


International Association of
Business Analytics Certification

# Data Science Body of Knowledge

EDSF DS-BoK - Release 2

www.iabac.org




International Association of Business Analytics Certification

# Data Science Body of Knowledge (DS-BoK)

EDSF DS-BoK - Release 2

EDISON DATA SCIENCE FRAMEWORK
EUROPEAN COMMISION GRANT AGREEMENT NO: 675419

IABAC is a registered B.V (equivalent of UK English Private Limited) company in Netherlands.

| | |
|---|---|
| RSIN | : 859414206 |
| Vestigingsnummer | : 000041293304 |
| Statutaire naam | : IABAC B.V. |
| Statutaire zetel | : Eindhoven, The Netherlands |



European Commission

Content from EDISON project is licensed under the Creative Commons Attribution 4.0 International License (CC BY).

©2019 IABAC™ B.V. *A Guide to the Data Science Body of Knowledge – Version 2*

.

| 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 Data Science Body of Knowledge (DS-BoK) is designed to create a foundation for establishing a profession of Data Scientist for the industry, across the world. The 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 groups (KGA) 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 be a subject for further development in tight cooperation with the domain specialists.

The intended 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 the community forum and by presentation at community oriented workshops and conferences.

TABLE OF CONTENTS

1 Introduction ..... 5
2 EDISON Data Science Framework ..... 6
3 Overview of BoKs relevant to DS-BoK ..... 8
3.1 ACM Computer Science Body of Knowledge (CS-BoK) ..... 8
3.2 ICT professional Body of knowledge ICT-BoK ..... 9
3.3 Software Engineering Body of Knowledge (SWEBOK) ..... 9
3.4 Business Analysis Body of Knowledge (BABOK) ..... 10
3.5 Data Management Body of Knowledge (DM-BoK) by DAMAI ..... 11
3.6 Project Management Professional Body of knowledge (PM-BoK) ..... 11
4 Data Science Body of Knowledge (DS-BoK) definition ..... 13
4.1 General Approach and Structure of DS-BoK ..... 13
4.2 Data Analytics Knowledge Area ..... 14
4.3 DS-BoK Knowledge Area Groups ..... 14
4.4 Data Science Body of Knowledge Areas and Knowledge Units ..... 17
5 Conclusion and further developments ..... 28
5.1 Summary of findings ..... 28
5.2 Further developments to formalize CF-DS and DS-BoK ..... 28
6 References ..... 29
Acronyms ..... 30
Appendix A. Overview of Bodies of Knowledge relevant to Data Science ..... 31
A.1. ICT Professional Body of knowledge ..... 31
A.2. Data Management Professional Body of knowledge ..... 32
A.3. Project Management Professional Body of knowledge ..... 34
Appendix B. Subset of ACM/IEEE CCS2012 for Data Science (as defined in DS-BoK Release 1) ..... 37
B.1. ACM Classification Computer Science (2012) structure and Data Science related Knowledge Areas ..... 37
Appendix C. Data Science Competence Framework (CF-DS) Excerption ..... 41
C.1. Identified Data Science Competence Groups ..... 41
C.2. Knowledge required to support identified competences ..... 45
C.3. Identified Data Science Skills ..... 47

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

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

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

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

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

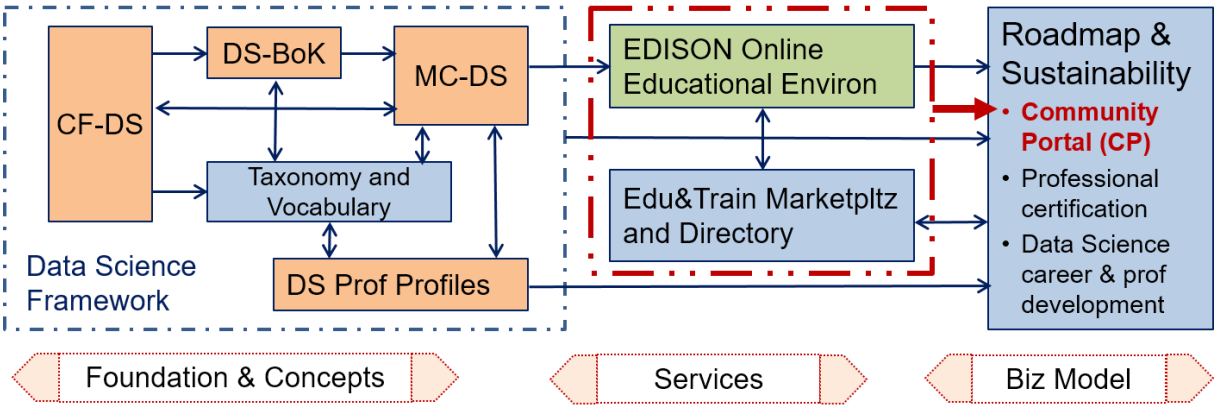



**Figure 1 EDISON Data Science Framework components.**

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

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

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK follows the same approach to collect community feedback and contribution: Open Access CC-BY community discussion document is published on the project website. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs defined based on the Classification Computer Science (CCS2012), components taken from other BoKs and proposed new KA to incorporate new technologies used in Data Science and their recent developments.

The MC-DS is built based on CF-DS and DS-BoK where Learning Outcomes are defined based on CF-DS competences and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. The proposed Learning outcomes are enumerated to have direct mapping to the enumerated competences in CF-DS. The preliminary version of MC-DS has been discussed at the first EDISON Champions Conference in June 2016 and collected feedback is incorporated in current version of MC-DS.

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

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

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

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

# 3 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) [6, 7, 8]
- ICT professional Body of Knowledge (ICT-BoK) [9]
- Business Analytics Body of Knowledge (BABOK) [10]
- Software Engineering Body of Knowledge (SWEBOK) [11]
- Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [12]
- Project Management Professional Body of Knowledge (PM-BoK) [13]

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 [7, 8] 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

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

programmes and specialisations. To be complete a curriculum should cover in addition to the topics of Core 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 [9] 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<br>• Software design<br>• Software construction<br>• Software testing<br>• Software maintenance<br>• Software configuration management<br>• Software engineering management<br>• Software engineering process<br>• Software engineering models and methods<br>• Software quality<br>• Software engineering professional practice<br>• Software engineering economics<br>• Computing foundations<br>• Mathematical foundations<br>• Engineering foundations | • Computer engineering<br>• Systems engineering<br>• Project management<br>• Quality management<br>• General management<br>• Computer science<br>• 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 [14]).

## 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).

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

This section presents summary of the Data Science Body of Knowledge definition given in separate DS-BoK document [2]. The presented DS-BoK definition is based on overview and analysis of existing bodies of knowledge that are relevant to 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 five Knowledge Area Groups (KAG) that are linked to the identified competence groups: KGA1-DSDA Data Analytics; KGA2-DSENG Data Science Engineering, KGA3-DSDM Data Management, KGA4-DSRMP Research Methods and Project Management; and KGA5-DSBA Business Analytics that represents one of the most active domain area empowered by Data Science. Defining the domain knowledge groups KAG*-DSDK both for science and business will be a subject for further development in tight cooperation with the domain specialists.

## 4.1 General Approach and Structure 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-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRMP: *Research Methods and Project Management*
- KAG5-DSBA: Business Analytics
- KAG*-DSDK: Placeholder for the Data Science Domain Knowledge groups to include domain specific knowledge

The subject domain related knowledge group (scientific or business) KAG*-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 (as discussed in section 3.8.3) and corresponding data analysis methods and models. These knowledge areas will be a subject for future development in tight cooperation with subject domain specialists.

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 proposed DS-BoK re-uses where possible or provides links to existing BoK's taking necessary KA definitions and 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:
ACM Computer Science CS-BoK [7, 8]
Business Analysis BABOK [10]
Software Engineering SWEBOK [11]
Data Management DMBOK by DAMA [12],
Project Management PM-BoK [13],
Classification Computer Science (CCS2012) [6] for Computer Science related knowledge areas.

## 4.2 Data Analytics Knowledge Area

Data Analytics includes different methods and algorithms, primarily statistical, to enable data processing, modelling, analysis and inspection with the goal of discovering useful information, providing insight and recommendations, and supporting decision-making. The following are commonly defined the Data Science Analytics Knowledge Areas:

- KA01.01 (DSDA.01/SMA) 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;
- KA01.02 (DSDA.02/ML) Machine learning and related methods for information search, image recognition, decision support, classification;
- KA01.03 (DSDA.03/DM) *Data mining* is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes;
- KA01.04 (DSDA.04/TDM) Text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data;
- KA01.05 (DSDA.05/PA) Predictive analytics focuses on application of statistical models for predictive forecasting or classification.
- KA01.06 (DSDA.06/BA) Business Analytics and Business Intelligence covers data analysis that relies heavily on aggregation and different data sources and focusing on business information;
- KA01.07 (DSDA.07/MSO) Computational modelling, simulation and optimisation

## 4.3 DS-BoK 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-DSDA, KAG2-DSENG, KAG3-DSDM groups that are well supported by existing BoK's and academic materials. General suggestions are provided for KAG4-DSRMP, KAG5-DSBA 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-DSENG 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.

Knowledge Units (KU) corresponding to suggested KAs are defined from different sources: existing BoK, CCS2012, and from practices in designing academic curricula and corresponding courses by universities and professional training organisations.

For the detailed definition of the KA and KU refer to the DS-BoK document [2]. The DS-BoK document contains detailed definition of the KAG1-DSA, KAG2-DSE, KAG3-DSDM, KAG4-DSRM, KAG5-DSBA 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.

Table 4.1. DS-BoK Knowledge Area Groups and corresponding Knowledge Areas

| KA Groups | Suggested DS Knowledge Areas (KA) | Knowledge Areas from existing BoK and CCS2012 scientific subject groups |
|---|---|---|
| KAG1-DSDA: Data Science Analytics | KA01.01 (DSDA.01/SMDA) Statistical methods for data analysis<br>KA01.02 (DSDA.02/ML) Machine Learning<br>KA01.03 (DSDA.03/DM) Data Mining<br>KA01.04 (DSDA.04/TDM) Text Data Mining<br>KA01.05 (DSDA.05/PA) Predictive Analytics<br>KA01.06 (DSDA.06/MODSIM) Computational modelling, simulation and optimisation | There is no formal BoK defined for Data Analytics.<br><br>Data Science Analytics related scientific subjects from CCS2012:<br>CCS2012: Computing methodologies<br>CCS2012: Mathematics of computing<br>CCS2012: Computing methodologies |
| KAG2-DSENG: Data Science Engineering | KA02.01 (DSENG.01/BDI) Big Data Infrastructure and Technologies<br>KA02.02 (DSENG.02/DSIAPP) Infrastructure and platforms for Data Science applications<br>KA02.03 (DSENG.03/CCT) Cloud Computing technologies for Big Data and Data Analytics<br>KA02.04 (DSENG.04/SEC) Data and Applications security<br>KA02.05 (DSENG.05/BDSE) Big Data systems organisation and engineering<br>KA02.06 (DSENG.06/DSAPPD) Data Science (Big Data) applications design<br>KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making) | ACM CS-BoK selected KAs:<br>AL - Algorithms and Complexity<br>AR - Architecture and Organization (including computer architectures and network architectures)<br>CN - Computational Science<br>GV - Graphics and Visualization<br>IM - Information Management<br>PBD - Platform-based Development (new)<br>SE - Software Engineering (can be extended with specific SWEBOK KAs)<br><br>SWEBOK selected KAs<br>• Software requirements<br>• Software design<br>• Software engineering process<br>• Software engineering models and methods<br>• Software quality<br><br>Data Science Analytics related scientific subjects from CCS2012:<br>CCS2012: Computer systems organization<br>CCS2012: Information systems<br>CCS2012: Software and its engineering |
| KAG3-DSDM: Data Management | KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation<br>KA03.02 (DSDM.02/DMS) Data management systems<br>KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure<br>KA03.04 (DSDM.04/DGOV) Data Governance<br>KA03.05 (DSDM.05/BDST0R) Big Data storage (large scale) | DM-BoK selected KAs<br>(1) Data Governance,<br>(2) Data Architecture,<br>(3) Data Modelling and Design,<br>(4) Data Storage and Operations,<br>(5) Data Security,<br>(6) Data Integration and Interoperability,<br>(7) Documents and Content,<br>(8) Reference and Master Data,<br>(9) Data Warehousing and Business Intelligence,<br>(10) Metadata, and<br>(11) Data Quality. |

| KA Groups | Suggested DS Knowledge Areas (KA) | Knowledge Areas from existing BoK and CCS2012 scientific subject groups |
|---|---|---|
| | KA03.06 (DSDM.05/DLIB) Digital libraries and archives | Data Science Analytics related scientific subjects from CCS2012:<br>CCS2012: Information systems |
| KAG4-DSRM: Research Methods and Project Management | KA04.01 (DSRMP.01/RM) Research Methods<br>KA04.01 (DSRMP.02/PM) Project Management | There are no formally defined BoK for research methods<br><br>PMI-BoK selected KAs<br>• Project Integration Management<br>• Project Scope Management<br>• Project Quality<br>• Project Risk Management |
| KAG5-DSBPM: Business Analytics | KA05.01 (DSBA.01/BAF) Business Analytics Foundation<br>KA05.02 (DSBA.02/BAEM) Business Analytics organisation and enterprise management | BABOK selected KAs *)<br>• Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts.<br>• Requirements Analysis and Design Definition.<br>• Requirements Life Cycle Management (from inception to retirement).<br>• Solution Evaluation and improvements recommendation. |

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

## 4.4 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) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| KAG1-DSDA: Data Science Analytics | KA01.01 DSDA.01/SMDA Statistical methods for data analysis | KU1.01.01 | Probability & Statistics | **CCS2012: Mathematics of computing**<br>• Discrete mathematics<br>○ Graph theory<br>○ Probability and statistics<br>○ Probabilistic representations<br>○ Probabilistic inference problems<br>○ Probabilistic reasoning algorithms<br>○ Probabilistic algorithms<br>• Statistical paradigms<br>• Mathematical software<br>• Information theory<br>• Mathematical analysis |
| | | KU1.01.02 | Statistical paradigms (regression, time series, dimensionality, clusters) | |
| | | KU1.01.03 | Probabilistic representations (causal networks, Bayesian analysis, Markov nets) | |
| | | KU1.01.04 | Frequentist and Bayesian statistics | |
| | | KU1.01.05 | Probabilistic reasoning | |
| | | KU1.01.06 | Exploratory and confirmatory data analysis | |
| | | KU1.01.07 | Quantitative analytics | |
| | | KU1.01.08 | Performance analysis | |
| | | KU1.01.09 | Markov models, Markov networks | |
| | | KU1.01.10 | Operations research | |
| | | KU1.01.11 | Information theory | |
| | | KU1.01.12 | Discrete Mathematics and Graph Theory | |
| | | KU1.01.13 | Mathematical analysis | |
| | | KU1.01.14 | Mathematical software and tools | |
| KAG1-DSDA: Data Science Analytics | KA01.02 DSDA.02/ML Machine Learning | KU1.02.01 | Machine Learning theory and algorithms | **CCS2012: Computing methodologies**<br>• Artificial intelligence<br>○ Machine learning<br>○ Learning paradigms<br>▪ Supervised learning<br>▪ Unsupervised learning<br>▪ Reinforcement learning<br>▪ Multi-task learning<br>• Machine learning approaches<br>○ Machine learning algorithms |
| | | KU1.02.02 | Supervised Machine Learning | |
| | | KU1.02.03 | Unsupervised Machine Learning | |
| | | KU1.02.04 | Reinforced learning | |
| | | KU1.02.05 | Classification methods | |
| | | KU1.02.06 | Design and Analysis of Algorithms | |
| | | KU1.02.07 | Game Theory & Mechanism design | |
| | | KU1.02.08 | Artificial Intelligence | |
| | | KU1.01.02 | Statistical paradigms (regression, time series, dimensionality, clusters) | |
| | | KU1.01.03 | Probabilistic representations (causal networks, Bayesian analysis, Markov nets) | |
| | | KU1.01.04 | Frequentist and Bayesian statistics | **CCS2012: Theory of computation** |
| | | KU1.01.05 | Probabilistic reasoning | |
| | | KU1.01.08 | Performance analysis | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | | | • Design and analysis of algorithms<br>○ Data structures design and analysis<br>• Theory and algorithms for application domains<br>○ Machine learning theory<br>○ Algorithmic game theory and mechanism design<br>• Semantics and reasoning |
| KAG1-DSDA: Data Science Analytics | KA01.03 DSDA.03/DM Data Mining | KU1.01.08 | Performance analysis | **CCS2012: Theory of computation**<br>• Design and analysis of algorithms<br>○ Data structures design and analysis<br>• Theory and algorithms for application domains<br>○ Machine learning theory<br>○ Algorithmic game theory and mechanism design<br>• Semantics and reasoning |
| | | KU1.02.01 | Machine Learning theory and algorithms | |
| | | KU1.02.02 | Supervised Machine Learning | |
| | | KU1.02.03 | Unsupervised Machine Learning | |
| | | KU1.02.04 | Reinforced learning | |
| | | KU1.02.05 | Classification methods | |
| | | KU1.03.01 | Data mining and knowledge discovery | |
| | | KU1.03.02 | Knowledge Representation and Reasoning | |
| | | KU1.03.03 | CRISP-DM and data mining stages | |
| | | KU1.03.04 | Anomaly Detection | |
| | | KU1.03.05 | Time series analysis | |
| | | KU1.03.06 | Feature selection, Apriori algorithm | |
| | | KU1.03.07 | Graph data analytics | |
| KAG1-DSDA: Data Science Analytics | KA01.04 DSDA.04/TDM Text Data Mining | KU1.04.01 | Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data | **CCS2012: Computing methodologies**<br>• Artificial intelligence<br>○ Natural language processing<br>○ Knowledge representation and reasoning<br>○ Search methodologies |
| | | KU1.04.02 | Data mining and text analytics | |
| | | KU1.04.03 | Natural Language Processing | |
| | | KU1.04.04 | Predictive Models for Text | |
| | | KU1.04.05 | Retrieval and Clustering of Documents | |
| | | KU1.04.06 | Information Extraction | |
| | | KU1.04.07 | Sentiments analysis | |
| | KA01.05 DSDA.05/PA | KU1.05.01 | Predictive modeling and analytics | |
| | | KU1.05.02 | Inferential and predictive statistics | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| KAG1-DSDA: Data Science Analytics | Predictive Analytics | KU1.05.03 | Machine Learning for predictive analytics | |
| | | KU1.05.04 | Regression and Multi Analysis | |
| | | KU1.05.05 | Generalised linear models | |
| | | KU1.05.06 | Time series analysis and forecasting | |
| | | KU1.05.07 | Deploying and refining predictive models | |
| KAG1-DSDA: Data Science Analytics | KA01.06 DSDA.06/MODSIM Computational modelling, simulation and optimisation | KU1.06.01 | Modelling and simulation theory and techniques (general and domain oriented) | **CCS2012: Computing methodologies**<br>• Modeling and simulation<br>○ Model development and analysis<br>○ Simulation theory<br>○ Simulation types and techniques<br>○ Simulation support systems |
| | | KU1.06.02 | Operations research and optimisation | |
| | | KU1.06.03 | Large scale modelling and simulation systems | |
| | | KU1.06.04 | Network oprtimisation | |
| | | KU1.06.05 | Risk simulation and queueing | |
| | | | | |
| KAG2-DSENG: Data Science Engineering | KA02.01 DSENG.01/BDI Big Data Infrastructure and Technologies | KU2.01.01 | Computer systems organisation for Big Data applications, CAP, BASE and ACID theorems | **CCS2012: Computer systems organization**<br>• Architectures<br>○ Parallel architectures<br>○ Distributed architectures<br>• Networks *)<br>○ Network Architectures<br>○ Network Services<br>○ Cloud Computing |
| | | KU2.01.02 | Parallel and Distributed Computer Architecture | |
| | | KU2.01.03 | High Performance and Cloud Computing | |
| | | KU2.01.04 | Clouds and scalable computing | |
| | | KU2.01.05 | Cloud based Big Data platforms and services | |
| | | KU2.01.06 | Big Data (large scale) storage and filesystems (HDFS, Ceph, etc) | |
| | | KU2.01.07 | NoSQL databases | |
| | | KU2.01.08 | Computer networks for high-performance computing and Big Data infrastructure | |
| | | KU2.01.09 | Computer networks: architectures and protocols | |
| | | KU2.01.10 | Big Data Infrastructure management and operation | |
| KAG2-DSENG: Data | KA02.02 DSENG.02/DSIAPP | KU2.02.01 | Big Data Infrastructure: services and components, including data storage infrastructure | • Proposed new KA for DS-BoK |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| Science Engineering | Infrastructure and platforms for Data Science applications | KU2.02.02 | Big Data analytics platforms and tools (including Hadoop, Spark, and cloud based Big Data services) | • Infrastructure and platforms for Data Science applications group:<br>• CCENG - Cloud Computing Engineering (infrastructure and services design, management and operation)<br>• CCAS - Cloud based applications and services development and deployment<br>• BDA – Big Data Analytics platforms (including cloud based)<br>• BDI - Big Data Infrastructure services and platforms, including data storage infrastructure |
| | | KU2.02.03 | Large scale cloud based storage and data management | |
| | | KU2.02.04 | Cloud based applications and services operation and management | |
| | | KU2.02.05 | Big Data and cloud based systems design and development | |
| | | KU2.02.06 | Data processing models (batch, steaming, parallel) | |
| | | KU2.02.07 | Enterprise information systems | **CCS2012: Information systems**<br>• Information storage systems<br>• Information systems applications |
| | | KU2.02.08 | Data security and protection | |
| | | | | |
| KAG2-DSENG: Data Science Engineering | KA02.03 DSENG.03/CCT Cloud Computing technologies for Big Data and Data Analytics | KU2.03.01 | Cloud Computing architecture and services | **DSDA Extension group for CCS201**<br>**Theory of computation**<br>• DSA Extension point: Algorithms for Big Data computation<br>**Mathematics of computing**<br>• DSA Extension point: Mathematical software for Big Data computation<br>**Computing methodologies**<br>• DSA Extension point: New DSA computing<br>**Information systems** |
| | | KU2.03.02 | Cloud Computing Engineering (infrastructure and services design, management and operation) | |
| | | KU2.03.03 | Cloud based applications and services operation and management | |
| KAG2-DSENG: Data Science Engineering | KA02.04 DSENG.04/SEC Data and Applications security | KU2.04.01 | Infrastructure, applications and data security | |
| | | KU2.04.02 | Data encryption and key management, bockchain based technologies | |
| | | KU2.04.03 | Access Control and Identiy Management | |
| | | KU2.04.04 | Security services management, including compliance and certification | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | KU2.04.05 | Data anonymisation | • DSA Extension point: Big Data systems (e.g. cloud based)<br>**Information systems applications**<br>• DSA Extension point: Big Data applications<br>DSA Extension point: Doman specific Data applications |
| | | KU2.04.06 | Data privacy | |
| KAG2-DSENG: Data Science Engineering | KA02.05 DSENG.05/BDSE Big Data systems organisation and engineering | KU2.05.01 | Big Data systems organisation and design | **CCS2012: Software and its engineering**<br>• Software organization and properties<br>○ Software system structures<br>• Software architectures<br>○ Software system models<br>○ Distributed systems organizing principles<br>▪ Cloud computing<br>▪ Grid computing<br>• Software notations and tools<br>○ General programming languages<br>○ Software creation and management |
| | | KU2.05.02 | Big Data algorithms for large scale data processing | |
| | | KU2.05.03 | Big Data Analytics | |
| | | KU2.05.04 | Big Data analytics platforms and tools (including Hadoop, Spark, and cloud based Big Data services) | |
| | | KU2.05.05 | Big Data algorithms for data ingest, pre-processing, and visualisation | |
| | | KU2.05.06 | Big Data systems for application domains | |
| | | KU2.05.07 | Big Data software (systems) architectures | |
| | | KU2.05.08 | Requirements engineering and software systems development | |
| | | KU2.05.09 | Large and ultra-large scale software systems organisation | |
| | | KU2.05.10 | DevOps and cloud enabled applications development | |
| | | KU2.05.11 | Big Data Infrastructure management and operation | |
| | | | | |
| KAG2-DSENG: Data Science Engineering | KA02.06 DSENG.06/DSAPPD Data Science (Big Data) applications design | KU2.06.01 | Data analytics, data handling software requirements and design | **SWEBOK selected KAs**<br>• Software requirements<br>• Software design<br>• Software construction<br>• Software testing<br>• Software maintenance |
| | | KU2.06.02 | Applications engineering management | |
| | | KU2.06.03 | Software engineering models and methods | |
| | | KU2.06.04 | Software quality assurance | |
| | | KU2.06.05 | Programming languages for Big Data analytics: R, python, Pig, Hive, others | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | KU2.06.06 | Models and languages for complex interlinked data presentation and visualisation | • Software configuration management<br>• Software engineering management<br>• Software engineering process<br>• Software engineering models and methods<br>• Software quality<br>• Agile development technologies<br>• Methods, platforms and tools<br>• DevOps and continuous deployment and improvement paradigm |
| | | KU2.06.07 | Agile development methods, platforms and tools | |
| | | KU2.06.08 | DevOps and continuous deployment and improvement paradigm | |
| | | | | |
| KAG2-DSENG: Data Science Engineering | KA02.07 DSENG.07/IS Information systems (to support data driven decision making) | KU2.07.01 | Decision Analysis and Decision Support Systems | **CCS2012: Information systems**<br>• Information systems applications<br>  ○ Decision support systems<br>  ▪ Data warehouses<br>  ▪ Expert systems<br>  ▪ Data analytics<br>  ▪ Online analytical processing<br>  ○ Multimedia information systems<br>  ○ Data mining |
| | | KU2.07.02 | Predictive analytics and predictive forecasting | |
| | | KU2.07.03 | Data Analysis and statistics | |
| | | KU2.07.04 | Data warehousing and Data Mining | |
| | | KU2.07.05 | Data Mining | |
| | | KU2.07.06 | Multimedia information systems | |
| | | KU2.07.07 | Enterprise information systems | |
| | | KU2.07.08 | Collaborative and social computing systems and tools | |
| | | | | **CCS2012: Information systems**<br>• Information systems applications<br>  ○ Enterprise information systems<br>  ○ Collaborative and social computing systems and tools |
| KAG3-DSDM: Data Management | KA03.01 DSDM.01/DMORG General principles and concepts in | KU3.01.01 | Data type registries, PID, metadata | **Proposed new KA for DS-BoK**<br>General Data Management KA's |
| | | KU3.01.02 | Data Lifecycle Management | |
| | | KU3.01.03 | Data infrastructure and Data Factories | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | Data Management and organisation | KU3.01.04 | Research data infrastructure, Open Science, Open Data, Open Access, ORCID | • Data Lifecycle Management<br>• Data archives/storage compliance and certification |
| | | KU3.01.05 | Data infrastructure compliance and certification | New KAs to support RDA recommendations and community data management models (Open Access, Open Data, etc)<br>• Data type registries, PIDs<br>• Data infrastructure and Data Factories<br>• New KAs to follow RDA and ERA community developments |
| | | KU3.01.06 | Ethical principle and data privacy | |
| | | KU3.01.07 | FAIR (Findable, Accessible, Interoperable, ) principles in Data Management | |
| | | | | |
| KAG3-DSDM: Data Management | KA03.02 DSDM.02/DMS Data management systems | KU3.02.01 | Data architectures (OLAP, OLTP, ETL) | **CCS2012: Information systems**<br>• Data management systems<br>○ Database design and models<br>○ Data structures<br>○ Database management system engines<br>○ Query languages<br>○ Database administration<br>○ Middleware for databases<br>○ Information integration<br>• CCS2012: Theory of computation<br>○ Database theory |
| | | KU3.02.02 | Data Modelling, Databases and Database Management Systems | |
| | | KU3.02.03 | Data structures | |
| | | KU3.02.04 | Data Models and Query Languages | |
| | | KU3.02.05 | Database design and models | |
| | | KU3.02.06 | Database administration | |
| | | KU3.02.07 | Data warehouses | |
| | | KU3.02.08 | Middleware for databases | |
| | | | | |
| KAG3-DSDM: Data Management | KA03.03 DSDM.03/EDMI Data Management and Enterprise data infrastructure | KU3.03.01 | Data management, including Reference and Master Data | **DM-BoK selected KAs**<br>(1) Data Governance,<br>(2) Data Architecture,<br>(3) Data Modelling and Design,<br>(4) Data Storage and Operations,<br>(5) Data Security, |
| | | KU3.03.02 | Data Warehousing and Business Intelligence | |
| | | KU3.03.03 | Data storage and operations | |
| | | KU3.03.04 | Data archives/storage compliance and certification | |
| | | KU3.03.05 | Metadata, linked data, provenance | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | KU3.03.06 | Data infrastructure, data registries and data factories | (6) Data Integration and Interoperability,<br>(7) Documents and Content,<br>(8) Reference and Master Data,<br>(9) Data Warehousing and Business Intelligence,<br>(10) Metadata, and<br>(11) Data Quality. |
| | | KU3.03.07 | Data security and protection | |
| | | KU3.03.08 | Data backup | |
| | | KU3.03.09 | Data anonymisation | |
| | | KU3.03.10 | Data Privacy | |
| | | | | |
| KAG3-DSDM: Data Management | KA03.04 DSDM.04/DGOV Data Governance | KU3.04.01 | Data governance, data quality, data Integration and Interoperability | DM-BoK (as above) |
| | | KU3.04.02 | Data Management Planning | |
| | | KU3.04.03 | Data Management Policy | |
| | | KU3.04.04 | Data interoperability | |
| | | KU3.04.05 | Data curation | |
| | | KU3.04.06 | Data provenance | |
| | | KU3.04.07 | Responsible data use, data privacy, ethical principles, IPR, legal issues | |
| KAG3-DSDM: Data Management | KA03.05 DSDM.05/BDST0R Big Data storage (large scale) | KU3.05.01 | Big Data storage infrastructure and operations | New DSENG Knowledge area:<br>Big Data Storage<br>• Distributed file systems<br>• Data Lakes<br>• Data Factories |
| | | KU3.05.02 | Storage architectures, distributed files systems (HDFS, Ceph, Lustre, Gluster, etc) | |
| | | KU3.05.03 | Data storage redundancy and backup | |
| | | KU3.05.04 | Data factories, data pipelines | |
| | | KU3.05.05 | Cloud based storage, Data Lakes | |
| KAG3-DSDM: Data Management | KA03.06 DSDM.05/DLIB Digital libraries and archives | KU3.06.01 | Digital libraries and archives organisation | **CCS2012: Information systems**<br>• Information systems applications<br>  ○ Digital libraries and archives |
| | | KU3.06.02 | Information Retrieval | |
| | | KU3.06.03 | Data curation and provenance | |
| | | KU3.06.04 | Search Engines technologies | |
| | | | | |
| KAG4-DSRMP: Research Methods and Project Management | KA04.01 DSRMP.01/RM Research Methods | KU4.01.01 | Research methodology, paradigms and research cycle | **Proposed new KA for DS-BoK for DSRM related competences:**<br>• Research methodology, research cycle (e.g. 4 steps model Hypothesis – |
| | | KU4.01.02 | Modelling and experiment planning | |
| | | KU4.01.03 | Data selection and quality evaluation | |
| | | KU4.01.04 | Data lifecycle | |
| | | KU4.01.05 | Use cases analysis: research infrastructures and projects | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | KU4.01.06 | Research data management plan and ethical issues | Research Methods – Artefact – Validation)<br>• Modelling and experiment planning<br>• Data selection and quality evaluation<br>• Use cases analysis: research infrastructures and projects |
| KAG4-DSRMPM: Research Methods and Project Management | KA04.01 DSRMP.02/PM Project Management | KU4.02.01 | Project Integration Management | **PMI-BoK selected KAs**<br>• Project Integration Management<br>• Project Scope Management<br>• Project Quality<br>• Project Risk Management |
| | | KU4.02.02 | Project Scope Management | |
| | | KU4.02.03 | Project Quality | |
| | | KU4.02.04 | Project Risk Management | |
| KAG5-DSBPM: Business Analytics | KA05.01 DSBA.01/BAF Business Analytics Foundation | KU5.01.01 | Business Analytics and Business Intelligence: Data, Models (statistical) and Decisions | **BABOK selected KAs**<br>• Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts.<br>• Requirements Analysis and Design Definition.<br>• Requirements Life Cycle Management (from inception to retirement).<br>• Solution Evaluation and improvements recommendation. |
| | | KU5.01.02 | Data driven Customer Relations Management (CRP), User Experience (UX) requirements and design | |
| | | KU5.01.03 | Operations Analytics | |
| | | KU5.01.04 | Business Process Optimization | |
| | | KU5.01.05 | Data Warehouses technologies, data integration and analytics | |
| | | KU5.01.06 | Data driven marketing technologies | |
| | | KU5.01.07 | Business Analytics Capstone | |
| | | KU5.01.08 | Econometrics methods and application for Business Analytics | |
| | | KU5.01.09 | Cognitive technologies for Business Analytics | |
| KAG6-DSBA: Business Analytics | KA05.02 DSBA.02/BAEM Business Analytics organisation and enterprise management | KU5.02.01 | Business processes and operations | **Proposed new KA/KU for DS-BoK**<br>• General Business processes and operations KAs<br>• Business processes and operations<br>• Agile Data Driven methodologies, processes and enterprises |
| | | KU5.02.02 | Project scope and risk management | |
| | | KU5.02.03 | Business Analysis Planning and Monitoring | |
| | | KU5.02.04 | Requirements Analysis and Design Definition | |
| | | KU5.02.05 | Requirements Life Cycle Management (from inception to retirement) | |

| Knowledge Area Groups (KAG) | Knowledge Areas (KA) | Knowledge Unit (KU) | Suggested Knowledge Units (KU) | Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK) |
|---|---|---|---|---|
| | | KU5.02.06 | Solution Evaluation and improvements recommendation | • Use cases analysis: business and industry |
| | | KU5.02.07 | Agile Data Driven methodologies, processes and enterprises | |
| | | KU5.02.08 | Use cases analysis: business and industry | |

# 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 experts via standardisation bodies, professional communities and directly via the project network.

## 5.1 Summary of findings

The provided document contains ongoing 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 Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- Domain Knowledge and Expertise (Subject/Scientific domain related)
- Data Management and Governance (including data stewardship, curation, and preservation)
- *Research Methods for research related professions and Business Process Management for business related professions*

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

- KAG1-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRMP: *Research Methods and Project Management*
- KAG5-DSBA: Business Analytics
- KAG*-DSDK: Placeholder for the Data Science Domain Knowledge groups to include domain specific knowledge

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

It is anticipated that the presented ongoing development will require practical validation by experts and communities of practice that will include the following specific tasks and activities:

- Continue validating and improving the currently proposed knowledge areas and knowledge units 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 suggestions for new knowledge areas and classifications classes.
- 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] Data Science Competence Framework(CF-DS) [online] http://edison-project.eu/data-science-competence-framework-cf-ds

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

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

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

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

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

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

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

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

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

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

[12] 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

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

[14] 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:<br>***1.*** ICT Strategy & Governance<br>***2.*** Business and Market of ICT<br>***3.*** Project Management<br>***4.*** Security Management<br>***5.*** Quality Management<br>***6.*** Architecture<br>***7.*** Data and Information Management<br>***8.*** Network and Systems Integration<br>***9.*** Software Design and Development<br>***10.*** Human Computer Interaction<br>***11.*** Testing<br>***12.*** Operations and Service Management<br><br>Each Knowledge Area is defined by;<br>• List of items required as foundational knowledge necessary under this Knowledge Area;<br>• List of references to the e-Competence Framework (dimension 4: knowledge);<br>• List of possible job profiles that require having an understanding of the Knowledge Area;<br>• 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;<br>• Professional Associations: to promote the Body of Knowledge to their members, ICT professionals;<br>• 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<br>• http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf<br>DAMA International Guide to Data Management Body of Knowledge (on purchase)<br>• https://technicspub.com/dmbok/<br>Other resources<br>DAMA International Dictionary of Data Management Terms (on purchase)<br>• https://technicspub.com/dmbok/ |
| 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.<br>The 11 Data Management Knowledge Areas are:<br>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.<br>2. **Data Architecture** – the overall structure of data and data-related resources as an integral part of the enterprise architecture<br>3. **Data Modelling &Design** – analysis, design, building, testing, and maintenance (was Data Development in the DAMA - DMBOK 1st edition)<br>4. **Data Storage & Operations** – structured physical data assets storage deployment and management (was Data Operations in the DAMA-DMBOK 1st edition)<br>5. **Data Security** – ensuring privacy, confidentiality and appropriate access<br>6. **Data Integration & Interoperability** – acquisition, extraction, transformation, movement, delivery, replication, federation, virtualization and operational support ( a Knowledge Area new in DMBOK2)<br>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.<br>9. **Data Warehousing & Business Intelligence** – managing analytical data processing and enabling access to decision support data for reporting and analysis<br>10. **Metadata** – collecting, categorizing, maintaining, integrating, controlling , managing, and delivering metadata<br>11. **Data Quality** – defining, monitoring, maintaining data integrity, and improving data quality<br>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.<br>Each context diagram includes:<br>• *Definition*: a concise description of the Knowledge Area.<br>• *Goals*: he desired outcomes of the Knowledge Area within this Topic.<br>• *Process*: the list of discrete activities and sub-activities to be performed, with activity group indicators.<br>• *Inputs*: what documents or raw materials are directly necessary for a Process to initiate or continue<br>• *Supplier roles*: roles and/or teams that supply the inputs to the process.<br>• *Responsible roles:* roles and/or teams that perform the process.<br>• *Stakeholder roles*: roles and/or teams Informed or consulted on the process execution.<br>• *Tools*: technology types used by the process to perform the function.<br>• *Deliverables* : what is directly produced by the processes<br>• *Consumer roles*: roles and/or teams that expect and receive the Deliverables.<br>• *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.<br>• Helping build consensus within the data management community.<br>• Helping data stewards, data owners, and data professionals understand their responsibilities.<br>• Providing the basis for assessments of data management effectiveness and maturity.<br>• Guiding efforts to implement and improve data management knowledge areas.<br>• Educating students, new hires, practitioners and executives on data management knowledge areas<br>• Guiding the development and delivery of data management curriculum content for higher education.<br>• Suggesting areas of further research in the field of data management.<br>• Helping data management professionals prepare for Certified Data Management Professional (CDMP) data exams.<br>• Assisting organizations in defining their enterprise data strategy. |
| Certification promoted | Certified Data Management Professional (CDMP) in four levels:<br>• Associate (https://www.dama.org/content/cdmp-associate),<br>• Practitioner (https://www.dama.org/content/cdmp-practitioner),<br>• Master (https://www.dama.org/content/cdmp-master),<br>• 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.<br>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<br>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.<br>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)<br>other resources<br>• Lexicon of PM terms (http://www.pmi.org/PMBOK-Guide-and-Standards/PMI-lexicon.aspx - free for members)<br>• PMBoK in other 11 languages (Arabian, Italian, Korean, Russian, Hindi, Japanese, Portuguese, Spanish, German, French, Chinese);<br>• **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)<br>External sites:<br>http://www.projectmanagement.com/Practices/PMI-Standards/ |
| Year/Edition | 2014/5th edition |
| Structure of BoK | The Five Process Groups<br>*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.<br>*Executing* - Processes to complete the work documented within the Project Management Plan.<br>*Monitoring and Controlling* - Processes to track and review the project progress and performance. This group contains the Change Management.<br>*Closing* - Processes to formalize the project or phase closure.<br><br>The Nine Knowledge Areas<br>*Project Integration Management* - Processes to integrate various parts of the Project Management.<br>*Project Scope Management* - Processes to ensure that all of the work required is completed for a successful Project and manages additional "scope creep".<br>*Project Time Management* - Processes to ensure the project is completed in a timely manner.<br>*Project Cost Management* - Processes to manage the planning, estimation, budgeting and management of costs for the duration of the project.<br>*Project Quality Management* - Processes to plan, manage and control the quality and to provide assurance the quality standards are met.<br>*Project Human Resource Management* - Processes to plan, acquire, develop and manage the project team.<br>Project Communications Management - Processes to plan, manage, control, distribute and final disposal of project documentation and communication.<br>*Project Risk Management* - Processes to identify, analyse and management of project risks.<br>*Project Procurement Management* - Processes to manage the purchase or acquisition of products and service, or result to complete the project.<br>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):<br>CAPM – Certified Associate Project Management<br>PMP – Project Management Professional<br>PgMP – Program Management Professional<br>PfMP – Portfolio Management Professional<br>PMI–PBA – PMI-Professional Business Analyst<br>PMI-ACP – PMI Agile Certified Professional<br>PMI-RMP – PMI Risk Management Professional<br>PMI-SP – Scheduling Professional<br><br>*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 (as defined in DS-BoK Release 1)

This Appendix provides historical information about subset of the ACM/IEEE CCS2012 taxonomy used in the DS-BoK Release 1. This information is provided for those who build their Data Science curriculum definition on the previous DS-BoK version. The new DS-BoK Release 2 version has the whole set of the generically defined Knowledge areas and knowledge units that can be partly mapped to CCS2012 but primarily based on the knowledge subjects defined in CF-DS definition.

The defined below subset of ACM CCS (2012) classification can provide a basis for its future extension with a new classification group related to Data Science and individual disciplines that are missing in the current ACM/IEEE 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) [6] 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 DS-BoK Release 1, although compatible with DS-BoK Release 2:

- KAG1-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: 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-DSRMP: Research Methods and Project Management 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 Release 1 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<br>Design and analysis of algorithms<br>Data structures design and analysis<br>Theory and algorithms for application domains<br>Machine learning theory<br>Algorithmic game theory and mechanism design<br>Database theory<br>Semantics and reasoning |
| Data Science Analytics (DSDA) | Mathematics of computing<br>Discrete mathematics<br>Graph theory<br>Probability and statistics<br>Probabilistic representations<br>Probabilistic inference problems<br>Probabilistic reasoning algorithms<br>Probabilistic algorithms<br>Statistical paradigms<br>Mathematical software<br>Information theory<br>Mathematical analysis |
| Data Science Analytics (DSDA) | Computing methodologies<br>Artificial intelligence<br>Natural language processing<br>Knowledge representation and reasoning<br>Search methodologies<br>Machine learning<br>Learning paradigms<br>Supervised learning<br>Unsupervised learning<br>Reinforcement learning<br>Multi-task learning<br>Machine learning approaches<br>Machine learning algorithms |
| Data Science Analytics (DSDA) | Information systems<br>Information systems applications<br>Decision support systems<br>Data warehouses<br>Expert systems<br>Data analytics<br>Online analytical processing<br>Multimedia information systems |

| DS-BoK Knowledge Groups *) | ACM (2012) Classification facets related to Data Science |
|---|---|
| | Data mining |
| Data Science Analytics (DSDA) | Theory of computation |
| | DSA Extension point: Algorithms for Big Data computation |
| | Mathematics of computing |
| EXTENSION POINT | 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 |
| 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) | Information systems |
| | Data management systems |
| | Data types and structures description |
| EXTENSION POINT | 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 |

| DS-BoK Knowledge Groups *) | ACM (2012) Classification facets related to Data Science |
|---|---|
| | Collaborative and social computing systems and tools |
| Data Science Engineering (DSENG) | Software and its engineering |
| | Software organization and properties |
| | DSE Extension point: Big Data applications design |
| EXTENSION POINT | 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 |
| | DSA Extension point: Doman specific Data applications |
| DS Domain Knowledge (DSDK) | Applied computing |
| | Physical sciences and engineering |
| | Life and medical sciences |
| EXTENSION POINT | 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 [1] that is required for understanding of the presented in this document the DS-BoK. The excerption includes identified Data Science competences and corresponding knowledge units required to support those 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 the job market study and analysis for Data Science and Data Science enabled vacancies, conducted at the initial stage of the project, provided a basis and justification for defining the main competence groups that are commonly required by companies, including identification such skills as Data Management and Research methods that were not required formerly required for data analytics jobs.

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

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

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

Additional common competence groups demanded by organisations

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

Data management, curation and preservation competences are already attributed to the existing (research) data related professions such as data archivist, data manager, digital data librarian, and others. Data management is also important component of European Research Area and Open Data policies. It is extensively addressed by the Research Data Alliance and supported by numerous projects, initiatives and training programmes[4].

Knowledge of the scientific research methods and techniques is something that makes Data Scientist profession different from all previous professions.

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

The identified 5 Data Science related competence groups provide a better basis for defining consistent and balanced education and training programmes for Data Science related jobs, re-skilling and professional certification.

Table C.1 provides the proposed Data Science competences definition for different groups supported by the data extracted for the collected information. The presented competences definition has been reviewed by a number of expert groups and individual experts (see Section 7 for details). The presented competences are required for different professional profiles, organisational roles and throughout the whole data lifecycle, but not necessary to be provided y a single role or individuum. The presented competences are enumerated to allow easy use and linking between all EDSF document.

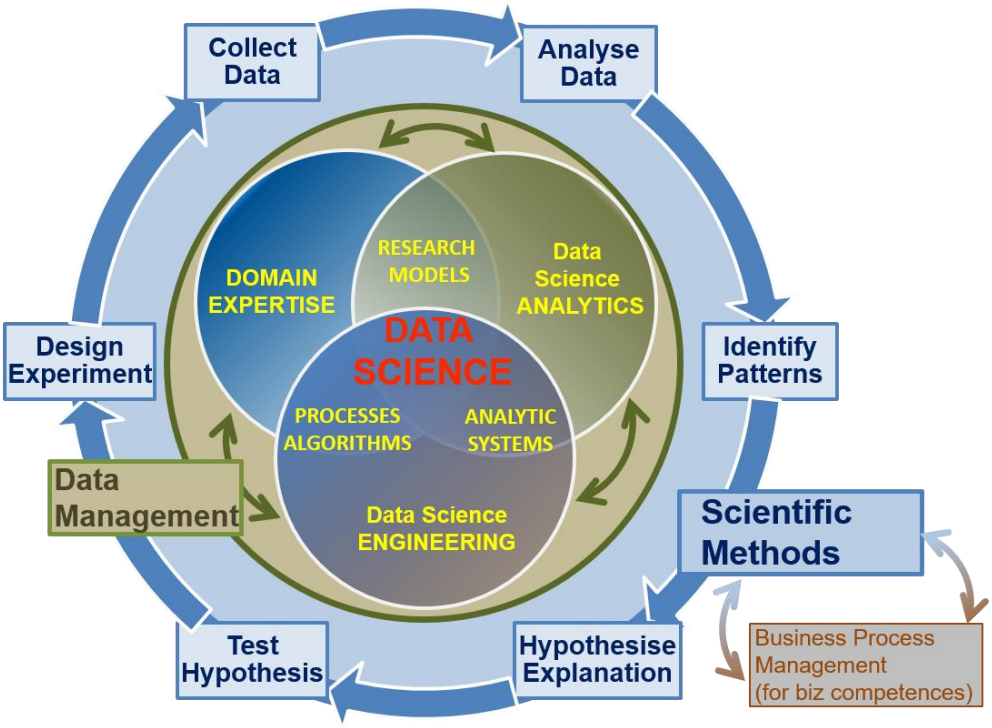
[4] Research Data Alliance Europe https://europe.rd-alliance.org/

Table C.1. Competences definition for different Data Science competence groups

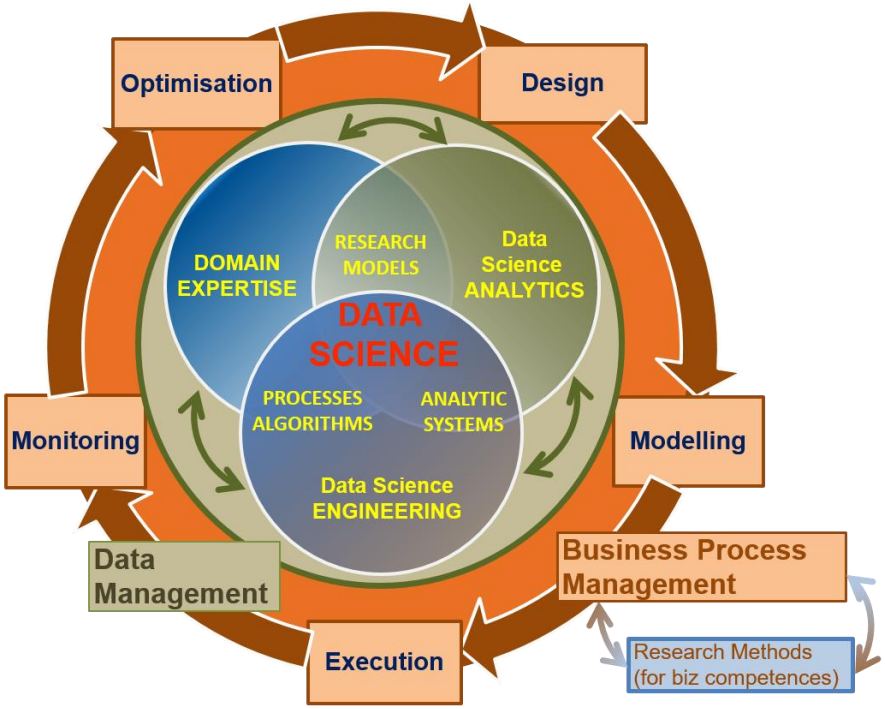
| Data Analytics (DSDA) | Data Science Engineering (DSENG) | Data Management (DSDM) | Research Methods and Project Management (DSRM) | Domain related Competences (DSDK): Applied to Business Analytics (DSBA) |
|---|---|---|---|---|
| DSDA<br>Use appropriate data analytics and statistical techniques on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

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

Figures C.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 the Data Management competences and skills and Research Methods or Business Process Management knowledge for all categories and profiles of Data Scientists.




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




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

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

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

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

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

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

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

## C.2. Knowledge required to support identified competences

Table C.2. Knowledge required to support identified competences

| **KDSDA Data Science Analytics** | **KDSENG Data Science Engineering** | **KDSDM Data Management** | **KDSRM Research Methods and Project Management** | **KDSBA Business Analytics** |
|---|---|---|---|---|
| KDSDA01<br>Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others | KDSENG01<br>Systems Engineering and Software Engineering principles, methods and models, distributed systems design and organisation | KDSDM01<br>Data management and enterprise data infrastructure, private and public data storage systems and services | KDSRM01<br>Research methods, research cycle, hypothesis definition and testing | KDSBA01<br>Business Analytics (BA) and Business Intelligence (BI); methods and data analysis; cognitive technologies |

| | | | | |
|---|---|---|---|---|
| KDSDA02<br>Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA) | KDSENG02<br>Cloud Computing, cloud based services and cloud powered services design | KDSDM02<br>Data storage systems, data archive services, digital libraries, and their operational models | KDSRM02<br>Experiment design, modelling and planning | KDSBA02<br>Business Processes Management (BPM), general business processes and operations, organisational processes analysis/modelling |
| KDSDA03<br>Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms) | KDSENG03<br>Big Data technologies for large datasets processing: batch, parallel, streaming systems, in particular cloud based | KDSDM03<br>Data governance, data governance strategy, Data Management Plan (DMP) | KDSRM03<br>Data lifecycle and data collection, data quality evaluation | KDSBA03<br>Agile Data Driven methodologies, processes and enterprises |
| KDSDA04<br>Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering) | KDSENG04<br>Applications software requirements and design, agile development technologies, DevOps and continuous improvement cycle | KDSDM04<br>Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.) | KDSRM04<br>Use cases analysis: research infrastructure and projects | KDSBA04<br>Econometrics: data analysis and applications |
| KDSDA05<br>Text Data Mining: statistical methods, NLP, feature selection, apriori algorithm, etc. | KDSENG05'<br>Systems and data security, data access, including data anonymisation, federated access control systems | KDSDM05<br>Data lifecycle and organisational workflow, data provenance and linked data | KDSRM05<br>Research Data Management Plan (DMP) and data stewardship | KDSBA05<br>Data driven Customer Relations Management (CRP), User Experience (UX) requirements and design |
| KDSDA06<br>Predictive Analytics | KDSENG06<br>Compliance based security models, privacy and IPR protection | KDSDM06<br>Data curation and data quality, data integration and interoperability | KDSRM06<br>Project management: scope, planning, assessment, quality and risk management, team management | KDSBA06<br>Use cases analysis: business and industry |
| KDSDA07<br>Prescriptive Analytics | KDSENG07<br>Relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets | KDSDM07<br>Data protection, backup, privacy, IPR, ethics and responsible data use | | KDSBA07<br>Data Warehouses technologies, data integration and analytics |
| KDSDA08<br>Graph Data Analytics: (path analysis, connectivity analysis, community analysis, centrality analysis, sub-graph isomorphism, etc. | KDSENG08<br>Big Data infrastructures, high-performance networks, infrastructure and services management and operation | KDSDM08<br>Metadata, PID, data registries, data factories, standards and com0liance | | KDSBA08<br>Data driven marketing technologies |
| KDSDA09<br>Qualitative analytics | KDSENG09<br>Modeling and simulation, theory and systems | KDSDM09<br>Open Data, Open Science, research data archives/repositories, Open Access, ORCID | | |
| KDSDA10<br>Natural language processing | KDSENG10<br>Information systems, collaborative systems | | | |

| | | | | |
|---|---|---|---|---|
| KDSDA11<br>Data preparation and pre-processing | | | | |
| KDSDA12<br>Performance and accuracy metrics | | | | |
| KDSDA13<br>Operations Research | | | | |
| KDSDA14<br>Optimisation | | | | |
| KDSDA15<br>Simulation | | | | |

## C.3. Identified Data Science Skills

For identified Data Science skills and technical platforms knowledge refer to the original CF-DS document [1].




EDISON DATA SCIENCE FRAMEWORK
EUROPEAN COMMISION GRANT AGREEMENT NO: 675419

IABAC is a registered B.V (equivalent of UK English Private Limited) company in Netherlands.
RSIN (Registration number) : 859414206
Registered Name : IABAC B.V ,
Place: Eindhoven, The Netherlands

Content from EDISON project is licensed under the Creative Commons Attribution 4.0 International License (CC BY).
©2019 IABACTM B.V. A Guide to the Data Science Body of Knowledge – Version 2