Development of a Job Advertisement Analysis for Assessing Data Science Competencies

RESEARCH PAPER

JAN VOGT
THILO VOIGT
ANNIKA NOWAK
JAN M. PAWLOWSKI

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data science competencies receive rising attention. How can competency profiles be built for data scientists? This question is encountered by synthesizing various competency frameworks from the literature and by conducting a job advertisement analysis. The 'Skills and Recruitment Ontology' is used as the underlying ontology for the job advertisement analysis. Therefore, over 5000 job postings were crawled from the job platform 'Jooble' and the results evaluated in a focus group. This work provides a competency data set for data science jobs. It points out newly found competencies and also provides design principles for competency frameworks.

CORRESPONDING AUTHOR:
Jan Vogt
Institute of Computer Science, University of Applied Science Ruhr West, Bottrop, Germany
contact@j-vogt.com

KEYWORDS:
Job Advertisement Analysis; Skills and Recruitment Ontology; Text Mining; Data Science Competencies; Design principles for competency profiles

TO CITE THIS ARTICLE:
Vogt, J, Voigt, T, Nowak, A and Pawlowski, JM. 2023. Development of a Job Advertisement Analysis for Assessing Data Science Competencies. *Data Science Journal*, 22: 33, pp. 1–16. DOI: https://doi.org/10.5334/dsj-2023-033

DOI: 10.5334/dsj-2023-033

I. INTRODUCTION

The data scientist's job and the field of data science is growing in industry (Shirani 2016) and academia (NASEM 2018). Data science is a combination of multiple disciplines (ibid.) and consequently a transdisciplinary field (Cao 2017). The economy uses data to design products, for data-driven decision making (Mandinach et al. 2015) and for business value creation (Breitfuss et al. 2019). New demands in the field of data science result in new competency requirements for employees. But how can the skill demand be described for data scientist's appropriately? At the moment, skill demands for data science are described through university curricula (Saltz, Armour, & Sharda 2018) and expert-driven competency frameworks in a static manner (Donoho 2017; Hattingh et al. 2019; Ridsdale et al. 2015). But it is essential to have competency frameworks that are up-to-date, as workforce skill needs are changing, e.g. through upcoming trends (Dadzie et al. 2018). Therefore, a more dynamical approach is needed. A suitable approach is the job advertisement analysis approach, which makes it possible to continuously record new competency requirements of the workforce. Job advertisement analysis approaches are gaining attention (Almaleh et al. 2019; Boselli et al. 2018; Khaouja et al. 2019; Khobreh et al. 2015; Lima, Bakhshi, et al. 2018; Wowczko 2015; Zhao et al. 2015) especially in the field of analyzing data science competencies (Dadzie et al. 2018; Djumalieva, Lima, Sleeman, et al. 2018; Debortoli, Müller, and Brocke 2014; Murawski and Bick 2017; E. Sibarani et al. 2020; Silveira et al. 2020; Shirani 2016).

Consequently, the following research questions arise:

RQ 1: Which data science competencies are needed by the workforce?

RQ 2: Which data science competencies can be extracted from job advertisement analysis?

The first question is answered through a broad literature review of data science competency frameworks, while the second research question is answered by the conduction of our own developed job advertisement analysis of data scientists' jobs.

II. RELATED WORK

i. COMPETENCY FRAMEWORKS

Donoho (2017) did a review of the current data science movement and distinguishes between 'lesser data science' and the larger field of 'greater data science' (GDS). The larger field of data science 'cares about every step that the professional must take from getting acquainted with the data all the way to delivering results based upon it' (Donoho 2017). Moreover, (ibid.) is concerned with the development of data science from statistics, and thus provides an overview of the most important aspects of the scientific data research. On the basis of GDS, he formulated the following six dimensions of greater data science (ibid.):

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science

Hattingh et al. (2019) conducted a literature review of data science competencies by scanning 139 titles and creating a 'unified data science competency model' (ibid.). The findings were grouped into the following competency themes: organizational; technical analytical; ethical and regulatory; cognitive and social (ibid.).

The SFIA global skills and competency framework is an expert-driven model for the digital world and data science skills (SFIA 2018):

- Data governance skills
- Data culture and capability skills
- Data lifecycle management skills
- Data security quality skills

Furthermore, it (ibid.) is also recommended as a 'holistic approach' for information and communication technology (ICT) curriculum design and management (Konsky, Miller, & Jones 2016).

As an intermediate conclusion, comprehensive approaches for the competency requirements of data scientist jobs exist (Donoho 2017; Hattingh et al. 2019; SFIA 2018), but often do not reflect the most current requirements from companies. The topics and skill areas for data scientists are defined, but the description of important tools and products are missing. Therefore, the job advertisement analysis approach is an interesting one to consider and will be described in the following chapter.

ii. JOB ADVERTISEMENT ANALYSIS APPROACHES FOR EXTRACTING DATA SCIENCE COMPETENCIES

Da Silveira et al. (2020) conducted a job advertisement analysis for data scientist jobs with data from the social network service LinkedIn. Soft skills and technical skills should be analyzed through a qualitative and a quantitative typology (Silveira et al. 2020). Furthermore, according to their research, 'most companies do not care about the degree and education level of the candidate, but about the necessary soft skills and technical competences' (ibid.).

They also identified a trend towards multidisciplinary skill profiles for data scientists (ibid.). Nevertheless, the following competencies were identified as the most important competencies: 'Communication', 'Team Player', 'Problem Solver', 'Python', 'English' and 'SQL'. The results were sorted according to the 'six Dimensions,' as outlined by Donoho (2017). A weakness of Silveira et al.'s (2020) approach is that no source code is provided. Another issue is the comparability to other studies is limited, as the data was collected in Brasilia; therefore an analysis of international job advertisements should be done to ensure comparability.

Shirani (2016) used a job advertisement analysis approach with the tool 'RapidMiner' as a textmining software, while 'Kdnuggets.com' and 'r-bloggers.com' were used as job sources. The text analysis and competency identification were done according to the following steps (ibid.):

1. RapidMiner used as textmining software
2. Removal of stop-words
3. Lower case to avoid case sensitivity
4. One-word terms were first identified, then expanded by n-gram 2 and n-gram 3 to obtain more meaningful capabilities

At the end, (ibid.) did a formulation of a taxonomy for data science and analytic competencies consisting of the columns 'data science and analytics', 'big and relational data analytics,' and the rows 'advanced skills', 'introductory to intermediate skills', 'foundational competencies' and 'soft skill'. The aim of the study done by Shirani (2016) is to identify data science competencies based on the industry demand. An interesting outcome is that employers placed as much emphasis on soft skills as on hard skills during the recruitment process (ibid.). A weakness of their approach of is that no correlations or connections among the competencies are identified, and the source code is also not provided. Due to the fact that the source code is not open, no updated skill set can be provided for 2021.

Djumalieva et al. (2018) developed an open and data-driven skills taxonomy, which is independent of ESCO and O*NET (expert-derived ontologies). The taxonomy captures interrelationships between skills, job titles and the salaries of UK job adverts. Skills are modeled by means of a graph where skills are represented as vertices, co-occurrences as edges. Forty-one million job adverts were collected from the platform Burning Glass Technologies between January 2012 and December 2017 (ibid.). For each job, its title, salary, education and experience requirements were extracted from the adverts. This procedure consists of five steps: count co-occurences of skills in all job adverts, train a word2vec model, build a skills graph, remove transversal skills, and create hierarchical clusters (ibid.). Djumalieva et al. (2018) defined a concise skill set with a good clustering. Although the results of the data set are open, there is no possibility of updating the results, as the job advert data becomes rapidly outdated; there is also no open source code.

Sibarani E et al. (2020) quantified skill needs in the industry through an analysis of online job adverts using an ontology-based information extraction (OBIE) method (Sibarani E et al. 2020; Sibarani E M et al. 2017). Therefore, the Skills and Recruitment Ontology (SARO) was developed

to capture important terms and skill relationships. A job advertisement analysis was conducted with data from 'Adzuna' with 872 job adverts between August 2015 and November 2015. Also, the SARO-guided OBIE method was used, and 184 keywords (skills) were extracted. A co-word analysis was conducted based on this set of keywords and their co-occurences in job-adverts (Sibarani E M et al. 2017). The study from Sibarani (2016) has a strong approach with an underlying ontology, but does not provide a complete and comprehensive skill set for the data scientist.

Murawski and Bick (2017) determined Big Data competencies by conducting a job advertisement analysis ofnthe job portal 'Monster.com,' using the Latent Dirichlet Allocation (LDA). In their research, they considered which the necessary competencies of a data professional in the UK, and which of these are imparted in Higher Education programs. Therefore, they used a topic model approach and a deductive content analysis. Two independent analyses were conducted: required competencies (online job adverts) and imparted competencies in curricula. As a data source, the portal 'monster.co.uk' was used, and the search terms were 'Data Analyst' and 'Data Analytics'. Thereby, 500 jobs were analyzed with a topic modeling algorithm. An important key finding was that the companies need 'allrounders' with technical, analytical, and business management competencies as employees, but Master's programs rarely teach business management competencies. Murawski & Bick (2017) identified five roles but rather put a focus on data competencies than on data scientists' jobs. Furthermore, no open data set was provided to conduct the analysis again.

Debortoli, Müller, & Brocke (2014) extracted information about the knowledge and skills requirements for big data and business intelligence professionals. The data source was 'Monster.com' and two data single-day snapshots were taken in September 2013 and March 2014. Moreover, a Latent Semantic Analysis (LSA) was applied to the data. As a result, a competence taxonomy for business intelligence and big data was available. Furthermore, the similarities and differences between business intelligence and big data were marked (Debortoli, Müller, & Brocke 2014). A weakness of the Debortoli et al. (2014) approach is that no information about the connections between competencies was provided, and no tools or products were considered, though soft skills were taken into consideration.

Dadzie et al. (2018) made a visual exploratory analysis of skill demand. An analytical methodology was used to conduct the exploration and analysis. An interactive visualization was used as a working tool. For the analysis, SARO was used as underlying ontology. The OBIE-Pipeline and the wikifier pipeline were used to track skill demands across time and location (Dadzie et al. 2018). The visualization of data makes it possible to recognize patterns and trends in the skill market demand of data science competencies. A weakness of their approach, however, was that no open data skill set nor an open source code was provided.

iii. ANALYSIS OF CURRENT APPROACHES

Silveira et al. (2020) also used other search terms for the data scientist, including 'computer scientist' and 'big data consultant'. The field of data scient is very broad and diverse. They provide an overview of competencies but do not provide any information about the connections and correlations between the competencies. Moreover, Murawski and Bick (2017) used the search terms 'Data Analyst' and 'Data Analytics'. So they limited their research to a specific area in the field of data science. Therefore, there is no broad view of data scientists' competencies. Shirani (2016) provides a top 50 list of relevant terms, ordered by rank, and also a taxonomy of data science and data analytics competencies. But the analyzed data set is from 2016 and thus considered old in the context of data science, as the required competencies change quickly. Additionally, Djumalieva et al. (2018) used a hierarchical representation for the competencies and no ontology. They provided an open data set but a small competency set with 26 competencies. Furthermore, Sibarani E et al. (2020) analyzed job adverts by using the OBIE-method. But the SARO is still dependent on ESCO and extends the ESCO with 1000 further competencies. Nevertheless, no open data set of the trend analysis results was provided. Dadzie et al. (2018) do not provide their results open-source, and additionally, a second weakness of their work is that their data set is from October 2015 and is no longer up to date. On the other hand, Debortoli, Müller, and Brocke (2014) focused on the difference between Business Intelligence and Big Data competencies. Although a big data competence taxonomy is provided, no information on the connections among competencies was included.

Nevertheless, their data is from September 2013 and March 2014, and therefore also no longer up to date.

For our approach, SARO (Sibarani E M et al. 2017) was used as underlying ontology, partitioning products, topics, skills, and tools. Moreover, the representation done by Shirani (2016) was adapted to represent top competencies. The data science competencies from the frameworks of Djumalieva et al. (2018), Sibarani E M et al. (2017), and SFIA (2018) were used for evaluating our job adverts analysis approach. Furthermore, our code was provided open-source, so that: the analysis could be conducted quickly for further research; and competency maps could be updated quickly. The tool is available at https://github.com/jvogt2306/job-advertise-analysis. This step is needed because data science competencies change rapidly (Shirani 2016) and therefore competency maps need to be updated regularly.

III. RESEARCH METHOD

To meet the needs of the industry, we have chosen an explorative approach: an analysis of data-science-related job advertisements, since job advertisements reflect the needs of the industry. The analysis works with a named entity recognition (NER) algorithm and a descriptive frequency evaluation.

In the beginning, we looked into the existing literature to find a suitable ontology for the field and also searched for a job portal that provides the necessary data. After that, we implemented this tool for the job advertisement analysis. During the first observation of the results, a lack of competencies was found in the ontology (SARO). A whitelist with the missing competencies was developed, so that the NER algorithm was able to consider them too. The whitelist used in the study was based on the output of the unassigned words of the parser with SARO that was created after the first application. Once the job descriptions were collected and information extracted using SARO and the whitelist, the data was examined. In the next step, the terms (of the SARO ontology and the whitelist) were analyzed with respect to occurrences in job adverts. Based on this, conclusions were drawn about the degree to which additional found terms (whitelist) offered added value. The outcome of the tool was evaluated by related competency frameworks and by a focus group. Through this evaluation, a theoretical and practical contribution could be made. The process is illustrated in Figure 1.

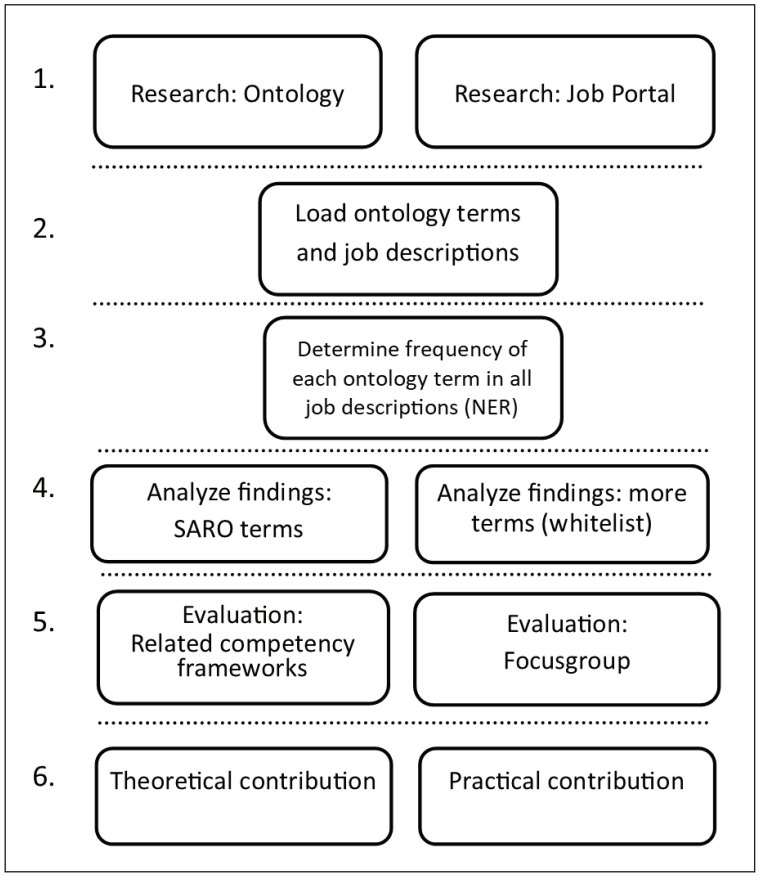

Figure 1 Structure of the research approach.

IV. JOB ADVERTISEMENT ANALYSIS

i. ARCHITECTURE

To determine data science competencies, a job advertisement analysis on available jobs on the job portal 'Jooble' was performed. The project contains different dependencies, such as data acquisition, data preparation and information extraction. Figure 2 shows the systems architecture. Furthermore, the structure and tasks are described below.

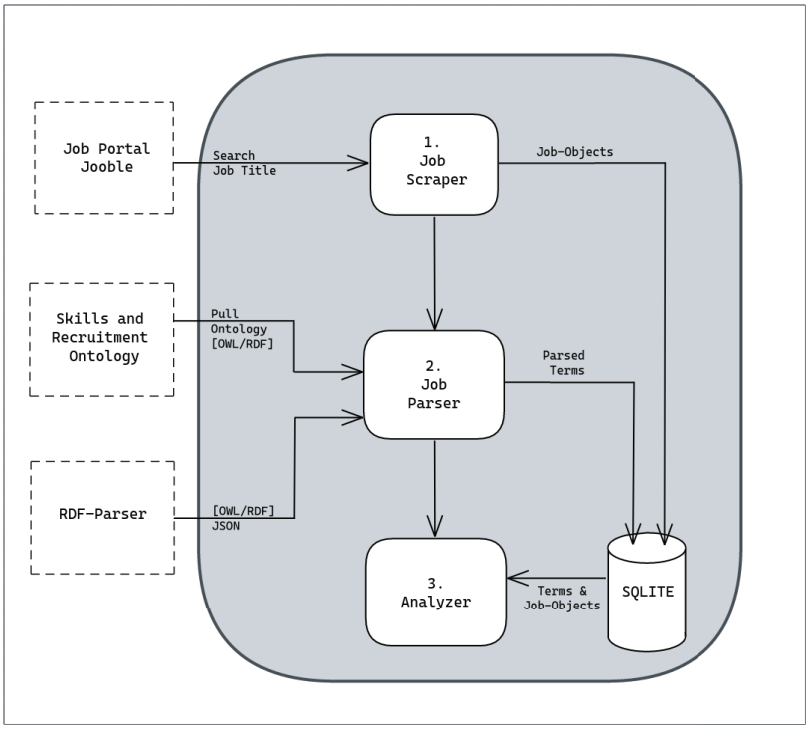

Figure 2 System architecture.

This tool is specifically made for job advertisement analyses and is composed of three independent Python scripts that communicate via a single database. The 'Scraper' collects all information about a job instance. An API and a custom HTML DOM scraper are used to get all necessary information. The results are persisted in a database. The 'Parser' is the second script. Its task is to combine the job information with an ontology. This means that the parser first queries all labels of the ontology. These are first extracted and stored in the database. Then, the parser pulls the job information and the ontology labels from the database and iterates through them in sequence, with the goal of finding the frequencies of each label from the ontology within each job posting. The information regarding frequency is also stored in the database in a further table. The third script, 'Analyzer' analyzes the previously collected data. The 'Analyzer' accesses the frequency information stored by the parser and relates and evaluates them.

ii. IMPLEMENTATION

ii.1 Scraper

In the context of this study, the scraper is used to determine and store job descriptions of previously determined, relevant job titles. The scraper is structured in two parts. The scraper uses the Jooble API as a REST interface to obtain advertised job descriptions with the search terms from the job list. Using the REST interface, the following information can already be obtained: title, location, company, salary, type, link, updated date.

Within the interface, the job description is delimited as text, so that necessary information is truncated. To get a complete description, the full HTML DOM of each job description is extracted and filtered. Thus, each individual advertisement can be found and loaded via the ID. Subsequently, the HTML DOM is filtered to obtain necessary components of the described job posting. At the end, the necessary data has been determined and stored in the database. After all job advertisements were collected, an inspection of the stored job titles was performed by examining random samples of the job descriptions in more detail to ensure the quality.

DOI: 10.5334/dsj-2023-033

ii.2. Parser

In this study, the goal of the parser is to determine how many times each term in the ontology occurs within each job advert. In general, the parser checks each job ad for how often each competence of the ontology occurs. This is done by checking for each sentence to explore whether one of the competence labels from the ontology also occurs in the job advert. When a competence is found, the counter for that competence is incremented, and the competence label is removed from the sentence.

Algorithmic 1 Parser.

```
1: function PARSE
2:     Get all skills from ontology.
3:     Sort list of skills by length descending.
4:     Get jobs.
5:     for each job do
6:         Get job description.
7:         Get all sentences of description.
8:         for each skill in ontology do
9:             Initialize each skill_occurency_counter to 0.
10:            for each sentence do
11:                for each occurence of skill in sentence do
12:                    skill_occurency_counter += 1
13:                    Remove skill string from sentence.
14:        Save job information in database.
```

First, the parser loads all ontology-related skills from the database. For the algorithm, it is necessary that the list of all skills is ordered by length, so that the longest term is the first and the shortest term is the last in the list. The reason for this is that there are whole terms that can occur as a partial string in another term. For example, in SARO, there are the terms, 'sql', 'sql server' and 'sql server 2008'. If the term 'sql server 2008' occurs in a job posting, but the term 'sql' was queried first, then this would be problematic because firstly, the term 'sql server 2008' would never be considered and second, the term 'sql' would be counted incorrectly.

Next, the jobs are taken from the database. These are then handled iteratively. First, a record tokenization is performed using the Python package 'nltk'. With the tokenization, the entire job description is divided from an entire contiguous string into individual sentences and afterwards stored in a list. Then, the parser iterates through all the sentences, checking for each to see if a term from the ontology occurs. When a term is identified, the term's counter is incremented and the term is removed from the sentence, so that other terms do not mistakenly recognize it as well (see 'sql' example above).

The parser is designed by its architecture not to be dependent on SARO. Although it was built for SARO, it can easily be used with other ontologies through optimization. For example, a whitelist with additional terms can also be added to the database without complications. The sole restriction is that it must have the structure of SARO.

ii.3 Whitelist

The whitelist used in the study is based on the output of the unassigned words of the parser with SARO that was created after the first application. To determine additional concepts, the most frequent n-grams were determined with the stagger of three, two, and one and sorted in descending order of frequency. Subsequently, the n-grams were peer-reviewed. All concepts that were found were noted in a JSON structure (whitelist) and afterwards applied. The additional found terms (output of the unassigned words) are noted in the whitelist. Later in the database, the 'class' describes from which original ontology the term originates; e.g. SARO or CUSTOM (own whitelist). This forms the basis for further analysis and evaluation in order to compare the procedures and their information content. The additional competencies are saved in the whitelist and can be seen in Table 3.

ii.4 Analyzer

For the analysis a 'Jupyter notebook' is used. In the beginning of the analysis, the collected information is loaded from the database and stored in a dataframe. For this purpose, the data is divided into four different dataframes (product, topic, tool, and skill) according to the SARO. An evaluation was developed and provides the following information:

- **Maximum frequency (max. freq):** The most frequent mention of the search term within a job description.
- **Occurrence:** Occurrence of the search term across all job descriptions.
- **Average occurrence:** Average occurrence of the search term within the job descriptions were it was found.
- **Total Average occurrence %:** Average occurrence of the search term across all job descriptions.

For the later interpretation of the analysis, the dataframes are sorted in descending order according to the percentage of occurrences across all job descriptions. As noted earlier, the parser searches for terms from the SARO in combination with the whitelist. Many of these terms have a strong similarity to each other. To give an example, the terms 'problem solving', 'problem-solving' or 'problem-solving skills' are all grouped in the cluster 'problem solving'. Furthermore, there are terms that have similar meanings and could be combined into one field. An example would be 'ab testing', 'automated testing', 'integration testing', 'regression testing', 'testing', and 'unit testing,' which can all be assigned to 'testing'. Subsequently, the terms were manually clustered and were used as a basis for further evaluation. This assignment was done by the authors.

V. EVALUATION

i. RESULTS OF THE JOB ADVERTISEMENT ANALYSIS

From the underlying 1258 terms of the SARO, 632 could be found within the job descriptions. The whitelist, which was already described in chapter ii.3, consists of an additional 206 terms. As a result, the evaluation is based on 315 Topics, 166 Skills, 206 Tools and 151 Products. Similar meanings or descriptions were combined in a cluster. If terms are not combined with other terms, then these form their own cluster. Thus, a total of 214 clusters was formed. The clustering was carried out exclusively in 'Topics' and 'Skills'. This was not necessary within 'Products' and 'Tools,' as there were no overlapping competencies.

The top 25 Topics consist of eight SARO clusters, nine clusters from the custom whitelist, and eight more that were combined into one cluster (SARO and CUSTOM). By far the most frequently mentioned cluster within a job description is 'security,' with 38 mentions. The cluster includes the following terms: 'security', 'network security', 'pki', 'web security'.

Furthermore, the cluster is referenced an average of 3166 times in relation to all of the 850 job adverts found. The most referenced Topic, with a total of 3529 references, with a result of 65% as the most frequently found cluster, is 'Programming/software development'. The cluster includes the following keywords:

> 'development', 'software development', 'tdd', 'software engineering', 'scripting language', 'bdd', 'object oriented', 'shell scripting', 'web services', 'developer', 'web', 'develop solutions', 'software solutions', 'web technologies', 'oop', 'software engineer', 'sdk', 'mobile development', 'web applications', 'web frontend', 'web application', 'web frameworks', 'android developer', 'webforms', 'reverse engineering'

The top 25 Topics are shown below, sorted in descending order by average occurrence across all job descriptions:

> 'Programming/software development', 'support', 'it', 'management', 'tool', 'engineering', 'financial experience', 'research', 'marketing experience', 'reporting', 'testing', 'network', 'sales', 'security', 'platforms', 'relationship', 'monitoring', 'frontend', 'programming', 'machine learning', 'service', 'devops', 'maintenance', 'data science', 'software architecture'

The top 25 Skills consist of two clusters of the SARO, 20 clusters that are formed from the whitelist and three more that are represented in both sources (SARO and CUSTOM). The cluster 'creative' is represented within the job descriptions, found with an average frequency of 1528. The following terms make up the cluster: 'creative', 'develop creative ideas'.

Furthermore, it can be observed that 'communication' was by far the most frequent Skill that was found within the job adverts. Of the underlying 5398 job adverts, 'communication' is mentioned 2627 times. The following terms are included in the cluster:

> 'communicate analytical insights', 'written communication', 'communications', 'communicate', 'communication', 'oral communication', 'verbal communication', 'communicate with customers'

In addition, the term 'communications' appears as the most frequent term with a maximum frequency of 32. The cluster could be identified in 48.4% of the jobs and thus forms the top one. The top 25 Skills are shown below, sorted in descending order by average occurrence across all job descriptions:

> 'communication', 'developing', 'implementation', 'problem solving', 'analyt', 'innovative', 'creative', 'collaborate', 'access', 'leadership', 'responsibility', 'technical basics', 'proactive', 'english', 'work knowledge', 'relationship', 'data management skills', 'decision making', 'mathematics', 'industry experience', 'accountability', 'commercial expirience', 'ability to manage', 'evaluation', 'optimising'

The top 25 Tools for data science are divided into 23 SARO and two custom terms. As already mentioned, clustering was not performed in tools. The free programming language for statistical computing 'R', was mentioned most often within a description with a frequency of 50. Furthermore, on average, the cloud computing platform 'Microsoft Azure' appeared most frequently with 1862 mentions per referenced job advert. The most frequently mentioned programming language, and thus also the top one, is Python with 1086 references found, which was described in one out of five job advertisements (20%). The top 25 Tools are shown below, sorted in an descending order by the average occurrence across all job descriptions.

> 'python', 'aws', 'reports', 'sql', 'azure', 'javascript', 'java', 'react', 'databases', 'api', 'rest', 'git', 'scripting', 'jenkins', 'r', 'ux', 'css', 'html', 'typescript', 'ansible', 'etl', 'jira', 'c', 'nosql', 'github'

The top 25 products for data science are divided into 20 SARO and five custom terms. The online CRM software system 'Salesforce' was mentioned with a maximum frequency of 14 references within one job description. Within 159 occurrences, the Product was mentioned on average 2082 times. The spreadsheet program 'Microsoft Excel' was identified most frequently within job descriptions. With 587 mentions, this 'Product' appeared in about one out of ten cases across all job descriptions (10.8%), making it the top product. The top 25 Products are shown below, sorted in descending order by average occurrence across all job descriptions:

> 'excel', 'go', 'docker', 'linux', 'gcp', 'fintech', 'windows', 'powerpoint', 'spark', 'salesforce', 'tableau', 'sql server', '.net', 'power bi', 'oracle', 'adobe', 'angular', 'elasticsearch', 'hadoop', 'ms office', 'mongodb', 'graphql', 'node', 'vmware', 'lambda'

ii. FOCUS GROUP

This job advertisement analysis approach is evaluated by a focus group consisting of six experts. For the focus group, a whiteboard and the SARO categories 'Products', 'Tools', 'Topics' and 'Skills' were used. Next to the results of our own job advertisement analysis approach, two further data sources were considered.

The task of the focus group was to classify each given competence in a grid regarding their importance and their trend in future. Each participant rated the competencies regarding their importance for the job role and its trendiness for the future by positioning stickies (competencies) on a grid on a digital whiteboard. One example of an output can be seen in Figure 3.

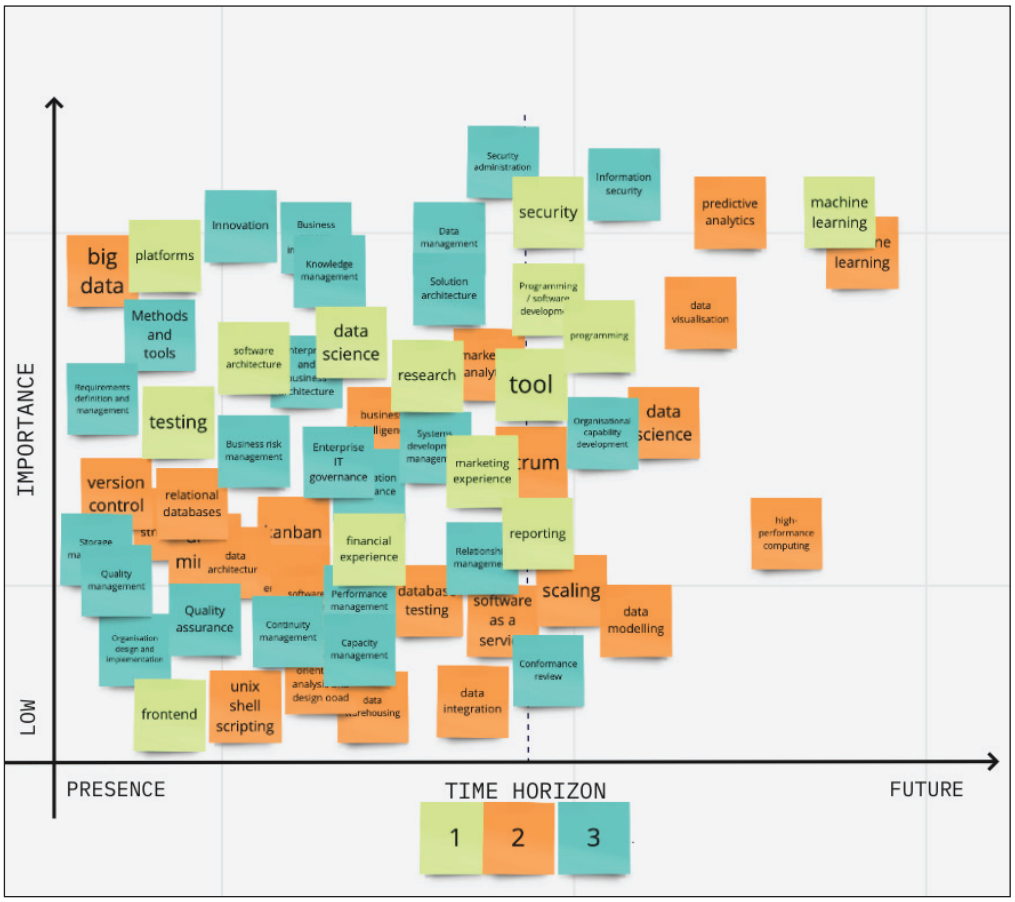

Figure 3 Experts' rating of topics' importance and trendiness.

For the focus group, the following questions were answered by the participants for each competence category/class (products, skills, topics, skills):

- What is the favorite data source?
- How comprehensible are the competencies?
- Are the most important competencies covered?

To keep the focus on each category, the classification process was split into two same segments. In the following the procedure of the focus group can be seen:

1. Introduction
2. Classification of the sticky notes on the whiteboard
3. Discussion of the results
4. Classification of the sticky notes on the whiteboard
5. Discussion of the results
6. Final discussion

The focus group consists of several members who have different backgrounds. In Table 1, an overview of the interviewees is provided. Informed consent was declared by the participants.

NO.	DESCRIPTION OF THE PARTICIPANT
E1	Professor in business information systems
E2	Professor in artificial intelligence
E3	Academic researcher, focus: positive computing
E4	Academic researcher, focus: project management, digital transformation
E5	Academic researcher, focus: e-commerce, competency model
E6	Academic researcher, focus: makerspace, digital innovation

Table 1 Participants of the focus group.

ii.1. Used data sources for evaluation

The Skill Taxonomy by Djumalieva et al. (2018) has a strength in the area of 'Products' and 'Tools' and is considered a reference in these categories. The 'Topics' and 'Skills' of this study are also considered, as an overall competence approach is taken.

Skills Framework for the Information Age (SFIA) (2018) offers a comprehensive description of competencies and has been used as a comparison in the categories 'Skills' and 'Topics'. In the case of SFIA and the Skill Taxonomy by Djumalieva et al. (2018), there was no need to shorten the list, as the number of topics and skills were appropriate and the number of topics and skills shown was equal.

Our own job advertisement analysis has hundreds of results for 'Topics', 'Skills', 'Products' and 'Tools'. The Top 25 of each category were chosen. The term 'Top' refers to the most frequently mentioned competencies.

Within the results of the focus group, some experts mentioned the different data sources by numbers that were introduced to them. The data sources used are shown in Table 2.

Table 2 Data Sources for evaluation in the focus group.

NO.	DATA SOURCE
1	This job advertisement analysis
2	Djumalieva et al. (2018)
3	SFIA (2018)

ii.2. Results

A lot of important aspects were noted by the experts. Through the evaluation, it is possible to see the strengths and weaknesses of this job advertisement analysis approach. The according expert is mentioned after each statement. First of all, the distinction between the categories 'Tools' and 'Products' is sometimes unclear (E2, E3). Furthermore, for some competencies, the context was missing (E3). In addition, the unassigned competencies were sometimes unclear or the context was missing (E1). With regard to data source one, some of the 'Products' were not comprehensible. The coverage and the scope of the competencies were assessed throughout positively (E1–E6). In addition, some new competences were added to the categories 'Products', 'Skills', and 'Tools'. For 'Products', a combination of data sources one and two is recommended by the experts, but no preference is communicated (E2, E3). Regarding the available time, it was mentioned that this was very difficult to classify and that more sticky notes were needed for this purpose (E1, E2). Moreover, timing is difficult to assess, because some competencies remain important, while others are currently important or only will be important in the future (E1). Regarding the graphical representation, there should be more space in future examinations so that the competencies are not crowded in the end (E6). Furthermore, there was a difficulty to distinguish between the categories 'Skills' and 'Topic', since it is possible to make a 'Topic' out of every 'Skill' by reformulating a competence and vice versa (E1). In addition, the 'Tools' were more difficult to classify than the 'Products' (E4), and some 'Skills' were simply 'key words' (E6). Data source one was named as the preferred one, but a combination of data sources would be more efficient (E2). In the area of 'Topics', data source two was seen as a favorite (E5) because the competencies for data source one were not clearly formulated (E5). In addition, data source two and three were seen as favorites for 'Topics' and 'Skills' (E6). All in all, a combination of data sources seems most suitable (E1). Sticky notes are a good keyword generator, but not a comprehensive description of competencies (E1). For the description of the competencies, one would need several sticky notes next to each other with 'Topic', 'Tools' and an activity verb (for example according to Bloom (1956)), so that one can describe the level of the activity (E1).

Figure 4 shows the competencies classified in the focus group according to high relevance, medium relevance and low relevance, graded by color coding of the areas 'Topic', 'Tool', 'Skill', and 'Product'. As already mentioned, the given structure did not allow any statement to be made about the chronological (time-related) classification.

Table 3 lists the competencies that were found through the whitelist and were also ranked by the focus group. These competencies were not mentioned in the frameworks SARO, SFIA or the skill taxonomy by Djumalieva et al. (2018). This implies that new competencies were found by this work.

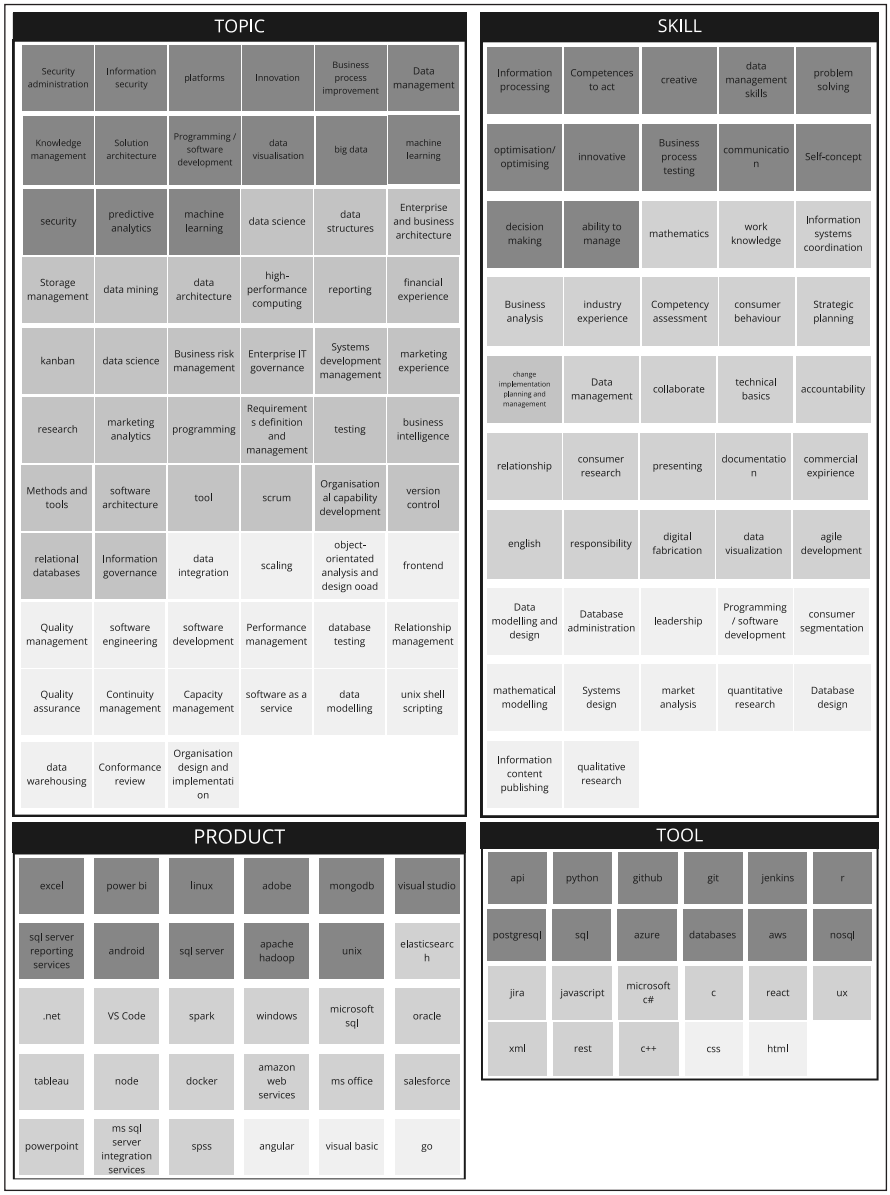

Figure 4 High, middle and low ranked relevance of competencies from focus group.

TYPE	COMPETENCY
topic	research
	marketing experience
	platforms
skill	developing
	implementation
	problem solving
	innovative
	creative
	collaborate
	leadership
	responsibility
	english
	work knowledge
	relationship
	decision making
	industry experience
	accountability
	commercial experience
	ability to manage
tools	databases
product	powerpoint
	power bi
	node

Table 3 Additional competencies found through this approach (whitelist).

VI. DISCUSSION

i. THEORETICAL CONTRIBUTION

There are different methods to extract competencies. In addition to quantitative surveys or qualitative expert interviews, there is also the procedure of job advertisement analysis, which has proven to be a very promising procedure in this work. The job advertisement analysis assumes that the truth of the necessary competencies lies within the job adverts. Since there are several thousand job advertisements in a job portal in one occupational field, it is very time-consuming to process them manually. In this respect, machine-supported algorithms are used. They allow, as this work showed, to analyze several thousand advertisements in a few hours. In the context of this study, information extraction based on an ontology was applied by means of the NER. The procedure offers the possibility to fall back on already available vocabulary in order to extract content from the texts. The resulting weakness, that only known terms can be found, can be remedied by expanding on the terms through a whitelist. The evaluation of the focus group shows that the approach serves as a basis for the development of competences.

ii. PRACTICAL CONTRIBUTION

The practical contribution is provided in the form of a base for the creation of data scientists' competency profiles. Furthermore, first steps are taken towards a foundation for the further development of curricula. The results of this job advertisement analysis can be seen as a useful 'keyword-provider' and a guideline for curriculum development. Especially for academic courses like 'data science' or 'computer science,' it is very important to prepare students for the future and to design the curricula so it is up-to-date. Through identified trending tools and products for the future, students can be optimally prepared.

iii. GENERALIZATION OF FINDINGS

This work also provides some design principles derived from feedback from experts in the evaluation. The following design principles can be formulated for the development of data scientists' competency profiles:

- **Principle of completeness**: Due to the principle of completeness, a competence should not only consist of a keyword, but ideally of an activity verb, a topic or a tool, and an associated competency level, e.g. Bloom's (1956) taxonomy, so that one can better describe and express the level of competence.
- **Principle of triangulation**: The principle of triangulation states that research should be based on different sources of data (Harper 2012). With regard to the creation of competency profiles, this means that different approaches (expert-based approach, job advertisement analysis approach, curriculum-based approach) should be combined to have a whole competency set for the role of the data scientist.
- **Principle of non-singleness**: The principle of non-singleness refers to the triangulation and can be seen as a sub-principle, especially for the job advertisement analysis approaches. 'One-word' keywords as 'IT' are often presumed as confusing, as there appears to be a lack of context. Therefore, n-gram 2 or n-gram 3 should be used instead of n-gram 1.
- **Principle of idea generation**: Competencies from job adverts can provide a good base as keyword references. The detection of missing or blind spots allows for the creation of a broader competency frame.
- **Principle of structure**: The structure of an ontology should be clear to the evaluation group, as it prevents confusion. An underlying meaningful ontology can facilitate the evaluation process and its quality.
- **Principle of understandability**: Stems should not be used as a competence. Therefore, careful handling should be applied with stemmed words, as they are not understandable, e.g. 'analyst'. The context of a competency should also be clear to make it understandable.

VII. SUMMARY

i. CONCLUSION

In this paper, a job advertisement analysis was used to extract competencies for the field of data science. Through the analysis, it was possible to highlight the most required skills, topics, tools, and products that are requested by the industry. The first step in the analysis was a NER to match over 1200 terms from the SARO in over 5000 job postings. After that, the findings were analyzed by frequency and occurrence, so it became possible to rank all competencies. The most commonly occurring competencies were evaluated by a focus group next to the competency frameworks SFIA and the findings of Djumalieva et al. (2018).

Through the result of the focus group, a list of competencies for the professional field of 'Data Science' was created. This competence profile provides an overview of the competences required by the industry in the categories 'Products', 'Tools', 'Topics,' and 'Skills'. The results of our own created job advertisement analysis can be seen as a useful 'keyword-provider' and a guideline for curriculum development. Especially for the courses 'data science' or 'computer science,' it is important to prepare students for the future and make curricula up-to-date. Through identified trending tools and products for the future, students can be optimally prepared. It should be noted that this competence profile is only a snapshot in time. Any statement about changes, trends, or forecasts cannot be made by this analysis alone.

Another result besides the competence profile are six design principles: 'Principle of completeness', 'Principle of triangulation', 'Principle of non-singleness', 'Principle of ide generation', 'Principle of structure,' and 'Principle of understandability'. The design principles should be taken into account in the development of competency models. Furthermore, they can contribute to the development of a competence theory.

ii. FUTURE WORK

For the future, the conduction of job advertisement analyses with different job-portals is planned. This should be done within the same time frame to address issues of comparability. Also, further development of the SARO or the development of an own ontology with clusterings and categorizations should be done. Another option for future works is to use some of the semantics provided by the SARO ontology (relation between concepts), instead of taking the concepts as they are. Additionally, more focus groups with a focus on free economy should be conducted. Therefore, more job advertisement analyses must be conducted in different regions and job portals. Furthermore, through the conduction of a trend analysis, the skill-need-differences throughout the years can be better tracked. Thereby, developments can be identified through trend analyses. At this point, a combination of data sources could be helpful.

Lastly, the development of an own competency framework based on results from the job advertisement analysis, feedback from experts, and expert-based approaches is conceivable. Therefore, competency levels should be considered to have a complete competency framework.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Jan Vogt
Institute of Computer Science, University of Applied Science Ruhr West, Bottrop, Germany

Thilo Voigt
Institute of Computer Science, University of Applied Science Ruhr West, Bottrop, Germany

Annika Nowak
Institute of Computer Science, University of Applied Science Ruhr West, Bottrop, Germany

Jan M. Pawlowski
Institute of Computer Science, University of Applied Science Ruhr West, Bottrop, Germany

REFERENCES

Almaleh, A, et al. 2019. Align My Curriculum: A Framework to Bridge the Gap between Acquired University Curriculum and Required Market Skills. *Sustainability*, 11(9): 2607. DOI: https://doi.org/10.3390/su11092607

Bloom, BS. 1956. *Taxonomy of educational objectives: The classification of educational goals*. New York: David McKay Company.

Boselli, R, et al. 2018. WoLMIS: a labor market intelligence system for classifying web job vacancies. *Journal of Intelligent Information Systems*, 51(3): 477–502. DOI: https://doi.org/10.1007/s10844-017-0488-x

Breitfuss, G, et al. 2019. The Data-Driven Business Value Matrix-A Classification Scheme for Data-Driven Business Models. *Bled eConference*, 19. DOI: https://doi.org/10.18690/978-961-286-280-0.42

Cao, L. 2017. Data science: challenges and directions. *Communications of the ACM*, 60(8): 59–68. DOI: https://doi.org/10.1145/3015456

Dadzie, AS, et al. 2018. Structuring visual exploratory analysis of skill demand. *Journal of Web Semantics*, 49: 51–70. DOI: https://doi.org/10.1016/j.websem.2017.12.004

Debortoli, S, Müller, O and **vom Brocke, J.** 2014. Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, 6(5): 289–300. DOI: https://doi.org/10.1007/s12599-014-0344-2

Djumalieva, J, Lima, A and **Sleeman, C,** et al. 2018. Classifying occupations according to their skill requirements in job advertisements. *Economic Statistics Centre of Excellence Discussion Paper*, 4: 2018.

Donoho, D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766. DOI: https://doi.org/10.1080/10618600.2017.1384734

Harper, R. 2012. The collection and analysis of job advertisements: A review of research methodology. *Library and Information Research*, 36(112): 29–54. DOI: https://doi.org/10.29173/lirg499

Hattingh, M, et al. 2019. Data Science Competency in Organisations: A Systematic Review and Unified Model. *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*, 1–8. DOI: https://doi.org/10.1145/3351108.3351110

Khaouja, I, et al. 2019. Building a soft skill taxonomy from job openings. *Social Network Analysis and Mining*, 9(1): 1–19. DOI: https://doi.org/10.1007/s13278-019-0583-9

Khobreh, M, et al. 2015. An ontology-based approach for the semantic representation of job knowledge. *IEEE Transactions on Emerging Topics in Computing*, 4(3): 462–473. DOI: https://doi.org/10.1109/TETC.2015.2449662

Von Konsky, B, Miller, C and **Jones, A.** 2016. The skills framework for the information age: Engaging stakeholders in curriculum design. *Journal of Information Systems Education*, 27(1): 37.

Lima, A, Bakhshi, B, et al. 2018. Classifying occupations using web-based job advertisements: an application to STEM and creative occupations. *Economic Statistics Centre of Excellence Discussion Paper*, 8: 2018.

Mandinach, EB, et al. 2015. Ethical and appropriate data use requires data literacy. *Phi Delta Kappan*, 96(5): 25–28. DOI: https://doi.org/10.1177/0031721715569465

Murawski, M and **Bick, M.** 2017. Demanded and imparted big data competences: towards an integrative analysis. In *Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal, June 5–10, 2017, 1375–1390.

National Academies of Sciences, Engineering, Medicine (NASEM). 2018. Envisioning the data science discipline: the undergraduate perspective: interim report. National Academies Press.

Ridsdale, C, Rothwell, J, Smit, M, Ali-Hassan, H, Bliemel, M, Irvine, D, Kelley, D, Matwin, S and **Wuetherick, B.** 2015. Strategies and best practices for data literacy education. Knowledge synthesis report. SSHRC. DOI: https://doi.org/10.13140/RG.2.1.1922.5044

Saltz, J, Armour, F and **Sharda, M.** 2018. Data science roles and the types of data science programs. *Communications of the Association for Information Systems*, 43(1): 33. DOI: https://doi.org/10.17705/1CAIS.04333

SFIA. 2018. The global skills and competency framework for a digital world. Available at: https://sfia-online.org/en/sfia-7 [Last accessed on 4 April 2021].

Shirani, A. 2016. Identifying Data Science and Analytics Competencies Based on Industry Demand. *Issues in Information Systems*, 17(4).

Sibarani, E, et al. 2020. *Skills and Recruitment Ontology*.

Sibarani, EM, et al. 2017. Ontology-guided job market demand analysis: a cross-sectional study for the data science field. *Proceedings of the 13th International Conference on Semantic Systems*, 25–32. DOI: https://doi.org/10.1145/3132218.3132228

Silveira, CC, et al. 2020. What is a Data Scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias–IPTEC*, 8(1): 25–39. DOI: https://doi.org/10.5585/iptec.v8i1.17263

Wowczko, IA. 2015. Skills and vacancy analysis with data mining techniques. *Informatics*, 2(4): 31–49. DOI: https://doi.org/10.3390/informatics2040031

Zhao, M, et al. 2015. SKILL: A system for skill identification and normalization. *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*, 4012–4017. DOI: https://doi.org/10.1609/aaai.v29i2.19064

TO CITE THIS ARTICLE:
Vogt, J, Voigt, T, Nowak, A and Pawlowski, JM. 2023. Development of a Job Advertisement Analysis for Assessing Data Science Competencies. *Data Science Journal*, 22: 33, pp. 1–16. DOI: https://doi.org/10.5334/dsj-2023-033

Submitted: 25 August 2021
Accepted: 11 April 2023
Published: 07 September 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.