to allow easy use and linking between all EDSF document.

Table 4.1. Competences definition for different Data Science competence groups

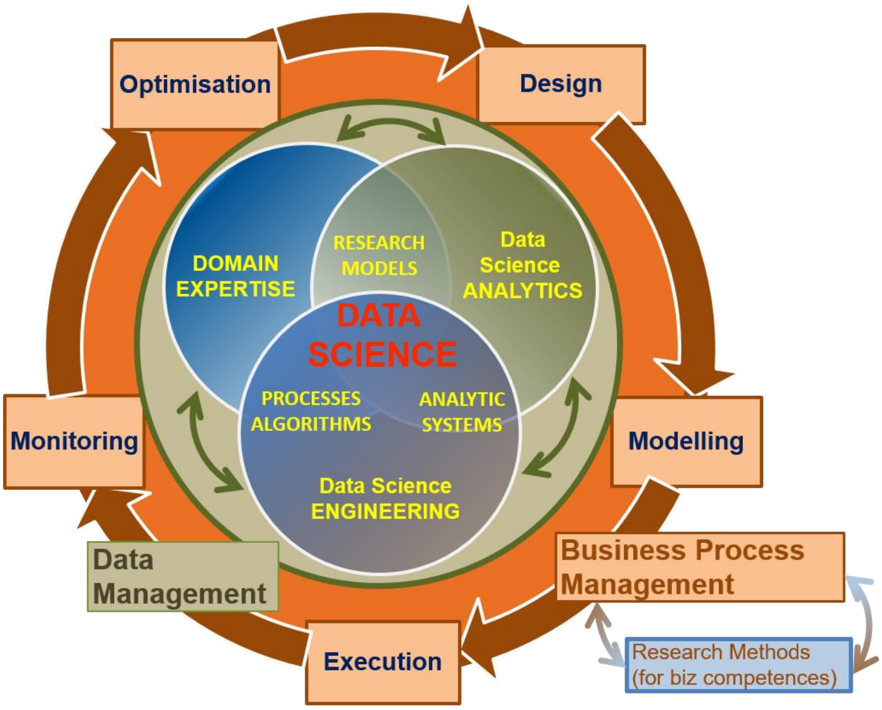
Data Analytics (DSDA)	**Data Science Engineering (DSENG)**	**Data Management (DSDM)**	**Research Methods and Project Management (DSRM)**	**Domain related Competences (DSDK): Applied to Business Analytics (DSBA)**
DSDA Use appropriate data analytics and statistical techniques on available data to discover new relations and deliver insights into research problem or organizational processes and support decision-making.	DSENG Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.	DSDM Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	DSRM Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	DSDK Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSDA01 Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle	DSENG01 Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation	DSDM01 Develop and implement data strategy, in particular, in a form of data management policy and Data Management Plan (DMP)	DSRM01 Create new understandings by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods	DSBA01 Analyse information needs, assess exisitng data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources
DSDA02 Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction	DSENG02 Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with the special focus on Big Data technologies for large datasets and cloud based data analytics platforms	DSDM02 Develop and implement relevant data models, define metadata using common standards and practices, for different data sources in variety of scientific and industry domains	DSRM02 Direct systematic study toward understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSBA02 Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges
DSDA03 Identify, extract, and pull together available and pertinent heterogeneous data, including modern data sources such as social media data, open data, governmental data	DSENG03 Develop and prototype specialised data analysis applicaions, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and streaming processing platforms, including online and cloud based solutions for on-demand provisioned and scalable services	DSDM03 Integrate heterogeneous data from multiple source and provide them for further analysis and use	DSRM03 Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate sound hypothesis	DSBA03 Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends

DSDA04 Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	DSENG04 Develop, deploy and operate large scale data storage and processing solutions using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, Hbase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	DSDM04 Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)	DSRM04 Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications, contribute to the development of organizational objectives	DSBA04 Analyse opportunity and suggest use of historical data available at organisation for organizational processes optimization
DSDA05 Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	DSENG05 Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	DSDM05 Ensure data quality, accessibility, interoperability, compliance to standards, and publication (data curation)	DSRM05 Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	DSBA05 Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
DSDA06 Visualise results of data analysis, design dashboard and use storytelling methods	DSENG06 Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	DSDM06 Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management	DSRM06 Develop and guide data driven projects, including project planning, experiment design, data collection and handling	DSBA06 Analyse multiple data sources for marketing purposes; identify effective marketing actions

Figures 5 (a) and (b) provide graphical presentation of relations between identified competence groups as linked to Scientific Methods or to Business Process Management. The figure illustrates importance of the Data Management competences and skills and Research Methods or Business Process Management knowledge for all categories and profiles of Data Scientists.



(a) Data Science competence groups for general or research oriented profiles.



(b) Data Science competence groups for business oriented profiles.

Figures 5. Relations between identified Data Science competence groups for (a) general or research oriented and (b) business oriented professions/profiles: Data Management and Scientific/Research Methods or Business Processes Management competences and knowledge are important for all Data Science profiles.

The Research Methods typically include the following stages (see Appendix C for reference to existing Research Methods definitions):

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Important part of the research process is the theory building but this activity is attributed to the domain or subject matter researcher. The Data Scientist (or related role) should be aware about domain related research methods and theory as a part of their domain related knowledge and team or workplace communications. See example of Data Science team building in the Data Science Professional Profiles definition provided as a separate document [4].

There is a number of the Business Process Operations models depending on their purpose but typically they contain the following stages that are generally similar to those for Scientific methods, in particular in collecting and processing data (see reference to exiting definitions (see Appendix C for reference to existing Business Process Management stages definitions):

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

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 the 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 of many master programs.

4.3 Identified Data Science Skills and their mapping to Competences

Required Data Science skills are defined based on the job market study of the current analysis of Data Science job market, extended with the numerous blog articles analysis[4] published by Data Science practitioners which provide valuable information in such new emerging area as Data Science.

The identified skills can be organised in the following groups:

- Data Science skills related to the main competence groups that cover knowledge and experience related to effectively realise defined competences and related organisational functions;
- Data analytics and data handling languages, tools, platforms and applications, including SQL based applications and data management tools;
- Knowledge and experience with the Big Data infrastructure platforms and tools.

Separately defined are personal and attitude skills also referred to as the 21st century skills and Data Science professional skills those that define specific (personal) skills that the Data Scientist need to develop to successfully work as a Data Scientist in different organisational roles and along their career. The analysis and identified Data Science soft skills are described in section 4.6.

[4] It is anticipated that for such new technology domain as Data Science the blog articles constitutes valuable source of information. Information extracted from them can be correlated with other sources and in many cases provides valuable expert opinion. Opinion based research is one of basic research methods and can produce valid results.

4.3.1 Data Science skills related to the main competence groups

Table 4.2 lists identified Data Science skills related to the main competence groups

- Data Science Analytics covering extensive skills related to using different Machine Learning, Data Mining, statistical methods and algorithms;
- Data Science Engineering skills related to design, implementation and operation of the Data Science (or Big Data) infrastructure, platforms and applications
- Data Management and governance (including both general data management and research data management)
- Research Methods and Project Management
- Business Analytics as an example of domain related skills

The Data Science Analytics group is the most populated what reflects wide spectrum of required skills in this group as a core for the Data Science. It is followed by the Data Science Engineering skills that are important for the Data Scientist to have ability to implement the effective data analytics solutions and applications.

Table 4.2. Identified Data Science skills related to the main Data Science competence groups

SDSDA **Data Science Analytics**	**SDSENG** **Data Science Engineering**	**SDSDM** **Data Management**	**SDSRM** **Research Methods and Project Management**	**SDSBA** **Business Analytics**
SDSDA01 Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	SDSENG01 Use systems and software engineering principles to organisations information system design and development, including requirements design	SDSDM01 Specify, develop and implement enterprise data management and data governance strategy and architecture, including Data Management Plan (DMP)	SDSRM01 Use research methods principles in developing data driven applications and implementing the whole cycle of data handling	SDSBA01 and Business Intelligence (BI) methods for data analysis; apply cognitive technologies and relevant services
SDSDA02 Use Data Mining techniques	SDSENG02 Use Cloud Computing technologies and cloud powered services design for data infrastructure and data handling services	SDSDM02 Data storage systems, data archive services, digital libraries, and their operational models	SDSRM02 Design experiment, develop and implement data collection process	SDSBA02 Apply Business Processes Management (BPM), general business processes and operations for organisational processes analysis/modelling
SDSDA03 Use Text Data Mining techniques	SDSENG03 Use cloud based Big Data technologies for large datasets processing systems and applications	SDSDM03 Define requirements to and supervise implementation of the hybrid data management infrastructure, including enterprise private and public cloud resources and services	SDSRM03 Apply data lifecycle management model to data collection and data quality evaluation	SDSBA03 Apply Agile Data Driven methodologies, processes and enterprises
SDSDA04 Apply Predictive Analytics methods	SDSENG04 Use agile development technologies, such as DevOps and continuous improvement cycle, for data driven applications	SDSDM04 Develop and implement data architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)	SDSRM04 Apply structured approach to use cases analysis	SDSBA04 Use Econometrics for data analysis and applications
SDSDA05 Apply Prescriptive Analytics methods	SDSENG05' Develop and implement systems and data security, data access, including data anonymisation, federated access control systems	SDSDM05 Implement data lifecycle support in organisational workflow, support data provenance and linked data	SDSRM05 Develop and implement Research Data Management Plan (DMP), apply data stewardship procedures	SDSBA05 Develop data driven Customer Relations Management (CRP), User Experience (UX) requirements and design
SDSDA06 Use Graph Data Analytics for organisational network analysis, customer relations, other task	SDSENG06 Apply compliance based security models, in particular for privacy and IPR protection	SDSDM06 Consistently implement data curation and data quality controls, ensure data integration and interoperability	SDSRM06 Consistently apply project management workflow: scope, planning, assessment, quality and risk management, team management	SDSBA06 Apply structured approach to use cases analysis in business and industry

SDSDA07 Use Qualitative analytics	SDSENG07 Use relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for structured and unstructured data	SDSDM07 Implement data protection, backup, privacy, mechanisms/ services, comply with IPR, ethics and responsible data use		SDSBA07 Use Data Warehouses technologies for data integration and analytics, including use open data and social media data
SDSDA08 Apply analytics and statistics methods for data preparation and pre-processing	SDSENG08 Effectively use Big Data infrastructures, high-performance networks, infrastructure and services management and operation	SDSDM08 Use and implement metadata, PID, data registries, data factories, standards and compliance		SDSBA08 Use data driven marketing technologies
SDSDA09 Be able to use performance and accuracy metrics for data analytics assessment and validation	SDSENG09 Use and apply modeling and simulation technologies and systems	SDSDM09 Adhere to the principles of the Open Data, Open Science, Open Access, use ORCID based services		SDSBA09 Mechanism Design and/or Latent Dirichlet Allocation
SDSDA10 Use effective visualiation and storytelling methods to create dashboards and data analytics reports	SDSENG10 Use and integrate with the organisational Information systems, collaborative system			
SDSDA11 Use Natural Language Processing methods	SDSENG11 Design efficient algorithms for accessing and analysing large amounts of data, including API to different databases and data sets			
SDSDA12 Operations Research	SDSENG12 Use of Recommender or Ranking system			
KDSDA13 Optimisation				
SDSDA14 Simulation				

It is important to mention that the whole complex of Data Science related competences, skills and knowledge are strongly based on the mathematical foundation that should include knowledge of mathematics (including linear algebra, calculus, etc), statistics and probability theory.

4.3.2 Data Science skills related to the Data Analytics languages, tools, platforms and Big Data infrastructure

Table 4.3 lists identified skills related to the Data Analytics languages, tools, platforms and Big Data infrastructure that are split on the following sub-groups:

- DSDALANG - Data Analytics and Statistical languages and tools
- DSADB - Databases and query languages
- DSVIZ- Data/Applications visualization
- DSADM - Data Management and Curation platform
- DSBDA - Big Data Analytics platforms
- DSDEV - Development and project management frameworks, platforms and tools

It is also important for Data Scientist to be familiar with multiple data analytics languages and demonstrate proficiency in one or few most popular languages (what should be supported with several years of practical experience)[5],

- R including extensive data analysis libraries
- Python and related data analytics libraries
- Julia
- SPSS
- KNIME, Orange, WEKA, others

Data Science practitioner must be familiar and have experience with the general programming languages, software versioning and projects management environments such as

- Java, JavaScript and/or C/C++ as general applications programming languages
- Git versioning system as a general platform for software development
- Scrum agile software development and management methodology and platform

It is essential to mention that all modern Big Data platforms and general data storage and management platforms are cloud based. The knowledge of Cloud Computing and related platforms for applications deployment and data management are included in the table. The use of cloud based data analytics tools is growing and most of big cloud services providers provide whole suites of platforms and tools for enterprise data management from Enterprise Data Warehouses, data backup and archiving to business data analytics, data visualization and content streaming

[5] Consider proposed here list as examples and refer to other more focused and extended research and discussions such as for example blog article “Data Scientist Core Skills”, Blog article by Mitchell Sanders, posted on August 27, 2013 [online] http://www.datasciencecentral.com/profiles/blogs/data-scientist-core-skills

Table 4.3. Required skills related to analytics languages, tools, platforms and Big Data infrastructure [6]

DSDALANG **Data Analytics and Statistical languages and tools**	**DSADB** **Databases and query languages**	**DSVIZ** **Data/Applications visualization**	**DSADM** **Data Management and Curation platform**	**DSBDA** **Big Data Analytics platforms**	**DSDEV** **Development and project management frameworks, platforms and tool**
DSDALANG01 R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)	DSADB01 SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)	DSVIZ01 Data visualization Libraries (mathpoltlib, seaborn, D3.js, FusionCharts, Chart.js, other)	DSADM01 Data modelling and related technologies (ETL, OLAP, OLTP, etc.)	DSBDA01 Big Data and distributed computing tools (Spark, MapReduce, Hadoop, Mahout, Lucene, NLTK, Pregel, etc.)	DSDEV01 Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
DSDALANG02 Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)	DSADB02 SQL and relational databases (proprietary: Oracle, MS SQL Server, others)	DSVIZ02 Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)	DSADM02 Data Warehouse platform and related tools	DSBDA02 Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	DSDEV02 Python, Java or C/C++ Development platforms/IDE (Eclipse, R Studio, Anaconda/Jupyter Notebook, Visual Studio, Jboss, Vmware, others)
DSDALANG03 SAS	DSADB03 NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)	DSVIZ03 Online visualization tools (Datawrapper, Google Visualisation API, Google Charts, Flare, etc)	DSADM03 Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)	DSBDA03 Real time and streaming analytics systems (Flume, Kafka, Storm)	DSDEV03 Git versioning system as a general platform for software development
DSDALANG04 Julia	DSADB 04 Hive (query language for Hadoop)		DSADM04 Backup and storage management (iRODS, XArch, Nesstar, others)	DSBDA04 Hadoop Ecosystem/platform	DSDEV04 Scrum agile software development and management methodology and platform
DSDALANG05 IBM SPSS	DSADB 05 Data Modeling (UML, ERWin, DDL, etc)			DSBDA05 Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	
DSDALANG06 Other Statistical computing and languages (WEKA, KNIME, Scala, Stata, Orange, etc)				DSBDA06 Amazon Data Analytics platform (Kinesis, EMR, etc)	

[6] The presented here Big Data platforms and tools are examples of the most popular platforms and tools and are not exhaustive. Please search for general and domain specific other general and domain specific reviews and inventories, for example: Data Science Knowledge Repo https://datajobs.com/data-science-repo/

DSDALANG07 Scripting language, e.g. Octave, PHP, Pig, others				DSBDA07 Other cloud based Data Analytics platforms (IBM Watson, HortonWorks, Vertica LexisNexis HPCC System, etc)	
DSDALANG08 Matlab Data Analytics				DSBDA08 Cognitive platforms (such as IBM Watson, Microsoft Cortana, others)	
DSDALANG09 Analytics tools (R/R Studio, Python/Anaconda, SPSS, Matlab, etc)				DSBDA09 Kaggle competition, resources and community platform	
DSDALANG10 Data Mining tools: RapidMiner, Orange, R, WEKA, NLTK, others					
DSDALANG11 Excel Data Analytics (Analysis ToolPack, PivotTables, etc)					

*) The majority of the presented and used Big Data and analytic platforms are cloud based and online data analytics and data management platforms that are becoming increasingly popular for enterprise and business applications and provide important features such as scalability and on demand resources allocation. The cloud based services and applications are typically well supported by the providers' deployment and monitoring

4.4 Knowledge required to support identified competences

Table 4.4 provides enumerated list of knowledge topics/units that are required to support corresponding competence groups. There is no direct mapping between individual competences and knowledge units, singe competence may be mapped to multiple knowledge units.

Table 4.4. Knowledge required to support identified competences

KDSDA **Data Science Analytics**	**KDSENG** **Data Science Engineering**	**KDSDM** **Data Management**	**KDSRM** **Research Methods and Project Management**	**KDSBA** **Business Analytics**
KDSDA01 Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	KDSENG01 Systems Engineering and Software Engineering principles, methods and models, distributed systems design and organisation	KDSDM01 Data management and enterprise data infrastructure, private and public data storage systems and services	KDSRM01 Research methods, research cycle, hypothesis definition and testing	KDSBA01 Business Analytics (BA) and Business Intelligence (BI); methods and data analysis; cognitive technologies
KDSDA02 Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	KDSENG02 Cloud Computing, cloud based services and cloud powered services design	KDSDM02 Data storage systems, data archive services, digital libraries, and their operational models	KDSRM02 Experiment design, modelling and planning	KDSBA02 Business Processes Management (BPM), general business processes and operations, organisational processes analysis/modelling
KDSDA03 Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)	KDSENG03 Big Data technologies for large datasets processing: batch, parallel, streaming systems, in particular cloud based	KDSDM03 Data governance, data governance strategy, Data Management Plan (DMP)	KDSRM03 Data lifecycle and data collection, data quality evaluation	KDSBA03 Agile Data Driven methodologies, processes and enterprises
KDSDA04 Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	KDSENG04 Applications software requirements and design, agile development technologies, DevOps and continuous improvement cycle	KDSDM04 Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)	KDSRM04 Use cases analysis: research infrastructure and projects	KDSBA04 Econometrics: data analysis and applications
KDSDA05 Text Data Mining: statistical methods, NLP, feature selection, apriori algorithm, etc.	KDSENG05’ Systems and data security, data access, including data anonymisation, federated access control systems	KDSDM05 Data lifecycle and organisational workflow, data provenance and linked data	KDSRM05 Research Data Management Plan (DMP) and data stewardship	KDSBA05 Data driven Customer Relations Management (CRP), User Experience (UX) requirements and design

KDSDA06 Predictive Analytics	KDSENG06 Compliance based security models, privacy and IPR protection	KDSDM06 Data curation and data quality, data integration and interoperability	KDSRM06 Project management: scope, planning, assessment, quality and risk management, team management	KDSBA06 Use cases analysis: business and industry
KDSDA07 Prescriptive Analytics	KDSENG07 Relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	KDSDM07 Data protection, backup, privacy, IPR, ethics and responsible data use		KDSBA07 Data Warehouses technologies, data integration and analytics
KDSDA08 Graph Data Analytics: (path analysis, connectivity analysis, community analysis, centrality analysis, sub-graph isomorphism, etc.	KDSENG08 Big Data infrastructures, high-performance networks, infrastructure and services management and operation	KDSDM08 Metadata, PID, data registries, data factories, standards and com0liance		KDSBA08 Data driven marketing technologies
KDSDA09 Qualitative analytics	KDSENG09 Modeling and simulation, theory and systems	KDSDM09 Open Data, Open Science, research data archives/repositories, Open Access, ORCID		
KDSDA10 Natural language processing	KDSENG10 Information systems, collaborative systems			
KDSDA11 Data preparation and pre-processing				
KDSDA12 Performance and accuracy metrics				
KDSDA13 Operations Research				
KDSDA14 Optimisation				
KDSDA15 Simulation				

4.5 Proficiency levels

It is essential to mention that for such complex professional domain as Data Science the practical experience of working with the data analytics languages, tools and platforms is essential and typically required from minimum 1 to 3 years to be able to develop complex analytics applications necessary to solve critical organisational needs. minimum required experiences with related methods. Although many companies explicitly require experience up to 5 years, the current shortage of skilled Data Scientists will demand novel approaches on targeted competences and skills development that should combine individual competences assessment, design of tailored training for deficient skills development and personalised workplace (self-)training.

Definition of the proficiency levels of individual competes is an important dimension in the CF-DS definition. The CF-DS will follow e-CF3.0 approach in defining the proficiency levels of individual competences. e-CF defines 5 proficiency levels that are mapped to levels 4-8 of the EQF (European Qualification Framework) [10]. At this stage of development, the CF-DS will intend to define 3 levels of the Data Science competences:

- Associate: basic or entry level that defines minimum competences and skills to be able to work in a Data Science team under supervision
- Professional that indicates ability to solve major tasks independently, use multiple languages, tools and platforms and develop specialised applications
- Expert that require wide knowledge experience with the multiple Data Analytics, engineering and data management areas, and related tools. platforms and Big Data infrastructure services. Expert level is typically required from the lead Data Scientist, manager of the Data Science team, or similar.

Examples of the proficiency levels definition foe Data Science Analytics competences is provided in section 5. It is essential that all Data Science competences are strongly based on the common required competences and skills that include basic competences in mathematics, statistics, statistical languages, general computation skills, visualisation as defined in the previous section.

4.6 Data Scientist Personal skills (aka “soft” skills)

Although it is commonly agreed on the importance of the soft skills for Data Scientist, the job market analysis clearly confirmed importance of personal skills and identified a number of specific Data Science professional skills (what means “Thinking and action like a Data Scientist”) that are required for the Data Scientist to effectively work in the modern agile data driven organisations and project teams. These should be also complemented with the general personal skills referred to as 21st century skills. Importance of such skills for Data Scientist is defined by their cross-organisational functions and responsibilities in collecting and analysing organisational data to provide insight for decision making. In such a role the Data Scientist is often reports to executive level or to other departments and teams. These skills extend beyond traditionally required communication or team skills. In addition, the ideal Data Scientist is expected to bring and spread new knowledge to organisation and ensure that all benefit and contribute to the processes related to data collection, analysis and exploitation.
The importance of the Data Science professional and soft skills is confirmed by the DARE project.

4.6.1 Data Science Professional or Attitude skills (Thinking and acting like Data Scientist)

The Data Science is growing as a distinct profession and consequently will need professional identification via definition of the specific professional skills and code of conduct that can be defined as “Thinking and acting like Data Scientist”. Understanding, recognising and acquiring such skills is essential for the Data Scientist to successfully progress along their career. It is also important for team leaders to correctly build relations in the team of project group.

Table 4.5 lists the Data Science professional or attitude skills which are identified by the Data Science practitioners and educators. Although some of the skills are common the 21st century skills, it is important to provide the whole list of skills that can provide a guidance for future Data Scientists what skills are expected from them and need to be developed along their career.

Table 4.5. Data Science Professional or Attitude skills (Thinking and acting like Data Scientist)

Skill ID	Skill definition
DSPS	General group definition: Thinking and acting like a Data Scientist
DSPS01	Accept/be ready for iterative development, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
DSPS02	Ask the right questions
DSPS03	Recognise what things are important and what things are not important
DSPS04	Respect domain/subject matter knowledge in the area of data science
DSPS05	Data driven problem solver and impact-driven mindset
DSPS06	Recognise value of data, work with raw data, exercise good data intuition
DSPS07	Good sense of metrics, understand importance of the results validation, never stop looking at individual examples
DSPS08	Be aware about power and limitations of the main machine learning and data analytics algorithms and tools
DSPS09	Understand that most of data analytics algorithms are statistics and probability based, so any answer or solution has some degree of probability and represent an optimal solution for a number variables and factors
DSPS10	Working in agile environment and coordinate with other roles and team members
DSPS11	Work in multi-disciplinary team, ability to communicate with the domain and subject matter experts
DSPS12	Embrace online learning, continuously improve your knowledge, use professional networks and communities
DSPS13	Story Telling: Deliver actionable result of your analysis
DSPS14	Attitude: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
DSPS15	Ethics and responsible use of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

4.6.2 21st Century workplace skills

21st Century skills comprise a set of workplace skills that include critical thinking, communication, collaboration, organizational awareness, ethics, and others. The presented list is defined based on the DARE project recommendations (see Table 4.6).

Table 4.6. The 21st Century workplace skills

Skill ID	**Skill definition**
SK21C	General group definition: Critical thinking, communication, collaboration, organizational awareness, attitude, etc.
SK21C01	1. Critical Thinking: Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
SK21C02	2. Communication: Understanding and communicating ideas
SK21C03	3. Collaboration: Working with other, appreciation of multicultural difference
SK21C04	4. Creativity and Attitude: Deliver high quality work and focus on final result, initiative, intellectual risk
SK21C05	5. Planning & Organizing: Planning and prioritizing work to manage time effectively and accomplish assigned tasks
SK21C06	6. Business Fundamentals: Having fundamental knowledge of the organization and the industry
SK21C07	7. Customer Focus: Actively look for ways to identify market demands and meet customer or client needs
SK21C08	8. Working with Tools & Technology: Selecting, using, and maintaining tools and technology to facilitate work activity
SK21C09	9. Dynamic (self-) re-skilling: Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adopt to changes
SK21C10	10. Professional network: Involvement and contribution to professional network activities
SK21C11	11. Ethics: Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation

4.6.3 Data Scientist in the modern agile data driven organisation

Companies intending to implement data driven business methods and benefit from available data or data that can be collected expect that Data Scientist will provide necessary expertise and insight to achieve the company's goals. In these cases, Data Scientist will face and will need to cope with the expectations to his or her role in organisation which are in some cases far beyond ordinary analyst, engineer or programmer. The following list od expected Data Scientist's contribution is compiled from the collected information and other studies:

- Optimise, improve what related to organizational mission, goals, performance
- Support, advise what related to organizational processes, roles
- Develop, implement and operate data driven services
- Prepare insightful report, targeted analysis
- Monitor processes and services with smart data
- Discover new relations and realise new possibilities
- Use scientific/research methods to discover new relations and solve problems
- Translate business/organizational needs to computational tasks
- Manage data: collect, aggregate, curate, search, visualize

Following observation that organisations expect that Data Scientist will bring general Big Data and Data Science knowledge to organisation, we can see a need for the general Data Science literacy in organisation along competences and skills listed in section 4.5. This means that management and all workers would need to obtain general knowledge on data analytics methods, visualisation, data management, data presentation and structures, understand data analytics and other tools.

This should motivate the general Data Science literacy training in organisations what should be a responsibility of the management. Such training should also focus on a general data presentation and visualisation to enable effective communication of the results and clear definition of the data analytics tasks.

4.7 Data Science Literacy: Commonly required competences and skills for Data Science related and enabled occupations/roles

Data Scientist can do data analytics work and provide important insight into organisational, process or events related data. However, for the Data Scientists or data analytic team to work effectively, there is a need for common knowledge and understanding of the data analysis process and its place in the whole data lifecycle and organisational data driven workflow. This can be achieved by defining a common required knowledge and skills in data handling and data analytics. The goal of this is to enable all workers and roles correctly handle data, collect and present them to analysis, understand the outcome the analysis and provide possible feedback from the domain expertise point of view.

Following outcome and recommendations by the DARE project we define the basic Data Science Analytics competences and skills (also can be referred as literacy) that must be required from all roles working in the Data Science team or interacting with the data analytics teams.

The following competences and skills are defined as basic or common literacy level:

- **Statistical techniques:** General statistical analysis techniques and their use for data inspection, exploration, analysis and visualisation (as supporting activity for more complex data analysis).
- **Computational thinking and programming with data:** Apply information technology, computational thinking, and utilize programming languages and software and hardware solutions for data analysis.
- **Programming languages and tools for data analysis:** Use general and specialised statistical and data analysis programming languages and tools to develop specialised data analysis processes and applications
- **Data visualization languages and tools:** Create and communicate compelling and actionable insights from data using visualization and presentation tools and technologies.
- **Data Management:** Data collection, data entry and annotation, data preparation, data and files versioning, Data Management Plan (DMP), metadata, Open Data, data repositories

4.8 Relation between Data Scientist and Subject Domain specialist

Data Scientist by definition is playing assistant role to the main organisational management (decision making) role or a subject domain scientific/researcher role to help them with organizing data management and data processing to achieve their specific management or research role. However, Data Scientist has also an opportunity to play a leading role in some data driven projects or functions because of their potentially wider vision of the organisational processes or influencing factors.

To understand this, we need to look closer at relation between Data Scientist and subject domain specialist. The subject domain is generally defined by the following components:

- Model (and data types)
- Methods (and additionally theory)
- Processes
- Domain specific data types and presentation, including visualization methods
- Organisational roles and relations

Data Scientist as an assistant to the subject domain specialist will do the following work that should bring benefits to organisation or facilitate scientific discovery:

- Translate subject domain Model, Methods, Processes into abstract data driven form

- Implement computational models in software, build required infrastructure and tools
- Do (computational) analytic work and present it in a form understandable to subject domain
- Discover new relations originated from data analysis and advice subject domain specialist
- Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Figure 6 illustrates relations between subject domain components and those mapped to Data Science domain which is abstract, formalised and data driven.

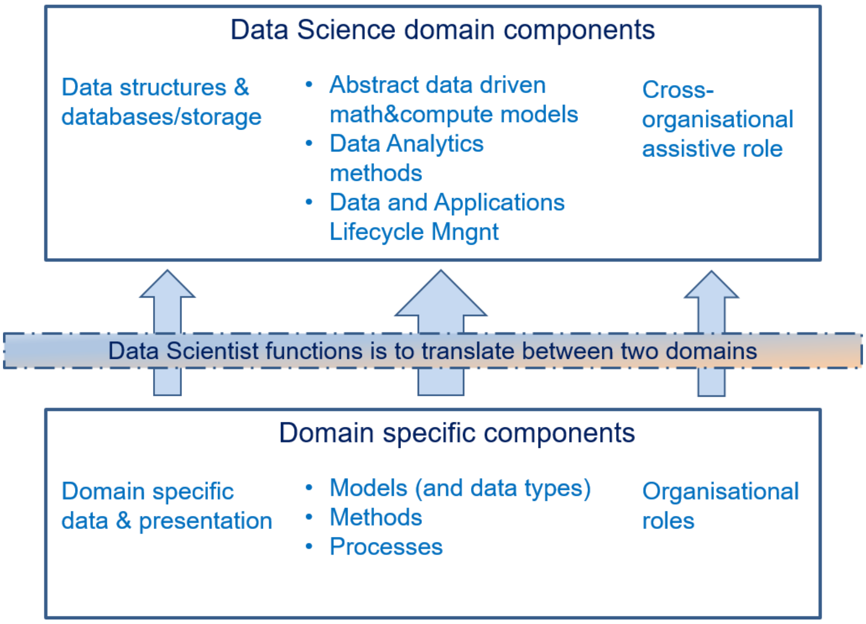

Data Science domain components
Data structures & databases/storage
• Abstract data driven math&compute models
• Data Analytics methods
• Data and Applications Lifecycle Mngnt
Cross-organisational assistive role
Data Scientist functions is to translate between two domains
Domain specific components
Domain specific data & presentation
• Models (and data types)
• Methods
• Processes
Organisational roles

Figure 6. Relations between subject domain and Data Science domain and role of Data Scientist.

Formalisation of the relations between the components and work activities of the subject domain specialist/scientist and Data Science domain provides additional arguments to the discussion about the Data Scientist contribution to the scientific research and discovery that has been recently disputed in many forums: Should Data Scientist be treated as an author of the potential scientific discovery, or just be acknowledged for contribution as assistant role?

5 Example of Data Science Analytics Competences definition

This section provides examples of the detailed competences definition for the Data Science competence group in a format similar to e-CF3.0. This includes the definition of the proficiency levels, mapping to identified skills and knowledge subjects.

Note: The presented definition of the Data Science Analytics competences is done to the best of authors knowledge and is provided as an example and a request for comments.

<table>
<tr><td>Dimension 1 Competence Group</td><td>DSDA</td><td colspan="2">Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA01</td><td colspan="2">Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle</td></tr>
<tr><td></td><td></td><td colspan="2"></td></tr>
<tr><td rowspan="2">Dimension 3 Proficiency level</td><td>Level 1 (Entry/Associate)</td><td>Level 1 (Professional)</td><td>Level 1 (Expert)</td></tr>
<tr><td>Understand and be able to select an approach to analyzing selected datasets.
Demonstrate understanding and perform statistical hypothesis testing, explain statistical significance.</td><td>Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction</td><td>Develop and plan required data analytics for organizational tasks, including: evaluating requirements and specifications of problems to recommend possible analytics-based solutions</td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td colspan="2">Knowledge unit definition</td></tr>
<tr><td rowspan="8">Knowledge</td><td>KDSDA01</td><td colspan="2">Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others</td></tr>
<tr><td>KDSDA02</td><td colspan="2">Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)</td></tr>
<tr><td>KDSDA03</td><td colspan="2">Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)</td></tr>
<tr><td>KDSDA04</td><td colspan="2">Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)</td></tr>
<tr><td>KDSDA06</td><td colspan="2">Predictive Analytics</td></tr>
<tr><td>KDSDA07</td><td colspan="2">Prescriptive Analytics</td></tr>
<tr><td>KDSDA11</td><td colspan="2">Data preparation and pre-processing</td></tr>
<tr><td>KDSDA12</td><td colspan="2">Performance and accuracy metrics</td></tr>
<tr><td></td><td>Skill ID</td><td colspan="2">Skills definition</td></tr>
<tr><td rowspan="5">Skills Data Analytics methods and algorithms</td><td>SDSDA01</td><td colspan="2">Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)</td></tr>
<tr><td>SDSDA02</td><td colspan="2">Use Data Mining techniques</td></tr>
<tr><td>SDSDA04</td><td colspan="2">Apply Predictive Analytics methods</td></tr>
<tr><td>SDSDA05</td><td colspan="2">Apply Prescriptive Analytics methods</td></tr>
<tr><td>SDSDA06</td><td colspan="2">Use Graph Data Analytics for organisational network analysis, customer relations, other tasks</td></tr>
<tr><td rowspan="9">Skills Data Analytics languages, tools and platforms</td><td>DSALANG01</td><td colspan="2">R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)</td></tr>
<tr><td>DSALANG02</td><td colspan="2">Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)</td></tr>
<tr><td>DSADB01</td><td colspan="2">SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)</td></tr>
<tr><td>DSADB03</td><td colspan="2">NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)</td></tr>
<tr><td>DSAVIZ02</td><td colspan="2">Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)</td></tr>
<tr><td>DSABDA02</td><td colspan="2">Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)</td></tr>
<tr><td>DSABDA03</td><td colspan="2">Real time and streaming analytics systems (Flume, Kafka, Storm)</td></tr>
<tr><td>DSABDA09</td><td colspan="2">Kaggle competition, resources and community platform</td></tr>
<tr><td>DSADEV03</td><td colspan="2">Git versioning system as a general platform for software development</td></tr>
</table>

<table>
<tr><td>Dimension 1 Competence Group</td><td>DSDA</td><td>Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA02</td><td>Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>Dimension 3 Proficiency level</td><td colspan="2"><table><tr><th>Level 1 (Entry/Associate)</th><th>Level 1 (Professional)</th><th>Level 1 (Expert)</th></tr><tr><td>Be familiar and use related methods and tools.
Work under supervision or guidance</td><td>Independent work and development.
Knowledge and experience with multiple techniques ad tools.
Full applications development and deployment</td><td>Expert knowledge and experience with multiple data analytics techniques, tools and platforms,
Architecture level development, assessment and selection of appropriate solution. Suggestions for new approaches and applications, including relevant data collection.</td></tr></table></td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td>Knowledge unit definition</td></tr>
<tr><td>Knowledge</td><td>KDSDA01</td><td>Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others</td></tr>
<tr><td></td><td>KDSDA02</td><td>Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)</td></tr>
<tr><td></td><td>KDSDA04</td><td>Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)</td></tr>
<tr><td></td><td>KDSDA06</td><td>Predictive Analytics</td></tr>
<tr><td></td><td>KDSDA14</td><td>Optimisation</td></tr>
<tr><td></td><td>KDSDA15</td><td>Simulation</td></tr>
<tr><td></td><td>Skill ID</td><td>Skills definition</td></tr>
<tr><td>Skills Data Analytics methods and algorithms</td><td>SDSDA01</td><td>Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)</td></tr>
<tr><td></td><td>SDSDA02</td><td>Use Data Mining techniques</td></tr>
<tr><td></td><td>SDSDA04</td><td>Apply Predictive Analytics methods</td></tr>
<tr><td></td><td>SDSDA13</td><td>Apply oprtimisation methods</td></tr>
<tr><td></td><td>SDSDA14</td><td>Use computer simulation methods</td></tr>
<tr><td>Skills Data Analytics languages, tools and platforms</td><td>DSALANG02</td><td>Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)</td></tr>
<tr><td></td><td>DSADB01</td><td>SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)</td></tr>
<tr><td></td><td>DSADB03</td><td>NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)</td></tr>
<tr><td></td><td>DSABDA02</td><td>Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)</td></tr>
<tr><td></td><td>DSABDA03</td><td>Real time and streaming analytics systems (Flume, Kafka, Storm)</td></tr>
<tr><td></td><td>DSADEV01</td><td>Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others</td></tr>
<tr><td></td><td>DSADEV03</td><td>Git versioning system as a general platform for software development</td></tr>
</table>

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA03	Identify, extract, and pull together available and pertinent heterogeneous data, including modern data sources such as social media data, open data, governmental data
Dimension 3 Proficiency level	Level 1 (Entry/Associate) Collect data from multiple sources, apply data quality check, use corresponding APIs to access different data sources. Be able to write SQL and ETL scripts. Level 1 (Professional) Collect and integrate necessary data sources. Define necessary transformations and data preparation procedures, write necessary pipelines. Level 1 (Expert) Identify existing and suggest new data required for organisational analytics tasks to deliver maximum insight. Verify data quality and veracity. Define policy and manage IPR issues.	
Dimension 4	Knowledge ID	Knowledge unit definition
Knowledge	KDSDA10	Natural language processing
	KDSDA11	Data preparation and pre-processing
	KDSDM04	Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)
	KDSDM05	Data lifecycle and organisational workflow, data provenance and linked data
	KDSDM06	Data curation and data quality, data integration and interoperability
	KDSDM08	Metadata, PID, data registries, data factories, standards and com0liance
	KDSDM09	Open Data, Open Science, research data archives/repositories, Open Access, ORCID
	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA08	Apply analytics and statistics methods for data preparation and pre-processing
	SDSENG03	Use cloud based Big Data technologies for large datasets processing systems and applications
	SDSENG07	Use relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for structured and unstructured data
	SDSDM06	Consistently implement data curation and data quality controls, ensure data integration and interoperability
Skills Data Analytics languages, tools and platforms	DSALANG02	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)
	DSADB03	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)
	DSADM02	Data Warehouse platform and related tools
	DSADM05	Big Data and cloud based storage platforms and services
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)

Dimension 1 Competence Group	DSDA	Data Science Analytics	
Dimension 2 Competence	DSDA04	Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	
Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 1 (Professional)	Level 1 (Expert)
	Be familiar and be able to use different performance and accuracy metrics as part of used data analytics platforms	Select appropriate performance metrics and apply them for specific analytics applications. Develop new metrics and use it for fine tuning the used analytics solutions.	Not specifically defined. Advanced knowledge and experience.
Dimension 4	Knowledge ID	Knowledge unit definition	
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	
	KDSDA06	Predictive Analytics	
	KDSDA11	Performance and accuracy metrics	
	KDSDA14	Optimisation	
	Skill ID	Skills definition	
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	
	SDSDA04	Apply Predictive Analytics methods	
	SDSDA09	Be able to use performance and accuracy metrics for data analytics assessment and validation	
Skills Data Analytics languages, tools and platforms	DSALANG01	R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)	
	DSALANG02	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)	
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	
	DSABDA09	Kaggle competition, resources and community platform	

Dimension 1 Competence Group	DSDA	Data Science Analytics	
Dimension 2 Competence	DSDA05	Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	
Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 1 (Professional)	Level 1 (Expert)
	Develop analytics solutions for specific tasks and pre-defined data sets. Ensure correct interaction with other components of the application.	Develop organisational analytics applications that support the whole organisational data lifecycle. Integrate and deploy all components. Integrate analytics application with the enterprise information system.	Plan, design, develop, implement analytics for organizational tasks. Develop the whole data processing workflow and integrate it with organisational workflow. Use research methods principles in developing data driven applications and implementing the whole cycle of data handling
Dimension 4	Knowledge ID	Knowledge unit definition	
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	
	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	
	KDSDA06	Predictive Analytics	
	KDSDA07	Prescriptive Analytics	
	KDSENG04	Applications software requirements and design, agile development technologies, DevOps and continuous improvement cycle	
	KDSENG07	Relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	
	Skill ID	Skills definition	
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	
	SDSDA04	Apply Predictive Analytics methods	
	SDSDA05	Apply Prescriptive Analytics methods	
	SDSENG01	Use systems and software engineering principles to organisations information system design and development, including requirements design	
	SDSENG02	Use Cloud Computing technologies and cloud powered services design for data infrastructure and data handling services	
	SDSRM01	Use research methods principles in developing data driven applications and implementing the whole cycle of data handling	
Skills Data Analytics languages, tools and platforms	DSALANG01	R and data analytics libraries	
	DSALANG02	Python and data analytics libraries	
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)	
	DSADB03	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)	
	DSAVIZ02	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)	
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	
	DSADEV01	Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others	
	DSADEV03	Git versioning system as a general platform for software development	

<table>
<tr><td>Dimension 1 Competence Group</td><td colspan="3">DSDA
Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA06</td><td colspan="2">Visualise results of data analysis, design dashboard and use storytelling method</td></tr>
<tr><td rowspan="2">Dimension 3 Proficiency level</td><td>Level 1 (Entry/Associate)</td><td>Level 1 (Professional)</td><td>Level 1 (Expert)</td></tr>
<tr><td>Use visualisation techniques and tools for existing data set and applications. Develop simple dashboards</td><td>Use multiple visualisation techniques, languages for existing and new analytics applications and processes.
Develop new visualisation solutions and advanced dashboards.</td><td>Define best visualisation approach and solutions for specific business bases. Use multiple techniques to create interactive dashboards.</td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td colspan="2">Knowledge unit definition</td></tr>
<tr><td rowspan="3">Knowledge</td><td>KDSDA04</td><td colspan="2">Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)</td></tr>
<tr><td>KDSDA06</td><td colspan="2">Predictive Analytics</td></tr>
<tr><td>KDSDA07</td><td colspan="2">Prescriptive Analytics</td></tr>
<tr><td></td><td>Skill ID</td><td colspan="2">Skills definition</td></tr>
<tr><td rowspan="3">Skills Data Analytics methods and algorithms</td><td>SDSDA01</td><td colspan="2">Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)</td></tr>
<tr><td>SDSDA02</td><td colspan="2">Use Data Mining techniques</td></tr>
<tr><td>SDSDA10</td><td colspan="2">Use effective visualiation and storytelling methods to create dashboards and data analytics reports</td></tr>
<tr><td rowspan="7">Skills Data Analytics languages, tools and platforms</td><td>DSALANG01</td><td colspan="2">R and data analytics libraries for data visualisation</td></tr>
<tr><td>DSALANG02</td><td colspan="2">Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)</td></tr>
<tr><td>DSAVIZ01</td><td colspan="2">Data visualization Libraries (mathpoltlib, seaborn, D3.js, FusionCharts, Chart.js, other)</td></tr>
<tr><td>DSAVIZ02</td><td colspan="2">Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)</td></tr>
<tr><td>DSAVIZ03</td><td colspan="2">Online visualization tools (Datawrapper, Google Visualisation API, Google Charts, Flare, etc)</td></tr>
<tr><td>DSABDA02</td><td colspan="2">Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)</td></tr>
<tr><td>DSABDA05</td><td colspan="2">Azure Data Analytics platforms (HDInsight, APS and PDW, etc)</td></tr>
</table>

6 Alignment with other competence frameworks and suggested extensions

6.1 Proposed e-CF3.0 extension with the Data Science related competences

The proposed new competence groups provide a basis for defining new competences related to Data Science that can be added to the existing e-CF3.0. In particular, this report suggests the following additional e-competences related to Data Scientist functions as listed in Table 3.4 (assigned numbers are continuation of the current e-CF3.0 numbering). When defining individual professional profile or role the presented competences can be combined with those generic listed in original e-CF3.0 because normally Data Scientist need to have basic or advanced knowledge and skills in general ICT domain.

Table 6.1. Proposed e-CF3.0 extension with the Data Science related Competences

Competence group	Competences related to Data Science	Corresponding CF-DS competence groups
A. PLAN (and Design)	A.10* Organisational workflow/processes model definition/formalization A.11* Data models and data structures	DSDA DSENG
B. BUILD (Develop and Deploy/ Implement)	B.7* Apply data analytics methods (to organizational processes/data) B.8* Data analytics application development B.9* Data management applications and tools B.10* Data Science infrastructure deployment (including computing, storage and network facilities)	DSDA DSENG DSDM
C. RUN (Operate)	C.5* User/Usage data/statistics analysis C.6* Service delivery/quality data monitoring	DSDM DSENG
D. ENABLE (Use/Utilise)	D10. Information and Knowledge Management (powered by Data Science Analytics) - *refactored* D.13* Data analysis, insight or actionable information extraction, visualisation D.14* Support business processes/roles with data analytics, visualisation and reporting (support to D.5, D.6, D.7, D.12) D.15* Data management, curation, preservation, provenance	DSDA DSDK/DSBA
E. MANAGE	E.10* Support Management and Business Improvement with data and insight (data driven organisational processes management) (support to E.5, E.6) E.11* Data analytics for (business) Risk Analysis/Management (support to E.3) E.12* ICT and Information security monitoring and analysis (support to E.8)	DSDA DSENG DSDM

Analysis of the demanded Data Scientist functions and responsibilities in relations to typical organisational workflow revealed that Data Scientist roles and functions can be treated as rather cross-organisational and crossing-multiple competence area (as defined by e-CF3.0); they are rather linked to research or business process management lifecycle than to organisational structure.

6.2 Mapping Data Science competences to CRISP-DM model

Although initially proposed in 1990s, CRISP-DM (Cross Industry Standard Process for Data Mining) [26] model is still used in defining Data Mining and Data Analytics workflows and processes. It is also used for defining common Data Mining and Data Analytics processes and stages, however not limited to data analytics or data management. Figure 7 illustrates CRISP-DM stages. It is important to mention that modern agile technologies and agile business technologies engage the main data handling and data analytics processes into continuous development and continuous improvement cycle.

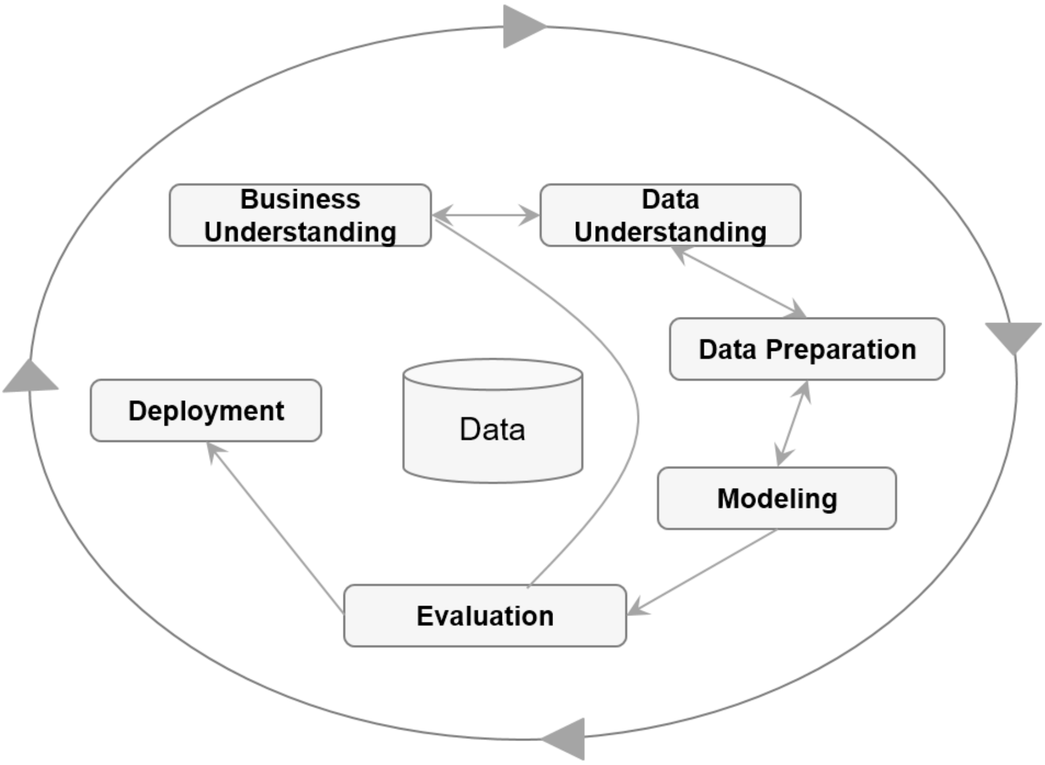


Figure 7. CRISP-DM (Cross Industry Standard Process for Data Mining)

Table 6.2. provides example of initial mapping CF-DS competences to the CRISP-DM processes and stages that will ned to undergo cross-checking with the corresponding knowledge subjects in DS-BoK.

Table 6.2. Mapping CF-DS competences to the CRISP-DM processes and stages

CRISP-DM Processes and Stages	Description	Mapping to CF-DS
Business Understanding	General Business understanding, role of data and required actionable information	DSBAxx DSRMPxx
Determine Business Objectives	Business Objectives (SMART approach). Specific, Measurable, Attainable (in principle), Relevant and Timely. This is performed by Business Stakeholders!	DSBA01
	Business Success Criteria (or benchmark or threshold values)	
Assess Situation	Inventory of Resources, Requirements, Assumptions and Constraints	DSBA01
	Risks and Contingencies	
	Costs and Benefits	
Determine Data Mining Goals	Data Mining Goals	DSRMP05, DSRMP06
	Data Mining Success Criteria	
Produce Project Plan	Project Plan	DSRMP05, DSRMP06

	Initial Assessment of Tools and Techniques	
Data Acquisition and Understanding	Collect data, assign metadata, explore data, run ETL processes	DSDAxx DSDMxx
Collect Initial Data	Acquire access to data from internal and external sources (API, webscrapping). In a steady state, data extraction and transfer routines would be in place.	DSDA03
Describe Data	Describe data, add metadata	DSDA03 DSDM04
Explore Data	Checking on definitions and meaning of data acquired. This requires Business Knowledge (Business Analyst/ or business stakeholder)	DSDA03 DSBA01
	Examine the ´surface´ properties of the acquired data	
	Understand distribution of Key attributes, perform initial visualisations, understand initial relationships between small number of attributes, perform simple aggregations	
Verify Data Quality	Checking if data is up-to-date	DSDA03
	Checking if data is complete, correct, error-free	
Data Preparation (select and cleanse)	Data preprocessing, cleaning, reduction, sampling	DSDAxx
Select Data	Decide on data to be used for analysis	DSDA03
Clean Data	Increase data quality by substitution, imputation (estimating of missing data) / insertion of suitable defaults. Identification of outliers, anomalies and patterns	DSDA03, DSDM05
Construct Data	Transform data set, produce derived values, produce new (composed) records	DSDA03
Integrate Data	Merging of tables (joins) or aggregations of data	DSDA03, DSDM03
Format data	Syntactic modifications (not changing meaning but produce format required by modeling tool (e.g. Convert dataset to JSON)	DSDA03, DSDM03
Hypothesis and Modelling		DSDAxx
Select Modelling Techniques	Decide on techniques to be used, depending on type of problem (ML, Decision Tree, Neural Nets, etc.)	DSDA01
Generate Test Design	Generate procedure or mechanism to test model quality and validity. Separate data set in train and validation sets and test sets	DSDA01
Build Model	Run the modeling tool on the prepared dataset to create on or more models. Perform parameter selection (e.g. hyper param)	DSDA01, DSDA02
Assess Model & Revise Parameters	Judge success of the application of modeling and discovery technically: contact business analysts and domain experts in order to discuss model.	DSDA04

	Summarize qualities of generated models (i.e. Accuracy, etc).	
	Revise parameter settings and tune them for the next run in the Build Model task	
Evaluate Results	Summarize assessment results in terms of business success criteria. This involves Business Stakeholders. This is not to evaluate the models accuracy/ generalizaton: this is already done in the previous step	DSDA04
Review Process and determine improvement	Formally assess data analytics process.	DSDA04, DSRM01
Deployment, Operations & Maintenance	Deploy application or process, maintain, prepare	DSRM06
Plan deployment	Determine strategy for deployment, determine how information will be propagated to users, decide how use of result will be monitored and benefits measured	DSRM06
	Plan potential re-coding (e.g. from python to Java for production environment)	
Plan Monitoring & Maintenance	Check for dynamic aspects, decide how accuracy will be monitored, determine threshold below which result can not be used anymore or should be updated/recalibrated. Monitor and measure performance of model.	DSDA05, DSRMP06
	Plan for DEVOPS or Continuous delivery/ Agile development	DSENGxx
Produce Final Report	Produce final report, create dashboard for proper visualisation.	DSDA06, DSRM06

6.3 Process Groups in Data Management and their mapping to CF-DS competences

Data handling includes multiple stages and processes that can defined as the data lifecycle that can be related to data management processes or to more general project management processes.

The following Process Groups can be identified based on existing Data Lifecycle Management models (as reviewed in Appendix C) and corresponding processes definitions in Data Management BoK (DMBOK) [27] and Project Management BoK (PMBOK) [28]:

1. **Data Identification and Creation**: how to obtain digital information from in-silico experiments and instrumentations, how to collect and store in digital form, any techniques, models, standard and tools needed to perform these activities, depending from the specific discipline.
2. **Data Access and Retrieval**: tools, techniques and standards used to access any type of data from any type of media, retrieve it in compliance to IPRs and established legislations.
3. **Data Curation and Preservation:** includes activities related to data cleansing, normalisation, validation and storage.
4. **Data Fusion (or Data integration)**: the integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation.
5. **Data Organisation and Management**: how to organise the storage of data for various purposes required by each discipline, tools, techniques, standards and best practices (including IPRs management and compliance to laws and regulations, and metadata definition and completion) to set up ICT solutions in order to achieve the required Services Level Agreement for data conservation.
6. **Data Storage and Stewardship**: how to enhance the use of data by using metadata and other techniques to establish a long term access and extended use to that data also by scientists and researchers from other disciplines and after very long time from the data production time.
7. **Data Processing**: tools, techniques and standards to analyse different and heterogeneous data coming from various sources, different scientific domains and of a variety of size (up to Exabytes) – it includes notion of programming paradigms.
8. **Data Visualisation and Communication**: techniques, models and best practices to merge and join various data sets, techniques and tools for data analytics and visualisation, depending on the data significant and the discipline.

Majority of the Data Management processes can be mapped to the DSDM competence group but Data Processing and Data Visualisations will require also DSDA competences, while development, deployment and operation corresponding tools may require DSENG competences.

Note, the defined Data Management processes are linked to but don’t substitute the research or business processes management lifecycle which are focused on the delivery of value to scientific research or business.

7 Practical uses of CF-DS

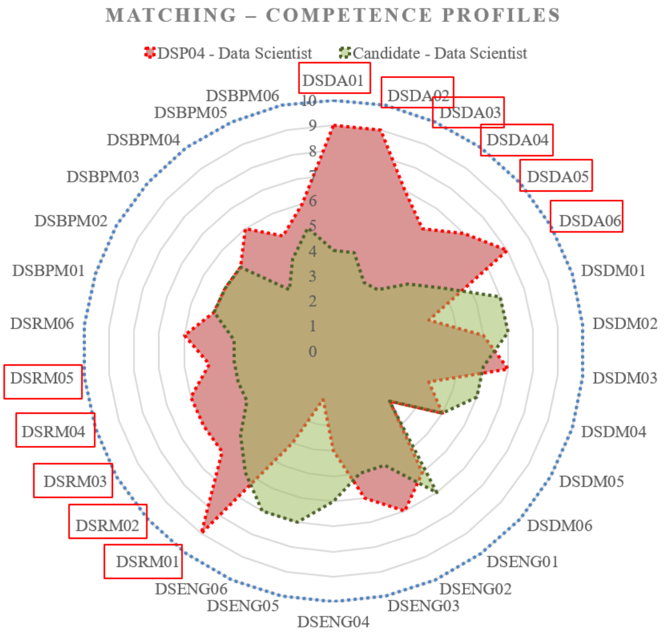
The presented CF-DS provides basis for the definition of all other EDSF components: Data Science Body of Knowledge, Model Curriculum and Data Science Professional Profiles. Competences are used to define required knowledge and learning outcomes is applied to the curriculum design. Data Science professional Profiles are defined based on the set of competences required for each professional profiles or groups of profiles.

Other practical uses include but not limited to:

- Assessment of individual and team competences, as well as balanced Data Science team composition
- Developing tailored curriculum for academic education or professional training, in particular to bridge skills gap and staff up/re-skilling
- Professional certification and self-training.

7.1 Usage example: Competences assessment

Figure 7 illustrates example of the individual competences assessment that maybe used for one of the general use cases: the Data Science practitioner competences assessment against the target/desirable competence profile or role; or competences matching between the job vacancy and the candidate's competence profile.



Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
    - *DSDA01 – DSDA06 Data Science Analytics*
    - *DSRM01 – DSRM05 Data Science Research Methods*
- Can be use for team skills match marking and organisational skills management

Figure 7. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

The intended professional profile or job vacancy are defined in the radial coordinates based on CF-DS competences required for the profiles or vacancy. The candidate's profiles can be defined based on a self-assessment or using simple test. The illustrated competences mismatch can be used either for deciding on the suitability of the candidate or suggesting necessary training program.

Using enumerated set of competences, skills and knowledge units can be used for different applications dealing with competences assessment, knowledge assessment, job vacancy design and candidate assessment.

8 Conclusion and further developments

The presented Data Science Competence Framework Release 2 summarises the framework development since the published Release 1 in October 2016 that presented the set of 4 documents Data Science Competence Framework, Body of Knowledge, Model Curriculum, and Profession Profiles – defined as the EDISON Data Science Framework. The initial CF-DS versions have been created based on extensive analysis of available information that includes Data Science job market study (primarily demand side, i.e. job advertisement), existing standards, best practices, academic publications and blog articles that are posted by experts, practitioners and enthusiasts of the new technology domain and profession of Data Scientist.

The focused work on defining all the foundational components of the whole EDISON framework for consistent Data Science profession definition have been done with wide consultation and engagement of different stakeholders, primarily from research community and Research Infrastructures, but also involving industry via standardisation bodies, professional communities and directly via the project network.

The proposed EDSF Release 2 documents has been updated and extended based on review and contribution from the EDISON Liaison Groups (ELG), individual experts, discussions at a number of workshops and conferences where the EDSF development has been presented, it also incorporated feedback from practitioners that assessed usability and practically used EDSF for curriculum design or review, as well as by organisations used EDSF for defining their skills management and training needs.

8.1 Summary of the recent developments

This document presents ongoing results of the Data Science Competence Framework definition based on the analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for Data Scientist profession in industry and research. The presented CF-DS Release 2 is significantly extended with the skills and knowledge subjects/units related to competences groups. The document also contains the Data Science professional (workplace) skills definition and provides reference to the general "soft" skills often referred to as 21st century skills.

- The presented CF-DS defines five groups of competences for Data Science that include the commonly recognised groups Data Analytics, Data Science Engineering, Domain Knowledge (as defined in the NIST definition of the Data Scientist) and extend them with the two additional groups *Data Management* and *Research Methods* (or Business Process management for business related occupations) that are recognised to be important for the successful work of Data Scientist.
- The document provides example of the individual competences mapping to identified skills and knowledge for the Data Science Analytics competence group.
- The identified competences, skills and knowledge subjects are provided as enumerated lists to allows easy use in applications and developing compatible APIs.
- The report suggests possible extensions to e-CF3.0 on the Data Science related competences.

All defined competences, skills and knowledge units are enumerated and in the future releases will be provided as an API what should simplify the EDSF based applications developments.

8.2 Further developments to formalize CF-DS

The presented CF-DS Release 2 present already mature framework for the Data Science competences definition that has been validated via numerous practical uses. However, further development and community contribution will be required to achieve it wider use and become a standard-de-facto for the Data Science competences definition and curriculum design.

The following CF-DS related developments are suggested:

- Define a taxonomy and classification for Data Science competences, skills and knowledge areas as a basis for more formal CF-DS definition including other components of the intended CF-DS such as proficiency levels (mapped to the expected role of Data Scientist and position seniority), skills and knowledge definition and mapping to individual competences. The intended taxonomy and classification should be compatible with existing frameworks and taxonomies such as e-CF3.0 and ACM CCS2013.

- Run surveys among target EDISON, European Research Infrastructures, Data Science communities, and industry sectors, to validate the proposed CF-DS and collect information from multiple scientific and industry domains. Design a adjustable questionnaire covering the main CF-DS competences
- Use the constructed questionnaire to run few targeted interviews, Data Science experts and top executives at universities and companies to understand intended use of CF-DS, identify case studies, and consequently suggest necessary extensions.
- Provide guidelines for intended CF-DS and EDSF usage in general: job profile and vacancy description generation; individual competences and skills assessment, certification profiles, and others.

Special attention and activity will be undertaken to contribute to the ongoing standardisation process on European ICT professional profiles and competences to extend them with specific Data Science competences and overall taxonomy of the Data Science related competences, skills, and occupations, in particular:

- Provide suggestions to the CEN TC428 for the e-CF extension with the Data Science related and specific data related competences
- Finalise the taxonomy definition for the Data Science related occupations by consulting ESCO committee and practitioners from research and industry on their Human Resource management practices. Provide suggestion for ESCO extension with the Data Science and data related occupations
- Finalise the taxonomy of Data Science related knowledge areas and scientific disciplines based on ACM CCS (2012), provide suggestion for new knowledge areas and classifications classes
- Work with the IEEE and ACM curriculum workshop to define Data Science Curriculum and extend current CCS2012 (Classification Computer Science 2012)

9 References

[1] Data Science Competence Framework(CF-DS) [online] http://edison-project.eu/data-science-competence-framework-cf-ds

[2] Data Science Body of Knowledge (DS-BoK) [online] http://edison-project.eu/data-science-body-knowledge-ds-bok

[3] Data Science Model Curriculum (MC-DS) [online] http://edison-project.eu/data-science-model-curriculum-mc-ds

[4] Data Science Professional Profiles (DSPP) [online] http://edison-project.eu/data-science-professional-profiles-definition-dsp

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

[6] European eCompetences Framework http://www.ecompetences.eu/

[7] 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-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[8] User guide for the application of the European e-Competence Framework 3.0. CWA 16234:2014 Part 2. [online] http://ecompetences.eu/wp-content/uploads/2014/02/User-guide-for-the-application-of-the-e-CF-3.0_CEN_CWA_16234-2_2014.pdf

[9] 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

[10] European Skills, Competences, Qualifications and Occupations (ESCO) [online] https://ec.europa.eu/esco/portal/home

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

[12] Information Technology Competency Model of Core Learning Outcomes and Assessment for Associate-Degree Curriculum(2014) http://www.capspace.org/uploads/ACMITCompetencyModel14October2014.pdf

[13] The U.S. Department of Labor IT Competency Model is available at www.careeronestop.org/COMPETENCYMODEL/pyramid.aspx?IT=Y

[14] Bloom's taxonomy: the 21st century version. [online] http://www.educatorstechnology.com/2011/09/blooms-taxonomy-21stcentury-version.html

[15] Harris, Murphy, Vaisman, Analysing the Analysers. O'Reilly Strata Survey, 2013 [online] http://cdn.oreillystatic.com/oreilly/radarreport/0636920029014/Analyzing_the_Analyzers.pdf

[16] Skills and Human Resources for e-Infrastructures within Horizon 2020, The Report on the Consultation Workshop, May 2012. [online] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/report_human_skills.pdf

[17] Auckland, M. (2012). Re-skilling for research. London: RLUK. [online] http://www.rluk.ac.uk/files/RLUK%20Re-skilling.pdf

[18] Big Data Analytics: Assessment of demand for Labour and Skills 2013-2020. Tech Partnership publication, SAS UK & Ireland, November 2014 [online] https://www.e-skills.com/Documents/Research/General/BigData_report_Nov14.pdf

[19] Italian Web Association (IWA) WSP-G3-024. Date Scientist [online] http://www.iwa.it/attivita/definizione-profili-professionali-per-il-web/wsp-g3-024-data-scientist/

[20] LERU Roadmap for Research Data, LERU Research Data Working Group, December 2013 [online] http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

[21] EDSA Project Data Science skills classification and vocabulary

[22] ELIXIR community projects RITrain and CORBEL dealing with competences and skills definition for bioinformaticians as an example of Data Science enabled professions

[23] 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

[24] Burning Glass Technology, IBM, and BHEF report "The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market" (April 2017) http://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market

[25] DARE Project Recommended Data Science and Analytics Skills – To be published 2017, currently in work.

[26] Cross Industry Standard Process for Data Mining (CRISP-DM) Reference Model [online] http://crisp-dm.eu/reference-model/

[27] Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

[28] Project Management Professional Body of Knowledge (PM-BoK) [online] http://www.pmi.org/PMBOK-Guide-and-Standards/pmbok-guide.aspx

[29] Data Life Cycle Models and Concepts, CEOS Version 1.2. Doc. Ref.: CEOS.WGISS.DSIG, 19 April 2012

[30] European Union. A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe. Brussels : European Commission, 2012. Final Report. Contributing author. Internal identification SMART-Nr 2011/0056. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

[31] Demchenko, Yuri, Peter Membrey, Paola Grosso, Cees de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA. ISBN: 978-1-4673-6402-7; IEEE Catalog Number: CFP1316A-CDR.

[32] NIST SP 1500-6 NIST Big Data interoperability Framework (NBDIF): Volume 6: Reference Architetcure, September 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf

[33] E. Bright Wilson Jr., An Introduction to Scientific Research, Dover Publications; Rev Sub edition, January 1, 1991

[34] Scientific Methods, Wikipedia [online] https://en.wikipedia.org/wiki/Scientific_method

[35] Research Methodology [online] https://explorable.com/research-methodology

[36] 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/

[37] Business process management, Wikipedia [online] https://en.wikipedia.org/wiki/Business_process_management

[38] Theodore Panagacos, The Ultimate Guide to Business Process Management: Everything you need to know and how to apply it to your organization Paperback, CreateSpace Independent Publishing Platform (September 25, 2012)

[39] Tobias Weigel, Timothy DiLauro, Thomas Zastrow, RDA PID Information Types WG: Final Report, Research Data Alliance, 2015/07/10 [online] https://b2share.eudat.eu/record/245/files/PID%20Information%20Types%20Final%20Report.pdf

Acronyms

Acronym	**Explanation**
ACM	Association for Computer Machinery
BABOK	Business Analysis Body of Knowledge
CCS	Classification Computer Science by ACM
CF-DS	Data Science Competence Framework
CODATA	International Council for Science: Committee on Data for Science and Technology
CS	Computer Science
DM-BoK	Data Management Body of Knowledge by DAMAI
DS-BoK	Data Science Body of Knowledge
EDSA	European Data Science Academy
EOEE	EDISON Online E-Learning Environment
ETM-DS	Data Science Education and Training Model
EUDAT	http://eudat.eu/what-eudat
EGI	European Grid Initiative
ELG	EDISON Liaison Group
EOSC	European Open Science Cloud
ERA	European Research Area
ESCO	European Skills, Competences, Qualifications and Occupations
EUA	European Association for Data Science
HPCS	High Performance Computing and Simulation Conference
ICT	Information and Communication Technologies
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
LERU	League of European Research Universities
LIBER	Association of European Research Libraries
MC-DS	Data Science Model Curriculum
NIST	National Institute of Standards and Technologies of USA
PID	Persistent Identifier
PM-BoK	Project Management Body of Knowledge
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
SWEBOK	Software Engineering Body of Knowledge

Appendix A. Data used in the study of demanded Data Science competences and skills

The proposed study has used data collected from job advertisements on such popular job search and employment portals as IEEE Jobs portal and LinkedIn Jobs advertised that provided rich information for defining Data Science competences, skills and required knowledge of Big Data tools and data analytics software. The IEEE Jobs portal posts job advertisements predominantly from US companies and universities. LinkedIn posts vacancies related to region or country from where the request is originated and many job ads are posted in national language. In particular case of this study, the job advertisements were collected for positions available in Netherlands that appeared to be quite extensive and representing the whole spectrum of required competences and skills.

In this initial stage we used set of Data Science job openings from IEEE Jobs portal (around 120) and LinkedIn Netherlands (around 140) collected in period of mid September to beginning of October 2015. A number of Data Science related key words were used like Data Science, Big Data, Data Intensive technologies, data analytics, machine learning. Initial analysis of collected information allowed to make assumption that collected information from more than 250 samples was sufficiently representative for initial study, taking into account that Netherlands is one of leading countries in relation to Big Data and Data Science technologies acceptance and development.

The following are general characteristics of the collected data.

- Total number of advertisements collected: IEEE Jobs – 120; Linkedin Jobs – 140
- Number of advertisement selected for analysis IEEE Jobs – 28; Linkedin Jobs – 30
- Number of companies posted Data Science related jobs – more than 50
- The most active recruiting companies: Booking.com, Scandia, etc.

All working data are available in the shared project storage area on Google Drive and are available on demand.

A.1. Selecting sources of information

To verify existing frameworks and potentially identify new competences, different sources of information have been investigated:

- First of all, job advertisements that represent demand side for Data Scientist specialists and based on practical tasks and functions that are identified by organisations for specific positions. This source of information provided factual data to define demanded competences and skills.
- Structured presentation of Data Science related competences and skills produced by different studies as mentioned above, in particular NIST definition of Data Science that provided a basis for definition of initial 3 groups of skills, namely Data Analytics, Data Science Engineering, and Domain expertise. This information was used to correlate with information obtained from job advertisements.
- Blog articles and community forums discussions that represented valuable community opinion. This information was specifically important for defining practical skills and required tools.

It appeared that the richest information can be collected from job advertisements on such popular job search and employment portals such as IEEE Jobs portal and LinkedIn Jobs advertised. Important to admit that although IEEE Jobs designed to post international job openings, the advertisements are mostly from US companies and universities. LinkedIn posts vacancies related to region or country from where the request is originated and many job ads are posted in national language. In particular case of this report, it was possible to collect information from LinkedIn for Netherlands, however it was quite representative due to a large number of advertisements. This means that at the following stage, the information needs to be collected by EDISON partners in their own countries. The same relates to collecting information from different scientific, technology and industry domains what should take place at next stage of this study.

- If referred to the category of job openings such as academic positions or industry and business related positions, the academic positions didn’t provide valuable information as they don’t specify detailed competences and skills but rather search for candidates who are capable to teach, create or support new academic courses on Data Science.

- In this initial stage we used set of Data Science job openings from IEEE Jobs portal (around 120) and LinkedIn Netherlands (around 140) collected in period of mid-September to beginning of October 2015. A number of Data Science related key words were used like Data Science, Big Data, Data Intensive technologies, data analytics, machine learning. Initial analysis of collected information allowed to make assumption that collected information from more than 250 samples was sufficiently representative for initial study, taking into account that Netherlands is one of leading countries in relation to Big Data and Data Science technologies acceptance and development. See Appendix B for more details about collected data.

A.2. EDISON approach to analysis of collected information

1) Collect data on required competences and skills
2) Extract information related to competences, skills, knowledge, qualification level, and education; translate and/or reformulate if necessary
3) Split extracted information on initial classification or taxonomy facets, first of all, on required competences, skills, knowledge; suggest mapping if necessary
4) Apply existing taxonomy or classification: for purpose of this study we used skills and knowledge groups as defined by the NIST Data Science definition (i.e. Data Analytics, Domain Knowledge, and Engineering) [5]
5) Identify competences and skills groups that don't fit into initial/existing taxonomy and create new competences and skills groups
6) Do clustering and aggregations of individual records/samples in each identified group
7) Verify the proposed competences groups definition by applying to originally collected and new data
8) Validate the proposed CF-DS via community surveys and individual interviews[7].

The Data Science competences and skills defined in this way will be used to provide input to existing professional competence frameworks and profiles:

- Map to e-CF3.0 if possible, suggest new competences
- Map to CWA ICT profiles where possible, suggest new profiles if needed
- Identify inconsistencies in using current e-CF3.0 and CWA ICT profiles and explore alternative frameworks if necessary.

The outlined above process has been applied to the collected information and all steps are tracked in the two Excel workbooks provided as supplementary materials to this report that are available on the project shared storage and later to be available via project wiki

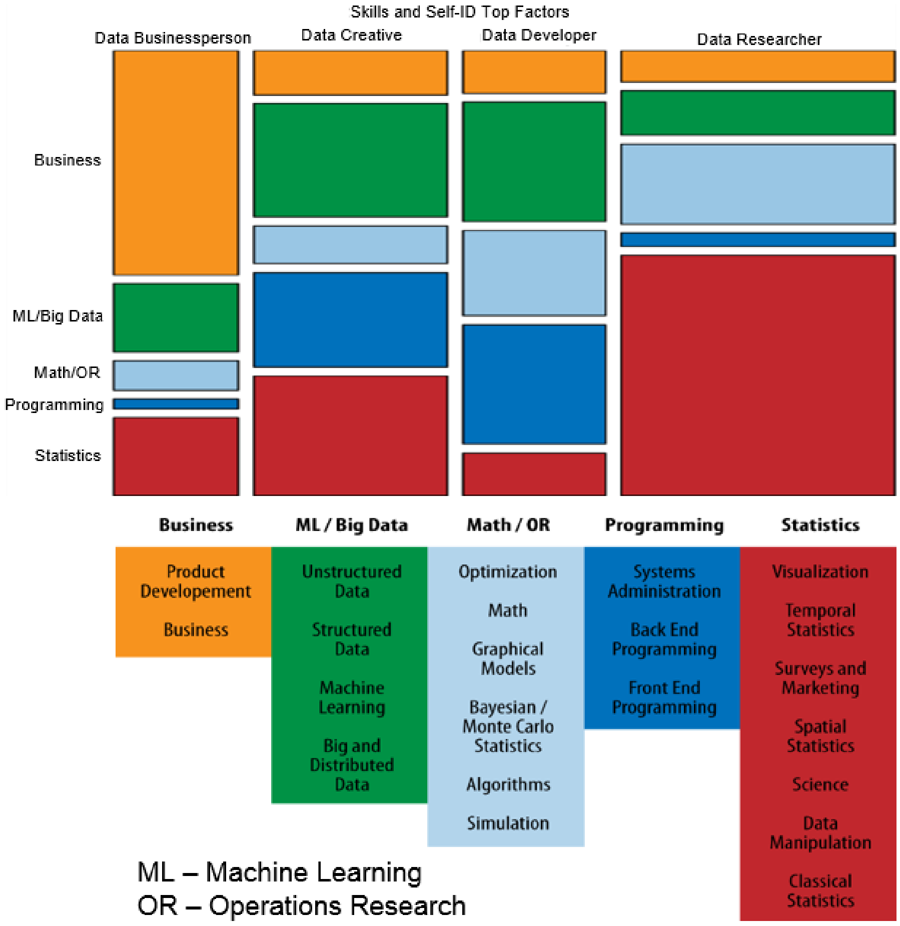
[7] This activity will be done at the next stage and results will be reported in the final deliverable D2.3

Appendix B. Overview: Studies, reports and publications related to Data Science competences and skills

B.1. O'Reilly Strata Survey (2013)

O'Reilly Strata industry research [15] defines the four Data Scientist profession profiles and their mapping to the basic set of technology domains and competencies as shown in Figure A.1. The four profiles are defined based on the Data Scientists practitioners self-identification:

- Data Businessperson
- Data Creative
- Data Developer
- Data Researcher



Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

ML – Machine Learning
OR – Operations Research

Figure A.1. Data Scientist skills and profiles according to O'Reilly Strata survey [15]

Table A.1 below lists skills for Data Science that are identified in the study. They are very specific in technical sense but provide useful information when mapped to the mentioned above Data Science profiles. We will refer to this study in our analysis of CF-DS and related competence groups.

Table A.1. Data Scientist skills identified in the O'Reilly Strata study (2013)

Data Science Skills	Examples -> Knowledge and skills
Algorithms	computational complexity, CS theory
Back-End Programming	JAVA/Rails/Objective C

Bayesian/Monte-Carlo Statistics	MCMC, BUGS
Big and Distributed Data	Hadoop, Map/Reduce
Business	management, business development, budgeting
Classical Statistics	general linear model, ANOVA
Data Manipulation	regexes, R, SAS, web scraping
Front-End Programming	JavaScript, HTML, CSS
Graphical Models	social networks, Bayes networks
Machine Learning	decision trees, neural nets, SVM, clustering
Math	linear algebra, real analysis, calculus
Optimization	linear, integer, convex, global
Product Development	design, project management
Science	experimental design, technical writing/publishing
Simulation	discrete, agent-based, continuous)
Spatial Statistics	geographic covariates, GIS
Structured Data	SQL, JSON, XML
Surveys and Marketing	multinomial modeling
Systems Administration	*nix, DBA, cloud tech.
Temporal Statistics	forecasting, time-series analysis
Unstructured Data	noSQL, text mining
Visualization	statistical graphics, mapping, web-based data visualisation

B.2. Skills and Human Resources for e-Infrastructures within Horizon 2020

The Report on the Consultation Workshop (May 2012) “Skills and Human Resources for e-Infrastructures within Horizon 2020” [16] summarises the outcomes of a consultation workshop that was organised by DG INFSO "GÉANT and e-Infrastructures" unit to consult the stakeholders on their views of approaching these challenges. The workshop discussions highlighted cross-cutting challenges of

i) new and changed skills needs which combine technical and scientific skills and require interdisciplinary thinking and communication;

ii) recognizing new job profiles and tasks rising from the emergence of computing intensive and data-driven science with integral role of e-infrastructures;

iii) need for effective European level collaboration and coordination to avoid duplication of efforts and join the forces for developing high quality human capital for e-infrastructures

Several concrete recommendations for supporting the suggested development aspects with e-Infrastructures activities under Horizon 2020 were devised. It was considered important to have both specific and integrated activities to support skills and human resources aspects within the e-Infrastructures projects.

The report defined three perspectives for e-infrastructure related skills needs:

- Development
  - Create new tools, further develop e-Infrastructure
  - Technological innovations
- Operation
  - Support users, maintain and operate services
  - Process/service innovations
- Scientific use
  - Use ICT tools, apply e-science methods
  - Scientific innovations

The nine areas identified as having potentially the most significant skills gap according to RLUK report “Re-skilling for Research” (2012) [17]

- Ability to advise on preserving research outputs (49% essential in 2-5 years; 10% now)
- Knowledge to advise on data management and curation, including ingest, discovery, access, dissemination, preservation, and portability (48% essential in 2X-5 years; 16% now)
- Knowledge to support researchers in complying with the various mandates of funders, including open access requirements (40% essential in 2-5 years; 16% now)
- Knowledge to advise on potential data manipulation tools used in the discipline/subject (34% essential in 2-5 years; 7% now)
- Knowledge to advise on data mining (33% essential in 2-5 years; 3% now)
- Knowledge to advocate, and advise on, the use of metadata (29% essential in 2-5 years; 10% now)
- Ability to advise on the preservation of project records e.g. correspondence (24% essential in 2-5 years; 3% now)
- Knowledge of sources of research funding to assist researchers to identify potential funders (21% essential in 2-5 years; 8% now)
- Skills to develop metadata schema, and advise on discipline/subject standards and practices, for individual research projects (16% essential in 2-5 years; 2% now)

B.3. UK Study on demand for Big Data Analytics Skills (2014)

The study “Big Data Analytics: Assessment of demand for Labour and Skills 2013-2020” [18] provided extensive analysis of the demand side for Big Data specialists in UK in forthcoming year. Although majority of roles are identified as related to Big Data skills, it is obvious that all these roles can be related to more general definition of the Data Scientist as an organisational role working with Big Data and Data Intensive Technologies.

The report lists the following Big Data roles:

- Big Data Developer
- Big Data Architect
- Big Data Analyst
- Big Data Administrator
- Big Data Consultant
- Big Data Project Manager
- Big Data Designer
- Data Scientist

B.4. IWA Data Science profile

Italian Web Association (IWA) published the WSP-G3-024. Date Scientist Profile for web related projects [19]. It provide a god example of domain specific definition of the Data Science competences, skills and organisational responsibilities, it suggests also mapping to e-CF3.0 competences.

The Data Scientist is defined as “Professional that owns the collection, analysis, processing, interpretation, dissemination and display of quantitative data or quantifiable organization for analytical, predictive or strategic.”

The profile contains the following sections:

- Concise definition
- Mission
- Documentation produced
- Main tasks
- Mapping to e-CF competences
- Skills and knowledge
- Application area of KPI
- Qualifications and certifications (informational)
- Personal attitudes (informational)

- Reports and reporting lines (informational)

For reference purpose, it is worth to mention that IWA Data Scientist profile maps its competences and skills to the following e-CF3.0 competences:

A.6. Application design: Level e-3
A.7. Monitoring of technological Bertrand: Level e-4
B.1. Development of applications: Level e-2
B.3. Testing: Level e-3
B.5. Production of documentation: Level e-3
C.1. User assistance: Level e-3
C.3. Service Delivery: Level e-3
C.4. Management Problem: Levels e-3, e-4.

B.5. Other studies, reports and ongoing projects on defining Data Science profiles and skills

The following reports and studies and ongoing works to define the Data Science skills profiles and needs for European Research Area and industry are considered relevant to current study and will be used to finalise the DS-CF definition:

- LERU Roadmap for Research Data (2013) [20]
- EDSA Project Data Science skills classification and vocabulary [21]
- ELIXIR community projects RITrain and CORBEL dealing with competences and skills definition for bioinformaticians as an example of Data Science enabled professions [22]
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017) [23]
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017) [24]

Appendix C. Concepts and models related to CF-DS definition

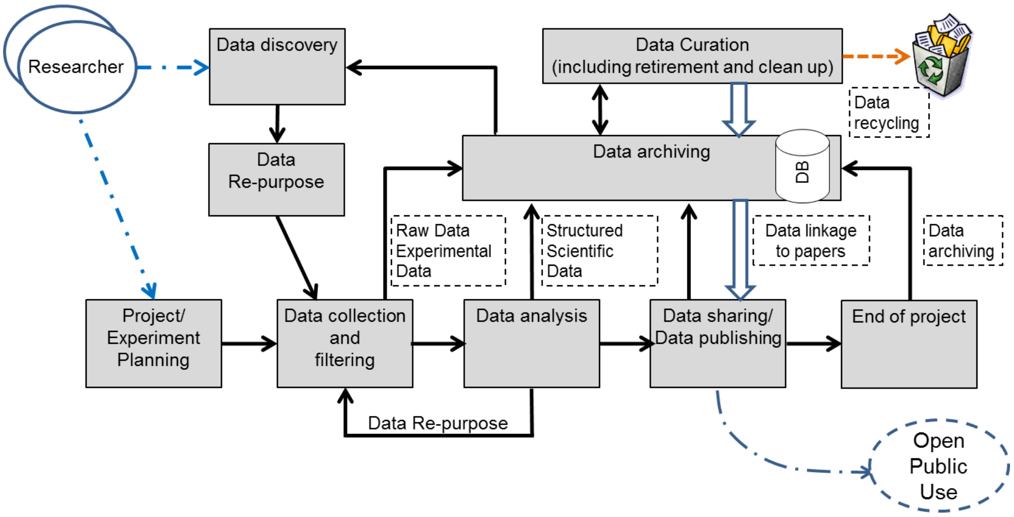
This section provides important definitions that are needed for consistent CF-DS definition in the context of organisational and business processes, e-Infrastructure and scientific research. First of all, this includes definition of typical organisational processes and scientific workflow or research data lifecycle.

C.1. Scientific Data Lifecycle Management Model

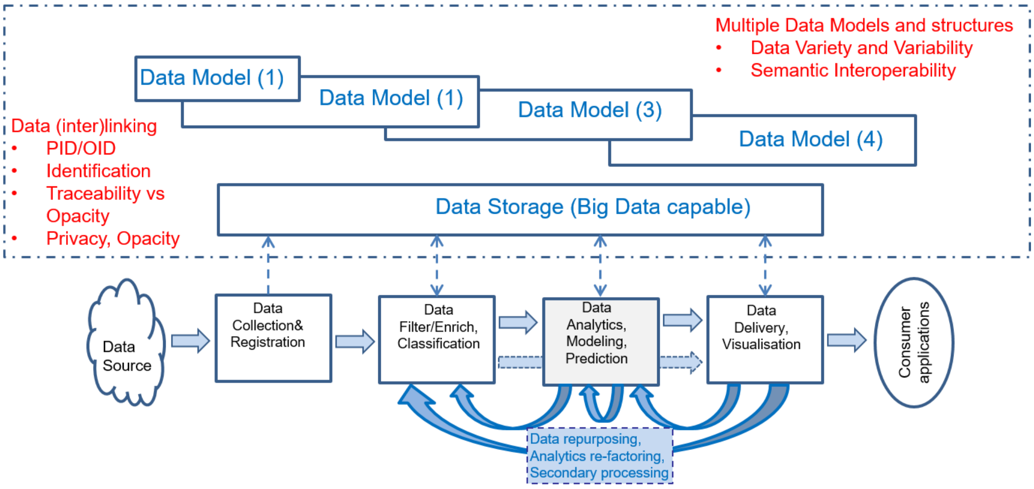
Data lifecycle is an importance component of data centric applications, which Data Science and Big Data applications belong to. Data lifecycle analysis and definition is addressed in many domain specific projects and studies. Extensive compilation of the data life cycle models and concepts is provided in the CEOS.WGISS.DSIG document [29].

For the purpose of defining the major groups of competences required for Data Scientist working with scientific applications and data analysis we will use the Scientific Data Lifecycle Management (SDLM) model [30] shown in Figure 6 (a) defined as a result of analysis of the existing practices in different scientific communities. Figure C.1 (b) illustrates the more general Big Data Lifecycle Management model (BDLM) involving the main components of the Big Data Reference Architecture defined in NIST BDIF [5, 31, 32]. The proposed models are sufficiently generic and compliant with the data lifecycle study results presented in [29].

The generic scientific data lifecycle includes a number of consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding). SDLM reflects complex and iterative process of the scientific research that is also present in Data Science analytics applications.


Researcher
Data discovery
Data Re-purpose
Data Curation (including retirement and clean up)
Data recycling
Data archiving
DB
Raw Data Experimental Data
Structured Scientific Data
Data linkage to papers
Data archiving
Project/ Experiment Planning
Data collection and filtering
Data analysis
Data sharing/ Data publishing
End of project
Data Re-purpose
Open Public Use

(a) Scientific data lifecycle management - e-Science focused


Multiple Data Models and structures
• Data Variety and Variability
• Semantic Interoperability
Data Model (1)
Data Model (1)
Data Model (3)
Data Model (4)
Data (inter)linking
• PID/OID
• Identification
• Traceability vs Opacity
• Privacy, Opacity
Data Storage (Big Data capable)
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer applications
Data repurposing, Analytics re-factoring, Secondary processing

(b) Big Data Lifecycle Management model (compatible with the NIST NBDIF definition)

Figure C.1. Data Lifecycle Management in (a) e-Science and (b) generic Big Data Lifecycle Management model.

Both SDLM and BDLM require data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in Scientific Data Infrastructure (SDI). Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed data lifecycle models and must also be done in a secure and trustworthy way. The research data management and handling issues are extensively addressed in the work of the Research Data Alliance[8].

C.2. Scientific methods and data driven research cycle

For Data Scientist that is dealing with handling data obtained in the research investigation understanding of the scientific methods and the data driven research cycle is essential part knowledge that motivate necessary competences and skills for the Data Scientists for successfully perform their tasks and support or lead data driven research.

The scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge [33, 34, 35]. Traditional steps of the scientific research were developed over time since the time of ancient Greek philosophers through modern theoretical and experimental research where experimental data or simulation results were used to validate the hypothesis formulated based on initial observation or domain knowledge study. The general research methods include: observational methods, opinion based methods, experimental and simulation methods.

The increased power of computational facilities and advent of Big Data technologies created a new paradigm of the data driven research that enforced ability of researchers to make observation of the research phenomena based on bigger data sets and applying data analytics methods to discover hidden relations and processes not available to deterministic human thinking. The principles of the data driven research were formulated in the seminal work “The Fourth Paradigm: Data-Intensive Scientific Discovery” edited by Tony Hey [36].

The research process is iterative by its nature and allows scientific model improvement by using continuous research cycle that typically includes the following basic stages:

- Define research questions
- Design experiment representing initial model of research object or phenomena
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation

[8] Research Data Alliance https://rd-alliance.org/

- Test Hypothesis
- Refine model and start new experiment cycle

The traditional research process may be concluded with the scientific publication and archiving of collected data. Data driven and data powered/driven research paradigm allows research data re-use and combining them with other linked data sets to reveal new relations between initially not linked processes and phenomena. As an example, biodiversity research when studying specific species population can include additional data from weather and climate observation, solar activity, other species migration and technogenic factor.

The proposed CF-DS introduces research methods as an important component of the Data Science competences and knowledge and uses data lifecycle as an approach to define the data management related competences group.

C.3. Business Process Management lifecycle

New generation Agile Data Driven Enterprises (ADDE) use Data Science methods to continuously monitor and improve their business processes and services. The data driven business management model allows combining different data sources to improve predictive business analytics what allows making more effective solutions, faster adaptation of services, and more specifically target different customer groups as well as do optimal resources allocations depending on market demand and customer incentives.

Similarly, to the research domain the data driven methods and technologies change how the modern business operates attempting to benefit from the new insight that big data can give into business improvement including internal processes organisation and relation with customers and market processes. Understanding Business Process Management lifecycle [37, 38] is important to identify necessary competences and knowledge for business oriented Data Science profiles.

The following are typical stages of the Business Process Management lifecycle:

- Define the business target: both services and customers
- Design the business process
- Model/Plan
- Deploy and Execute
- Monitor and Control
- Optimise and Re-design

The need for the Business Process management competences and knowledge for business oriented Data Science profiles is reflected in the proposed CF-DS definition.