EDISON Data Science Framework:
Part 2. Data Science Body of Knowledge (DS-BoK)
Release 1

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

Date	10 October 2016
Document Author/s	Yuri Demchenko, Andrea Manieri, Adam Belloum
Version	Release 1, version 0.3
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)
0.0	22/01/2016	Overview of existing BoKs related to Data Science	AM
0.1	18/03/2016	Updated version for ELG discussion	YD, AM, AB
0.2	4/07/2016	Updated version (after deliverable D2.2)	YD, TW
0.3	9/09/2016	Updated based on feedback from MC-DS implementation	YD
Release 1	10/10/2016	Release 1 after ELG03 meeting discussion	YD

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

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 will be enabled through the creation of a comprehensive framework for Data Science education and training that includes such components as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS). This will provide a formal basis for Data Science professional certification, organizational and individual skills management and career transferability.

This document presents initial results of the Data Science Body of Knowledge (DS-BoK) definition that is linked to and based on the Data Science Competence Framework described in another document. The presented DS-BoK definition is based on overview and analysis of existing bodies of knowledge that are relevant to intended frameworks for Data Science and required to fulfil identified in CF-DS competences and skills.

The definition of the Data Science Body of Knowledge provides a basis for defining the Data Science Model Curriculum and further for the Data Science professional certification.

The presented DS-BoK defines six groups of knowledge areas (KA) that are linked to the identified competence groups: KGA-DSA Data Analytics; KGA-DSDM Data Management, KGA-DSE Data Science Engineering, KGA-DSRM Research Methods; and KGA-DSBM Business Process Management. Defining the domain knowledge groups both for science and business will a subject for further development in tight cooperation with domain specialists.

The intended EDISON framework comprising of 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.

Further work will be required to develop consistent DS-BoK that can be accepted by academic community and professional training community. The DS-BoK is presented to the academic and research community and will undergo wide community discussion via EDISON community forum and by presentation at community oriented workshops and conferences.

TABLE OF CONTENTS

1 Introduction

This document presents initial results of the Data Science Body of Knowledge (DS-BoK) definition that is linked to and based on the Data Science Competence Framework described in another document. The presented DS-BoK definition is based on overview and analysis of existing bodies of knowledge that are relevant to intended frameworks for Data Science and required to fulfil identified in CF-DS competences and skills.

The intended EDSION framework comprising of 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 definition of the Data Science Body of Knowledge will provide a basis for defining the Data Science Model Curriculum and further for the Data Science professional certification.

The presented DS-BoK defines six groups of Knowledge Areas (KA) that are linked to the identified competence groups: KGA-DSA Data Analytics; KGA-DSDM Data Management, KGA-DSE Data Science Engineering, KGA-DSRM Research Methods; and KGA-DSBM Business Process Management. Defining the domain knowledge groups both for science and business will a subject for further development in tight cooperation with domain specialists.

The DS-BoK definition is based on the proposed CF-DS that defines the 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 Data Science) and extends them with the two new groups *Data Management* and *Scientific Methods* (or Business Process management for business related occupations) that are recognised to be important for a successful work of Data Scientist but are not explicitly mentioned in existing frameworks.

Further work will be required to develop consistent DS-BoK that can be accepted by academic community and professional training community. The proposed initial version will be used to initiate community discussion and solicit contribution from the subject matter expects and practitioners. The DS-BoK will be presented to EDISON Liaison Group for feedback and will undergo wide community discussion via EDISON community forum and by presentation at community oriented workshops and conferences.

The presented document has the following structure. Section 2 provides an overview of the EDISON Data Science framework and related project activities that coordinate the framework components development and pilot implementation. Section 3 provides overview of existing BoKs related to Data Science knowledge areas. Section 3 also includes other important components for the DS-BoK definition such as data lifecycle management models, scientific methods, and business process management lifecycle models. Section 4 described the proposed DS-BoK structure and provides the initial definition of the DS-BoK. Section 5 provides summary of the achieved results and section 5 suggests questions for discussion to collects community feedback and experts opinion.

Appendices to this document contain important supplementary information: detailed information about reviewed bodies of knowledge related to identified Data Science knowledge areas; taxonomy of the Data Science knowledge areas and scientific disciplines built as a subset of the ACM CCS (2012) classification; and a short summary of the proposed CF-DS that includes identified competence groups and skills, required technical knowledge of relevant Big Data platforms, analytics and data management tools, and programming languages.

2 EDISON Data Science Framework

The EDISON project is designated 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 will be enabled through the creation of a comprehensive framework for Data Science education and training that includes such components as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS).

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

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- Data Science Professional profiles and occupations taxonomy (DSP)
- 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 Marketplace and Directory
- Data Science Community Portal (CP) that also includes tools for individual competences benchmarking and personalized educational path building
- Certification Framework for core Data Science competences and professional profiles

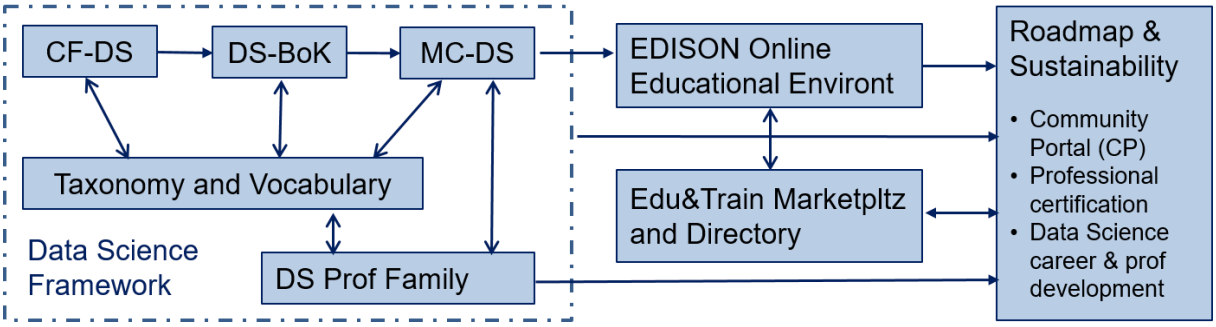


Figure 1 EDISON Data Science Framework components.

The CF-DS includes common competences required for successful work of Data Scientists in different work environments in industry and in research and through the whole career path. 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 required Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs defined based on the Classification Computer Science (CCS2012), components taken from other BoKs and proposed new KAs to incorporate new technologies used in Data Science and their recent developments.

The MC-DS is built based on CF-DS and DS-BoK where Learning Outcomes are defined based on CF-DS competences and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles.

The DSP profiles and Data Science occupations taxonomy are defined based on and as an extension to the European Skills, Competences, Qualifications and Occupations (ESCO). DSP profiles definition will create an important instrument to define effective organisational structures and corresponding roles. DSP can also be used for building individual career path and corresponding competences and skills transferability between organisations and economy sectors.

The Data Science Taxonomy and Scientific Disciplines Classification will serve to maintain consistency between four core components of EDSF. To ensure easy navigation and mapping between the EDSF components, all attributes and properties are enumerated: competences in CF-DS, KAGs and KAs in DS-BoK, LOs and LUs in MC-DS, professional profiles in DSP.

The EDISON Data Science professional ecosystem illustrated in Figure 1 uses core EDSF components to shape and profile the offered services and ensure the EDISON project sustainability. In particular, CF-DS and DS-BoK are used for individual competences and knowledge benchmarking and they are instrumental for constructing personalised learning path and professional (up/re-) skilling based on MC-DS.

3 Overview of BoKs relevant to DS-BoK

The following BoK's have been reviewed to provide a basis for initial definition of the DS-BoK:

- ACM Computer Science Body of Knowledge (ACM CS-BoK) [10, 11]
- ICT professional Body of Knowledge (ICT-BoK) [35]
- Business Analytics Body of Knowledge (BABOK) [36]
- Software Engineering Body of Knowledge (SWEBOK) [37]
- Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [38]
- Project Management Professional Body of Knowledge (PM-BoK) [39]

In the following sections we give a short description of each body of knowledge. The presented analysis provides a motivation that the intended/proposed DS-BoK should incorporate two dimensions in constructing DS-BoK: using Knowledge Areas (KA) and organisational workflow often linked to Project Phases (PP) or industry business process management (BPM). The DS-BoK should also reflect the data-lifecycle management[1]. Each process is defined as one or more activity, related knowledge and tools needed to execute it, inputs from other processes and expected output.

The presented analysis allowed to identify what existing BoK's can be used in the DS-BoK definition or mapped to ensure knowledge transferability and education programmes combination. From this initial analysis the relevant best practices have been identified to structure the DS-BoK and provide basis for defining the EDISON certification scheme and sustainability model.

3.1 ACM Computer Science Body of Knowledge (CS-BoK)

In the ACM-CS2013-final report [8, 9] the Body of Knowledge is defined as a specification of the content to be covered in a curriculum as an implementation. The ACM-BoK describes and structures the knowledge areas needed to define a curriculum in Computer Science, it includes 18 Knowledge Areas (where 6 KAs are newly introduced in ACM CS2013):

AL - Algorithms and Complexity
AR - Architecture and Organization
CN - Computational Science
DS - Discrete Structures
GV - Graphics and Visualization
HCI - Human-Computer Interaction
IAS - Information Assurance and Security (new)
IM - Information Management
IS - Intelligent Systems
NC - Networking and Communications (new)
OS - Operating Systems
PBD - Platform-based Development (new)
PD - Parallel and Distributed Computing (new)
PL - Programming Languages
SDF - Software Development Fundamentals (new)
SE - Software Engineering
SF - Systems Fundamentals (new)
SP - Social Issues and Professional Practice

Knowledge areas should not directly match a particular course in a curriculum (this practice is strongly discouraged in the ACM report), often courses address topics from multiple knowledge areas. The ACM-CS2013-final report distinguish between two type of topics: Core topics subdivided into "Tier-1" (that are mandatory for each curriculum) and "Tier-2" (that are expected to be covered at 90-100% with minimum advised 80%), and elective topics. The ACM classification suggests that a curriculum should include all topics in Tier-1 and all or almost the topics in Tier 2. Tier 1 and Tier 2 topics are defined differently for different programmes and specialisations. To be complete a curriculum should cover in addition to the topics of Core

[1] Such assumption has been also confirmed by user experience and expert interviews conducted in the project task T4.4.

Tier 1 and 2 significant amount of elective material. The reason for such a hierarchical approach to the structure of the Body of Knowledge is a useful way to group related information, not as a structure for organizing material into courses.

The ACM for computing Education in Community Colleges [10, 11] defines a BoK for IT outcome-based learning/education which identifies 6 technical competency areas and 5 work-place skills. While the technical areas are specific to IT competences and specify a set of demonstrable abilities of graduates to perform some specific functions, the so called work-place skills describe the ability the student/trainee to:

(1) function effectively as a member of a diverse team,
(2) read and interpret technical information,
(3) engage in continuous learning,
(4) professional, legal, and ethical behaviour, and
(5) demonstrate business awareness and workplace effectiveness

The CS-BoK uses ACM Computing Classification System (CCS) which is standard and widely accepted what makes it a good basis for using it as a basis for building DS-BoK and providing necessary extensions/KAs related to identified Data Science competence groups (see section 3.4) which majority require background knowledge components from the general CS-BoK.

3.2 ICT professional Body of knowledge ICT-BoK

The ICT-BoK is an effort promoted by the European Commission, under the eSkills initiative (http://eskills4jobs.ec.europa.eu/) to defines and organises the core knowledge of the ICT discipline. In order to foster the growth of digital jobs in Europe and to improve ICT Professionalism a study has been conducted to provide the basis of a "Framework for ICT professionalism" (http://ictprof.eu/). This framework consists of four building blocks which are also found in other professions:

i) body of knowledge (BoK);
ii) competence framework;
iii) education and training resources; and
iv) code of professional ethics.

A competence framework already exists and consists in the e-Competence Framework (now in its version 3.0 and promoted by CEN). However, an ICT Body of Knowledge that provides the basis for a common understanding of the foundational knowledge an ICT professional should possess, is not yet available.

The ICT-BoK is suggested to be structured in 5 *Process Groups*, defining the various phases of the project development or organisational workflow: *Initiating*, *Planning*, *Executing*, *Monitoring and Controlling*, *Closing*.

The ICT-BoK aims at informing about the level of knowledge required to enter the ICT profession and acts as the first point of reference for anyone interested in working in ICT. Even if the ICT-BoK does not refer to Data Science competences explicitly the identified ICT processes can be applied to data management processes both in industry and academia in the context of well-defined and structured projects.

3.3 Software Engineering Body of Knowledge (SWEBOK)

The Software Engineering Body of Knowledge (SWEBOK) is an international standard ISO/IEC TR 19759:2015[2] specifying a guide to the generally accepted Software Engineering Body of Knowledge. The Guide to the Software Engineering Body of Knowledge (SWEBOK Guide) has been created through cooperation among several professional bodies and members of industry and is published by the IEEE Computer Society. The standard can be accessed freely from the IEEE Computer Society (http://www.computer.org/web/swebok/v3).[3]

[2] ISO/IEC TR 19759:2015 Software Engineering - Guide to the software engineering body of knowledge (SWEBOK)
[3] SWEBOK can be also accessed from http://www4.ncsu.edu/~tjmenzie/cs510/pdf/SWEBOKv3.pdf

The published version of SWEBOK V3 has the following 15 knowledge areas (KAs) within the field of software engineering: and 7 additional disciplines are recognized as linked and providing important background knowledge that are beneficial for Software engineering:

Table 4.1. SWEBOK Knowledge Areas and related disciplines

SWEBOK Knowledge Areas	Additional linked disciplines
• Software requirements • Software design • Software construction • Software testing • Software maintenance • Software configuration management • Software engineering management • Software engineering process • Software engineering models and methods • Software quality • Software engineering professional practice • Software engineering economics • Computing foundations • Mathematical foundations • Engineering foundations	• Computer engineering • Systems engineering • Project management • Quality management • General management • Computer science • Mathematics

3.4 Business Analysis Body of Knowledge (BABOK)

BABOK Guide was first published by International Institute of Business Analysis (IIBA) as a draft document version 1.4, in October 2005, for consultation with the wider business analysis and project management community, to document and standardize generally accepted business analysis practices. Current version 3 was released in April 2015.

The Business Analysis Body of Knowledge *provides interesting example of business oriented body of knowledge that covers important for Data Science knowledge domain. BABOK is published in a Guide to the Business Analysis Body of Knowledge* (*BABOK Guide*). It is the globally recognized standard for the practice of business analysis. *BABOK Guide* reflects the collective knowledge of the business analysis community and presents the most widely accepted business analysis practices.

BABOK Guide recognizes and reflects the fact that business analysis is continually evolving and is practiced in a wide variety of forms and contexts. It defines the skills, knowledge, and competencies required to perform business analysis effectively. It does not describe the processes that people will follow to do business analysis.

BABOK Guide includes chapters on:

- Business Analysis Key Concepts: define important terms that are the foundation of the practice of business analysis.
- Knowledge Areas: represents the core content of *BABOK Guide* and contain the business analysis tasks that are used to perform business analysis.
- Underlying Competencies: describes the behaviours, characteristics, knowledge, and personal qualities that help business analysts be effective in their job.
- Techniques: describes 50 of the most common techniques used by business analysts.
- Perspectives (new to version 3): describes 5 different views of business analysis (Agile, Business Intelligence, Information Technology, Business Architecture, and Business Process Management).

BABOK Guide organizes business analysis tasks within 6 knowledge areas. The knowledge areas logically organize tasks but do not specify a sequence, process, or methodology. Each task describes the typical

knowledge, skills, deliverables, and techniques that the business analyst requires to be able to perform those tasks competently.

The following knowledge areas of *BABOK Guid*e are defined:

- Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts.
- Elicitation and Collaboration: describes the tasks used to prepare for and conduct elicitation activities and confirm the results.
- Requirements Life Cycle Management: describes the tasks used to manage and maintain requirements and design information from inception to retirement.
- Strategy Analysis: describes the tasks used to identify the business need, address that need, and align the change strategy within the enterprise.
- Requirements Analysis and Design Definition: describes the tasks used to organize requirements, specify and model requirements and designs, validate and verify information, identify solution options, and estimate the potential value that could be realized.
- Solution Evaluation: describes the tasks used to assess the performance of and value delivered by a solution and to recommend improvements on increasing value.

BABOK knowledge areas organisation by tasks allows easy linking to Business Analysis competences what can be implemented in the intended DS-BoK.

3.5 Data Management Body of Knowledge (DM-BoK) by DAMAI

The Data Management Association International (DAMAI) has been founded in 1988 in US with the aim: (i) to provide a non-profit, vendor-independent association where data professionals can go for help and assistance; (ii) to provide the best practice resources such as the DM-BoK and DM Dictionary of Terms; (iii) to create a trusted environment for DM professionals to collaborate and communicate.

The DM-BoK version2 “Guide for performing data management” is structured in 11 knowledge areas covering core areas in data management:

(1) Data Governance,
(2) Data Architecture,
(3) Data Modelling and Design,
(4) Data Storage and Operations,
(5) Data Security,
(6) Data Integration and Interoperability,
(7) Documents and Content,
(8) Reference and Master Data,
(9) Data Warehousing and Business Intelligence,
(10) Metadata, and
(11) Data Quality.

Each KA has *section topics* that logically group activities and is described by a *context diagram*. There is also an additional Data Management section containing topics that describe the knowledge requirements for data management professionals. Each context diagram includes: *Definition*, *Goals*, *Process*, *Inputs*, *Supplier roles*, *Responsible, Stakeholder, Tools*, *Deliverables*, *and Metrics* (See Appendix B).

When using DM-BoK for defining Data Management knowledge area for DS-BoK (DS-DM) it needs to be extended with the recent data modelling technologies and Big Data management platforms that address generic Big Data properties such as Volume, Veracity, Velocity. New data security and privacy protections need to be addressed as well (see CSA Top 10 Big Security challenges [40]).

3.6 Project Management Professional Body of knowledge (PM-BoK)

The PM-BoK is maintained by the Project Management Institute (PMI) the provides research and education services to Project Managers through publications, networking-opportunities in local chapters, hosting

conferences and training seminars, and providing accreditation in project management. PMI, exploit volunteers and sponsorships to expand project management's body of knowledge through research projects, symposiums and surveys, and shares it through publications, research conferences and working sessions. The "A Guide to the Project Management Body of Knowledge" (PM-BoK), has been recognized by the American National Standards Institute (ANSI) and in 2012 ISO adapted the project management processes from the PMBOK Guide 4th edition (see Appendix B).

The PMI-BoK defines five Process Groups related to project management:

- Initiating - Processes to define and authorize a project or project phase
- Planning - Processes to define the project scope, objectives and steps to achieve the required results.
- Executing - Processes to complete the work documented within the Project Management Plan.
- Monitoring and Controlling - Processes to track and review the project progress and performance. This group contains the Change Management.
- Closing - Processes to formalize the project or phase closure.

The nine Knowledge Areas are linked to the Process Groups:

- Project Integration Management - Processes to integrate various parts of the Project Management.
- Project Scope Management - Processes to ensure that all of the work required is completed for a successful Project and manages additional "scope creep".
- Project Time Management - Processes to ensure the project is completed in a timely manner.
- Project Cost Management - Processes to manage the planning, estimation, budgeting and management of costs for the duration of the project.
- Project Quality Management - Processes to plan, manage and control the quality and to provide assurance the quality standards are met.
- Project Human Resource Management - Processes to plan, acquire, develop and manage the project team.
- Project Communications Management - Processes to plan, manage, control, distribute and final disposal of project documentation and communication.
- Project Risk Management - Processes to identify, analyse and management of project risks.
- Project Procurement Management - Processes to manage the purchase or acquisition of products and service, or result to complete the project.
- Project Stakeholder Management – Process to identify stakeholders, determine their requirements, expectations and influence

Each Process Group contains processes within some or all of the Knowledge Areas. Each of the 42 processes has Inputs, Tools and Techniques, and Outputs. (It is not the scope of this analysis enter into the details of each process).

3.7 Components and concepts related to CF-DS and DS-BoK definition

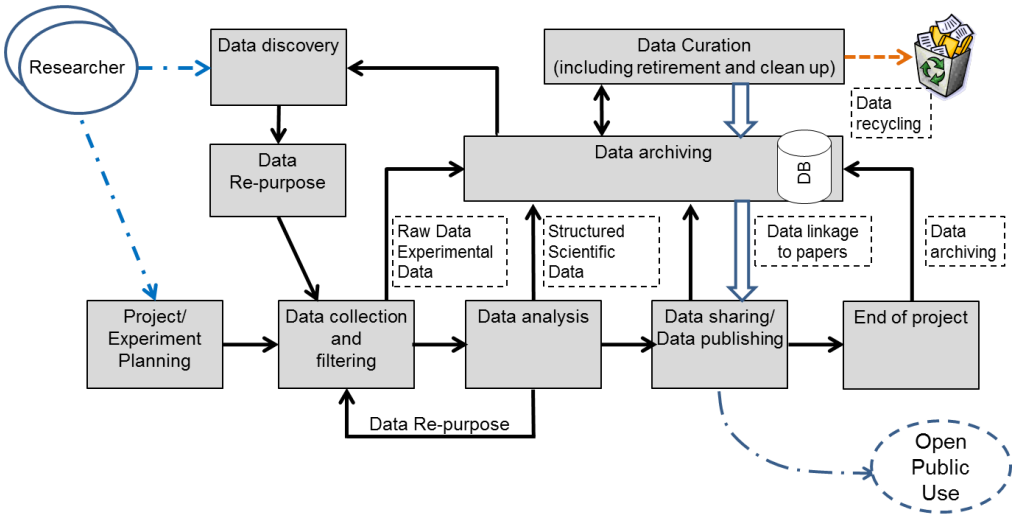
This section provides important definitions that are needed for consistent CF-DS definition in the context of organisational and business processes, e-Infrastructure and scientific research. First of all, this includes definition of typical organisational processes and scientific workflow or research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

3.7.1 Scientific Data Lifecycle Management Model

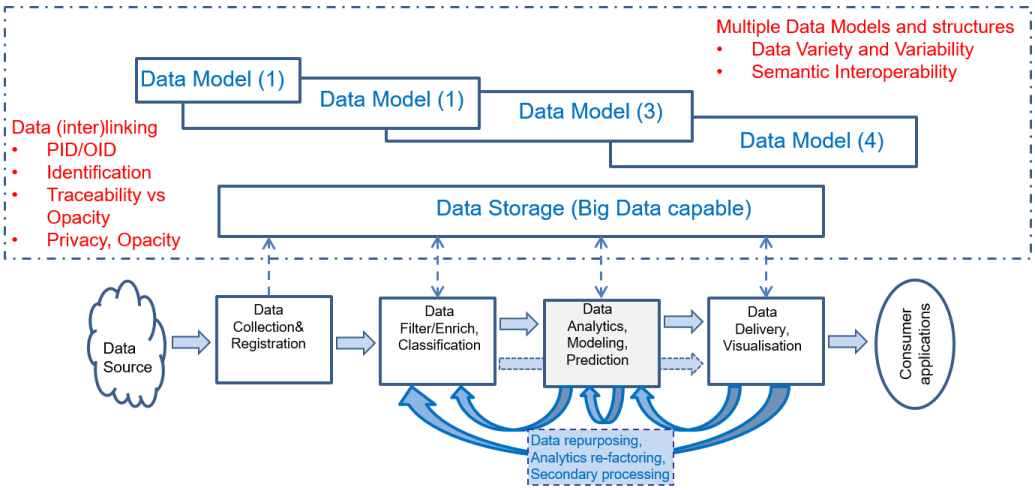
Data lifecycle is an importance component of data centric applications, which Data Science and Big Data applications belong to. Data lifecycle analysis and definition is addressed in many domain specific projects and studies. Extensive compilation of the data life cycle models and concepts is provided in the CEOS.WGISS.DSIG document [14].

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 [15] shown in Figure 2 (a) defined as a result of analysis of the existing practices in different scientific communities. Figure 2 (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 [1, 16, 17]. The proposed models are sufficiently generic and compliant with the data lifecycle study results presented in [14].

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.



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



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

Figure 2. 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[4].

3.7.2 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

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

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 [18, 19, 20]. 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 [21].

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
- 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. This is discussed in section 3.4 below.

3.7.3 Business Process Management lifecycle

New generation Agile Data Driven Enterprises (ADDE) [22] 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 [23, 24] 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 as described in a separate CF-DS document and in appendix.

4 Data Science Body of Knowledge (DS-BoK) definition

To define consistently the DS-BoK as a new emerging technology and knowledge domain, we need to understand the commonly accepted approaches to defining education and training programmes and put them in the context of the European education system and policies. Two approaches to education and training are followed in practice, the traditional approach which is based on defining the time students have to spend learning a given topics or concept like the European Credit Transfer and Accumulation System (ECTS) [33] or Carnegie unit credit hour [34]. The former is also known as competence-based education or outcomes-based learning (OBE), it is focusing on the outcome assessing whether students have mastered the given competences, namely the skills, abilities, and knowledge. There is no specified style of teaching or assessment in OBE; instead classes, opportunities, and assessments should all help students achieve the specified outcomes. In 2012, the EC has called for a rethinking of education towards OBE approach. The motivation for such a rethinking is to ensure that education is more relevant to the needs of students and the labour market, assessment methods need to be adapted and modernised. Not like the traditional BoK which is defined in term of Knowledge Areas (KA), in OBE the BoK is defined in term of the core learning outcomes which are grouped into technical competence areas and workplace skills.

The term "Body-of-Knowledge" (BoK) is commonly referred as the set (from the Latin corpus) of knowledge (intended as concepts, terms and activities) that constitutes the common ground of a discipline or science. The constitution of this common ground is essential to create a professional community around this knowledge since its creation is done by the community itself. In most cases the BoK is created and maintained by a designed profession organisation that ensures the BoK that is up to date and reflect the constituent community needs and vision. With the BoK the community usually creates a vocabulary or lexicon where all the terms are defined and clarified in their meaning within this community.

The DS-BoK is defined based on the proposed competences model CF-DS that identified five competence areas that should be mapped into corresponding knowledge areas and groups (refer to the recent CF-DS version online[5]., it is also included into Appendix C for details and visual presentation in Figure C.1 (a) and (b)). The DS-BoK definition will require combination and synthesis of different domain knowledge areas with necessary selection or adaptation of educational and instructional models and practices.

4.1 Defining the structure and content of DS-BoK

The intended DS-BoK can be used as a base for defining Data Science related curricula, courses, instructional methods, educational/course materials, and necessary practices for university post and undergraduate programs and professional training courses. The DS-BoK is also intended to be used for defining certification programs and certification exam questions. While CF-DS (comprising of competences, skills and knowledge) can be used for defining job profiles (and correspondingly content of job advertisements) the DS-BoK can provide a basis for interview questions and evaluation of the candidate's knowledge and related skills.

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

- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Scientific or Research Methods group*
- KAG5-DSBP: Business process management group
- KAG6-DSDK: Data Science Domain Knowledge group includes domain specific knowledge

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

[5] http://www.edison-project.eu/data-science-competence-framework-cf-ds

It is also anticipated that due to complexity of Data Science domain, the DS-BoK will require wide spectrum of background knowledge, first of all in mathematics, statistics, logics and reasoning as well as general computing and cloud computing in particular. Similar to the ACM CS2013 curricula approach, background knowledge can be required as an entry condition or must be studied as elective courses.

The ACM CS-BoK can be used as a basis for DS-BoK definition, in particular the following KAs from CS-BoK can be used:
AL - Algorithms and Complexity
AR - Architecture and Organization (including computer architectures and network architectures)
CN - Computational Science
GV - Graphics and Visualization
IM - Information Management
PBD - Platform-based Development (new)
SE - Software Engineering

It is important to understand that ACM CS-BoK is originally defined for native Computer Science curricula but for the DS-BoK mentioned above KAs need to be re-structured and may be combined with other Data Science oriented courses.

The DS-BoK will re-use where possible existing BoK's taking necessary KA definitions combining them into defined above DS-BoK knowledge area groups. The following BoK's can be used or mapped to the selected DS-BoK knowledge groups: Software Engineering SWEBOK, Business Analysis BABOK, and Project Management PM-BoK, and others

Building on the assumption that the Data Scientist will need to work across different stages of organizational workflow and understand the whole data lifecycle management model, it is reasonable to suggest at this stage that the DS-BoK incorporate the Process Group concept proposed in PM-BoK and business process support approach proposed in BABOK.

4.1.1 Process Groups in Data Management

The following Process Groups can be identified based on discussed in section 3.2.1 existing data lifecycle models that reflect the major functions or activities in Data Management:

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.

4.1.2 Data Analytics Knowledge Area

Data Analytics, although formally may be treated as a part of data lifecycle, form the purpose of DS-BoK needs is defined as a separate process group and knowledge area group. **Data analysis** includes the modelling and inspecting of data with the goal of discovering useful information, providing recommendations, and supporting decision-making. The following are commonly defined Data Analytics Knowledge Areas:

- Machine learning and related methods for information search, image recognition, decision support, classification;
- Predictive analytics focuses on application of statistical models for predictive forecasting or classification.
- *Data mining* is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes;
- Business intelligence covers data analysis that relies heavily on aggregation and different data sources and focusing on business information;
- Text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data;
- Statistical methods, including
    - Descriptive statistics, exploratory data analysis (EDA) focused on discovering new features in the data, and confirmatory data analysis (CDA) dealing with validating formulated hypotheses;

4.2 DS-BoK structure and Knowledge Area Groups

Presented analysis allows us to propose an initial version of the Data Science Body of Knowledge implementing the proposed DS-BoK structure as explained in previous section. Table 4.1 provides consolidated view of the identified Knowledge Areas in the Data Science Body of Knowledge. The table contains detailed definition of the KAG1-DSA, KAG2-DSE, KAG3-DSDM groups that are well supported by existing BoK's and academic materials. General suggestions are provided for KAG4-DSRM, KAG5-DSBP groups that corresponds to newly identified competences and knowledge areas and require additional study of existing practices and contribution from experts in corresponding scientific or business domains.

The KAG2-DSE group includes selected KAs from ACM CS-BoK and SWEBOK and extends them with new technologies and engineering technologies and paradigm such as cloud based, agile technologies and DevOps that are promoted as continuous deployment and improvement paradigm and allow organisation implement agile business and operational models.

The KAG3-DSDM group includes most of KAs from DM-BoK however extended it with KAs related to RDA recommendations, community data management models (Open Access, Open Data, etc.) and general Data Lifecycle Management that is used as a central concept in many data management related education and training courses.

The presented DS-BoK high level content is not exhaustive at this stage and will undergo further development based on feedback from the project Task 3.1 that will use the presented DS-BoK for developing Data Science Model Curriculum (MC-DS). The project will present the current version of DS-BoK to ELG to obtain feedback and expert opinion. Numerous experts will be invited to review and contribute to the specific KAs definition.

Table 4.1. Identified DS-BoK Knowledge Area Groups

KA Groups	Knowledge Areas (KA) from existing BoKs	Additional Knowledge Areas
KAG1-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics	BABOK selected KAs *) • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.	General Data Analytics and Machine Learning KAs • Machine learning and related methods • Predictive analytics and predictive forecasting • Classification methods • Data mining and knowledge discovery • Business intelligence covers data analysis that relies heavily on aggregation and different data sources and focusing on business information; • Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data • Statistical methods, including descriptive statistics, exploratory data analysis (EDA) and confirmatory data analysis (CDA)
KAG2-DSENG: Data Science Engineering group including Software and infrastructure engineering	ACM CS-BoK selected KAs: AL - Algorithms and Complexity AR - Architecture and Organization (including computer architectures and network architectures) CN - Computational Science GV - Graphics and Visualization IM - Information Management PBD - Platform-based Development (new) SE - Software Engineering (extended with SWEBOK KAs) SWEBOK selected KAs • Software requirements • Software design • Software construction • Software engineering process • Software engineering models and methods • Software quality	Infrastructure and platforms for Data Science applications group: CCENG - Cloud Computing Engineering (infrastructure and services design, management and operation) CCAS - Cloud based applications and services development and deployment BDA – Big Data Analytics platforms (including cloud based) BDI - Big Data Infrastructure services and platforms, including data storage infrastructure Data and applications security KAs: SEC - Applications and data security SSM – Security services management, including compliance and certification Agile development technologies • Methods, platforms and tools • DevOps and continuous deployment and improvement paradigm
KAG3-DSDM: Data Management group including data curation, preservation and data infrastructure	DM-BoK selected KAs (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content,	General Data Management KA's • Data Lifecycle Management • Data archives/storage compliance and certification New KAs to support RDA recommendations and community data

KA Groups	Knowledge Areas (KA) from existing BoKs	Additional Knowledge Areas
	(8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality.	management models (Open Access, Open Data, etc)[6] • Data type registries, PIDs • Data infrastructure and Data Factories • TBD – To follow RDA and ERA community developments
KAG4-DSRM: Scientific or Research Methods group	There are no formally defined BoK for research methods	Suggested KAs to develop DSRM related competences: • Research methodology, research cycle (e.g. 4 step model Hypothesis – Research Methods – Artefact – Validation) • Modelling and experiment planning • Data selection and quality evaluation • Use cases analysis: research infrastructures and projects • TBD further extensions
KAG5-DSBPM: Business process management group	PMI-BoK selected KAs • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management BABOK selected KAs *) • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.	General Business processes and operations KAs • Business processes and operations • Agile Data Driven methodologies, processes and enterprises • Use cases analysis: business and industry • TBD further extensions

*) BABOK KA are more business focused and related to KAG5-DSBP, however its specific topics related to data analysis can be reflected in the KAG1-DSDA

[6] Example courses provided by RDA community and shared between European Research Infrastructures https://europe.rd-alliance.org/training-programme

4.3 Data Science Body of Knowledge Areas and Knowledge Units

Presented analysis allows us to propose an initial version of the Data Science Body of Knowledge implementing the proposed DS-BoK structure as explained in previous section. Table 4.2 provides consolidated view of the identified Knowledge Areas in the Data Science Body of Knowledge. The table contains detailed definition of the KAG1-DSA, KAG2-DSE, KAG3-DSDM groups that are well supported by existing BoK's and academic materials. General suggestions are provided for KAG4-DSRM, KAG5-DSBP groups that corresponds to newly identified competences and knowledge areas and require additional study of existing practices and contribution from experts in corresponding scientific or business domains.

The KAG2-DSE group includes selected KAs from ACM CS-BoK and SWEBOK and extends them with new technologies and engineering technologies and paradigm such as cloud based, agile technologies and DevOps that are promoted as continuous deployment and improvement paradigm and allow organisation implement agile business and operational models.

The KAG3-DSDM group includes most of KAs from DM-BoK however extended it with KAs related to RDA recommendations, community data management models (Open Access, Open Data, etc) and general Data Lifecycle Management that is used as a central concept in many data management related education and training courses.

The presented DS-BoK high level content is not exhaustive at this stage and will undergo further development based on feedback from MC-DS implementation. The project will present the current version of DS-BoK to ELG to obtain feedback and expert opinion. Numerous experts will be invited to review and contribute to the specific KAs definition.

Table 4.2 Detailed definition of the DS-BoK and suggested Knowledge Units (KU)

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
KAG1-DSDA: Data Analytics group (including Machine Learning, statistical methods)	Theory of computation	Design and Analysis of Algorithms	CCS2012: Theory of computation Design and analysis of algorithms Data structures design and analysis Theory and algorithms for application domains Machine learning theory Algorithmic game theory and mechanism design Database theory Semantics and reasoning
		Machine Learning Theory	
		Game Theory & Mechanism design	
		EXTENSIBILITY Point: Theory of computation	
	Mathematics of computing	Discrete Mathematics and Graph Theory	**CCS2012: Mathematics of computing** Discrete mathematics Graph theory Probability and statistics Probabilistic representations Probabilistic inference problems Probabilistic reasoning algorithms Probabilistic algorithms Statistical paradigms Mathematical software Information theory Mathematical analysis
		Probability & Statistics	
		Probabilistic reasoning	
		Statistical methods, including descriptive statistics, exploratory data analysis (EDA) and confirmatory data analysis (CDA)	
		Information theory	
		Mathematical analysis	
		Mathematical software and tools	
		EXTENSIBILITY Point: Mathematics of Data Science (computing)	TBD
	Computing methodologies	Artificial Intelligence	**CCS2012: Computing methodologies** Artificial intelligence Natural language processing Knowledge representation and reasoning Search methodologies Machine learning Learning paradigms Supervised learning Unsupervised learning Reinforcement learning Multi-task learning Machine learning approaches Machine learning algorithms
		Natural Language Processing	
		Knowledge Representation and Reasoning	
		Data mining and knowledge discovery	
		Text analysis, Data mining	
		Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data	
		Machine Learning theory and algorithms	
		Classification methods	
		EXTENSIBILITY Point: Computing methodologies	TBD
	Information systems (to support Data Science applications)	Decision Analysis and Decision Support Systems	**CCS2012: Information systems** Information systems applications Decision support systems Data warehouses Expert systems Data analytics
		Data warehousing and Data Mining	
		Data Analysis and statistics	
		Multimedia information systems	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
		Data Mining	Online analytical processing Multimedia information systems Data mining
		Predictive analytics and predictive forecasting	
		EXTENSIBILITY Point: Information systems	TBD
	Big Data Technologies and Systems	Big Data algorithm for large scale data processing	**DSDA Extension group for CCS2012** Theory of computation DSA Extension point: Algorithms for Big Data computation Mathematics of computing DSA Extension point: Mathematical software for Big Data computation Computing methodologies DSA Extension point: New DSA computing Information systems DSA Extension point: Big Data systems (e.g. cloud based) Information systems applications DSA Extension point: Big Data applications DSA Extension point: Doman specific Data applications
		Big Data Analytics	
		Big Data systems	
		Big Data algorithms for data ingest, pre-processing, and visualisation	
		Big Data analytics platforms and tools (including Hadoop, Spark, and cloud based Big Data services)	
		Big Data systems for application domains	
		EXTENSIBILITY Point: Information systems	TBD
KAG2-DSENG: Data Science Engineering group including Software and infrastructure engineering	Computer systems organisation for Big Data applications (including high performance networks)	Parallel and Distributed Computer Architecture	**CCS2012: Computer systems organization** Architectures Parallel architectures Distributed architectures Networks *) Network Architectures Network Services Cloud Computing
		Computer networks: architectures and protocols	
		Computer networks for high-performance computing and Big Data infrastructure	
		EXTENSIBILITY Point:	TBD
	Big Data software organisation and engineering	Software (systems) architectures	**CCS2012: Software and its engineering** Software organization and properties Software system structures Software architectures Software system models Ultra-large-scale systems Distributed systems organizing principles Cloud computing Grid computing Abstraction, modeling and modularity Real-time systems software Software notations and tools General programming languages
		Requirements engineering and software systems development	
		Large and ultra-large scale software systems organisation	
		Cloud enabled applications development	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
			Software creation and management
		EXTENSIBILITY Point:	TBD
	Modelling and simulation	Modelling and simulation theory and techniques (general and domain oriented)	**CCS2012: Computing methodologies** Modeling and simulation Model development and analysis Simulation theory Simulation types and techniques Simulation support systems
		Large scale modelling and simulation systems	
		EXTENSIBILITY Point:	TBD
	Big Data systems organisation and management	Enterprise information systems	**CCS2012: Information systems** Information storage systems Information systems applications Enterprise information systems Collaborative and social computing systems and tools
		Large scale data storage and data management systems	
		Collaborative and social computing systems and tools	
		EXTENSIBILITY Point:	TBD
	Big Data (Data Science) applications design	Programming languages for Big Data analytics: R, python, others	**Proposed new KA for DS-BoK** Linked to KAG1-DSA DSDA: Big Data applications design DSDA: Data Analytics programming languages
		Models and languages for complex interlinked data presentation and visualisation	
		EXTENSIBILITY Point:	TBD
	Infrastructure and platforms for Data Science applications group:	Cloud Computing architecture and services	**Proposed new KA for DS-BoK** Infrastructure and platforms for Data Science applications group: CCENG - Cloud Computing Engineering (infrastructure and services design, management and operation) CCAS - Cloud based applications and services development and deployment BDA – Big Data Analytics platforms (including cloud based) BDI - Big Data Infrastructure services and platforms, including data storage infrastructure Data and applications security KAs: SEC - Applications and data security SSM – Security services management, including compliance and certification
		Cloud Computing Engineering (infrastructure and services design, management and operation)	
		Big Data and cloud based systems design and development	
		Cloud based applications and services operation and management	
		Big Data Analytics platforms (including cloud based)	
		Big Data Infrastructure: services and components, including data storage infrastructure	
		Data security and protection	
		EXTENSIBILITY Point:	TBD
	Software engineering and management	Software requirements and design	**SWEBOK selected KAs** • Software requirements • Software design • Software construction • Software testing • Software maintenance
		Software engineering models and methods	
		Software quality assurance	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
		Agile development methods, platforms and tools DevOps and continuous deployment and improvement paradigm	• Software configuration management • Software engineering management • Software engineering process • Software engineering models and methods • Software quality Agile development technologies • Methods, platforms and tools DevOps and continuous deployment and improvement paradigm
		EXTENSIBILITY Point:	TBD
KAG3-DSDM: Data Management group (including data curation, preservation and data infrastructure)	Data management systems	Database management systems Database design and models Data Modelling, Databases and Database Management Systems Data Models and Query Languages Database administration	**CCS2012: Information systems** Data management systems Database design and models Data structures Database management system engines Query languages Database administration Middleware for databases Information integration
		EXTENSIBILITY Point:	TBD
	Digital libraries and archives	Digital libraries and archives organisation Information Retrieval Data curation and provenance Search Engines technologies	**CCS2012: Information systems** Information systems applications Digital libraries and archives Information retrieval Document representation Retrieval models and ranking Search engine architectures and scalability Specialized information retrieval
		EXTENSIBILITY Point:	TBD
	Data Management and Enterprise data infrastructure	Data management, including Reference and Master Data Data Warehousing and Business Intelligence Data storage and operations Data archives/storage compliance and certification Metadata, linked data, provenance Data infrastructure, data registries and data factories Data security and protection Data governance, data quality, data Integration and Interoperability Data Management Planning Responsible data use, data privacy, ethical principles, legal issues	**DM-BoK selected KAs** (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality.
		EXTENSIBILITY Point:	TBD
	General principles and concepts in	Data type registries, PID, metadata	**Proposed new KA for DS-BoK** General Data Management KA's

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
	Data Management and organisation	Research data infrastructure, Open Science, Open Data, Open Access, ORCID	• Data Lifecycle Management • Data archives/storage compliance and certification New KAs to support RDA recommendations and community data management models (Open Access, Open Data, etc) • Data type registries, PIDs • Data infrastructure and Data Factories • New KAs to follow RDA and ERA community developments
		Data infrastructure compliance and certification	
		Ethical principle and data privacy	
		EXTENSIBILITY Point:	TBD
KAG4-DSRM: Scientific and Research Methods group	Scientific/Research Methods	Research methodology, paradigms and research cycle	**Proposed new KA for DS-BoK** To develop DSRM related competences: • Research methodology, research cycle (e.g. 4 step model Hypothesis – Research Methods – Artefact – Validation) • Modelling and experiment planning • Data selection and quality evaluation • Use cases analysis: research infrastructures and projects • TBD further extensions
		Modelling and experiment planning	
		Data selection and quality evaluation	
		Use cases analysis: research infrastructures and projects	
		Research data management plan and ethical issues	
		EXTENSIBILITY Point:	TBD
KAG5-DSBPM: Business process management group	Business Process Management	Business processes and operations	**PMI-BoK selected KAs** • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management
		Project scope and risk management	
		EXTENSIBILITY Point:	TBD
	Business Analysis organisation and management	Business Analysis Planning and Monitoring	**BABOK selected KAs** • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.
		Requirements Analysis and Design Definition	
		Requirements Life Cycle Management (from inception to retirement)	
		Solution Evaluation and improvements recommendation	
		EXTENSIBILITY Point:	TBD
	Business analysis and enterprise organisation	Agile Data Driven methodologies, processes and enterprises	**Proposed new KA for DS-BoK** General Business processes and operations KAs • Business processes and operations • Agile Data Driven methodologies, processes and enterprises • Use cases analysis: business and industry TBD further extensions
		Use cases analysis: business and industry	
		EXTENSIBILITY Point:	TBD

5 Conclusion and further developments

The presented work on defining the DS-BoK and other foundational components of the whole EDISON framework 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.

5.1 Summary of findings

The provided document contains initials results of the Data Science Body of Knowledge definition that are based on the two other components Data Science Competence framework and Data Science knowledge area classification. In particular the DS-BoK uses the CF-DS competence and skills groups that include:

- Data Analytics (also referred to as Business Analytics or Machine Learning)
- Engineering or Programming
- Subject/Scientific Domain Knowledge
- Data Management, Curation, Preservation
- Scientific or Research Methods (for research profiles) and Business Process management (for business related profiles)

Consequently the CF-DS competence groups are presented in the DS-BoK Knowledge Area groups (KAG):

- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Scientific or Research Methods group*
- KAG5-DSBP: Business process management group
- KAG6-DSDK: Data Science Domain Knowledge group includes domain specific knowledge

5.2 Further developments to formalize CF-DS and DS-BoK

It is anticipated that the presented here the first versions of the Data Science Body of Knowledge will require further development and validation by experts and communities of practice that will include the following specific tasks and activities:

- Define specific knowledge areas related to the identified knowledge area groups by involving experts in the related knowledge areas, possibly also engaging with the specific professional communities such as IEEE, ACM, DAMA, IIBA, etc.
- 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.
- Provide input to the Data Science Model Curriculum development and obtain necessary feedback for DS-BoK improvement and extension.
- Engage with the partner and champion universities into pilot implementation of DS-BoK and collecting feedback from practitioners.

Validation is an important part of the products that could be widely accepted by community. Validation of the proposed DS-BoK will be done in two main ways (similar to CF-DS). First is presenting the proposed development to the communities of practice and soliciting feedback and contribution from the academic and professional community, including experts' interviews. The second way suggests involving the champion universities into validation and pilot implementation of the proposed DS-BoK and Model Curriculum.

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

6 References

[1] 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

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

[3] 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

[4] 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

[5] 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

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

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

[8] ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] http://dx.doi.org/10.1145/2534860

[9] ACM Curricula recommendations [online] http://www.acm.org/education/curricula-recommendations

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

[11] Computer Science 2013: Curriculum Guidelines for Undergraduate Programs in Computer Science http://www.acm.org/education/CS2013-final-report.pdf

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

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

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

[15] 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

[16] 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.

[17] 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

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

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

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

[21] 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/

[22] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.

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

[24] 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)

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

[26] 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

[27] 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

[28] Research Data Alliance Interest Group on Education and Training on Handline of Research Data (IG-ETRD) wiki [online] https://rd-alliance.org/node/971/all-wiki-index-by-group

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

[30] 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

[31] 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/

[32] 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

[33] European Credit Transfer and Accumulation System (ECTS) [online] http://ec.europa.eu/education/ects/users-guide/docs/year-2009/ects-users-guide-2009_en.pdf

[34] Carnegie unit credit hour [online] https://www.luminafoundation.org/files/resources/carnegie-unit-report.pdf

[35] ICT professional Body of Knowledge (ICT-BoK) [online] http://www.ictbok.eu/images/EU_Foundationa_ICTBOK_final.pdf

[36] Business Analytics Body of Knowledge (BABOK) [online] http://www.iiba.org/babok-guide.aspx

[37] Software Engineering Body of Knowledge (SWEBOK) [online] https://www.computer.org/web/swebok/v3

[38] 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

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

[40] Expanded Top Ten Big Data Security and Privacy Challenges, April 2013, Cloud Security Alliance [online] https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.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. Overview of Bodies of Knowledge relevant to Data Science

This section provides detailed information about existing Bodies of Knowledge relevant to the Data Science Body of Knowledge definition which can be linked to or mapped to the future DS-BoK.

A.1. ICT Professional Body of knowledge

Character	Explanation
Name of the Profession	ICT professional
Reference Community	(potentially) all ICT Professional
Leadership	Capgemini Consulting and Ernst & Young for the European Commission, Directorate General Internal Market, Industry, Entrepreneurship and SMEs
Organisation structure	N/A
Partners	N/A
Ethical Code	N/A
Estimated #members	N/A
Link to BoK	http://www.ictbok.eu/images/EU_Foundationa_ICTBOK_final.pdf
Year/Edition	2015/1st
Structure of BoK	There are 12 Knowledge Areas: ***1.*** ICT Strategy & Governance ***2.*** Business and Market of ICT ***3.*** Project Management ***4.*** Security Management ***5.*** Quality Management ***6.*** Architecture ***7.*** Data and Information Management ***8.*** Network and Systems Integration ***9.*** Software Design and Development ***10.*** Human Computer Interaction ***11.*** Testing ***12.*** Operations and Service Management Each Knowledge Area is defined by; • List of items required as foundational knowledge necessary under this Knowledge Area; • List of references to the e-Competence Framework (dimension 4: knowledge); • List of possible job profiles that require having an understanding of the Knowledge Area; • List of examples of specific Bodies of Knowledge, certification and training possibilities
Proposed use of BoK	• Education providers: as a source of inspiration for curricula design and development; • Professional Associations: to promote the Body of Knowledge to their members, ICT professionals; • HR Department and Managers within industry with a need to understand the range of knowledge and the entry level required by ICT professionals in order to improve recruiting and people development processes (together with skills and competencies).
Certification promoted	N/A

A.2. Data Management Professional Body of knowledge

Character	Explanation
Name of the Profession	Data Management Professional
Reference Community	Mainly US Data managers, professionals and scholars. Relevant chapters in UK and Australia.
Leadership	DAMAI a Volunteer US-based organization governed by an Executive Board of Directors. Directors are voted in for a 2 year term of office and may stand for re-election
Organisation structure	The members adhere through the nearest local chapter and through that (autonomous organisations affiliated with the central associations) participate to the life of the community
Partners	US-based organisation of medium relevance that provide educational resources (Dataversity, DEBtech, IRM UK, Technics Publications) or instruments and tools (VoltDB)
Ethical Code	Yes (available for members https://www.dama.org/content/chapter-kit-behind-login)
Estimated #members	Conferences are attended by a thousand people, 16 Chapters worldwide. No references about number of subscriptions
Link to BoK	BoK Framework • http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf DAMA International Guide to Data Management Body of Knowledge (on purchase) • https://technicspub.com/dmbok/ - [52€ on Amazon] Other resources DAMA International Dictionary of Data Management Terms (on purchase) • https://technicspub.com/dmbok/ - [54€ on Amazon]
Edition/version	2012/v.2
Structure of BoK	The document is structured in 11 knowledge areas covering core areas in the DAMA - DMBOK2 Guide for performing data management. The 11 Data Management Knowledge Areas are: 1. **Data Governance** – planning, oversight, and control over management of data and the use of data and data-related resources. Governance covers 'processes', not 'things', hence the common term for Data Management Governance is Data Governance. 2. **Data Architecture** – the overall structure of data and data-related resources as an integral part of the enterprise architecture 3. **Data Modelling &Design** – analysis, design, building, testing, and maintenance (was Data Development in the DAMA - DMBOK 1st edition) 4. **Data Storage & Operations** – structured physical data assets storage deployment and management (was Data Operations in the DAMA-DMBOK 1st edition) 5. **Data Security** – ensuring privacy, confidentiality and appropriate access 6. **Data Integration & Interoperability** – acquisition, extraction, transformation, movement, delivery, replication, federation, virtualization and operational support (a Knowledge Area new in DMBOK2) 7. **Documents & Content** – storing, protecting, indexing, and enabling access to data found in unstructured sources (electronic files and physical records), and making this data available for integration and interoperability with structured (database) data.

	8. **Reference & Master Data** – Managing shared data to reduce redundancy and ensure better data quality through standardized definition and use of data values. 9. **Data Warehousing & Business Intelligence** – managing analytical data processing and enabling access to decision support data for reporting and analysis 10. **Metadata** – collecting, categorizing, maintaining, integrating, controlling , managing, and delivering metadata 11. **Data Quality** – defining, monitoring, maintaining data integrity, and improving data quality Each KA has section topics that logically group activities and it is described by a context diagram. There Is also an additional Data Management section containing topics that describe the knowledge requirements for data management professionals. Each context diagram includes: • *Definition*: a concise description of the Knowledge Area. • *Goals*: he desired outcomes of the Knowledge Area within this Topic. • *Process*: the list of discrete activities and sub-activities to be performed, with activity group indicators. • *Inputs*: what documents or raw materials are directly necessary for a Process to initiate or continue • *Supplier roles*: roles and/or teams that supply the inputs to the process. • *Responsible roles:* roles and/or teams that perform the process. • *Stakeholder roles*: roles and/or teams Informed or consulted on the process execution. • *Tools*: technology types used by the process to perform the function. • *Deliverables* : what is directly produced by the processes • *Consumer roles*: roles and/or teams that expect and receive the Deliverables. • *Metrics* : Measurements That quantify the success of Processes based on the Goals
Proposed use of BoK	• Informing a diverse audience about the nature and importance of data management. • Helping build consensus within the data management community. • Helping data stewards, data owners, and data professionals understand their responsibilities. • Providing the basis for assessments of data management effectiveness and maturity. • Guiding efforts to implement and improve data management knowledge areas. • Educating students, new hires, practitioners and executives on data management knowledge areas • Guiding the development and delivery of data management curriculum content for higher education. • Suggesting areas of further research in the field of data management. • Helping data management professionals prepare for Certified Data Management Professional (CDMP) data exams. • Assisting organizations in defining their enterprise data strategy.
Certification promoted	Certified Data Management Professional (CDMP) in four levels: • Associate (https://www.dama.org/content/cdmp-associate), • Practitioner (https://www.dama.org/content/cdmp-practitioner), • Master (https://www.dama.org/content/cdmp-master), • Fellow (https://www.dama.org/content/cdmp-fellow)

	Cost per exam: vary depending on the examination (from $220 of Associate till the 1560 for Master). Fellow is an assigned through nomination by peers and Chapter. Requirements: member of local chapter, sign/adhere to Ethical code/ proven experiences verifiable on the CV and contributions to the Association at various level

A.3. Project Management Professional Body of knowledge

Character	Explanation
Name of the Profession	Project Management Professional
Reference Community	Industry-centred worldwide Project Managers
Leadership	Project Management Institute – www.pmi.org PMI is a worldwide not-for-profit professional membership association for the project, program and portfolio management profession. Founded in 1969, PMI delivers advocacy, collaboration, education and research to its members.
Organisation structure	PMI is governed by a 15-member volunteer Board of Directors. Each year PMI members elect five directors to three-year terms. Three directors elected by others on the Board serve one-year terms as officers. Day-to-day PMI operations are guided by the Executive Management Group and professional staff at the Global Operations Centre located near Philadelphia. Each member adhere through the nearest local chapter and through that (autonomous organisations affiliated with the central associations) participate to the life of the community
Partners	No specific partnership but some 1600 Registered Education Providers (R.E.P.s) and about 100 certified courses worldwide (http://www.pmi.org/learning/professional-development/global-accreditation-center.aspx)
Ethical Code	Yes (http://www.pmi.org/About-Us/Ethics/Code-of-Ethics.aspx#)
Estimated #members	700.000 in 195 countries (source www.pmi.org) [Estimated some 2,9 acting PM worldwide and some 1,5 million PM posts till 2020]
Link to BoK	http://www.pmi.org/PMBOK-Guide-and-Standards/pmbok-guide.aspx (on purchase - $46,17) other resources • Lexicon of PM terms (http://www.pmi.org/PMBOK-Guide-and-Standards/PMI-lexicon.aspx - free for members) • PMBoK in other 11 languages (Arabian, Italian, Korean, Russian, Hindi, Japanese, Portuguese, Spanish, German, French, Chinese); • **Software Extension to the PMBOK** Guide Fifth Edition (This standard, developed by PMI jointly with IEEE Computer Society, provides guidance on the management of software development projects, and bridges the gap between the traditional, predictive approach described in the PMBOK® Guide and iterative approaches such as agile more commonly used in software development) (on purchase – $37,07) External sites: http://www.projectmanagement.com/Practices/PMI-Standards/
Year/Edition	2014/5th edition
Structure of BoK	The Five Process Groups *Initiating* - Processes to define and authorize a project or project phase

	Planning - Processes to define the project scope, objectives and steps to achieve the required results. *Executing* - Processes to complete the work documented within the Project Management Plan. *Monitoring and Controlling* - Processes to track and review the project progress and performance. This group contains the Change Management. *Closing* - Processes to formalize the project or phase closure. The Nine Knowledge Areas *Project Integration Management* - Processes to integrate various parts of the Project Management. *Project Scope Management* - Processes to ensure that all of the work required is completed for a successful Project and manages additional "scope creep". *Project Time Management* - Processes to ensure the project is completed in a timely manner. *Project Cost Management* - Processes to manage the planning, estimation, budgeting and management of costs for the duration of the project. *Project Quality Management* - Processes to plan, manage and control the quality and to provide assurance the quality standards are met. *Project Human Resource Management* - Processes to plan, acquire, develop and manage the project team. Project Communications Management - Processes to plan, manage, control, distribute and final disposal of project documentation and communication. *Project Risk Management* - Processes to identify, analyse and management of project risks. *Project Procurement Management* - Processes to manage the purchase or acquisition of products and service, or result to complete the project. Each Process Group contains processes within some or all of the Knowledge Areas. Each of the 42 processes has Inputs, Tools & Techniques and Outputs. (It is not the scope of this analysis enter into the details of each process).
Proposed use of BoK	It provides project managers with the fundamental practices needed to achieve organizational results and excellence in the practice of project management. It's a competence framework to support the PM practices. It's used also as "one of the books" to pass the examination.
Certification promoted	Several certification other than the basic about Project Management Professional in correspondence of specific roles that the PM may adopt in the carrier or depending on the type of project (http://www.pmi.org/certification.aspx): CAPM – Certified Associate Project Management PMP – Project Management Professional PgMP – Program Management Professional PfMP – Portfolio Management Professional PMI–PBA – PMI-Professional Business Analyst PMI-ACP – PMI Agile Certified Professional PMI-RMP – PMI Risk Management Professional PMI-SP – Scheduling Professional *Cost*: it may vary from the $225 of CAPM till the $900 for PgMP and PfMP of non-Members;

	Requirements: general Education (Secondary school or Degree) + Experience on the field of certification + specific Education on the field of certification.

Appendix B. Subset of ACM/IEEE CCS2012 for Data Science

The presented taxonomy although based on ACM CCS (2012) classification can provide a basis and motivation for its extension with a new classification group related to Data Science and individual disciplines that are currently missing in the current ACM classification. This work will be a subject for future development and the results will be presented in other project deliverables.

B.1. ACM Classification Computer Science (2012) structure and Data Science related Knowledge Areas

The 2012 ACM Computing Classification System (CCS) [7] has been developed as a poly-hierarchical ontology that can be utilized in semantic web applications. It replaces the traditional 1998 version of the ACM Computing Classification System (CCS), which has served as the de facto standard classification system for the computing field for many years (also been more human readable). The ACM CCS (2012) is being integrated into the search capabilities and visual topic displays of the ACM Digital Library. It relies on a semantic vocabulary as the single source of categories and concepts that reflect the state of the art of the computing discipline and is receptive to structural change as it evolves in the future. ACM provides a tool within the visual display format to facilitate the application of 2012 CCS categories to forthcoming papers and a process to ensure that the CCS stays current and relevant.

However, at the moment none of Data Science, Big Data or Data Intensive Science technologies are reflected in the ACM classification. The following is an extraction of possible classification facets from ACM CCS (2012) related to Data Science what reflects multi-subject areas nature of Data Science:

As an example, the Cloud Computing that is also a new technology and closely related to Big Data technologies, currently is classified in ACM CCS (2012) into 3 groups:

> **Networks** :: Network services :: Cloud Computing
> **Computer systems organization** :: Architectures :: Distributed architectures :: Cloud Computing
> **Software and its engineering** :: Software organization and properties :: Software Systems Structures :: Distributed systems organizing principles :: Cloud Computing

Taxonomy is required to consistently present information about scientific disciplines and knowledge areas related to Data Science. Taxonomy is important component to link such components as Data Science competences and knowledge areas, Body of Knowledge, and corresponding academic disciplines. From practical point of view, taxonomy includes vocabulary of names (or keywords) and hierarchy of their relations.

The presented here initial taxonomy of Data Science disciplines and knowledge areas is based on the 2012 ACM Computing Classification System (ACM CCS (2012)). Refer to initial analysis of ACM CCS (2012) classification and subset of data related disciplines in section B.1 and Table B.1. The presented in Table B.2 taxonomy includes ACM CCS (2012) subsets/subtrees that contain scientific disciplines that are related to Data Science Knowledge Area groups as defined in chapter 4 Data Science Body of Knowledge definition:

- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: Data Management group including data curation, preservation and data infrastructure

Two other groups KAG4-DSRM: Scientific or Research Methods group and KAG5-DSBP: Business process management group cannot be mapped to ACM CCS (2012) and their taxonomy is not provided in this version. It is important to notice that ACM CCS (2012) provides a top level classification entry “Applied computing” that can be used as an extension point domain related knowledge area group KAG6-DSDK (see section 4.3 Knowledge Area groups definition).

The following approach was used when constructing the proposed taxonomy:

- ACM CCS (2012) provides almost full coverage of Data Science related knowledge areas or disciplines related to KAG1, KAG2, and KAG3. The following top level classification groups are used:
    - Theory of computation

  - Mathematics of computing
  - Computing methodologies
  - Information systems
  - Computer systems organization
  - Software and its engineering
- Each of KAGs includes subsets from few ACM CCS (2012) classification groups to cover theoretical, technology, engineering and technical management aspects.
- Extension points are suggested for possible future extensions of related KAGs together with their hierarchies.
- KAG3-DSDM: Data Management group is currently extended with new concepts and technologies developed by Research Data community and documented in community best practices.

Table 1 Data Science classification based on ACM Classification (2012)

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science
Data Science Analytics (DSDA)	Theory of computation Design and analysis of algorithms Data structures design and analysis Theory and algorithms for application domains Machine learning theory Algorithmic game theory and mechanism design Database theory Semantics and reasoning
Data Science Analytics (DSDA)	Mathematics of computing Discrete mathematics Graph theory Probability and statistics Probabilistic representations Probabilistic inference problems Probabilistic reasoning algorithms Probabilistic algorithms Statistical paradigms Mathematical software Information theory Mathematical analysis
Data Science Analytics (DSDA)	Computing methodologies Artificial intelligence Natural language processing Knowledge representation and reasoning Search methodologies Machine learning Learning paradigms Supervised learning Unsupervised learning Reinforcement learning Multi-task learning Machine learning approaches Machine learning algorithms
Data Science Analytics (DSDA)	Information systems Information systems applications Decision support systems Data warehouses Expert systems Data analytics Online analytical processing Multimedia information systems Data mining
Data Science Analytics (DSDA) EXTENSION POINT	Theory of computation DSA Extension point: Algorithms for Big Data computation Mathematics of computing DSA Extension point: Mathematical software for Big Data computation Computing methodologies DSA Extension point: New DSA computing Information systems

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science
	DSA Extension point: Big Data systems (e.g. cloud based) Information systems applications DSA Extension point: Big Data applications DSA Extension point: Doman specific Data applications
Data Science Data Management (DSDM)	Information systems Data management systems Database design and models Data structures Database management system engines Query languages Database administration Middleware for databases Information integration
Data Science Data Management (DSDM)	Information systems Information systems applications Digital libraries and archives Information retrieval Document representation Retrieval models and ranking Search engine architectures and scalability Specialized information retrieval
Data Science Data Management (DSDM) EXTENSION POINT	Information systems Data management systems Data types and structures description Metadata standards Persistent identifiers (PID) Data types registries
Data Science Engineering (DSE)	Computer systems organization Architectures Parallel architectures Distributed architectures
Data Science Engineering (DSENG)	Networks **) Network Architectures Network Services Cloud Computing
Data Science Engineering (DSENG)	Software and its engineering Software organization and properties Software system structures Software architectures Software system models Ultra-large-scale systems Distributed systems organizing principles Cloud computing Grid computing Abstraction, modeling and modularity Real-time systems software Software notations and tools General programming languages Software creation and management
Data Science Engineering (DSENG)	Computing methodologies Modeling and simulation Model development and analysis Simulation theory Simulation types and techniques Simulation support systems
Data Science Engineering (DSENG)	Information systems Information storage systems Information systems applications Enterprise information systems Collaborative and social computing systems and tools
Data Science Engineering (DSENG) EXTENSION POINT	Software and its engineering Software organization and properties DSE Extension point: Big Data applications design Data Analytics programming languages Information systems DSE Extension point: Big Data and cloud based systems design Information systems applications DSA Extension point: Big Data applications

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science
	DSA Extension point: Doman specific Data applications
DS Domain Knowledge (DSDK) EXTENSION POINT	Applied computing Physical sciences and engineering Life and medical sciences Law, social and behavioral sciences Computer forensics Arts and humanities Computers in other domains Operations research Education Document management and text processing

*) All Acronyms for classification groups and DS-BoK Knowledge Area Groups are brought in accordance to CF-DS-competence groups

**) Due to important role of the Internet and networking technologies, basic knowledge about networks are required. However, as a technology domain, Networks knowledge area group should be considered as a domain specific knowledge area in the general Data Science competences and knowledge definition.

Appendix C. Data Science Competence Framework (CF-DS) Excerption

This Appendix contains excerption from the original CF-DS document [??] that describes identified Data Science competences. The full CF-DS definition including both competences and skills is available in the CF-DS document.

C.1. Identified Data Science Competence Groups

The results of analysis presented here provides a basis and justification for defining two (new) competence areas that have not been explicitly defined in previous studies and frameworks. In particular, the proposed CF-DS competence and skills groups include:

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

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

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

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

Data Management, curation and preservation is already being conducted within existing (research) data related professions such as data archivist, data manager, digital data librarian, and others. Data management is also important component of European Research Area policy. It is extensively addressed by the Research Data Alliance and supported numerous projects, initiatives and training programmes[7].

Knowledge of the scientific research methods and techniques is something that makes Data Scientist profession different from all previous professions. Data Scientist is expected to have ability to find hidden value in raw data collected from scientific experiments or organisational activity. Tasks that are quite similar to researcher's work.

From the education and training point of view the identified competences can be treated or linked to expected learning or training outcomes. This aspect is discussed in details in other EDSF documents DS-BoK and MC-DS.

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

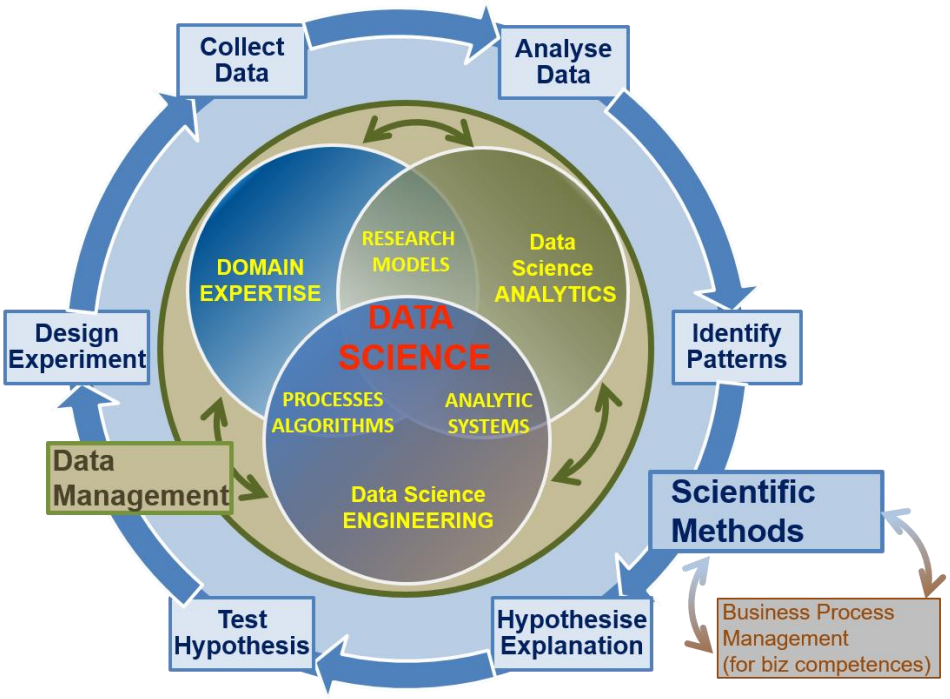
Table B.1 provides an example of competences definition for different groups that are extracted from the collected information. It indicates that all identified groups are demanded by companies and they have expected place in the corporate structure and activities.

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

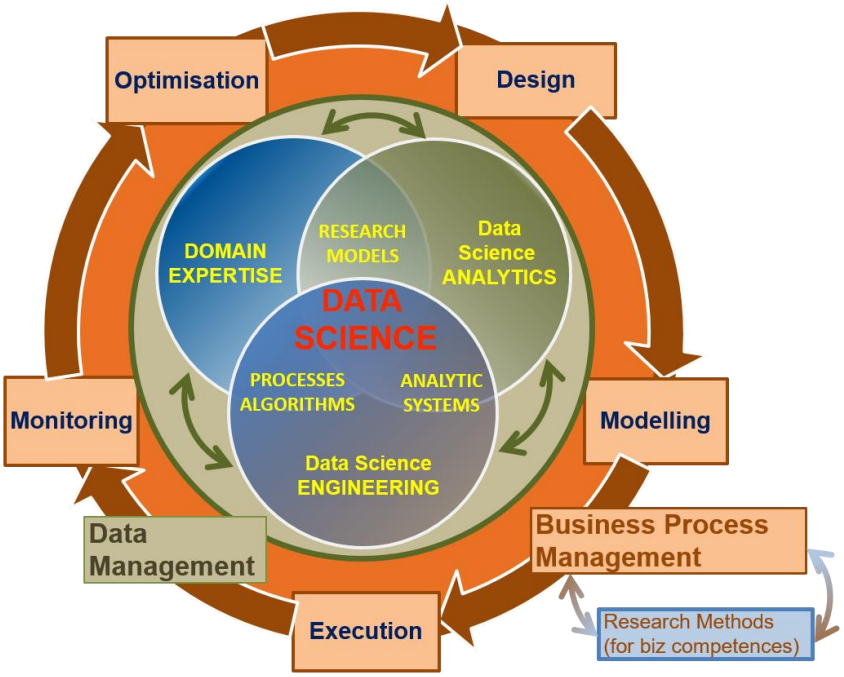
Table C.1. Enumerated competences definition for different Data Science competence groups

Data Science Analytics (DSDA)	**Data Management (DSDM)**	**Data Science Engineering (DSENG)**	**Scientific/ Research Methods (DSRM)**	**DS Domain Knowledge, e.g., Business Apps (DSDK)**
Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	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	Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSDA01 Use predictive analytics to analyse big data and discover new relations	DSDM01 Develop and implement data strategy, in particular, in a form of Data Management Plan (DMP)	DSENG01 Use engineering principles to research, design, prototype, data analytics applications, or develop structures, instruments, machines, experiments, processes, systems	DSRM01 Create new understandings and capabilities by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods	DSDK01 Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
DSDA02 Use appropriate statistical techniques on available data to deliver insights	DSDM02 Develop and implement relevant data models, including metadata	DSENG02 Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	DSRM02 Direct systematic study toward a fuller knowledge or understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSDK02 Use data to improve existing services or develop new services
DSDA03 Develop specialized analytics to enable agile decision making	DSDM03 Collect and integrate different data source and provide them for further analysis	DSENG03 Develops specialized data analysis tools to support executive decision making	DSRM03 Undertakes creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	DSDK03 Participate strategically and tactically in financial decisions that impact management and organizations
DSDA04 Research and analyse complex data sets, combine different sources and types of data to improve analysis.	DSDM04 Develop and maintain a historical data repository of analysis results (data provenance)	DSENG04 Design, build, operate relational non-relational databases	DSRM04 Ability to translate strategies into action plans and follow through to completion.	DSDK04 Provides scientific, technical, and analytic support services to other organisational roles
DSDA05 Use different data analytics platforms to process complex data	DSDM05 Ensure data quality, accessibility, publications (data curation)	DSENG05 Develop solutions for secure and reliable data access	DSRM05 Contribute to and influence the development of organizational objectives	DSDK05 Analyse customer data to identify/optimise customer relations actions
DSDA06 Visualise complex and variable data.	DSDM06 Manage IPR and ethical issues in data management	DSENG06 Prototype new data analytics applications	DERM06 Apply ingenuity to complex problems, develop innovative ideas	DSDK06 Analyse multiple data sources for marketing purposes

Figures B.1 (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 Data Management competences and skills and Scientific/Research Methods or Business Processes 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 C.1. 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 Scientific Methods typically include the following stages (see section 3.2.2 for reference to existing scientific methods definitions):

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

There are a number 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 in section 3.2.3):

- 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 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.

In summary, consolidation of the presented initial version of CF-DS was challenging task due to variety of information and required expertise. It is anticipated that it will undergo further improvement involving external and subject domain experts via EDISON Liaison Group and community involvement. Other project activities will provide feedback on necessary improvements and details.

It is a challenging task to include all required subjects and knowledge into education and training programs. It will require search for new approaches in Data Science education what will be a subject for subsequent EDISON project activities and work items.

C.2. Identified Data Science Skills

Another outcome of the current analysis of Data Science job advertisements and numerous blog articles[8] is the identification of required/demanded Data Science skills (also referred to as Big Data skills) that can split on 3 groups:

- General Data Science skills or required (provable) experience
- Knowledge and experience with Big Data hardware and software platforms
- Programming language: general and those having extended statistics libraries that are generally related to Data Science Engineering skills but in many cases are treated as a separate group of skills.

It is essential to mention that for such complex professional domain as Data Science the minimum required experiences with related methods, tools or platform is 2-3 years, some companies explicitly require experience up to 5 years.

Table C.2 lists identified Data Science skills related to the following groups

- Data Analytics and Machine Learning

[8] 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.

- Data Management/Curation (including both general data management and scientific data management)
- Data Science Engineering (hardware and software) skills
- Scientific/Research Methods
- Personal, inter-personal communication, team work (also called social intelligence or soft skills)
- Application/subject domain related (research or business)

The Data Analytics and Machine Learning group is the most populated what reflect real picture of required skills primary in this area as a basis for Data Science methods.

Table C.2. Identified Data Science skills

Skill Groups	Data Analytics and Machine Learning	Data Management/ Curation	Data Science Engineering (hardware and software)	Scientific/ Research Methods	Personal/Inter-personal communication, team work	Application/subject domain (research or business)
1	Artificial intelligence, machine learning	Manipulating and analysing complex, high-volume, high-dimensionality data from varying sources	Design efficient algorithms for accessing and analysing large amounts of data	Interest in data science	Communication skills	Recommender or Ranking system
2	Machine Learning and Statistical Modelling	for data improvement	Big Data solutions and advanced data mining tools	Analytical, independent, critical, curious and focused on results	Inter-personal intra-team and external communication	Data Analytics for commercial purposes
3	Machine learning solutions and pattern recognition techniques	Data models and datatypes	Multi-core/distributed software, preferably in a Linux environment	Confident with large data sets and ability to identify appropriate tools and algorithms	Network of contacts in Big Data community	Data sources and techniques for business insight and customer focus
4	Supervised and unsupervised learning	Handling vast amounts of data	Databases, database systems, SQL and NoSQL	Flexible analytic approach to achieve results at varying levels of precision		Mechanism Design and/or Latent Dirichlet Allocation
5	Data mining	Experience of working with large data sets	Statistical analysis languages and tooling	Exceptional analytical skills		Game Theory
6	Markov Models, Conditional Random Fields	(non)relational and (un)-structured data	Cloud powered applications design			Copyright and IPR
7	Logistic Regression, Support Vector Machines	Cloud based data storage and data management				
8	Predictive analysis and statistics (including Kaggle platform)	Data management planning				

9	(Artificial) Neural Networks	Metadata annotation and management				
10	Statistics	Data citation, metadata, PID (*)				
11	Natural language processing					
12	Computer Simulation					

(*) Persistent Identifier (PID), Data Types Registries, and Data Management Policies are outcome and products by Research Data Alliance (RDA) [27]

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, calculus, probability theory and statistics.

Table C.3 lists identified required skills and knowledge of the Big Data platforms (hardware and software) divided into groups:

- Big Data Analytics platforms
- Math& Stats tools
- Databases
- Data Management and Curation platform
- Data and applications visualisation

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

Table C.3. Required skills and knowledge of popular Big Data platforms and tools (hardware and software) [9]

	Big Data Analytics platforms	Math& Stats tools	Databases	Data/ applications visualization	Data Management and Curation platform
1	Big Data Analytics platforms	Advanced analytics tools (R, SPSS, Matlab, etc)	SQL and relational databases	Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)	Data modelling and related technologies (ETL, OLAP, OLTP, etc)
2	Big Data tools (Hadoop, Spark, etc)	Data Mining tools: RapidMiner, others	NoSQL Databases	Visualisation software (D3, Processing, Tableau, Gephi, etc)	Data warehouses platform and related tools
3	Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)	Matlab	NoSQL, Mongo, Redis	Online visualization tools (Datawrapper, Google Charts, Flare, etc)	Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
4	Real time and streaming analytics systems (like Flume, Kafka, Storm)	Python	NoSQL, Teradata		Backup and storage management (iRODS, XArch, Nesstar, others)

[9] 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/

5	Hadoop Ecosystem/platform	R, Tableau R	Excel		
6	Spotfire	SAS			
7	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	Scripting language, e.g. Octave			
8	Amazon Data Analytics platform (Kinesis, EMR, etc)	Statistical tools and data mining techniques			
9	Other cloud based Data Analytics platforms (HortonWorks, Vertica LexisNexis HPCC System, etc)	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

*) Highlighted are cloud based and online data analytics and data management platforms that are becoming increasingly popular for enterprise and business applications.

The following programming languages with extended data analysis and statistics libraries are identified as important for Data Scientist (and typically identified in job descriptions as requiring several years of practical experience)[10]:

- Python
- Scala
- pandas (Python Data Analysis Library)
- Julia
- Java 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

[10] 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