Data Scientist Professional Revisited: Competences Definition and Assessment, Curriculum and Education Path Design

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl

Mathijs Maijer
University of Amsterdam, The Netherlands
m.f.maijer@gmail.com

Luca Comminiello
University of Perugia, Italy
lucapio94@gmail.com

ABSTRACT

Data Science is maturing as a scientific and technology domain and creates a basis for new emerging technologies and data driven application domains. Educated and/or trained Data Scientist is becoming a critical component of the whole data driven science and technology ecosystem. It is important to revisit the Data Scientist Professional definition propose/identify effective approaches to Data Science competences and skills assessment that would allow developing customisable education and training curricula that would support organisational capacity building (effective HR management) and individual career development. The paper is discussing how the EDISON Data Science Framework can be used to solve these problems. New approaches to Data Science competences assessment is proposed that introduces the concept of acquired competence which is calculated based on the practitioner career path. Important aspect in targeted education and training for professionals is correct and effective education path building based on initial competences and knowledge assessment. The paper proposes a new approach in customised curriculum building by applying Bloom's Taxonomy to training courses sequence and timing/scheduling.

CCS CONCEPTS

• **Computing methodologies-Machine learning**; • **Social and professional topics -Professional topics -Computing education -Computing education programs -Software engineering education**;

KEYWORDS

Data Science, Data Scientist Professional, Big Data, EDISON Data Science Framework (EDSF), Data Science Competences Assessment, Data Science Curriculum Design, Bloom's Taxonomy

ACM Reference Format:
Yuri Demchenko, Mathijs Maijer, and Luca Comminiello. 2021. Data Scientist Professional Revisited: Competences Definition and Assessment, Curriculum and Education Path Design. In *2021 4th International Conference on Big Data and Education (ICBDE'21), February 03–05, 2021, London, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3451400.3451409

ICBDE'21, February 03–05, 2021, London, United Kingdom

ACM ISBN 978-1-4503-8938-9/21/02…$15.00
https://doi.org/10.1145/3451400.3451409

1 INTRODUCTION

Big Data technologies and availability of scalable cloud based Big Data platforms and tools that can be provisioned and used on demand created new opportunities to work with a variety of data produced by human activity and technological processes. Data driven technologies development facilitated emergence of Data Science as a scientific and technology domain focused on different aspects of data analysis to support data driven technologies and applications in all scientific, industry and human activity domains. Modern data driven research and industry created strong demand for new types of specialists that are capable of supporting all stages of the data lifecycle from data production and input to data processing and actionable results delivery, visualisation and reporting, as well and technological processes control and automation. With the growing importance of data in modern economy, data is becoming an important asset, understanding of importance to create the whole ecosystem for data management and governance is growing. Organisations moving to agile data driven model, need to redefine many organisational role and introduce new data related roles, in addition to commonly accepted importance of Data Scientists, which can be jointly defined as the Data Science professions family. Continuous technology evolution imposes new challenges to modern data driven organisations in technology change management and in managing organisational human/capacity resources in related data driven technologies. Effective Data Science education must combine theoretical and practical skills, while developing right attitude to continuous professional (self-) education. The fact that modern technologies are led by large technology companies who are interested in their technologies adoption, should be recognised and motivate universities and research community to cooperate with technology leaders in enriching academic education with using new available technology platform, especially in Big Data and cloud computing.

The proposed approaches to competences assessment and customisable educational and career path building are based on the EDISON Data Science Framework (EDSF), which was developed in the EU funded EDISON Project and currently is maintained by the EDISON Initiative [1], [2]. Since the first EDSF release in 2016, the framework has undergone significant development, EDSF Release 3 (2018) summarised the experience of numerous practitioners and educators that contributed to the definition of EDSF components. The new EDSF2020 (Release 4) incorporated recent technology

developments that confirmed the Data Science as a central component in the whole ecosystem of the data intensive and data driven technologies that include Machine Learning, Artificial Intelligence, Digital Twins, immersive technologies and IoT [3]. [4], [5], [6].

This paper presents new results in continuous research by the authors to improve the Data Science competences assessment based on the professional career path, and optimal learning path definition based on the competences gap. The paper refers to the previous authors' works that researched new approaches to building effective curricula in Cloud Computing, Big Data and Data Science [7], [8], [9], [10], [11] and based on a long time practical experience in developing both online and campus based education and training courses.

The paper is organized as follows. Section II revisits popular definitions of the Data Scientists and refers to important technologies related to Data Science to create a context to the proposed definition of the Data Scientist Professional and related Data Science competences. Section III describes the EDSF, its components and application domains. Section IV describes different uses of EDSF and functionality of the EDSF Toolkit. Section V describes the Data Science competences assessment, and describes the proposed method to assess acquired competences based on the career path. Section VI discusses how the EDSF can be used for designing customised curriculum based on competences assessment (or professional profile) while using Bloom's taxonomy and competences ranking for educational path construction, and Section VII provides summary and suggestions for future work.

2 DATA SCIENTIST PROFESSIONAL DEFINITION

2.1 Data Scientist Definition Evolution

There are multiple definitions of the Data Science discipline and technology, given in different contexts, that stress/put in the centre one of the four aspects of data analysis: Data Analytics, Data Science, Machine Learning/Deep Learning, and Artificial Intelligence:

- *Data Analytics* is a process of inspecting, transforming and modelling data with the goal to discover trends, patterns, relations that describe observable real-life phenomena and can be used for informed decision-making.
- *Data Science* makes the systematic study of the structure and behaviour of data in order to understand past and current occurrences, as well as predict the future behaviour of that data. Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.
- *Machine Learning* deals with the development of algorithms, some of them based on statistical models, with the objective that their computational implementation allows the computer not only to carry out the tasks without supervision but learn of the results for a continuous improvement. Within Machine Learning, *Deep Learning* is the set of predictive methodologies that use artificial neural networks to progressively extract higher level features from unstructured raw data. This class of methods is particularly effective for making predictions from big data generated by real life behavioural processes or sensors. Machine Learning and Deep Learning are considered as subfields of the Data Science focused on specific tasks, while Data Science provides a general methodology for working with wide variety of data using different methods and tools.
- *Artificial Intelligence* is a machine or application with the capability to autonomously execute upon predictions it makes from data, where prediction is made based on Data Science and analytics methods. Artificial Intelligence is strongly connected to Digital Twins and robotics that increase importance of the consistent industrial data management and quality assurance.

It is important to clarify the relation of the Data Science to other closely related scientific disciplines and technology domains such as Big Data, Artificial Intelligence, Machine Learning, and Statistics. Despite the fact that some authors may refer to historical facts of mentioning these terms 10s of years ago [12], we refer to the current data driven technologies development that made Data Science a central component of all other data related and data driven technologies development. We identify such technology fusion and consolidation took place in 2011-2013 with advents of Cloud Computing and Big Data what also aligned with the National Institute of Standards and Technologies, NIST definition of the Cloud Computing in 2011 [13] and Big Data definition in 2013 [14].

Big Data serves as a technology platform to allow the Data Science and Analytics solutions and applications to work with modern data, which are of the *Big Data 3V scale*: *Volume*, *Velocity*, and *Variety*. Big Data technology platform includes large scale computation, storage and network facilities, typically cloud based, such as Hadoop, Spark, NoSQL databases, data lakes, and others.

In the whole digital economy ecosystem, the Data Science integrates all multiple components from other scientific and technology domains to drive data intensive research and emerging digital technologies development. It is important to give Data Science definition as a scientific discipline to become a foundation for academic research and curricula development:

Data Science is a complex discipline that uses conceptual and mathematical abstractions and models, statistical methods, together with modern computational tools to obtain knowledge/derive insight from data to uncover correlations and causations in business data and support decision making in scientific research and business activity.

Data Scientist is defined as a professional practicing Data Science. Starting from the first years of the Data Science and Analytics technologies adoption there were many Data Scientist definitions proposed by practitioners in the new domain that reflected their personal professional development. The following competence areas and skills were included into Data Scientist competence profile: mathematics, statistics, computer skills, domain knowledge, and also hacking skills as the ability understand (undocumented) functionality of software and algorithms and effectively use them for practical purposes.

The experience of the EDISON Data Science Framework development and practical implementation supported by wide research and educational community discussions allowed us to propose an actionable definition of the Data Scientist Professional, which is

based on the NIST definition and extended with organisational role of the Data Scientist [14]:

A Data Scientist is a practitioner who has sufficient competences and knowledge in the overlapping regimes of expertise in data analytics skills, domain knowledge, business needs, 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.

2.2 Importance of Continuous and Self-education

It is commonly recognised that in such dynamically developing area as Big Data, Data Science, the continuous education and self-study plays a critical role. The proposed EDSF definition provides a good basis for defining Data Scientist professional development path, including knowledge acquisition, skills development, and career path building.

The recent OECD report [15] confirms the urgent need to address data and general digital skills for all types of workforce and economy sectors. An effective professional education should provide a foundation for future continuous professional self-development and mastering new emerging technologies, that can provide a basis for the life-long learning model adoption. Flexibility in providing education and training curricula and courses is key to adopting future skills management and capacity building models.

3 EDISON DATA SCIENCE FRAMEWORK (EDSF)

The EDISON Data Science Framework (EDSF), that is the product of the EDISON Project, provides a basis for Data Science education and training, curriculum design and competences management that can be customised for specific organisational roles or individual needs. EDSF can also be used for professional certification and career transferability.

The main EDSF components include:

- CF-DS – Data Science Competence Framework [3]
- DS-BoK – Data Science Body of Knowledge [4]
- MC-DS – Data Science Model Curriculum [5]
- DSPP - Data Science Professional profiles and occupations taxonomy [6]
- Data Science Taxonomy and Scientific Disciplines Classification

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path. The CF-DS is defined using the same approach as e-CFv3.0 [16] (competences defined as abilities supported by knowledge and skills with applied proficiency levels) but has competences structured according to the major identified functional groups (as explained below).

Figure 1 (a) and (b) provide a graphical presentation of relations between identified competence groups as linked to Research Methods or to Business Process Management. The figure illustrates the importance of the Data Management competences and skills and Research Methods or Business Process Management knowledge for all categories and profiles of Data Scientists.

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

An 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 of domain related research methods and theory as a part of their domain related knowledge and team or workplace communications. See an example of the Data Science team building in the Data Science Professional Profiles definition provided as a separate document [6].

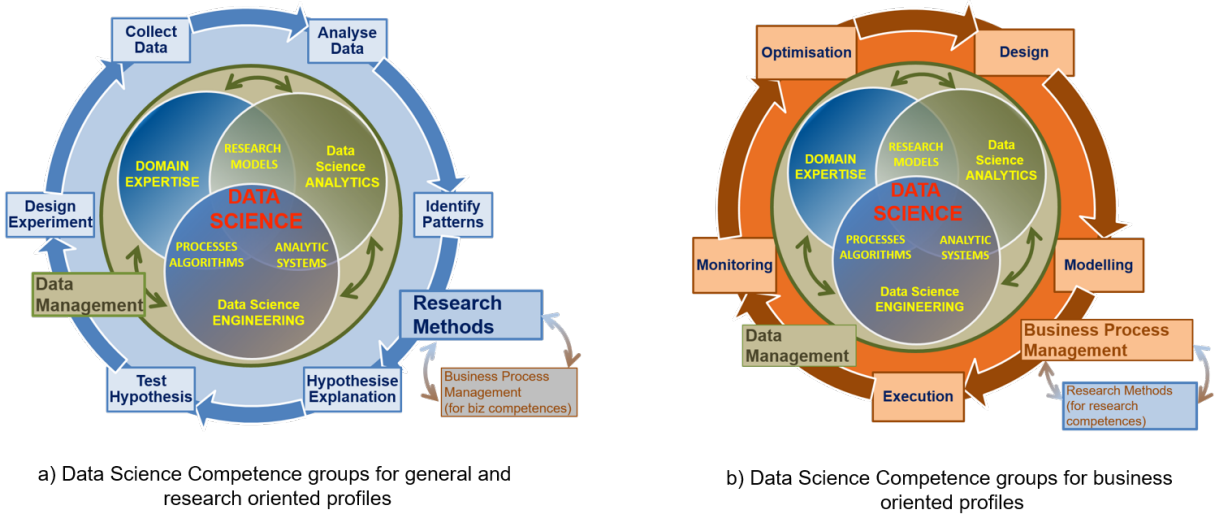
The following core CF-DS competence and skills groups are identified (refer to CF-DS specification [3] for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Big Data Infrastructure and Tools, Data Warehousing) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Management (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

In total, CF-DS includes 30 enumerated competences, 6 competences for each of competence groups. The Data Science competences must be supported by the knowledge that are defined primarily by education and training, and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills (refer to CF-DS [3] for the full definition of the identified knowledge and skills groups):

- Skills Type A which are built based on practicing major competences acquired based on education and training; depend on years of working as a Data Scientist or related roles,
- Skills Type B that are related to a wide range of practical computational skills, including using programming languages, development environment, and cloud based platforms.

The DS-BoK defines the Knowledge Areas (KA) and Knowledge Units (KU) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. It is important to note that the CF-DS defines knowledge topics linked to specific competences that can be mapped to KU and KA in the DS-BoK. The DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [17] and incorporates best practices in defining domain specific BoK's. It provides a reference to related existing BoK's and includes proposed



a) Data Science Competence groups for general and research oriented profiles

b) Data Science Competence groups for business oriented profiles

Figure 1: Relations between identified Data Science competence groups for (a) general or research oriented and (b) business oriented professions/profiles.

new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS [5] is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS competences (specifically skills type A), 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 practical curriculum should be supported by a corresponding educational environment for hands on labs and educational projects development.

The DSPP [6] defines a number of Data Science professional profiles in accordance with existing classifications, such as European standards ESCO [18] or EN 16234-1 "e-Competence Framework" [19].The DSPP includes important part the competences relevance (scores) to each defined profile in scale from 0 to 9 (from low to high) that be used defining targeted education and training and building an effective career path.

4 EDSF PRACTICAL USES AND EDSF TOOLKIT

The EDSF toolkit has being developed to support multiple practical applications for Data Science competences and skills management and to ensure their compatibility. It primarily contains enumerated competences, skills and knowledge topics/units definition supported by corresponding ontologies. Ongoing development includes API definition and creation of the reference datasets representing different components of the EDSF to support applications development. EDSF Toolkit is a community effort and available as an Open Source at the EDSF github project [2].

The following are the intended practical applications of EDSF facilitated by the EDSF toolkit:

- Academic curriculum design for general Data Science education and individual learning path construction for customizable training and career development
- Professional competence benchmarking, including a CV or organisational profiles matching
- Professional certification of Data Science Professionals
- Individual competences self-assessment and learning path advice tool
- Vacancy description construction tool for job advertisement (for HR) using controlled vocabulary and Data Science occupations taxonomy
- Data Science team building and organisational roles specification.

EDSF provides an example of the integrated competence and skills management framework that is being used in other technology domains and economy sectors, which examples include education and training framework for digital and data skills in maritime industry (as part of the MATES project [20]), Data Stewardship competences and curriculum definition (as part of the FAIRsFAIR project [21]), reference in the German Ministry of Economics study on Data Science competences and resources [22].

5 DATA SCIENCE COMPETENCES ASSESSMENT

The CF-DS and DSPP components provide a basis for the novel improved Data Science competences assessment in a quantified manner that takes into account the practitioner or candidate career path. This assessment is helpful to support the increasing demand for Data Scientists. For example, using CF-DS and DSPP, a desired profile can be constructed against which a curriculum vitae (CV) of a Data Scientist can be tested. Based on such a test, competence gaps can be identified between the desired profile and the assessed CV, which can help with the decision making if the candidate can

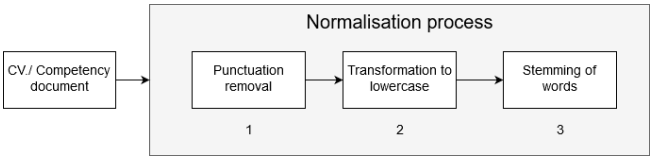


Figure 2: Processing steps of the CV and competency documents text

be hired, on which position or role, and also predict their possible career path.

As a part of the EDSF Toolkit development, the authors have tested different methods for CV and job vacancy/profile matching using Doc2Vec document embedding and PV-DBOW training algorithms (available in the genism Python libraries) [23], [24].

This section describes how the document similarity techniques combined with regular expressions were used on the CVs of Data Scientists to create insights into the competencies of the CVs' creators (later referred also as a job/vacancy candidate). This is done by computing three components: a timeline that indicates the career path of the data scientist, a graph showing the competency scores based on document similarity, and a graph showing competency scores based on the career path. This method can then be deployed as a tool that could, for example, be used by recruiters or practitioners that wish to assess their own competencies and to identify a potential path for professional development. It can also be used for a Data Science team composition and management as well as identification what kind of training the team requires. This method was used to develop a tool in the form of a web-application to provide easy access [25], [26], recent development has improved the gap identification to integrate it better with the customised curriculum design.

5.1 Pre-processing Steps

To create the three components mentioned that can be used to assess individual Data Science competences, several steps should be taken. First, to acquire the competency scores, both the CVs and the documents containing information about the required job competences (e.g. job vacancy typically using the CF-DS vocabulary) should be pre-processed, especially normalised. The punctuations are removed, all letters are transformed to lower case, and the Porter algorithm [27] is applied to remove suffixes of the words. This process is applied to compare similar words that appear in different forms; the steps are shown in figure 2. To extract the career path from a CV, a list of job names with corresponding DSPP classifications, a DSPP competence relevance scores list, and a CF-DS competence labels list were prepared. These lists were then linked to each other, so that every job found in the CV is classified with a profile that is listed in the DSPP list. This classification can then be used to obtain competence relevance scores belonging to that professional profile of the job. The relationship between the components is shown in figure 3

5.2 Implementation

5.2.1 CV matching and competence gap definition. The graph showing the competency scores extracted from a CV was acquired using

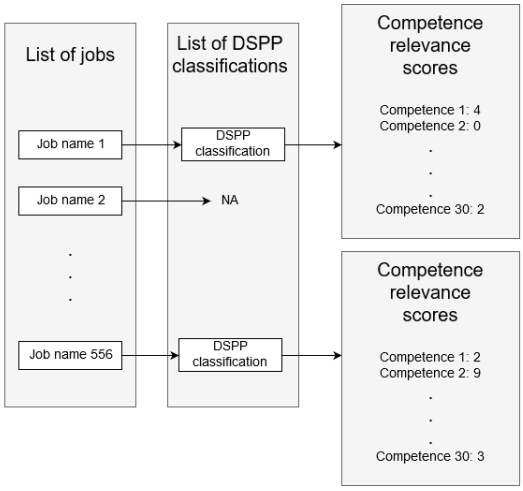


Figure 3: Relations between data components used for career path extraction

multiple steps. After pre-processing the CVs and the thirty competence documents representing the controlled vocabulary of the CF-DS, TF-IDF features were gathered. A matrix of TF-IDF features is acquired and can then be used to calculate cosine similarity between the CV and the competence documents. This is done by taking the dot product of the row-normalized vectors, which gives a similarity matrix that contains the similarity score. This score ranges from 0 to 1, however, because a CV will never contain the exact same text as a competence document, a score of 0.7 is seen as the maximum achievable value of a CV. These values then can be used to create a spider chart, to give an overview of the scores per competence.

When the competence scores have been acquired using a CV, it can be compared to a desired data scientist profile to identify gaps in lacking skills or knowledge. The competence relevance scores for a DSPP classification, also shown in Figure 3, can be mapped from a 0 – 9 range to a 0 – 1 range. After the mapping, we call these values rates, which then can be used to subtract the CV scores for each competence from the rates. If a negative result is acquired after the subtraction, the data scientist is deemed to have sufficient knowledge for that competence, and the result is set to 0.

After the subtractions, the differences are multiplied by the corresponding rates again, in order to weigh the proficiency level. Finally, the results are evaluated: all the competences that have scored a grade greater than 0.5 are identified as gaps. Figure 4 shows two graphs, where the first graph shows how example CV can be matched to a DSPP or a vacancy profile. As an example, the profiles DSP04 – Data Scientist and fictitious candidate CV are used. The second graph shows how the gaps are identified using the process described above.

5.2.2 Career path extraction. Using described above CV and job profile matching based on both documents similarity doesn't take into account acquired experience by a candidate. This can be resulted that two candidates using the same CV template or a CV design tool (popular service by online job search agencies), although

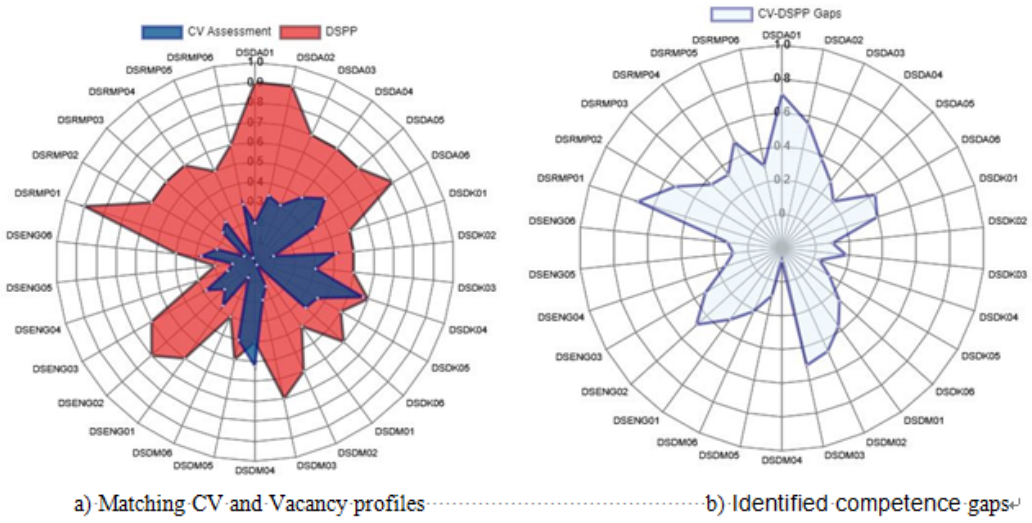


Figure 4: Comparing candidate's competences and target professional profile DSPP04 Data Scientist: (a) DSPP04 and candidates competences assessed; (b) competence gap.

having significantly different job experience, can be scored equally. To avoid this situation and make the CV and job matching correct, we introduced the acquired competence concept that defines the candidate's/professional's real competence as acquired competence amplified by years of working in relevant positions/roles.

To assess the acquired competences, next to defining the competence scores from a CV using document similarity techniques, the career path is also extracted from the CV. This information is then used to create another scores vector and the graph showing the competence relevance scores calculated together with a timeline indicating the career path of the candidate's CV. The first step in this process extracts all the mentioned jobs in the CV. Then, for each job, the position level is identified and years of work are extracted, a DSPP classification was made, based on this the DSPP competence relevance scores are assigned, and finally, the job is tagged as relevant or not. Different jobs in a CV are identified by testing whether each word of the CV appears in a list with known jobs using regular expressions. Multiple mentions of one job in a CV were separately handled to get a better overview.

Then, for all identified jobs, the position levels were acquired by looking at different position level names in the CV text. In this case, five different levels were used with their common alternatives: entry-, intermediate-, senior-, principal-, and lead-level. The location of where the job name was found is used to look for the job level by looking at words that appear near to the job name. Afterward, the amount of experience, the time that someone has practiced a job, is acquired based on the CV/career timeline. This is achieved using regular expressions to look for different date patterns near the occurrences of the job names mentioned in the CV. Multiple small regular expressions were used in combination with each other to be able to find date patterns in multiple formats, as there is no formal manner in which every CV lists its experience. The regular expressions are constructed to identify different date separators, months, time span indicators, years, and day patterns. These can be combined to identify dates in multiple formats that are often used. When data has been found for a job, the start date is subtracted from the end date, which is then divided by 365 and rounded down to get the amount of years that someone has practiced a job. These jobs are then classified using the DSPP. Then to tag whether identified jobs are relevant, they must have a DSPP classification, as well as other attributes such as start and end date, or a profession/job level.

To create a graph showing competence scores based on the extracted career path, two assumptions were made, namely: 1) when someone has practiced a job for a longer time, he/she becomes more competent at the relevant competencies that are listed, and 2) when someone quits a job, the competencies to perform relevant tasks from that job are not lost. Then, to calculate the competence scores using the career path, equation 1) was used:

$$Competence \leftarrow \min\left(\sum_{j \in J} c_{ij} \cdot multiplier_j, 100\right) \quad (1)$$

where i is the current competence, j is a job, J are all the jobs that were extracted, c_{ij} is the competence relevance score for the current competence and job, and multiplier is the multiplier that is used for the job.

For every competence in the CF-DS, the minimum of 100 and the sum of the values calculated using all the relevant jobs for the current competence is taken. The multiplier is acquired by either using the direct amount of job experience in years, or by using a common mapping from the position level to the amount of experience associated with the level. This will ensure final competence scores that have a range of 0 – 100.

Figure 5 provides an example of the candidate's competence profile with 8 years experience in positions related to Data Scientist calculate using simple documents similarity (like in Figure 4) and

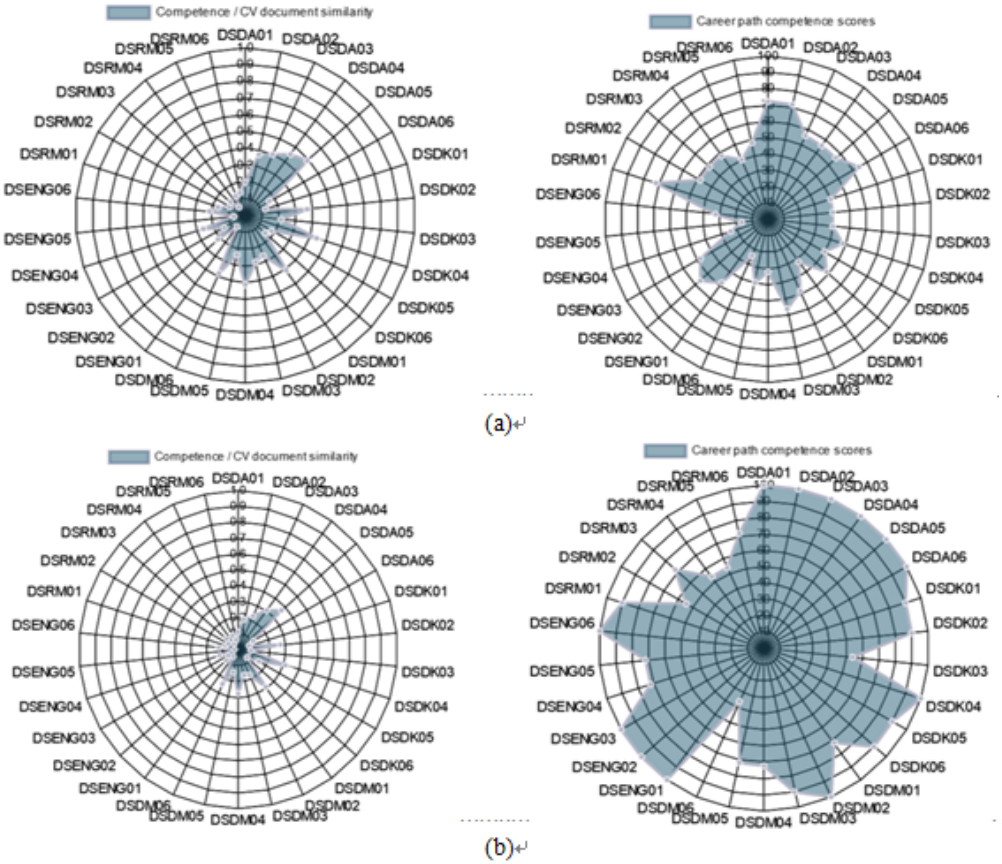


Figure 5: Comparison of competences calculated based on simple document similarity (left diagram) and using acquired competence concept (right diagram) for (a) Data Scientist with 8 years of experience, and (b) Hadoop developer with 15 years of experience.

using the proposed acquired competence algorithm explained in this section. Which then can be directly used in the same manner as shown in Figure 5, after remapping the range back to 0 – 1, which is in accordance with the earlier computed document similarity scores.

6 BUILDING LEARNING PATH FOR THE DESIGNED CURRICULA USING BOOM'S TAXONOMY

The next step after defining the set of intended competences (that can be either selected professional profile or assessed competence gap) is to design the effective curriculum and education or training path. It is rather a well defined process and a routing process to create the curriculum for the whole academic programme that should be delivered by an educational institution when a whole set of required competences and corresponding learning outcomes provide input to the curriculum definition. The general purpose curriculum can be created in this way. However, as technology develops fast and correspondingly required competences change, there is a need for educational and training institutions to react fast and offer as much as possible a customised curriculum design. This section presents the proposed approach that extends the curriculum design methods using EDSF ontology described in the previous authors' paper [11] by using knowledge units ranking and Bloom's taxonomy learning levels for customised learning path building.

6.1 Customised Curriculum Design using EDSF Ontology

The input for the (customised) curriculum design is the intended competences set together with competences relevance or ranking for the desirable/target professional profile or job position. For individual learning path building, the individual competences can be assessed based on CV matching against the intended job position or professional profile, certification exam, or just self-assessment questionnaire.

When a set of required competences is defined together with the relevance scores and required proficiency levels, the set of required knowledge topics can be extracted from individual competences (note, there exist multiple links from competence instances to single knowledge topic) and ordered according to required proficiency

Table 1: Core KUs identified forDSP04 Data Scientist

Data Analytics and Machine Learning (core)		
KU01.02.02	Supervised Machine Learning	48
KU01.02.03	Unsupervised Machine Learning	48
KA01.01	Statistical methods for data analysis	47
KU01.01.07	Quantitative analytics	38
KU01.01.08	Qualitative Analytics	33
KU01.02.04	Reinforced learning	32
KA01.05	Predictive Analytics	32
KU01.01.09	Data preparation and preprocessing	24
Data Management (core)		
KA03.02	Data management systems	33
KU03.02.01	Data architectures (OLAP, OLTP, ETL)	33
KA03.01	General principles and concepts in Data Management and organisation	26
KA03.03	Data Management and Enterprise data infrastructure	26
KU03.01.02	Data Lifecycle Management	24
Data Science Engineering (core)		
KU02.01.07	NoSQL databases	40
KA02.02	Infrastructure and platforms for Data Science applications	40
KA02.03	Cloud Computing technologies for Big Data and Data Analytics	22
Research Methods and Project Management (core)		
KU04.01.05	Use cases analysis: research infrastructures and projects	32
KA04.01	Research Methods	27
KU04.01.02	Modelling and experiment planning	26
KU04.01.03	Data selection and quality evaluation	26

level and relevance for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set specifies the structure of the curriculum that further can be mapped to the Model Curriculum Learning Units defined as individual courses and KAG related courses groups, otherwise, it can be used directly as advice for constructing curriculum by the programme or course manager.

At the same time, the required proficiency level is scored for each KA and KU, which will define mastery levels and corresponding learning outcomes for the targeted education or training curriculum. When using EDSF ontology, it is a routine task to extract all required knowledge topics, map them to KA/KU and define relevance score by querying ontology with a few lines of code using OwlReady2 Python module that allows manipulating ontology classes, instances and properties transparently.

The MC-DS provides a set of templates for designing general purpose curricula composed of specified Learning Units, together with mastery levels defined for different types of programmes: Introductory, bachelor. 3 master levels are defined based on Bloom's Taxonomy: Familiarity, Usage, Assessment (refer to MC-DS [5] for details). When using MC-DS for customised curricula design, the competence scores/relevance defined in DSPP using scale 0 to 9 can be easily mapped to MC-DS mastery levels [10]. Collected Skills type B linked to intended competences will provide advice on the required hands on training and practical project development tasks and development platform.

The EDSF Toolkit and its outcome provide advice on the suggested curriculum structure that can be adjusted to the real condition of the teaching or training institution depending on the available teaching staff and lab base. It is also important that the courses are correctly ordered, and necessary pre-requisite knowledge are specified. When using 3rd party educational platform providers and cloud based data labs, the presented approach can provide a specification for the required educational platform.

Table 1 below provides an example of KUs scores grouped into four DS-BoK Knowledge Area Groups for the competences defined for the DSP04 Data Scientist (only KU with the highest scores are included). Figure 6 illustrates the whole set of required KUs presented in the visual form of the tree map.

6.2 Defining Knowledge Units to include into the Curriculum

Once the suggested Knowledge Units have been obtained, it is possible to combine them into educational courses, map them to courses defined in MC-DS or to existing courses, which are typically defined according to DS-BoK Knowledge Areas or Knowledge Units. EDSF ontology defines for these purposes the Course class, whose instances are directly connected to the KUs through the object property course. This allows for collecting relative scores for all KUs linked to the required curriculum.

When moving to a practical curriculum and courses design, it is important to define the courses relevance and their priority or sequence. The suggested courses content can be defined by KUs

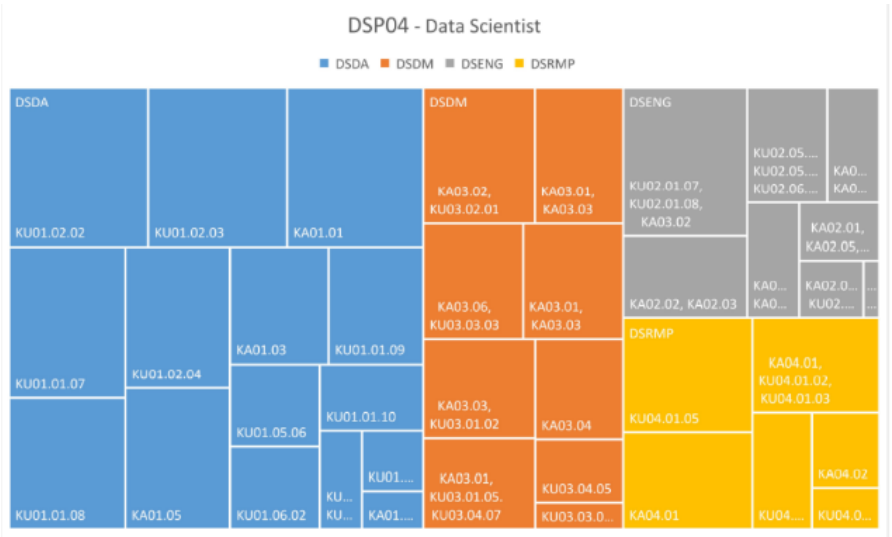


Figure 6: Example curriculum structure for DSP04 – Data Scientist

grouping based on their ontological similarity and difference. In a simple view, this defines the courses that need to be attended to achieve intended learning outcome and collect the necessary number of credits, in a classical education model. However, this doesn't solve the problem of the efficient programme planning or learning path design, what is especially important when designing a curriculum for workplace training, vocational education or self-education.

The Course class in the EDSF ontology can be used to calculate the course weight based on the integral score of the component KUs. This can be done by querying the ontology that will produce the list of associated courses for the required competence profile, sorted in descending order by weight. The course weight is calculated based on collecting all individual KU's scores linked to required competences, given the multiple relations and mapping between competences, knowledge topics in CF-DS, Knowledge Units in DS-BoK, and Learning Units in MC-DS. The course weights are normalized to 0-9 scale and aligned with the related competences relevance.

6.3 Applying Bloom's Taxonomy to Curriculum Structuring and Course Planning

Data Science programme structuring and courses planning is an important stage in the practical curriculum implementation. The EDSF Toolkit supports the interactive curriculum design approach and courses planning. At this stage, the course weights are used to assign the credit points to the planned courses in an academic curriculum. The course design application (programmed in Python) uses external csv files that contain the mapping between the courses and the related credit points. The association between weights and credits is assigned as follows, using the calculated course weight: low priority is given to courses that have weight less than 3, medium priority is given to courses whose weight is between 3 and 6, the higher priority is assigned to course intended for key competences. Figure 7 puts the candidate's competence profile and gaps (vertical bars) into the context of the target DSPP profile or vacancy (solid line); main competence gaps are marked with the circles.

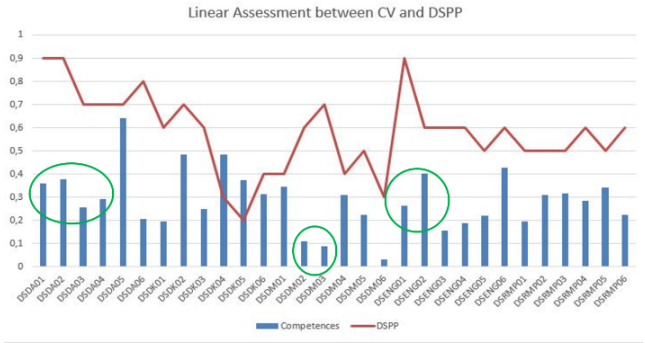


Figure 7: Candidate competences gap in the context of the target DSP04 profile (in a linear coordinate comparing to spider diagram). Main competence gaps are marked with the circles.

At this stage, the Bloom's Taxonomy learning (cognitive) levels are applied to the courses duration and planning [28]. Courses that correspond to both larger identified gaps and higher priority are suggested to have longer duration or even are recommended to split into multiple periods (in practice 2-or 3) to allow for the learners' reflection and practical skills acquisition, which are time dependent.

The number of credits for a course is linked to the course weight but limited to 3 or 6 credits. If the course has a weight greater than 6, then it is split into two parts, whose sum of credits is equivalent to the total expected, and it is ensured that the two modules appear in two consecutive semesters in case of two years master programme.

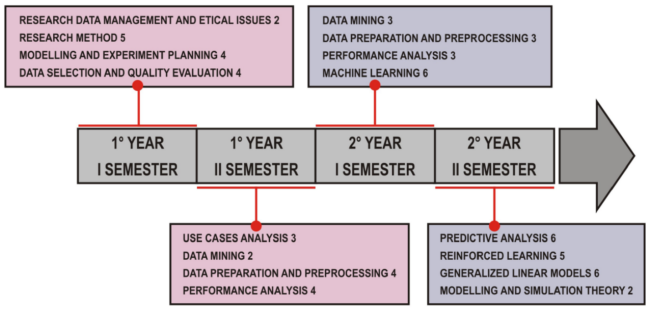


Figure 8: Example curriculum planning based on implied courses duration and DSPP profile proficiency level.

Furthermore, in defining the learning path, the number of courses for each semester is limited to a maximum of four.

Figure 8 provides an example of the two years Data Science master curriculum design that incorporates described above approach and methodology supported by the EDSF Toolkit. The first-year courses are targeted to create a strong background for core Data Science and Analytics courses. Core courses such as Data Mining and Machine Learning are split on two semesters and taught in two periods. Courses planning for vocational education and professional training will benefit from a similar approach, however using different time span.

7 CONCLUSION AND FURTHER DEVELOPMENTS

EDSF is a continuously evolving framework maintained by the community of educators and practitioners in Data Science and other data related technologies. EDSF provides a basis for defining Data Science competences, Body of Knowledge and Model Curriculum that can be used for designing customised curriculum for target competence profiles. With the publishing the new EDSF Release 4 (also referred to as EDSF2020), the framework became a mature product and currently counts multiple practical uses and is cited in multiple studies. The four EDSF parts describe Data Science Competence Framework, Body of Knowledge, Model Curriculum, and Data Science Professional Profiles. The new EDSF Part 5 (since current Release 4) is intended to provide a practical guidance for universities, training organisations, data management and data steward team, practitioners to define their Data Science curricula and courses selection, on one hand, and for companies to better define a set of required competences and skills for their specific industry domain in their search for Data Science talents, on the other hand.

The authors are involved into multiple ongoing developments related to Data Science programs definitions and courses development such as Vodafone Ukraine Data Science Academy, Data Science MBA programme for the Amsterdam Business School, as well as digital and data skills framework for the maritime industry in the framework of the EU funded MATES project. All such activity will contribute to further EDSF development and will facilitate the EDSF Toolkit development.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the Horizon2020 projects FAIRsFAIR (grant number 831558), MATES (grant number 591889) and EDISON (grant n. 675419).

REFERENCES

[1] EDISON Community wiki. [online] https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
[2] EDISON Data Science Framework (EDSF). [online] Available at https://github.com/EDISONcommunity/EDSF
[3] Data Science Competence Framework [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-competence-framework
[4] Data Science Body of Knowledge [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-body-of-knowledge
[5] Data Science Model Curriculum [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-model-curriculum
[6] Data Science Professional Profiles [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-professional-profile
[7] Demchenko, Yuri, *et al*, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. Proc. The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 Dec 2014, Singapore.
[8] Manieri, Andrea, *et al*, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, Proc. The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada
[9] Demchenko, Yuri, *et all*, EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry, Proc. The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 Dec 2016, Luxembourg.
[10] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.
[11] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03.
[12] David Donoho, 50 Years of Data Science, Journal of Computational and Graphical Statistics, Volume 26, 2017, Issue 4, pp 745-766, Published online: 19 Dec 2017 [online] https://doi.org/10.1080/10618600.2017.1384734
[13] SP 800-145, The NIST Definition of Cloud Computing, NIST 2011 [online] https://csrc.nist.gov/publications/detail/sp/800-145/final
[14] 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
[15] OECD Skills Outlook 2019, Thriving in a Digital World, Published on May 09, 2019 [online] https://www.oecd.org/education/oecd-skills-outlook-2019-df80bc12-en.htm
[16] e-CF3.0, 2016 European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. Available at http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf
[17] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012
[18] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at https://ec.europa.eu/esco/portal/#modal-one
[19] EN 16234-1 "e-Competence Framework", CEN Standard 2019.
[20] MATES Project: Maritime Alliance for fostering the European Blue Economy through a Marine Technology Skilling Strategy [online] https://www.projectmates.eu/
[21] FAIRsFAIR Project: Fostering FAIR data practices in Europe [online] https://www.fairsfair.eu/
[22] Data Science Lern- und Ausbildungsinhalte, Gesellschaft fur Informatik, BMBD und Plattform Lernende Systeme, Arbeitspapeir, Dezember 2019 [online] https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/GI_Arbeitspapier_Data-Science_2019-12_01.pdf
[23] Quoc Le and Tomas Mikolov, Distributed Representations of Sentences and Documents
[24] Phillip Lord (2010) Components of an Ontology. Ontogenesis.
[25] Matching CVs based on EDISON Data Science Competencies (CF-DS) [online] https://github.com/EDISONcommunity/EDSFapps/tree/edsfcv

[26] Maijer, Mathijs, Matching CVs based on EDISON Data Science Competencies (CF-DS) and advanced text analysis methods, Project report, 2018. [online] https://esc.fnwi.uva.nl/thesis/centraal/files/f1532411291.pdf
[27] Martin F Porter. "An algorithm for suffix stripping". In: Program 14.3 (1980), pp. 130-137.
[28] Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.