EDISON

Education for Data Intensive Science to Open New science frontiers

Project no. 675419
Coordination and Support Action
Funded by the Horizon 2020 Framework Programme of the European Union

Call identifier:	**H2020-ICT-2015-1**
Topic:	**INFRASUPP-4-2015 - New professions and skills for e-infrastructures**
Start date of project:	**1 September 2015 (24 months duration)**

Deliverable D2.3

EDISON Data Science Framework (final version)

Due date:	30/06/2017
Submission date:	03/07/2017
Deliverable leader:	UvA

Dissemination Level

☒	PU:	Public
□	PP:	Restricted to other programme participants (including the Commission Services)
□	RE:	Restricted to a group specified by the consortium (including the Commission Services)
□	CO:	Confidential, only for members of the consortium (including the Commission Services)

EDISON project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675419

Change history

Version	Date	Partners	Description/Comments
0.0	01/06/2017	UvA	Initial version and placeholders for contribution collection
0.1	18/06/2017	UvA	Chapter 2 and 3 content added
0.1	22/06/2017	UiS	CF-DS, DS-BOK, MC-DS and DSPP sections added
0.2	26/06/2017	UvA	Content of the deliverable finalised
0.3	30/06/2017	UvA	Final draft
0.4	03/07/2017	UvA	Final version

Contributors

Document Editors: Yuri Demchenko		
Contributors:		
Author Initials	Name of Author	Institution
YD	Yuri Demchenko	University of Amsterdam (UvA)
TW	Tomasz Wiktorski	University of Stavanger (UiS)
AB	Adam Belloum	University of Amsterdam (UvA)
AM	Andrea Manieri	Engineering (ENG)

Executive summary

The presented deliverable reports the final results of the EDISON Data Science Framework (EDSF) development that includes the following components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional Profiles (DSPP).

The deliverable itself provides only summary of the four EDSF documents sufficient for understanding the framework and its potential usability for different purposes and tasks related to Data Science competences and skills development, management, assessment, and shaping education and training programs.

The CF-DS is the core component of EDSF that provides a basis for all other components. CF-DS defines the five groups of competences for Data Science that include the commonly recognised groups Data Science Analytics, Data Science Engineering, Domain Knowledge and other groups *Data Management* and *Scientific Methods* that are recognised to be important for a successful work of Data Scientist but are not explicitly mentioned in existing frameworks. The CF-DS defines also a number of personal and attitude that are expected from the Data Scientist to successfully in the multi-disciplinary Data Science teams and agile data driven organisations.

The DS-BoK is built around the Knowledge Area Groups that are linked to the CF-DS competence groups. The presented DS-BoK defines related Knowledge Areas and Knowledge Units that can be directly used for constructing Data Science curricula with the MC-DS. DS-BoK re-uses where possible relevant knowledge areas from existing bodies of knowledge and includes new knowledge areas that are required to support identified Data Science competences. Further work will be required to define domain specific knowledge areas what will require involvement of the subject matter experts and/or corresponding communities.

Professional profiles defined in DSPP document can be used for customising education or training curricula, team building and candidate's competence profile assessment against the required job profile.

The EDSF can provide a basis and a guidance for universities and professional training organisations to define their Data Science curricula and training programmes, 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.

It is intended that EDSF provides a valuable contribution to defining common European Data Science competence framework and developing an effective model and approach to address growing demand for Data Science and data related competences and skills by European Digital Singe Market (DSM), European Open Science Cloud (EOSC) and other stakeholders and actors in the emerging digital data driven economy.

TABLE OF CONTENTS

1 Introduction

This deliverable presents the final results of the EDISON Data Science Framework (EDSF) development that includes the following components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional Profiles (DSPP).

The deliverable itself provides only summary of the four EDSF documents sufficient for understanding the framework and its potential usability for different purposes and tasks related to Data Science competences and skills development, management, assessment, and shaping education and training programs. The EDSF provides comprehensive taxonomy that can be used by organisation to develop their Data Science capacity building strategies starting from assessing the organisation's maturity on Data Science and related emerging technologies.

The CF-DS is the core component of EDSF that provides a basis for all other components. CF-DS defines the five groups of competences for Data Science that include the commonly recognised groups Data Science Analytics, Data Science Engineering, Domain Knowledge and other groups *Data Management* and *Scientific Methods* that are recognised to be important for a successful work of Data Scientist but are not explicitly mentioned in existing frameworks. The CF-DS defines also a number of personal and attitude that are expected from the Data Scientist to successfully in the multi-disciplinary Data Science teams and agile data driven organisations.

The DS-BoK is built around the Knowledge Area Groups that are linked to the CF-DS competence groups. The presented DS-BoK defines related Knowledge Areas and Knowledge Units that can be directly used for constructing Data Science curricula with the MC-DS. DS-BoK re-uses where possible relevant knowledge areas from existing bodies of knowledge and includes new knowledge areas that are required to support identified Data Science competences. Further work will be required to define domain specific knowledge areas what will require involvement of the subject matter experts and/or corresponding communities.

Professional profiles defined in DSPP document can be used for customising education or training curricula, team building and candidate's competence profile assessment against the required job profile.

The EDSF can provide a basis and a guidance for universities and professional training organisations to define their Data Science curricula and training programmes, 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.

It is intended that EDSF provides a valuable contribution to defining common European Data Science competence framework and developing an effective model and approach to address growing demand for Data Science and data related competences and skills by European Digital Singe Market (DSM), European Open Science Cloud (EOSC) and other stakeholders and actors in the emerging digital data driven economy.

The deliverable report has the following structure. Section 2 introduces the EDSF and describes its structure and components. Sections 3, 4, 5, 6 correspondingly provide short summaries of the EDSF main components: CF-DS, BS-BoK, MC-DS, and DSPP. Section 7 provides examples of possible EDSF use for competences assessment and Data Science team building. Section provide information about the mechanisms used to validate and collect feedback from expert community and academic and industry practitioners In Data Science. The report concludes with the summary of developments, suggestions for EDSF use and information about EDSF sustainability and maintenance after the project ends.

Refer to original EDSF documents for details and full EDSF Release 2 documentation.

2 EDISON Data Science Framework (EDSF)

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

Figure 2.1 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

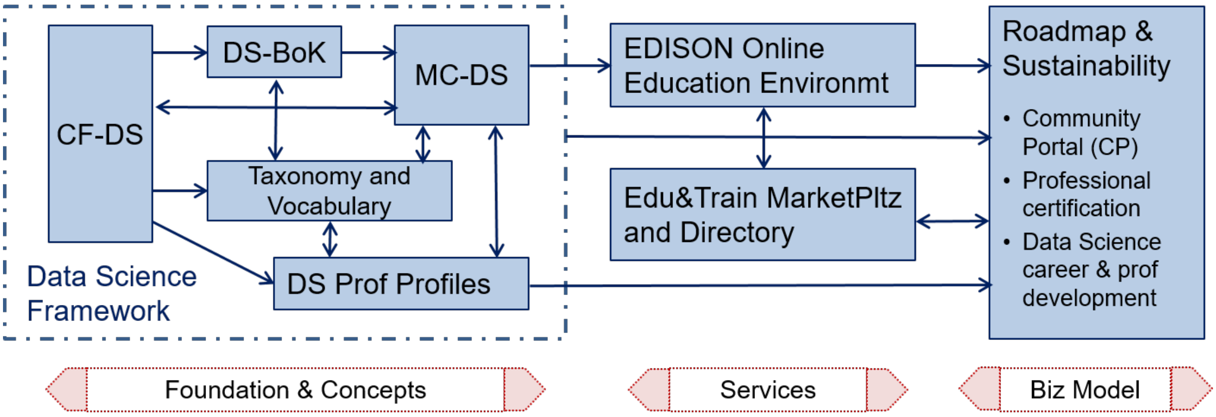


Figure 2.1. EDISON Data Science Framework components.

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 required 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 revised and updated DS-BoK version used in this deliverable is presented in Appendix C and will be published as a next version of the DS-BoK discussion document after discussion with the EDISON Liaison Group (ELG) experts.

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 will create 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 contacts with the European and international standardisation bodies and professional organisations. This work is being done by the project as a part of the Dissemination and communication activity.

The EDISON Data Science professional ecosystem illustrated in Figure 2.1 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 Data Science Competence Framework (CF-DS)

This section describes the final version of the Data Science Competence Framework (CF-DS) published as Release 2 EDSF document [1]. It serves as a foundation for the definition of the Data Science Body of Knowledge and Data Science Model Curriculum. The section provides a summary of the CF-DS sufficient to understand other sections describing other EDSF components. For details and full definition of the CF-DS refer to [1]

3.1 Definition of the Data Scientist

There is no well-established definition of the Data Scientist due to a number of competences and skills expected from these specialists. The proposed Data Scientist definition is based on the definition provided in the NIST SP1500-1 document [5] and extended with the need to deliver value to the organisation or to the project:

*"A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the **big data lifecycle**, till the delivery of an **expected scientific** and **business value** to science or industry."*

The NIST document defines the following groups of skills required from the Data Scientists: domain experience, statistics and data mining, and engineering skills [5]. The EDSF has proposed structured definition of the Data Scientist via definition of the related competences, skills, knowledge and proficiency level.

Initial attempt to define the Data Scientist has been made by O'Reilly Strata Survey (2013) (see [13] and Appendix A) which recognised creativity as an important feature of Data Scientist.

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

The qualified Data Scientist should be capable of working in different roles in different projects and organisations such as Data Engineer, Data Analyst or Data Architect, Data Steward, etc., and possess the necessary skills to effectively operate components of the complex data infrastructure and processing applications through all stages of the Data lifecycle till the delivery of expected scientific and business values to science and/or industry.

3.2 Relation to and use of existing framework and studies

The following describes how existing frameworks and documents were used for the analysis of initial competences and skills.

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

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

b) European e-Competence Framework (e-CFv3.0) [9]

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

c) European ICT profiles CWA 16458 (2012) [10]

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

d) European Skills, Competences, Qualifications and Occupations (ESCO) [11]

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

e) ACM Computing Classification System (ACM CCS2012) [12]

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

f) O'Reilly Strata Survey (2013) [13]

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

g) EC Report on the Consultation Workshop (May 2012) “Skills and Human Resources for e-Infrastructures within Horizon 2020” [14].

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

3.3 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[1].

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 3.1 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 8 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 by a single role or individuum. The presented competences are enumerated to allow easy use and linking between all EDSF document.

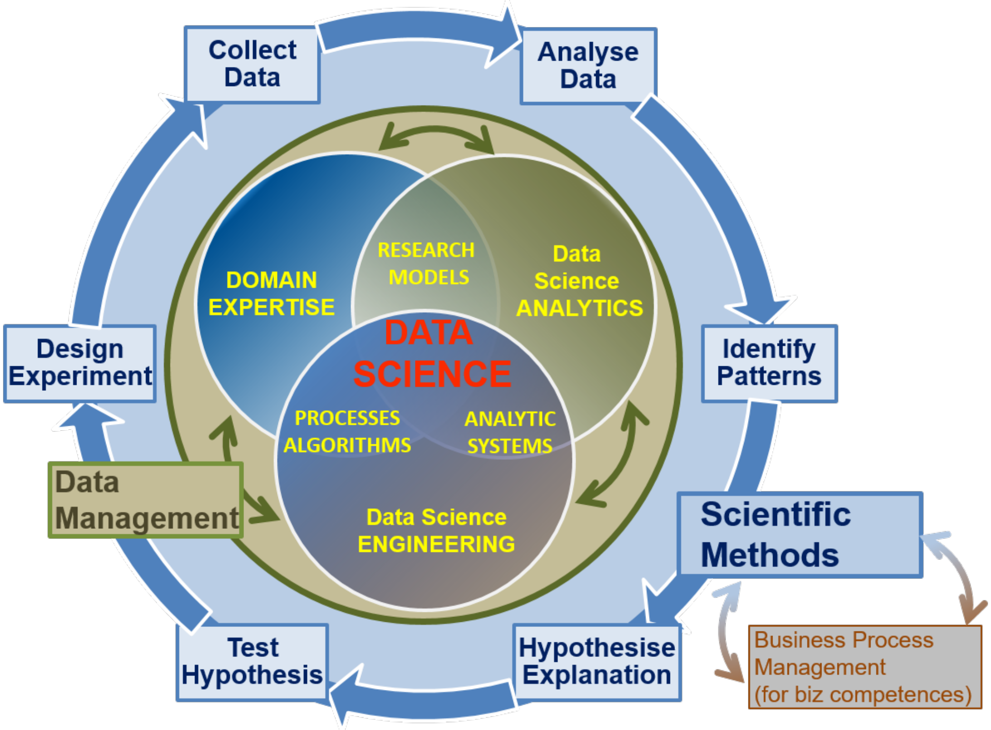
[1] Research Data Alliance Europe https://europe.rd-alliance.org/

Table 3.1. Competences definition for different Data Science competence groups

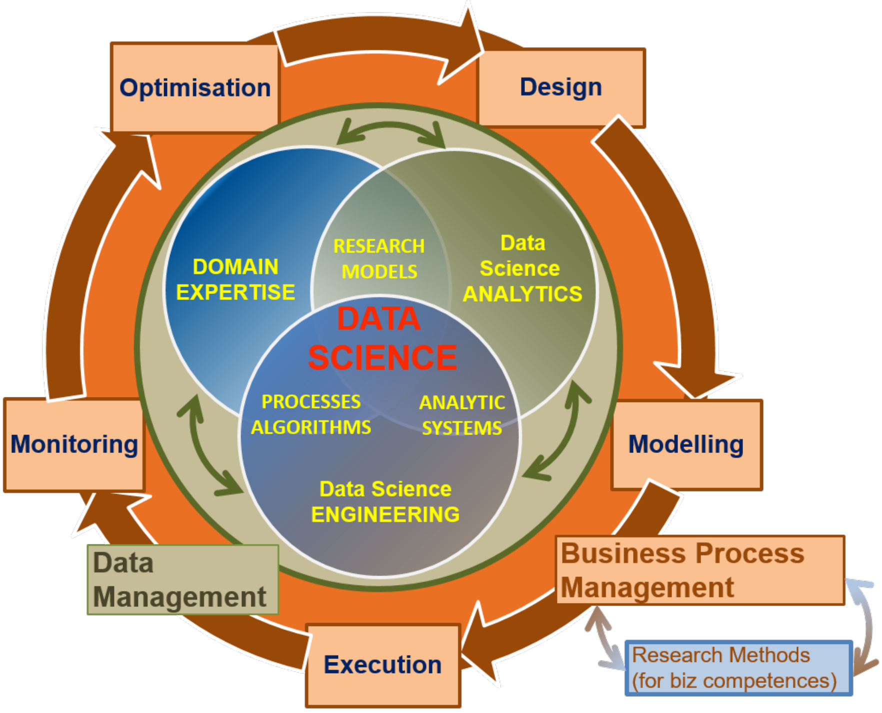
Data Analytics (DSDA)	**Data Science Engineering (DSENG)**	**Data Management (DSDM)**	**Research Methods and Project Management (DSRM)**	**Domain related Competences (DSDK): Applied to Business Analytics (DSBA)**
DSDA Use appropriate data analytics and statistical techniques on available data to discover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

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

Figures 3.1 (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 3.1. 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.

3.4 Identified Data Science Skills and their mapping to Competences

Required Data Science skills are defined based on the job market study of the current analysis of Data Science job market, extended with the numerous blog articles analysis[2] published by Data Science practitioners which provide valuable information in such new emerging area as Data Science.

The identified skills can be organised in the following groups:

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

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

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

3.4.1 Data Science skills related to the main competence groups

The following Data Science skills were identified to support competence groups (please refer to the CF-DS document for details [1]):

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

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

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

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

The following Data Science skill groups were identified related to the Data Analytics languages, tools, platforms and Big Data infrastructure that are split on the following sub-groups (please refer to the CF-DS document for details [1]):

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

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

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

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

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

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

3.5 Knowledge required to support identified competences

Table 3.2 provides enumerated list of knowledge topics/units that are required to support corresponding competence groups. There is no direct mapping between individual competences and knowledge units, singe competence may be mapped to multiple knowledge units.

Table 3.2. 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 Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	KDSENG01 Systems Engineering and Software Engineering principles, methods and models, distributed systems design and organisation	KDSDM01 Data management and enterprise data infrastructure, private and public data storage systems and services	KDSRM01 Research methods, research cycle, hypothesis definition and testing	KDSBA01 Business Analytics (BA) and Business Intelligence (BI); methods and data analysis; cognitive technologies
KDSDA02 Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	KDSENG02 Cloud Computing, cloud based services and cloud powered services design	KDSDM02 Data storage systems, data archive services, digital libraries, and their operational models	KDSRM02 Experiment design, modelling and planning	KDSBA02 Business Processes Management (BPM), general business processes and operations, organisational processes analysis/modelling
KDSDA03 Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)	KDSENG03 Big Data technologies for large datasets processing: batch, parallel, streaming systems, in particular cloud based	KDSDM03 Data governance, data governance strategy, Data Management Plan (DMP)	KDSRM03 Data lifecycle and data collection, data quality evaluation	KDSBA03 Agile Data Driven methodologies, processes and enterprises
KDSDA04 Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	KDSENG04 Applications software requirements and design, agile development technologies, DevOps and continuous improvement cycle	KDSDM04 Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)	KDSRM04 Use cases analysis: research infrastructure and projects	KDSBA04 Econometrics: data analysis and applications
KDSDA05 Text Data Mining: statistical methods, NLP, feature selection, apriori algorithm, etc.	KDSENG05’ Systems and data security, data access, including data anonymisation, federated access control systems	KDSDM05 Data lifecycle and organisational workflow, data provenance and linked data	KDSRM05 Research Data Management Plan (DMP) and data stewardship	KDSBA05 Data driven Customer Relations Management (CRP), User Experience (UX) requirements and design
KDSDA06 Predictive Analytics	KDSENG06 Compliance based security models, privacy and IPR protection	KDSDM06 Data curation and data quality, data integration and interoperability	KDSRM06 Project management: scope, planning, assessment, quality and risk management, team management	KDSBA06 Use cases analysis: business and industry

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

3.6 Proficiency levels

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

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

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

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

3.7 Data Scientist Personal skills (aka "soft" skills)

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

CF-DS document [1] provides detailed list of (1) Data Science Professional or Attitude skills (Thinking and acting like Data Scientist), and (2) 21st century workplace skills

Defining professional skills will contribute to the establishing the Data Science as a distinct profession and the Data Scientist as a recognisable role in organisation.

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

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

Following outcome and recommendations by the DARE project [3] we define the basic Data Science Analytics competences and skills (also can be referred as literacy) that must be required from all roles working in the Data Science team or interacting with the data analytics teams.

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

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

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

3.9 Example of Data Science Analytics Competences definition

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

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

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA01	Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle

Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 1 (Professional)	Level 1 (Expert)
	Understand and be able to select an approach to analysing selected datasets. Demonstrate understanding and perform statistical hypothesis testing, explain statistical significance.	Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction	Develop and plan required data analytics for organizational tasks, including: evaluating requirements and specifications of problems to recommend possible analytics-based solutions

Dimension 4	Knowledge ID	Knowledge unit definition
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA03	Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)
	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)
	KDSDA06	Predictive Analytics
	KDSDA07	Prescriptive Analytics
	KDSDA11	Data preparation and pre-processing
	KDSDA12	Performance and accuracy metrics
	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
	SDSDA02	Use Data Mining techniques
	SDSDA04	Apply Predictive Analytics methods
	SDSDA05	Apply Prescriptive Analytics methods
	SDSDA06	Use Graph Data Analytics for organisational network analysis, customer relations, other tasks
Skills Data Analytics languages, tools and platforms	DSALANG01	R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)
	DSALANG02	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)
	DSADB03	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)
	DSAVIZ02	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
	DSABDA03	Real time and streaming analytics systems (Flume, Kafka, Storm)
	DSABDA09	Kaggle competition, resources and community platform
	DSADEV03	Git versioning system as a general platform for software development

<table>
<tr><td>Dimension 1 Competence Group</td><td>DSDA</td><td colspan="2">Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA02</td><td colspan="2">Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction</td></tr>
<tr><td rowspan="2">Dimension 3 Proficiency level</td><td>Level 1 (Entry/Associate)</td><td>Level 1 (Professional)</td><td>Level 1 (Expert)</td></tr>
<tr><td>TBD
Be familiar and use related methods and tools.
Work under supervision or guidance</td><td>TBD.
Independent work and development.
Knowledge and experience with multiple techniques ad tools.
Full applications development and deployment</td><td>TBD.
Expert knowledge and experience with multiple data analytics techniques, tools and platforms,
Architecture level development, assessment and selection of appropriate solution. Suggestions for new approaches and applications, including relevant data collection.</td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td colspan="2">Knowledge unit definition</td></tr>
<tr><td rowspan="6">Knowledge</td><td>KDSDA01</td><td colspan="2">Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others</td></tr>
<tr><td>KDSDA02</td><td colspan="2">Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)</td></tr>
<tr><td>KDSDA04</td><td colspan="2">Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)</td></tr>
<tr><td>KDSDA06</td><td colspan="2">Predictive Analytics</td></tr>
<tr><td>KDSDA14</td><td colspan="2">Optimisation</td></tr>
<tr><td>KDSDA15</td><td colspan="2">Simulation</td></tr>
<tr><td></td><td>Skill ID</td><td colspan="2">Skills definition</td></tr>
<tr><td rowspan="5">Skills Data Analytics methods and algorithms</td><td>SDSDA01</td><td colspan="2">Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)</td></tr>
<tr><td>SDSDA02</td><td colspan="2">Use Data Mining techniques</td></tr>
<tr><td>SDSDA04</td><td colspan="2">Apply Predictive Analytics methods</td></tr>
<tr><td>SDSDA13</td><td colspan="2">Apply oprtimisation methods</td></tr>
<tr><td>SDSDA14</td><td colspan="2">Use computer simulation methods</td></tr>
<tr><td rowspan="7">Skills Data Analytics languages, tools and platforms</td><td>DSALANG02</td><td colspan="2">Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)</td></tr>
<tr><td>DSADB01</td><td colspan="2">SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)</td></tr>
<tr><td>DSADB03</td><td colspan="2">NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)</td></tr>
<tr><td>DSABDA02</td><td colspan="2">Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)</td></tr>
<tr><td>DSABDA03</td><td colspan="2">Real time and streaming analytics systems (Flume, Kafka, Storm)</td></tr>
<tr><td>DSADEV01</td><td colspan="2">Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others</td></tr>
<tr><td>DSADEV03</td><td colspan="2">Git versioning system as a general platform for software development</td></tr>
</table>

4 Data Science Body of Knowledge (DS-BoK)

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 [15]
Business Analysis BABOK [16]
Software Engineering SWEBOK [17]
Data Management DMBOK by DAMA [18],
Project Management PM-BoK [19],
Classification Computer Science (CCS2012) [12] for Computer Science related knowledge areas.

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

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

5 Data Science Model Curriculum (MC-DS)

The proposed MC-DS intends to provide guidance to universities and training organisations in the construction of Data Science programmes and individual courses selection that are balanced according to the requirements elicited from the research and industry domains. MC-DS can be used for assessment and improvement of existing Data Science programmes with respect to the knowledge areas and competence groups that are associated with specific professional profiles. When coupled with individual or group competence benchmarking, MC-DS can also be used for building individual training curricula and professional (self/up/re-) skilling for effective career management.

MC-DS follows the competence-based curriculum design approach grounded in the Data Science competences defined in CF-DS and correspondingly defined Learning Outcomes (LO). The DS-BoK provides a basis for structuring the proposed MC-DS by Knowledge Area Groups (KAG) and Knowledge Areas (KA) defined in correspondence with the CF-DS competence groups and individual competences. MC-DS design supports design of programs and courses that make use of best educational practices, such as Constructive Alignment, Problem- and Project-based Learning, Bloom’s Taxonomy.

This chapter presents a short overview of the MC-DS organization and its application to defining knowledge topics (knowledge units) and learning outcomes for two main Knowledge Area Groups: Data Science Analytics and Data Science Engineering. It also provides suggestions for ECTS points specification for main professional profiles group: Data Science Professionals DSP04-DSP09 (refer to section 6 or DSPP document [4]). Full MC-DS version is presented in the MC-DS document [3] and can be found on EDSF website. It contains MC-DS definitions for all Knowledge Area Groups, extended Learning Outcomes inventory, and ECTS points specification for all professional profile groups.

5.1 Organization and Application of Model Curriculum

In this section, we start by describing organization of MC-DS and relation between its elements and other elements of EDSF. Further, we explain how to use MC-DS together with EDSF to design a new education program in Data Science.

5.1.1 Organization of Model Curriculum

MC-DS organisation is based on Data Science Competence Framework, Professional Profiles and Body of Knowledge. For each enumerated competence, MC-DS defines Learning Outcome according to knowledge or mastery level (defined as Familiarity, Usage, Assessment). Each Knowledge Area Group of DS-BoK is mapped to existing academic subject classification groups that is primarily based on ACM Classification Computer Science CCS2012 [12] complemented with the domain or technology specific classifications such as defined in the existing BoK’s ACM CS-BOK [15], BABOK [16], SWEBOK [17], DM-BoK [18], PM-BOK [19], and others that should to be defined by subject matter experts. For each KAG, MC-DS specifies Learning Outcomes and mastery levels following Bloom’s Taxonomy verb usage. Learning Outcomes are also linked to a set of Learning Units, which are examples of practical application of Knowledge Units. ECTS points are provided for Professional Profile groups and divided into Tier-1, Tier-2, Elective and Prerequisite categories to help create detailed tracks and specializations for academic programs and professional training.

Figure 5.1 illustrates the relation between different EDSF components when defining specific academic or professional training programme that can be tailored for specific target Data Science professional group.

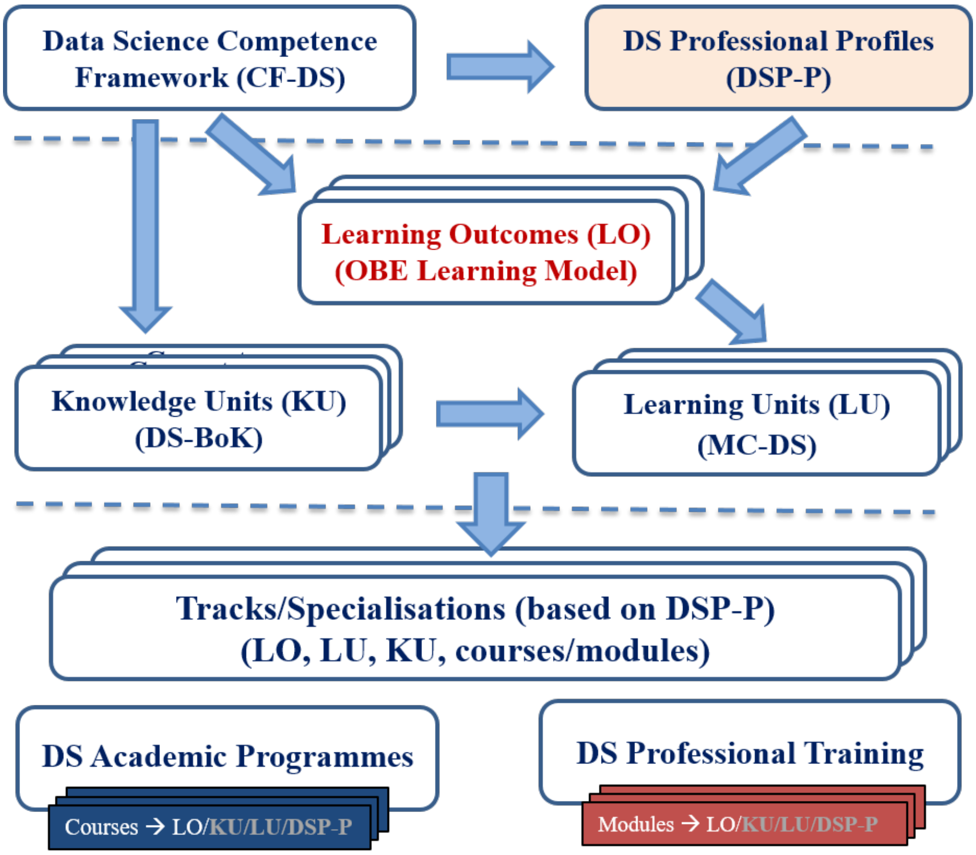

Data Science Competence Framework (CF-DS)
DS Professional Profiles (DSP-P)
Learning Outcomes (LO) (OBE Learning Model)
Knowledge Units (KU) (DS-BoK)
Learning Units (LU) (MC-DS)
Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)
DS Academic Programmes
Courses → LO/KU/LU/DSP-P
DS Professional Training
Modules → LO/KU/LU/DSP-P

Figure 5.1. Interaction between different components of EDSF when using Model Curriculum for defining academic of professional training programme for target professional group.

5.1.2 Application of Model Curriculum

This section describes a general approach to application of the Model Curriculum to create an educational program that is illustrated in Figure 5.2.

The work starts by deciding on a target Data Science professional profiles group the program should cover and the level of the program, usually Bachelor or Master. These elements allow to identify a set of competencies to be address in the program. To identify relevant Knowledge Units and to what extent they should be covered in the new program, the program designer can consult tables with ECTS point, which are defined for each Professional Profile. ECTS points specifications include a degree of flexibility to adjust to the particular needs. For each Knowledge Area, MC-DS defines a set of topics based on BoK and a set of learning outcomes based on Competence Framework. Topics and learning outcomes become a base for definition of new courses or use of existing courses. It is important to note that when designing a specific course, it may include elements from several Knowledge Areas to ensure consistency of the whole Data Science programme.

Adjustment of learning outcomes levels for different proficiency levels can be done based on the full MC-DS definition in [3] that defines learning outcomes for all CF-DS competences and for all mastery/proficiency levels. Learning outcomes can repeat between subgroups within the same KAG, however adjusted to a specific course and topics context.

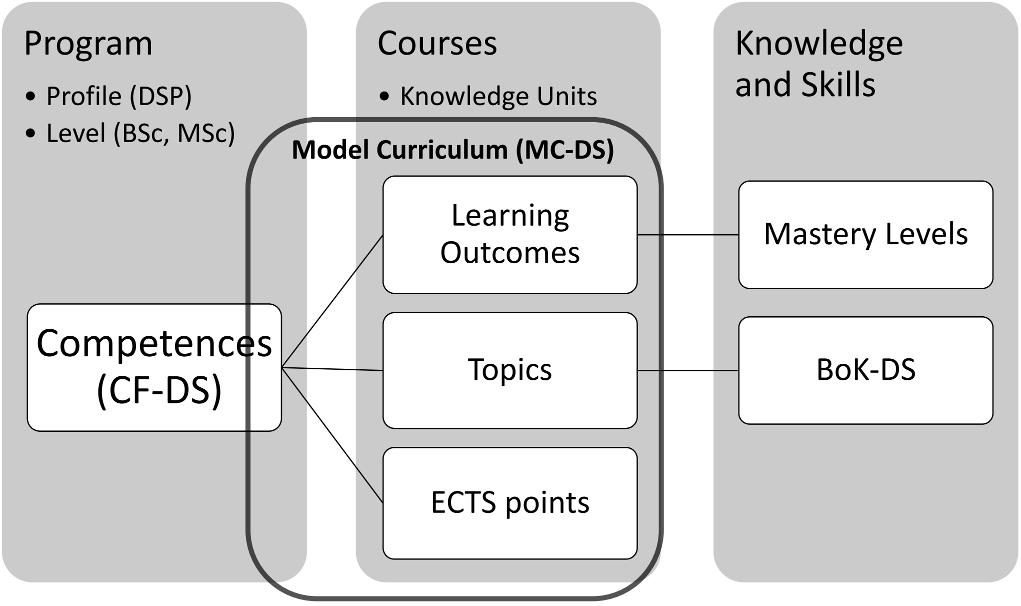


Figure 5.2. Visualization of Model Curriculum application for programs and courses.

5.2 Assignments of ECTS points to Competence Groups and Knowledge Areas

This section presents an example ECTS points specification for main professional profile group: Data Science Professionals. Table 5.1 contains example specification for a program on a Bachelor level, while Table 5.2 provides example specification for a program on Master level.

Points for each Knowledge Area are divided into four categories: Tier-1, Tier-2, Elective and Prerequisite. For each program 100% of Tier-1 should be covered, 80% of Tier-2 and 50% of Elective, with minor adjustments if necessary. Such system ensure that each program based on MC-DS covers basic competence and knowledge, but at the same time allowing for a necessary degree of flexibility. No prerequisites are expected for a Bachelor program, while for a Master program we set prerequisite at around 50% of combined Tier-1 and Tier-2. The goal is to ensure that students entering a program have at least basic competence necessary to succeed in Master education, but at the same time it allows students from relatively wide set of backgrounds to participate. Students that do not possess the required competences, should be able to make up the difference by engaging in additional courses or bootcamps. In case, program wants to accept student with a different profile, e.g. pure Computer Science or pure Statistics, we recommend that distribution of points in the program is adjusted to balance that. For instance, students with BSc in Computer Science come with a strong background in Software Development and Databases, but limited knowledge of statistics. In such a case ECTS points should be moved between these areas.

ECTS specification for Data Analytics and Data Science Engineering Knowledge Area groups is presented here. Points for Data Management and Research methods can be found in full specification of MC. They complement ECTS points from two groups presented here to provide 180 ECTS for Bachelor programs and 120 ECTS for Master programs.

Table 5.1. ECTS credit points for BSc program for profiles DSP04-09

Course related to DS-BoK Knowledge Areas	Tier - 1	Tier - 2	Elective	Prerequisite
DSDA/SMA (Statistical methods and data analysis)	7	4	6	NA
DSDA/ML (Machine learning)	9	8	8	NA
DSDA/DM (Data Mining)	5	4	3	NA
DSDA/TDM (Text Data Mining)	4	3	3	NA
DSDA/PA (Predictive analytics)	6	7	6	NA
DSDA/MSO (Modeling, simulation, and optimization)	5	3	4	NA

DSENG/BDI (Big Data infrastructure and technologies)	4	3	4	NA
DSENG/IPDS (Infrastructure and platforms for Data Science)	8	5	4	NA
DSENG/CCT (Cloud Computing technologies for BD and DA)	6	5	5	NA
DSENG/SEC (Data and Applications security)	2	2	2	NA
DSENG/BDSE (Big Data systems organization and engineering)	9	5	5	NA
DSENG/DSAD (Data Science/Big Data application design)	9	5	5	NA
DSENG/SE (Information Systems)	4	6	5	NA

Table 5.2. ECTS credit points for MSc program for profiles DSP04-09

Course related to DS-BoK Knowledge Areas	Tier - 1	Tier - 2	Elective	Prerequisite
DSDA/SMA (Statistical methods and data analysis)	6	2	4	6
DSDA/ML (Machine learning)	6	5	5	9
DSDA/DM (Data Mining)	4	2	4	5
DSDA/TDM (Text Data Mining)	3	2	4	4
DSDA/PA (Predictive analytics)	4	4	4	7
DSDA/MSO (Modeling, simulation, and optimization)	2	2	4	4
DSENG/BDI (Big Data infrastructure and technologies)	3	3	3	4
DSENG/IPDS (Infrastructure and platforms for Data Science)	5	3	4	7
DSENG/CCT (Cloud Computing technologies for BD and DA)	5	3	4	6
DSENG/SEC (Data and Applications security)	1	2	2	2
DSENG/BDSE (Big Data systems organization and engineering)	5	3	4	7
DSENG/DSAD (Data Science/Big Data application design)	5	3	4	7
DSENG/SE (Information Systems)	2	3	3	5

5.3 Data Science Data Analytics (KAG1 – DSDA) related courses

Data Science Analytics Knowledge Group builds the ability to use appropriate statistical and data analytics techniques on available data to deliver insights and discover information, providing recommendations, and supporting decision-making. It includes Knowledge Areas that cover: data mining, supervised and unsupervised machine learning, statistical modelling, and predictive analytics.

The following are commonly defined Data Science Analytics Knowledge Areas:

- KA01.01 (DSDA/SMDA) 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/ML) Machine learning and related methods for information search, image recognition, decision support, classification;
- KA01.03 (DSDA/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/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/PA) Predictive analytics focuses on application of statistical models for predictive forecasting or classification;
- KA01.06 (DSDA/MODSIM) Computational modelling, simulation and optimisation.

5.3.1 DSDA/SMDA - Statistical methods and data analysis

Statistics and probability theory are foundational components of data analytics and constitute a significant part of a Data Science competences and knowledge. This module provides an insight into major statistical and data analytics paradigms and schools of thought. They can be taught separately or as a part of other Data Analytics related modules or courses.

Topics:

- Statistical paradigms (regression, time series, dimensionality, clusters)

- Probabilistic representations (causal networks, Bayesian analysis, Markov nets)
- Frequentist and Bayesian statistics
- Exploratory and confirmatory data analysis
- Information theory
- Graph theory

Learning Outcomes:

- Choose and execute standard methods from existing statistical libraries to provide overview (LODA.02 L1)
- Select most appropriate statistical techniques and model available data to deliver insights (LODA.02 L2)
- Identify requirements and develop analysis approaches (LODA.01 L2)
- Assess and optimize organization processes using statistical techniques and simulation (LODA.02 L3)

5.3.2 DSDA/ML – Machine Learning

Data Scientists have a wide range of ready machine learning libraries available. Nevertheless, they also need to go beyond simple application of algorithms to achieve expected results. New problems they face might require in depth understanding of theoretical underpinning of both simple and advanced algorithms. This module covers the use, analyze and design of machine learning algorithms.

Topics:

- Machine learning theory (supervised, unsupervised, reinforced learning, deep learning, kernel methods, Markov decision processes)
- Design and analysis of algorithms (graph algorithms, data structures design and analysis, online algorithms, bloom filters and hashing, MapReduce algorithms)
- Game theory and mechanism design
- Classification methods
- Ensemble methods
- Cross-validation

Learning Outcomes:

- Choose and execute existing analytic techniques and tools (LODA.01 L1)
- Identify requirements and develop analysis approaches (LODA.01 L2)
- Develop specialized analytics to enable agile decision-making and integrate them into organizational workflows (L0DA.05 L2)
- Design and evaluate analysis techniques and tools to discover new relations (LODA.01 L3)

5.3.3 DSDA/DM - Data Mining

Mathematical and theoretical aspects of data analytics must be implemented in a computational form appropriate for both problem at hand and data size. This module builds familiarity with most relevant data mining algorithms and related methods for knowledge representation and reasoning.

Topics:

- Data mining and knowledge discovery
- Knowledge Representation and Reasoning
- CRISP-DM and data mining stages
- Anomaly Detection
- Time series analysis
- Feature selection, Apriori algorithm
- Graph data analytics

Learning Outcomes:

- Choose and execute standard methods from statistical libraries to provide overview (LODA.02 L1)
- Select most appropriate statistical techniques and model available data to deliver insights (LODA.02 L2)
- Analyze available data sources and develop tool that work with complex datasets (LODA.03 L2)

- Develop specialized analytics to enable agile decision-making and integrate them into organizational workflows (LODA.05 L2)
- Evaluate and recommend data analytics w.r.t. organizational strategy (LODA.05 L3)

5.3.4 DSDA/TDM - Text Data Mining

Text data mining can be considered a subset of data mining, but it is worth a separate consideration due to the amount of text data available and particular methods developed over the years to analyze it.

Topics

- Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data
- Data mining and text analytics
- Natural Language Processing
- Predictive Models for Text
- Retrieval and Clustering of Documents
- Information Extraction
- Sentiments analysis

Learning outcomes

- Choose and execute standard methods from statistical libraries to provide overview (LODA.02 L1)
- Analyze available data sources and develop tool that work with complex datasets (LODA.03 L2)
- Evaluate and recommend data analytics w.r.t. organizational strategy (LODA.05 L3)

5.3.5 DSDA/PA - Predictive Analytics

Predictive analytics are a commonly used to foresee future events in order to avoid them or act ahead. This module covers both traditional approaches based on time series and newer approaches based on deep learning. Anomaly detection is a particular focus since it is one of most common application areas.

Topics

- Predictive modeling and analytics
- Inferential and predictive statistics
- Machine Learning for predictive analytics
- Regression and Multi Analysis
- Generalised linear models
- Time series analysis and forecasting
- Deploying and refining predictive models

Learning outcomes

- Choose and execute existing analytic techniques and tools (LODA.01 L1)
- Identify requirements and develop analysis approaches (LODA.01 L2)
- Create stories and optimize visualizations to influence executive decisions (LODA.06 L3)

5.3.6 DSDA/MODSIM - Modelling, simulation and optimization

Modeling and simulation are essential approaches to handle complexity of some systems and event chains. This module provides an introduction in both theoretical and practical aspects of model development and simulation techniques.

Topics:

- Modelling and simulation theory and techniques (general and domain oriented)
- Operations research and optimisation
- Large scale modelling and simulation systems
- Network oprtimisation
- Risk simulation and queuing

Learning Outcomes:

- Describe and execute different performance and accuracy metrics (LODA.04 L1)
- Compare and choose performance and accuracy metrics (LODA.04 L2)
- Assess and optimize organization processes using statistical techniques and simulation (LODA.02 L3)

5.4 Data Science Engineering (KAG2-DSENG)

Data Science Engineering Knowledge Group builds the ability to use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management. It includes Knowledge Areas that cover: software and infrastructure engineering, manipulating and analysing complex, high- volume, high- dimensionality data, structured and unstructured data, Cloud based data storage and data management.

Data Science Engineering includes software development, infrastructure operations, and algorithms design with the goal to support Big Data and Data Science applications in and outside the Cloud. The following are commonly defined Data Science Engineering Knowledge Areas:

- KA02.01 (DSENG/BDI) Big Data infrastructure and technologies, including NOSQL databased, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approached to big data analysis and common MapReduce algorithms;
- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with focus on data warehouse and data centers.

5.4.1 DSENG/BDI - Big Data infrastructure and technologies

Big data infrastructures and technologies drive many of the Data Science applications. Systems and platforms behind big data differ significantly from traditional ones due to specific challenges of volume, velocity, and variety of data. This module addresses these aspects with focus on underlying storage technologies and distributed architectures.

Topics:

- Big Data Cloud platforms (Azure, AWS)
- Approaches to data ingestion at scale
- Parallel and distributed computer architectures (Cloud Computing, client/server, grid)
- Large scale storage systems, SQL and NoSQL databases
- Computer networks architectures and protocols
- Storage for big data infrastructures and high-performance computing (HDFS, Ceph)

Learning Outcomes:

- Find possible data storage and processing solutions including both traditional and NOSQL databases (LOENG.06 L1)
- Survey various specialized data-driven tools and identify the best option (LOENG.03 L2)
- Evaluate the difference in performance between various distribute and Cloud-based platforms and recommend a solution (LOENG.01 L3)

5.4.2 DSENG/DSIAPP - Infrastructure and platforms for Data Science applications

Deployment of Data Science applications is usually tied to one of most common platforms, such as Hadoop or Spark, hosted either on private or public Cloud. The application must be also tied to a whole data processing pipeline including ingestion and storage. This module covers these aspects with additional focus on handling most common types of data inputs at scale.

Topics:

- Big data frameworks (Hadoop, Spark, HortonWorks, others)
- Big data infrastructures (ingestion, storage, streaming, enabling analytics, Lambda Architecture)
- Data processing models (batch, streaming, parallelism)
- Large-scale data storage and management (data inputs: graph, text, image, table, time series)

Learning Outcomes:

- Define technical requirements for new distributed and Cloud-based application for a given high-level design (LOENG.04 L1)
- Apply existing data-driven solutions to data analytic platform (LOENG.02 L2)
- Evaluate the difference in performance between various distribute and Cloud-based platforms and recommend a solution (LOENG.04 L3)

5.4.3 DSENG/CCT - Cloud Computing technologies for Big Data and Data Analytics

Cloud Computing technologies are a most common way to deploy Big Data and Data Analytics applications. This module provides an introduction to various levels of Cloud Computing services, such as IaaS or PaaS on practical examples. It is also important to consider both private and public Cloud.

Topics

- Cloud Computing architecture and services
- Cloud Computing engineering (design, management, operation)
- Cloud-enabled applications development (IaaS, PaaS, SaaS, autoscaling)
- Capex vs Opex consideration

Learning outcomes

- Choose potential technologies to implement new applications for data collection and storage (LOENG.01 L1)
- Model a problem to apply distributed and Cloud-based platforms (LOENG.04 L2)
- Evaluate the difference in performance between various distribute and Cloud-based platforms and recommend a solution (LOENG.04 L3)

5.4.4 DSENG/SEC - Data and Applications security

Data Scientists should have a general understanding of data and application security aspects in order to properly plan and execute data-driven processing in the organization. This module provides an overview of the most important aspects, including sometime omitted concepts of accountability, compliance and certification.

Topics

- Data security, accountability, protection
- Blockchain
- Access control and Identity management
- Compliance and certification
- Data anonymization and privacy

Learning outcomes

- Identify security issues related to reliable data access (LOENG.05 L1)
- Analyze security threats and solve them using known techniques (LOENG.05 L2)

5.4.5 DSENG/BDSE - Big Data systems organization and engineering

Systems and platforms behind big data differ significantly from traditional ones due to specific challenges of volume, velocity, and variety of data. They require specialized approaches to data processing and algorithm engineering. This module addresses these aspects both in general and based on common MapReduce algorithms.

Topics

- Big data frameworks (Hadoop, Spark, HortonWorks, others)
- Algorithms for large scale data processing

- Methods for pre-processing data implemented in MapReduce, including problems of correct data spliting in clusters
- Approaches to Big Data analysis (Functional abstraction for data processing, MapReduce, Lambda Architecture)
- Algorithms for visualization of large data sets, including subsampling with different distributions
- Big Data systems for applications domains

Learning outcomes

- Choose potential technologies to implement new applications for data collection and storage (LOENG.01 L1)
- Find possible data storage and processing solutions including both traditional and NOSQL databases (LOENG.06 L1)
- Model data-driven application following engineering principles (LOENG.01 L2)
- Adapt and optimize existing data-driven solutions to better fit to a given data analytics platform (LOENG.02 L3)

5.4.6 DSENG/DSAPPD - Data Science (Big Data) application design

Data Scientists are often tasked with developing new applications and systems. Certain languages and tools are more suitable in a data scientific context than other. This module covers most common languages for data science and big data processing together with most common tools for data presentation.

Topics:

- Languages for big data (Python, R)
- Tools and models for data presentation and visualization (Jupyter, Zeppelin)
- Software requirements and design
- Software engineering models and methods
- Software quality assurance
- Agile development methods, platforms, tools
- DevOps and continuous deployment and improvement paradigm

Learning Outcomes:

- Identify a set of potential data analytics tools to fit specification (LOENG.03 L1)
- Define technical requirements for new distributed and Cloud-based application for a given high-level design (LOENG.06 L1)
- Model data-driven application following engineering principles (LOENG.01 L2)
- Apply existing techniques to develop new data analytics applications (LOENG.02 L2)
- Combine several techniques and optimize them to design new data analytic applications (LOENG.06 L3)

5.4.7 DSENG/IS - Information Systems

All organizations relay on some form of Information Systems to preserve knowledge and drive decision processes. This module focuses on basics of well-established data warehouse, expert systems and decision support systems. Big data influence on such systems is also of interest, but related technical details are covered by other KAs.

Topics:

- Decision support systems
- Data warehousing and expert systems
- Enterprise information systems (data centers, intra/extra-net)
- Multimedia information systems

Learning Outcomes:

- Identify a set of potential data-driven tools to fit specification (LOENG.03 L1)
- Model the problem to apply traditional or NOSQL database technology (LOENG.06 L2)
- Evaluate and recommend optimal data-driven tools to influence decision making (LOENG.03 L3)

6 Data Science Professional Profiles

This section presents the proposed definition of the Data Science Professional Profiles that includes also data handling and data infrastructure related profiles or occupations. The proposed profiles are defined in accordance with the ESCO (European Skills, Competences, Qualifications and Occupations) occupations taxonomy [11]. The proposed new occupations are placed in four top classification groups: Managers (for managerial roles); Professionals (for applications developers, for infrastructure engineers, and for data handling professionals); Technicians and associate professionals (for operators and technicians); and Clerical support workers (for data entry and field data collection workers).

6.1 Taxonomy of Data Science Occupations according to ESCO Hierarchy

The presented here taxonomy of Data Science Professional profiles and roles is based on the ESCO occupations classification, while their competences and organisational roles are defined similar to CWA 16458 ICT profiles.

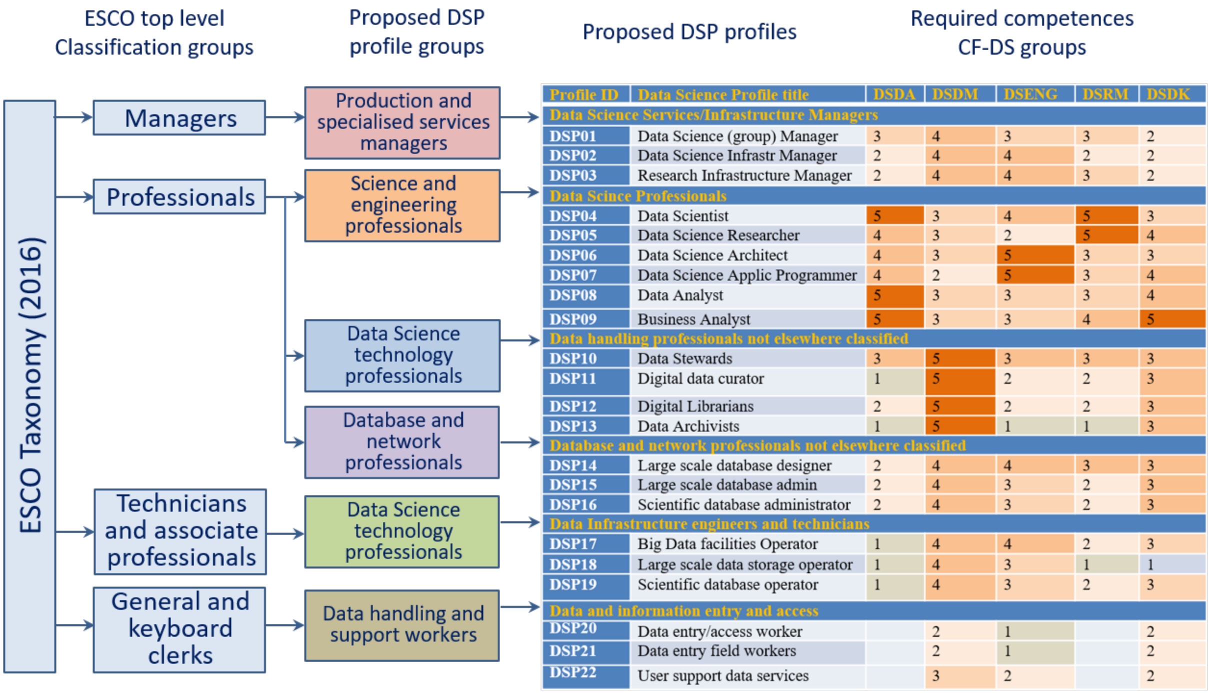
The following suggestions were used when constructing the proposed taxonomy:

- Data Science occupations depending on organisational role can be placed in the following top-level hierarchies:
    - Managers (for managerial roles);
    - Professionals (for analytics applications developers, for data handling professions, and for infrastructure and datacenter engineers);
    - Technicians and associate professionals (for operators and technicians)
- Correspondingly, new 3rd level occupation groups are proposed:
    - Data Science/Big Data Infrastructure Managers
    - Data Science Professionals
    - Data Science technology professionals
    - Data and information entry and access
- Group of occupations related to data stewardship, data curation, data libraries, data archives management are placed in the separate 4th level group
    - "*Professionals > Information and communications technology professionals > Data Science technology professionals > Data handling professionals not elsewhere classified*"
- Due to specifics of working with data, the new 2nd level group has been proposed
    - "*Clerical support workers > Data handling support workers (alternative)*"
    - Motivation for this is growing need for data support workers in all domains of human activities in the digital data driven economy.
- It is recognised that existing ESCO group "Database and network professionals" should be extended with new occupations (or professions) related to Big Data and specifically scientific data related profiles which examples are included in the table: *Large scale (cloud) database administrator/operator and Scientific database administrator/operator*, however further identification of such occupations need to be done.

Figure 6.1 graphically illustrates the existing ESCO hierarchy and the proposed new Data Science classification groups and corresponding new Data Science related profiles. The table in the figure illustrates mapping of different CF-DS competence groups to individual profiles where colour indicates relevance. Figure 8 provides visual presentation of the identified DSP profiles and their grouping by the proposed high-level classification groups.

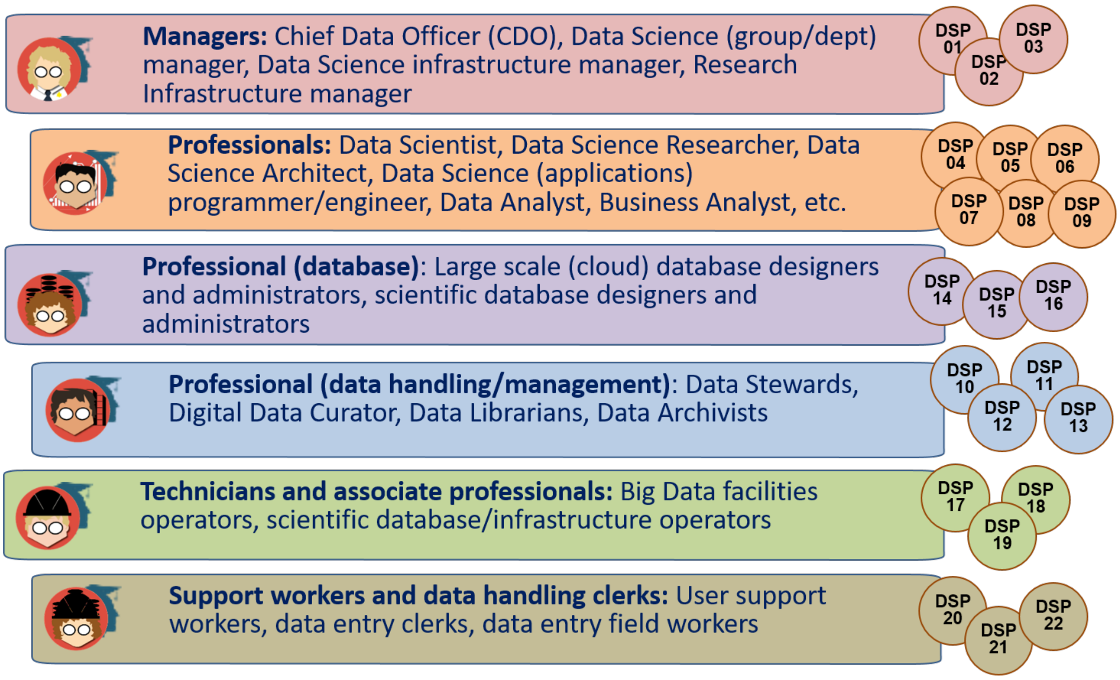
Table 4 provides an initial definition of the identified Data Science professional profiles collected from job advertisements, blogs and recent discussions at different forums, in particular, with the Research Data Alliance, and digital curation and data preservations communities.

For details refer to a separate DSPP document [4] the provides further details on the definition of the proposed Data Science professional profiles and suggests mapping between professional profiles and Data Science competence groups as they are defined in CF-DS [1] including the suggested ranking of the relevance of different competence groups to corresponding Data Science profiles.


ESCO top level Classification groups
Proposed DSP profile groups
Proposed DSP profiles
Required competences CF-DS groups
ESCO Taxonomy (2016)
Managers
Professionals
Technicians and associate professionals
General and keyboard clerks
Production and specialised services managers
Science and engineering professionals
Data Science technology professionals
Database and network professionals
Data Science technology professionals
Data handling and support workers

Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	5	3	4	5	3
DSP05	Data Science Researcher	4	3	2	5	4
DSP06	Data Science Architect	4	3	5	3	3
DSP07	Data Science Applic Programmer	4	2	5	3	4
DSP08	Data Analyst	5	3	3	3	4
DSP09	Business Analyst	5	3	3	4	5
Data handling professionals not elsewhere classified						
DSP10	Data Stewards	3	5	3	3	3
DSP11	Digital data curator	1	5	2	2	3
DSP12	Digital Librarians	2	5	2	2	3
DSP13	Data Archivists	1	5	1	1	3
Database and network professionals not elsewhere classified						
DSP14	Large scale database designer	2	4	4	3	3
DSP15	Large scale database admin	2	4	3	2	3
DSP16	Scientific database administrator	2	4	3	2	3
Data Infrastructure engineers and technicians						
DSP17	Big Data facilities Operator	1	4	4	2	3
DSP18	Large scale data storage operator	1	4	3	1	1
DSP19	Scientific database operator	1	4	3	2	3
Data and information entry and access						
DSP20	Data entry/access worker		2	1		2
DSP21	Data entry field workers		2	1		2
DSP22	User support data services		3	2		2

Figure 6.1. Proposed Data Science related extensions to the ESCO classification hierarchy and corresponding DSP profiles by classification groups (relevance of CF-DS competence groups to individual profiles is indicated by rank and colour: higher rank and darker colour means higher relevance).


Managers: Chief Data Officer (CDO), Data Science (group/dept) manager, Data Science infrastructure manager, Research Infrastructure manager

Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.

Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators

Professional (data handling/management): Data Stewards, Digital Data Curator, Data Librarians, Data Archivists

Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators

Support workers and data handling clerks: User support workers, data entry clerks, data entry field workers

Figure 6.2. Data Science Professional profiles and their grouping by the proposed new professional groups compliant with the ESCO taxonomy.

6.2 Definition of the Data Science Professional profiles

Table 6.1 provides an initial definition of the identified Data Science professional profiles collected from job advertisements, blogs and recent discussions at different forums, in particular, with the Research Data Alliance, and digital curation and data preservations communities.

For details refer to a separate DSPP document [4] the provides further details on the definition of the proposed Data Science professional profiles and suggests mapping between professional profiles and Data Science competence groups as they are defined in CF-DS [1] including the suggested ranking of the relevance of different competence groups to corresponding Data Science profiles.

Table 6.1. Data Science professional profiles definition

Profile ID	Data Science Profile title	Data Science Profile Summary statement	*Alternative titles and legacy titles*
Managers			
DSP01	Data Science (group) Manager	Proposes, plans and manages functional and technical evolutions of the data science operations within the relevant domain (technical, research, business).	Data analytics department manager
DSP02	Data Science Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the big data infrastructure within the relevant domain (technical, research, business).	Big Data Infrastructure Manager
DSP03	Research Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the research infrastructure within the relevant scientific domain.	Research Infrastructure data storage facilities manager
Professionals			
DSP04	Data Scientist	Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data.	Data Analyst
DSP05	Data Science Researcher	Data Science Researcher applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to scientific problem, business process, or reveal hidden relations between multiple processes.	Data Analyst
DSP06	Data Science Architect	Designs and maintains the architecture of Data Science applications and facilities. Creates relevant data models and processes workflows.	System Architect, Applications architect
DSP07	Data Science (Application) Programmer/Engineer	Designs/develops/codes large data (science) analytics applications to support scientific or enterprise/business processes.	Scientific Programmer
DSP08	Data Analyst	Analyses large variety of data to extract information about system, service or organisation performance and present them in usable/actionable form	
DSP09	Business Analyst	Analyses large variety of data Information System for improving business performance.	Business Development Manager (Data science role)

Professional (data handling/management)			
DSP10	Data Stewards	Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers	
DSP11	Digital data curators	Finds, selects, organises, shares (exhibits) digital data collections, maintains their integrity, up-to-date status and freshness, discoverability	Digital curator, digital archivist, digital librarian
DSP12	Data Librarians	Data librarians perform or support one or more of the following: acquisition (collection development), organization (cataloguing and metadata), and the implementation of appropriate user services. Data librarians apply traditional librarianship principles and practices to data management, including data citation, digital object identifiers (DOIs), ethics and metadata.	Digital data curator
DSP13	Data Archivists	Maintain historically significant collections of datasets, documents and records, other electronic data, and seek out new items for archiving.	Digital Archivists
Professional (database)			
DSP14	Large scale (cloud) database designer	Designs/develops/codes large scale data bases and their use in domain/subject specific applications according to the customer needs.	Large scale (cloud) database developer
DSP15	Large scale (cloud) database administrator	Designs and implements, or monitors and maintains large scale cloud databases	
DSP16	Scientific database administrator	Designs and implements, or monitors and maintains large scale scientific databases	Large scale (cloud) database administrator
Technicians and associate professionals			
DSP17	Big Data facilities Operator	Manages daily operation of facilities, resources, and responds to customer requests. Includes all operations related to data management and data lifecycle	
DSP18	Large scale (cloud) data storage operator	Manages daily operation of cloud storage, including related to data lifecycle, and responds to requests from storage users	
DSP19	Scientific database operator	Manages daily operation of scientific databases, including related to data lifecycle, and responds to requests from database users	Large scale (cloud) data storage operators
Clerical and support workers (general and keyboard workers)			
DSP20	Data entry/access worker	Enter data into data management systems directly reading them from source, documents or obtained from people/users	Data entry desk/terminal worker
DSP21	Data entry field workers	The same work done on field when collecting data from disconnected sensors or doing direct counting or reading	
DSP22	User support data services	Provides support to users to entry their data into governmental service and user facing applications	

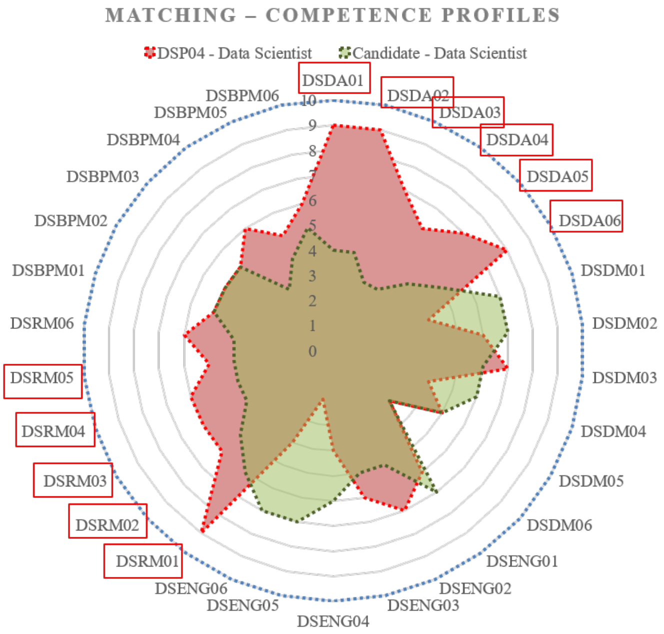
7 Examples practical use of the Data Science Professional profiles

The proposed EDSF provides a basis for multiple practical uses that include but not limited to:

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

7.1 Usage example: Competences assessment

Figure 7.1 illustrates example of the individual competences assessment that maybe used for one of the general use cases: the Data Science practitioner competences assessment against the target/desirable competence profile or role; or competences matching between the job vacancy and the candidate's competence profile.



Individual Education/Training Path based on Competence benchmarking

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

Figure 7.1. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

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

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

7.2 Data Science Team composition

Data Science team composition and competences matching is one of intended uses of the EDSF and DSPP in particular. Figure 7.2 illustrates a case of creating a Data Science team or group for an average size of the research organisation with affiliated number of researchers 200-300, what would require a Data Science team of 10-15 members whose responsibility would include supporting all main stages of data lifecycle: data collection, data input/ingest, data analysis, reporting, visualisation and storage. The figure also illustrates possible roles that may be assigned to perform different functions at different data workflow stages

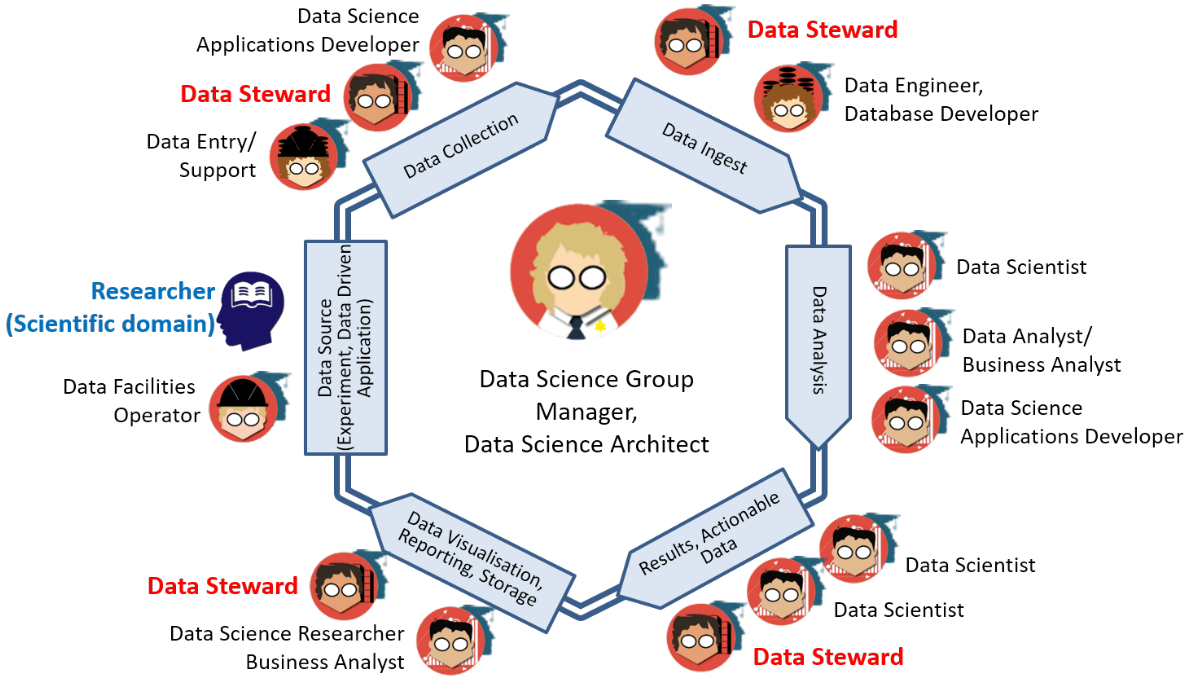


Figure 7.2. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

To support all data related research or production stages the following roles may be required (including suggested staffing for the team of 10-12 members):

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application architect/developer/programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- Data stewards, curators, archivists (3-5)

It is possible that some of the above roles can be re-defined and re-allocated to the Data Science team from the previous ICT and IT infrastructure groups or departments. In this case some basic Data Science training will be required for not initially data related professions.

It also suggested a distinct role of the Data Steward, a new emerging role for data driven research organisations and projects. Data Steward should play a bridging role between the subject domain researcher and the Data Science team or Data Scientist in particular cases to help to translate between subject domain and Data Science or data analytics domain. Data Stewards can have both backgrounds either ICT and computer or digital curation/librarian.

8 EDSF Adoption and Validation

This section provides information about mechanisms used by the project to ensure correctness and consistency of the proposed EDSF and its components. Short reference is also provided to examples of the EDSF adoption by universities, research community and industry. The section also provides summary of the community survey undertaken by the project to validate the proposed Data Science Competence as a core EDSF component.

8.1 Community review and contribution

Since the initial version of the Data Science Competence Framework defined in November 2015, the CF-DS and other documents' discussion and validation has been done at multiple workshops and events organised by the project or contributed by the project members. A designated role belongs to the EDISON Liaisons Groups (ELG) to which the main versions of CF-DS and other documents have been presented for review and comments. Formally, EDSF development was reported in deliverable D2.1 (M6, February 2016) and D3.1 (M9, May 2016). EDSF documents are published for public access and comments on the project website under the Creative Common Attribution (CC BY) license 4.0. This provided a basis for wide community contribution and ongoing adoption of EDSF from the early project stage.

The following are the main community involvement and contribution mechanisms and events where the EDSF has been presented and discussed:

1) EDSF progress and working draft have been presented to ELG and discussed at ELG meetings, starting from the first ELG01 meeting on 13 November 2016, and following ELG meeting in April 2016, September 2016, and the final ELG04 meeting on 29 June 2017 that has reviewed the final EDSF release presented in this deliverable.
2) EDISON Workshop on Data Infrastructure Competences and Skills Framework: a European and Global Challenge, 9th February, 2016, Brussels. The workshop gathered representatives from H2020 Research Infrastructure projects and other European stakeholders involved into building Data Infrastructure and dealing with building human capacities in Data infrastructure and Data Science. The proposed CF-DS was presented and discussed at the workshop. Feedback and comments have been incorporated into new CF-DS version and reported in Deliverable D2.1.
3) Admitting the fact that CF-DS development inspired by the e-CF approach and recommendations, the CF-DS and EDSF development has been presented regularly to the CEN e-CF workshop starting from 9th December 2015 workshop (AFNOR, Paris) and following workshop of which the last was on 4th May 2017 in Rome. The workshop confirmed that the EDSF approach is compliant with e-CF and will simplify possible formal adoption of CF-DS as an extension to e-CF3.0 or as a separate profile.
4) The EDEF development results were regularly presented at the RDA Interest Group on Education and Training on Research Data Handling (IF-ETHRD) that took place at biannual sequence. This allowed to collect feedback and contribution from wide international community and digital and data librarians' community in particular.

However, the most important and valuable contribution has been received from the Champion Universities who cooperated actively with the project. First of all, these are project partner universities: University of Amsterdam (that has 4 independent Data Science programmes all of which were influenced by EDSF), University of Stavanger and University of Southampton who are in the process of establishing new Data Science programmes. Other universities actively cooperating with EDISON on defining their Data Science programmes include EDISON internal champions: University of Perugia, University of Bedfordshire, Goethe University Frankfurt, Lucerne School of Information Technology, Lodz University of Technology, The Engineering Training School "Enrico della Valle", as well as the National Technical University of Ukraine "Kiev Polytechnic Institute" (see deliverable D3.1 for champion universities involvement). EDISON organised 3 Champion Universities conferences in July 2016, March 2017 and June 2017 as a venue to exchange information, practices and experience between champion universities and wider academic, research and industry community. The conferences and discussion with university practitioners provided valuable information for the DS-BoK and MC-DS development and maturity.

8.2 Community surveys on Data Science competences definition

8.2.1 Objectives and communities

The purpose of the EDISON survey was to identify the important competences related to Data Science profession and related knowledge and skills needed to proficiently perform the expected job and related roles.

The two rounds of the survey targeted European RI and the major stakeholder groups:

- Researchers and Scientists, from any discipline/market sector;
- Employers or Recruiters, from any discipline/market sector;
- Professional Data Scientists;

8.2.2 Survey results and CF-DS validation

The survey results in general confirmed the proposed CF-DS competences structure and ranked the importance of individual competences for the Research Infrastructures (RI) and community of practitioners. The survey was organised in two stages: first survey was proposed to EGI and associated RIs to identify need and relevance of Data Science competences and skills to organisational need; second stage was focused on defining how the RI and Data Science practitioners community ranks importance of the proposed Data Science competences and skills.

In general, the survey confirmed that a reference framework with a proper definition of data science and the required skills and competences is crucial to promote awareness about the Data Science profession and relevance of each competence group. The following are specific outcomes and recommendations obtained from the survey (detailed data are provided in Appendix B):

- The Data Scientist role must have a distinct set of functions and responsibilities that are primarily focused on the technical processing of data (analysis, visualization, apply Big Data techniques, operating and managing a data infrastructure). Majority admitted that Data Scientist occupies high level role reporting directly to middle level management.
- The majority of respondents indicated the Data Scientist must have a University degree (Bachelor/Engineer/Master), some respondents indicated a preference for a workplace training.
- Workplace training is considered as an important component of Data Scientist's skill management and it should be focused on new technologies.
- The following is the result of ranking CF-DS competences in each competence group:
    - (1) Data Science Analytics: Importance of using different statistics and data analytics techniques for complex datasets, with stress on statistical techniques; development of the specialised tools and visualisation were ranked with less importance.
    - (2) Data Science Engineering: Respondents admitted importance of more practical Data Science Engineering competences and skills, and less importance of the core computer and engineering knowledge and competences.
    - (3) Data management and data curation: Technical aspects of data management were admitted as more important than Data Management Plan and IPR related issues.
    - (4) Research Methods: Knowledge and ability to apply of Research Methods is treated as important but not critical with responses almost equally divided between relevance 3 and (4-5)
    - (5) Research Infrastructures: Knowledge about European RIs and their operation and policy are treat as average relevance.
    - (6) Communication and interdisciplinary competences: Inter-personal and interdisciplinary communication was admitted as important, however needs to understand organisational processes and have inter-professional knowledge was treated as less important.

In conclusion, the survey confirmed general Data Science competences structure and has been used to improve the CF-DS definition. It can be used for developing further actions to promote Data Science competences to surveyed community.

Appendix B provides details on the distribution of responses for each of individual competences.

9 Conclusion

This deliverable presents the final versions of the EDISON Data Science Framework that comprise of the four inter-related documents: Data Science Competence Framework, Data Science Body of Knowledge, Data Science Model Curriculum, and Data Science Professional Profiles.

The presented EDSF version is published as EDSF Release 2 on the project website and available for public use under CC BY 4.0 license.

The focused work on defining the Data Science Competence Framework as a core part of EDSF has been done with wide consultation and engagement of different stakeholders, primarily from research community and Research Infrastructures, but also involving industry via standardisation bodies, professional communities and directly via the project network.

The project established constructive cooperation with many e-Infrastructure projects and involved them into collaboration in addressing competences and skills gap for data related professions for European Research Area and European Digital Single Market.

The presented EDSF Release 2 fulfils the project workplan by month M22 and provides valuable contribution to defining common European Data Science competence framework and effective model and approach to address growing demand for Data Science and data related competences and skills by European Digital Singe Market (DSM), European Open Science Cloud (EOSC) and other stakeholders and actors in the emerging digital data driven economy.

The sustainability of the proposed EDSF after the project ends will be ensured by cooperation between the project partners, where the University of Amsterdam will maintain the future EDSF development and updates based on the community feedback and contribution.

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

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

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

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

[10] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

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

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

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

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

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

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

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

[18] 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

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

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
EGI	European Grid Initiative
ELG	EDISON Liaison Group
EOSC	European Open Science Cloud
ERA	European Research Area
ESCO	European Skills, Competences, Qualifications and Occupations
ICT	Information and Communication Technologies
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
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
RDA	Research Data Alliance
SWEBOK	Software Engineering Body of Knowledge

Appendix A. EDSF Document complementing this Deliverable

This document provides a short summary of the EDISON Data Science Framework Release 2 which is the product of the project development.

For details and full definition refer to the original documents that are published at the project website under CC BY 4.0 license to ensure maximum availability to the academic and professional community. The links below provides access to the recent version of published documents

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

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

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

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

Appendix B. Community Survey Summary

B.1. Objectives of the EDISON Survey and grouping of respondents

The purpose of the EDISON survey[4] is to analyse the Data Science (Data Scientist) demand side, that is the various potential employers of data scientists, in order to identify the important competences related to Data Science profession and related knowledge and skills needed to proficiently perform the expected job and related roles.

The first survey round (January – February 2016) collected information from researchers and educators in both academia and industrial environment including the job context, the career evolution and the role within the professional environment where graduated data scientist (will) work. A second survey round was conducted in November – December 2016, to further verify and detail the identified Data Science competences and skills requirements. In the second round of the survey, we have made some refinements based on the responses from the first survey and questions were adjusted to standards used in data collection in social sciences.

The survey helped to categorise the skills and competences by assigning knowledge level as provided by experts who are active in the field of Data Science. The quantification was done on the following subjects.

- Validate the list of skills/competences proposed for each domain by EDISON
- Identify the level [0-5][5] of knowledge for the proposed skills/competences
- Identify new skills/competences
- Identify respondents that will be interested to engage further with the project (interview, training, champion, certification)

B.1.1. Submission strategy and target groups

The survey aimed to capture all aspects needed to respond at the following questions:

- What are the common competences of all Data Scientists in any field of work (mainly public or private research, including eInfrastructures)?
- What are the specific competences that are required to a Data Scientist in each specific field of work (either discipline, eInfrastructure or market segment)?
- What are the career path(s) followed to become a Data Scientist?
- What are the specific competences requested by the employers for the Data Scientist profile and how these competences are valued/valuable?
- What are the trends in future Data Scientist positions?

Based on the respondents' occupation, the results were grouped by the main target groups (i.e. the Stakeholders):

- Researchers and Scientists, from any discipline/market sector;
- Employers or Recruiters, from any discipline/market sector;
- Professional Data Scientists;

B.1.2. The respondents

The first part of the EDISON survey aimed at collecting personal data needed to identify the background of the respondents:

- Respondents Affiliation
- Respondents Job Profile
- Diploma (PhD/MSc/Eng/Ba)
- Scientific background
- Respondents Expertise

[4] Online EDISON Survey (closed) https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

[5] Level of knowledge: (0 - no experience, 1- general knowledge about the topics, 2- general knowledge plus practical experience, 3- general knowledge plus practical experience plus practical experience 4 - advanced Theoretical knowledge, 5 - advanced Theoretical knowledge plus practical experience).

The respondents of the second survey were in majority (75%) of MSc level and higher, altogether they cover 13 different disciplines including humanities, social and behavioural sciences, journalism and information sciences, business and administration, life sciences, physical sciences, Mathematics and statistics, computer Science, manufacturing and construction engineering, and Health. In term of variety the respondents of the second survey were covered more scientific disciplines

Category	Respondents breakdown
Affiliation:	16.2% (6 respondents) National Research Center or Research Laboratory, 8.1% (3 respondents) European or International Research center, 40.5% (15 respondents) Higher Education Institute, 5.4% (2 respondents) Library, 13.5% (5 respondents) Small Medium, 13.5% (5 respondents) Large Industry (more than 250 employees).
Job profiles	• Data Science (group) Manager (3 respondents), • Research Infrastructure Manager (3 respondents), • Data Scientist (5 respondents), • Data Science Researcher (8 respondents) • Data Analyst (2 respondents), • Business Analyst (3 respondents), • Data Steward (1 respondent), • Digital Librarian (3 respondents), • Data Archivist (1 respondent), • User support data services (2 respondents), and • Scientific database administrator (1 respondent).
Diploma:	23 PhDs, 14MSc, 2 University degree
Education	humanities, social and behavioural sciences, journalism and information sciences, business and administration, life sciences, physical sciences, mathematics and statistics, computer Science, manufacturing and construction engineering, and Health.
Expertise:	• advanced theoretical knowledge in Scientific research methods, data management and Enterprise data infrastructure development • general knowledge plus practical experience digital library and archive, Data management systems, business process, • and only general knowledge about the Big Data software organisation and engineering, Business analysis and enterprise organisation

B.2. Survey data related to CF-DS competence groups

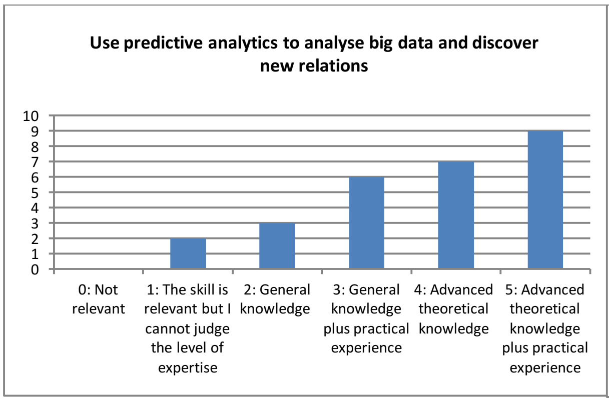
In this annex, you will find a summary of the ranking of the 39 competences and skill required for data Science. The skills are grouped in 7 competence categories: *(1) Data Analytics, (2) Data management and data curation, (3) Data Science Engineering, (4) Research Infrastructures, (5) Scientific and Research Methods, (6) Communication and interdisciplinary work* have been evaluated by the respondents by required level of expertise and knowledge[6].

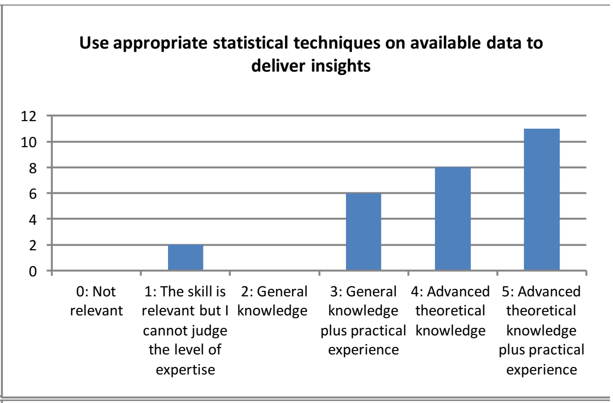
The figures presented below show the number of respondents (y-axes) and assessed required knowledge level (x-axes).

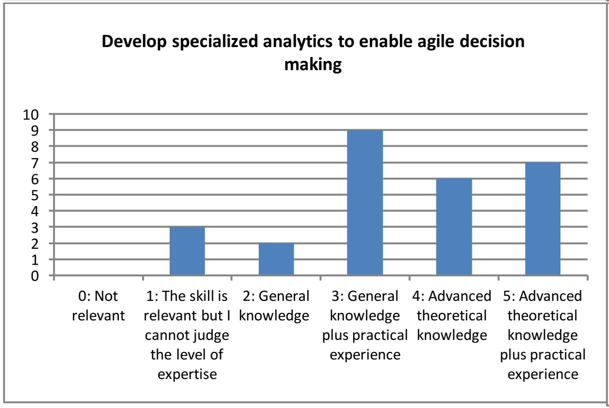
[6] Knowledge level are: 0 - no experience, 1- general knowledge about the topics, 2- general knowledge plus practical experience, 3- general knowledge plus practical experience plus practical experience 4 - advanced Theoretical knowledge, 5 - advanced Theoretical knowledge plus practical experience

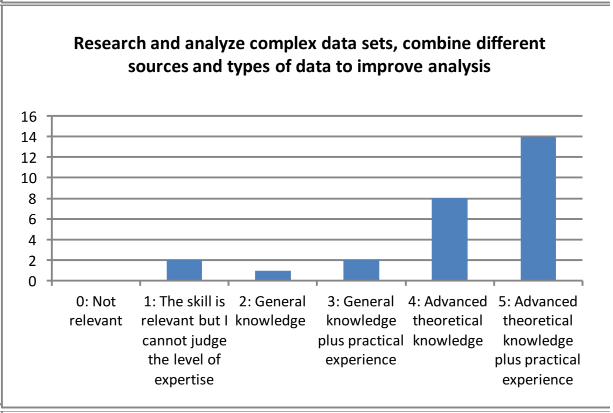
(1) Data Science Analytics competences/skills

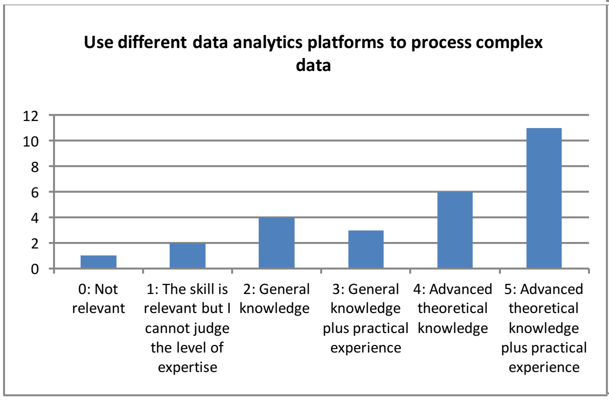
- Use predictive analytics to analyse big data and discover new relations
- Use appropriate statistical techniques on available data to deliver insights
- Develop specialized analytics to enable agile decision making
- Research/analyse complex data sets, combine sources and types of data to improve analysis
- Use different data analytics platforms to process complex data
- Visualise complex and variable data

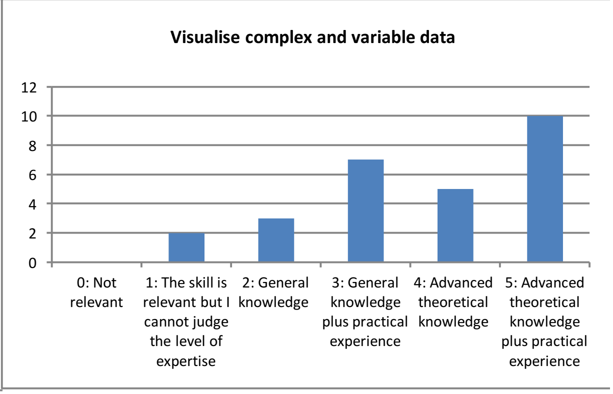






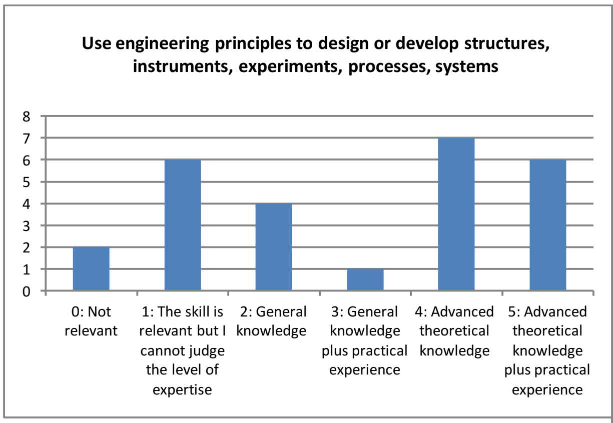


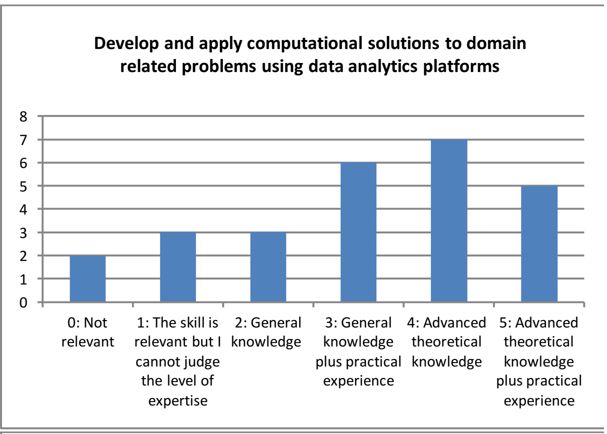


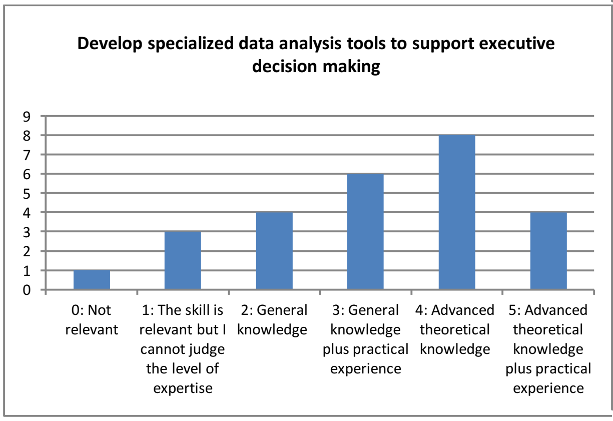


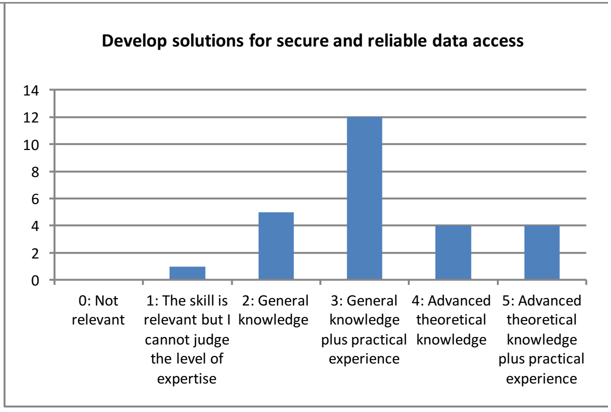
(2) Data Science Engineering competences/skills

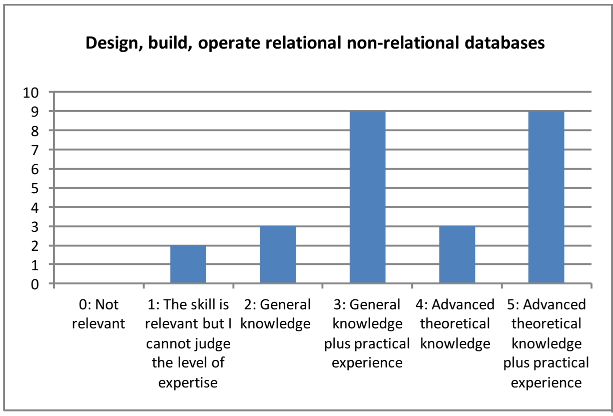
- Use engineering principles to design or develop structures, instruments, experiments, processes, systems
- Develop and apply computational solutions to domain related problems using data analytics platforms
- Develop specialized data analysis tools to support executive decision making
- Design, build, operate relational non-relational databases
- Develop solutions for secure and reliable data access
- Prototype new data analytics applications

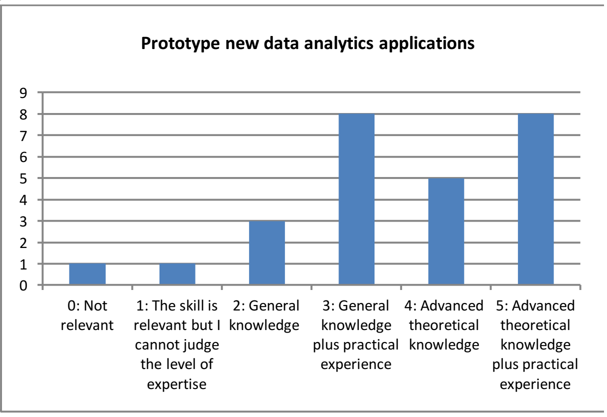

Use engineering principles to design or develop structures, instruments, experiments, processes, systems


Develop and apply computational solutions to domain related problems using data analytics platforms


Develop specialized data analysis tools to support executive decision making

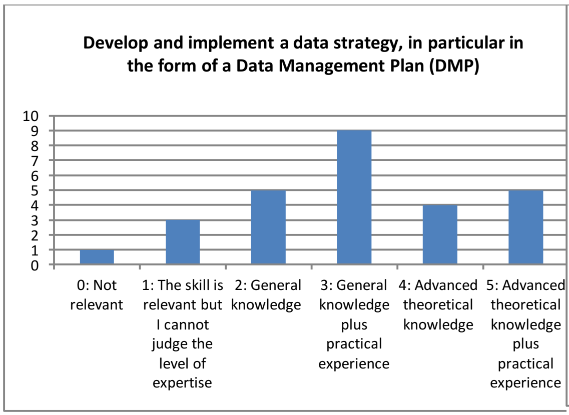

Develop solutions for secure and reliable data access

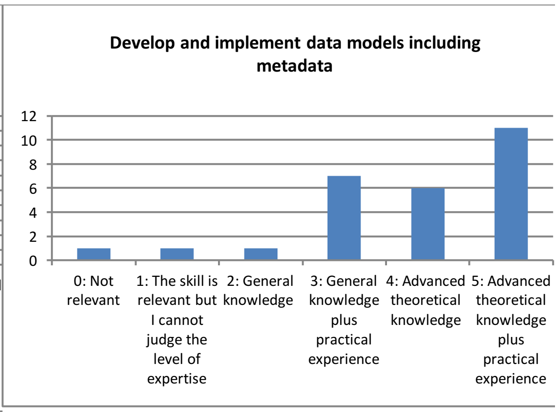

Design, build, operate relational non-relational databases

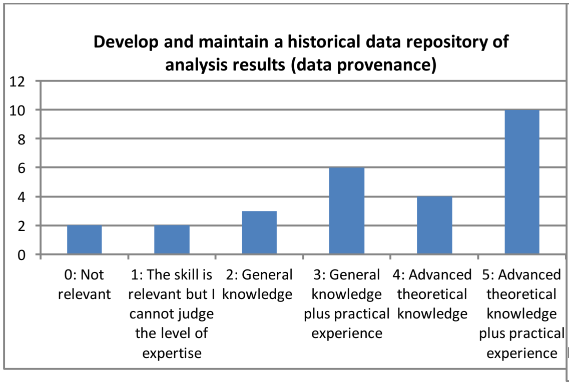

Prototype new data analytics applications

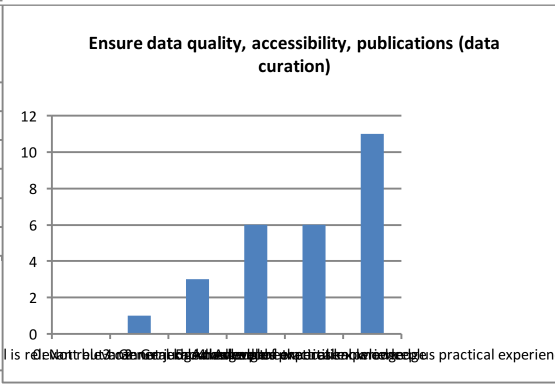
(3) Data management and data curation competences/skills

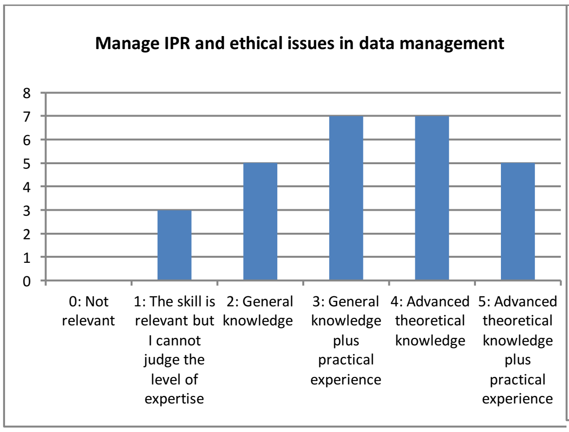
- Develop and implement a data strategy, in particular in the form of a Data Management Plan (DMP)
- Develop and implement data models including metadata
- Integrate different data sources and provide them for further analysis
- Develop and maintain a historical data repository of analysis results (data provenance)
- Ensure data quality, accessibility, publications (data curation)
- Manage IPR and ethical issues in data management

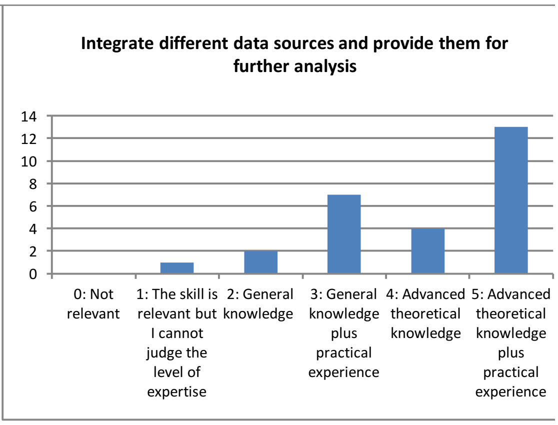

Develop and implement a data strategy, in particular in the form of a Data Management Plan (DMP)


Develop and implement data models including metadata


Develop and maintain a historical data repository of analysis results (data provenance)

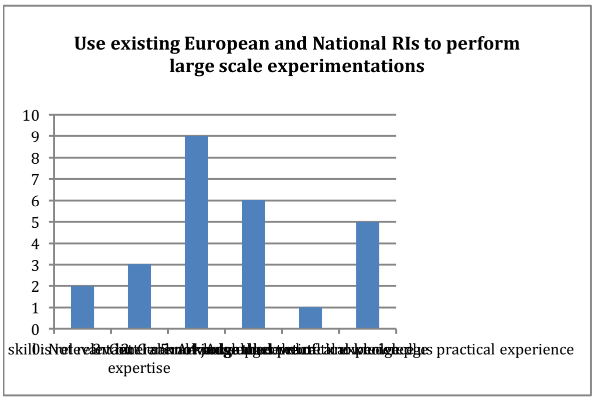

Ensure data quality, accessibility, publications (data curation)

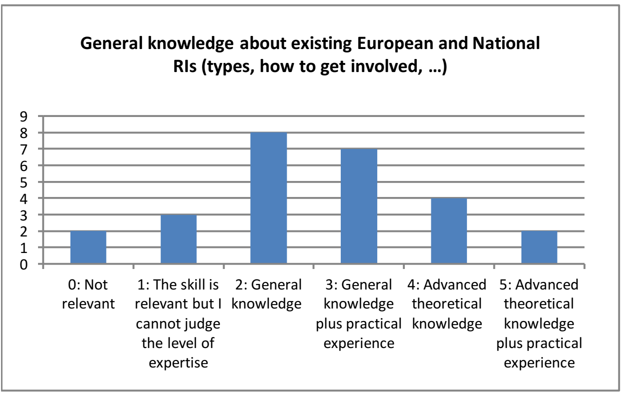

Manage IPR and ethical issues in data management

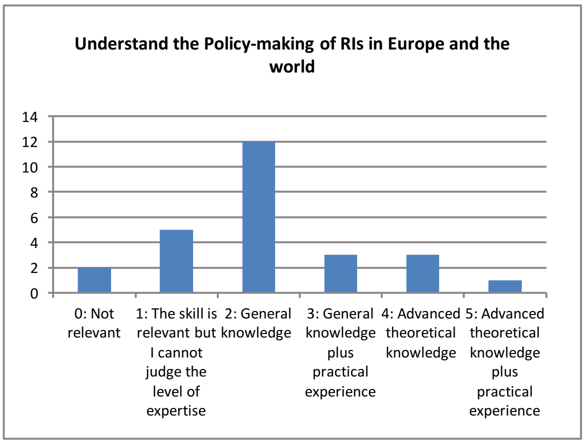

Integrate different data sources and provide them for further analysis

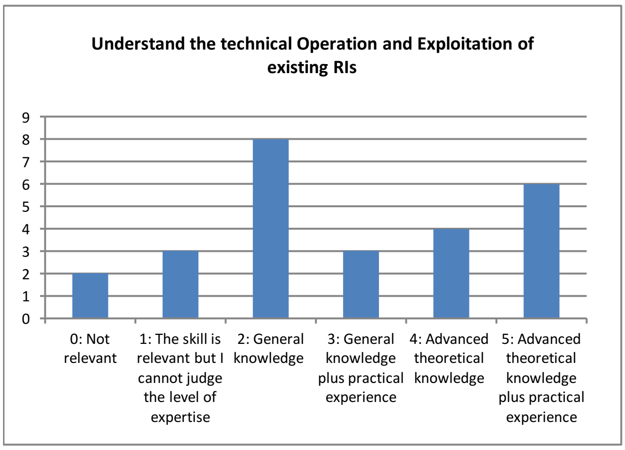
(4) Research Infrastructures competences/skills

- General knowledge about existing European and National RIs (types, how to get involved, …)
- Use existing European and National RIs to perform large scale experimentations
- Understand the technical Operation and Exploitation of existing RIs
- Understand the Policy-making of RIs in Europe and the world



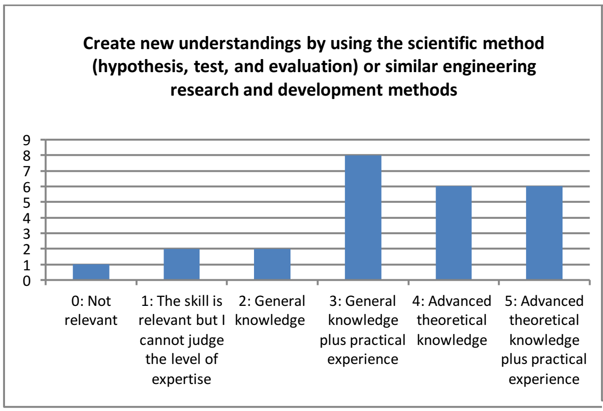




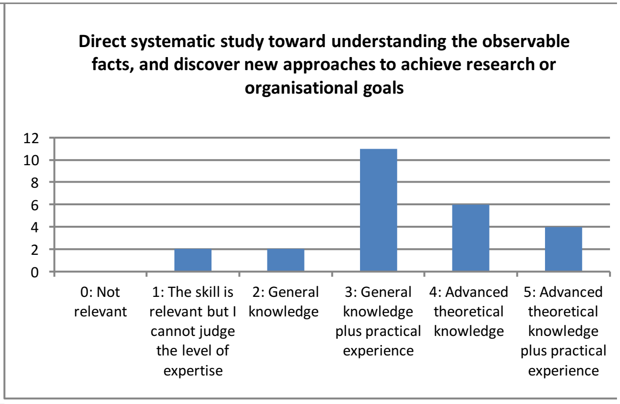


(5) Research Methods competences/skills

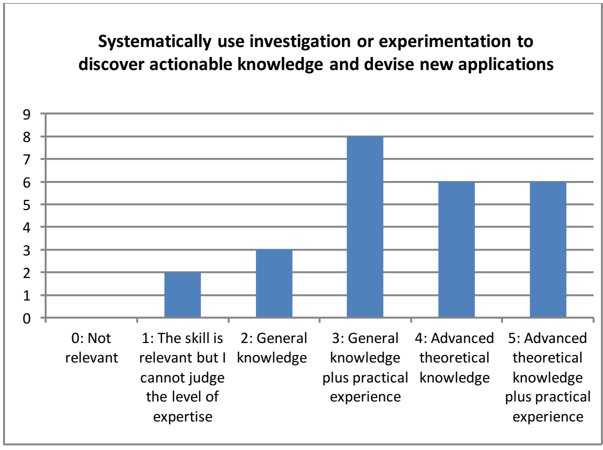
- Create new understandings by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods
- Direct systematic study toward understanding the observable facts, and discover new approaches to achieve research or organisational goals
- Systematically use investigation or experimentation to discover actionable knowledge and devise new applications
- Be able to translate strategies into action plans and follow through to completion
- Contribute to and influence the development of organizational objectives
- Apply ingenuity to complex problems, develop innovative ideas



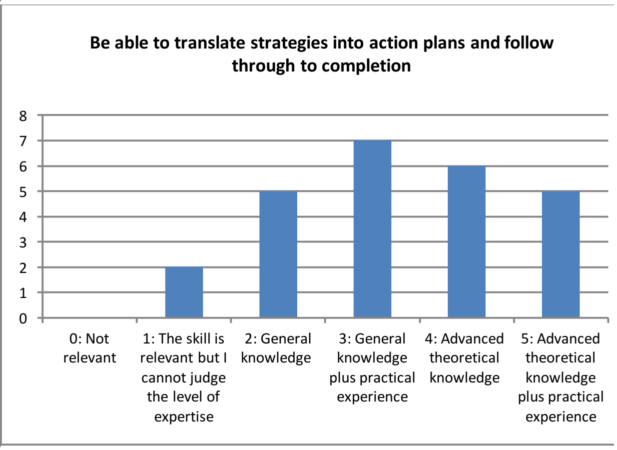
Create new understandings by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods



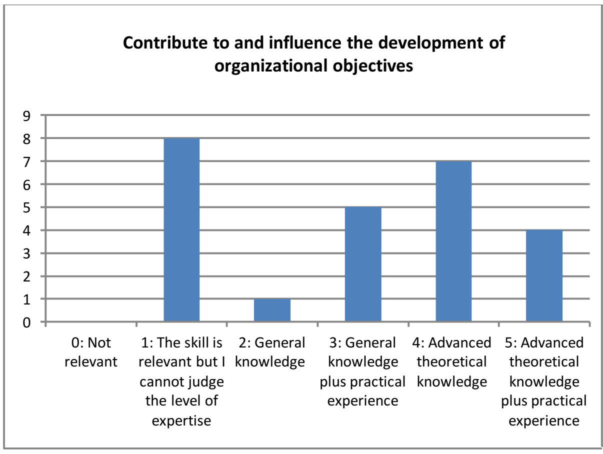
Direct systematic study toward understanding the observable facts, and discover new approaches to achieve research or organisational goals



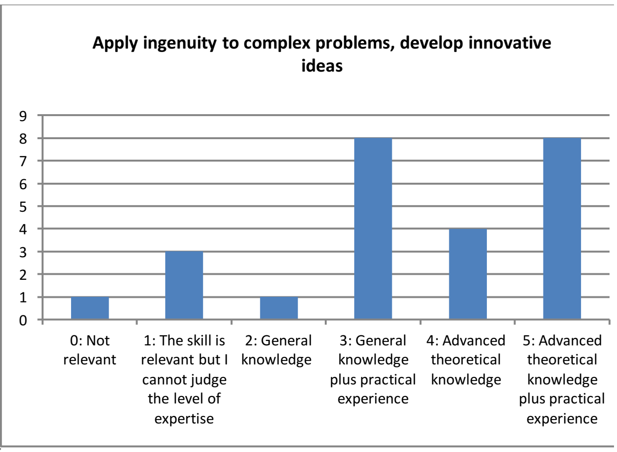
Systematically use investigation or experimentation to discover actionable knowledge and devise new applications



Be able to translate strategies into action plans and follow through to completion



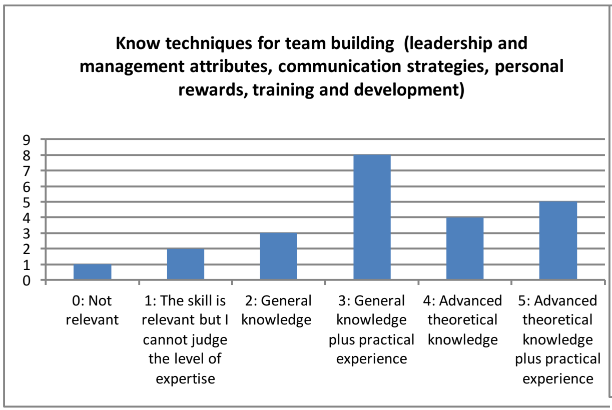
Contribute to and influence the development of organizational objectives

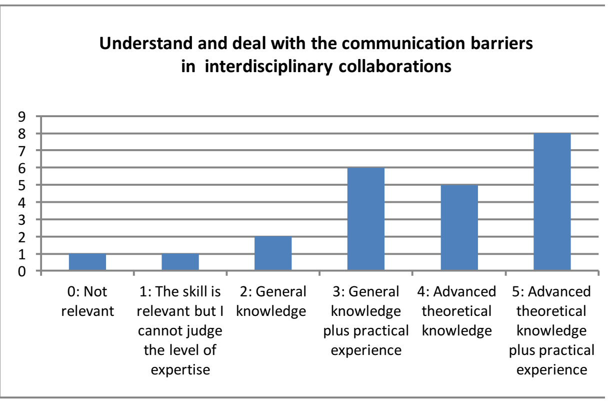


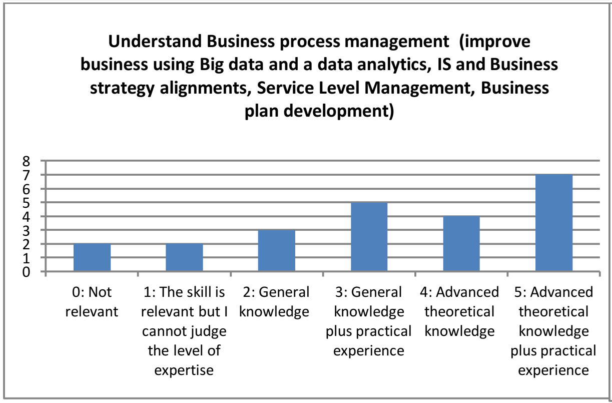
Apply ingenuity to complex problems, develop innovative ideas

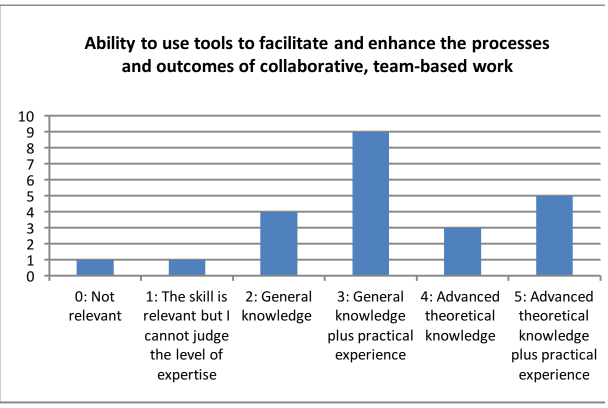
(6) Communication and interdisciplinary work competences/skills

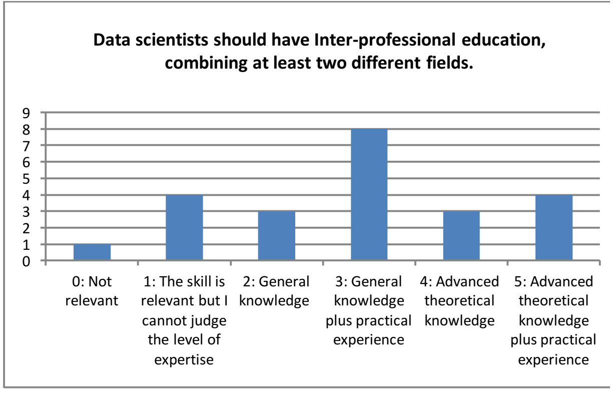
- Know techniques for team building (leadership and management attributes, communication strategies, personal rewards, training and development)
- Understand and deal with the communication barriers in interdisciplinary collaborations
- Understand Business process management (improve business using Big data and a data analytics, IS and Business strategy alignments, Service Level Management, Business plan development)
- Ability to use tools to facilitate and enhance the processes and outcomes of collaborative, team-based work
- Data scientists should have Inter-professional education, combining at least two different fields.


Know techniques for team building (leadership and management attributes, communication strategies, personal rewards, training and development)


Understand and deal with the communication barriers in interdisciplinary collaborations


Understand Business process management (improve business using Big data and a data analytics, IS and Business strategy alignments, Service Level Management, Business plan development)


Ability to use tools to facilitate and enhance the processes and outcomes of collaborative, team-based work


Data scientists should have Inter-professional education, combining at least two different fields.