Human Interaction and Emerging Technologies (IHIET 2023), Vol. 111, 2023, 80–90
https://doi.org/10.54941/ahfe1004012

The Impact of Automation Frameworks on Today's Data Science Competencies

Maria Potanin, Maike Holtkemper, and Christian Beecks

Chair of Data Science, University of Hagen, Germany

ABSTRACT

Among the many digital competences required in today's working environment, data scientific competencies like data-preprocessing, feature-engineering and model-generation are essential for the analysis of small-to-big data sets arising in different application scenarios. Since the combination of these competencies is in high demand on the job market, as a small number of highly talented employees cannot meet this demand, intelligent solutions are being sought that can support some of these competencies. The goal of this paper is to investigate whether modern automation frameworks have data science competencies in their design. It serves as an impulse for further discourse on whether modern automation frameworks can empower humans in acquiring data science competencies.

Keywords: Data science competencies, Automation frameworks, EDISON data science competence framework (CF-DS)

INTRODUCTION

The digital transformation has led to a rapid advancement of today's working environment (Hinz and Heinen, 2021). The increasing demand for AI solutions, especially in skilled professions, is deepening the relationships between humans and smart machines (Bakhshi et al., 2017). These new relationships require an evolution and transformation of the skills of today's workforce. Furthermore, technologies can be used to close potential skills gaps and thus make the interaction and collaboration between human and technology work in the best conceivable way (Berger and Frey, 2016). Among the many digital competences required in today's working environment, data scientific competencies like data-preprocessing, feature-engineering and model-generation are essential to manage and analyze small-to-big data sets arising in various data spaces (Jaiswal et al., 2022).

With the amount of data increasing every day, there is a growing demand for subject matter experts with data science (DS) skills to help companies to be competitive by making data-based decisions, identify future trends, identify patterns, optimize processes, or develop new products and services (Sturm, 2023). Starting with the ability to write in multiple programming languages, to apply mathematical and statistical concepts, to practice machine learning algorithms and techniques, to understand industry-specific requirements, to critically question and visualize results, to the ability to work in a team and to interact socially. These various requirements are met by employees, each of whom is not capable of handling such a wide range of competencies. Since the combination of these competencies is in high demand on the job market, intelligent solutions are being sought that can support or maybe covered some of these competencies completely (Necula, 2023). The goal of this paper is to investigate whether modern automation frameworks have data science competencies in their design. Based on the EDISON Data Science Competence Framework (Demchenko et al., 2017), we thus pose the following research question: *From a theoretical point of view, can modern automation frameworks be used to cover data science competencies?* First, we identify relevant automation frameworks which match the data science competencies of the EDISON Data Science Competence Framework (CF-DS) and second, we investigate which competencies of the CF-DS could theoretically be supported or covered by automation frameworks. We provide a subsequent overview of the identified competencies that practitioners can consult when addressing existing or foreseeable DS competency needs.

DATA SCIENCE COMPETENCIES AND AI

Today, the term competence is part of our everyday vocabulary. Which facts and constructs it includes, however, depends on the respective context and research discipline (Gogolin et al., 2008). In this research, competencies shall be considered as so-called job components (e-CF 2014), i.e. parts of the job profile of a data scientist. This is in line with the European e-Competence (e-CF) definition, where a competence is the "*demonstrated ability to apply knowledge, skills, and attitudes for achieving observable results*" and refers to job-related activities (e-CF 2014, p. 5). The authors relate this concept to human behavior, which is not the case in older competency definitions such as Butler (1978) or Gale and Pol (1975). In this work we thus assume that a competence refers to an actor (human or technology) in general.

The investigated competencies relate to the data scientist profession. Data scientists are defined as practitioners who have sufficient knowledge in the areas of "*business needs, domain knowledge, analytical skills, and software and systems engineering*" and manage the entire data process along the analysis lifecycle (Chang and Grady 2019, p. 28). A detailed overview of data science competencies is provided by the CF-DS from a European Union Project (Demchenko et al., 2017). It was developed from the analysis of existing frameworks for ICT and data science competencies and skills as well as the formulated demand for the profession of data scientist from industry and research (Demchenko et al., 2017). From these analyses, the authors identified three core competencies (DS Analytics, DS Engineering, Domain Knowledge) and two additional sets of competencies that (Data Management, Research Methods/BMP) (Demchenko et al., 2017). Table 1 lists the definitions of the five competence groups.

Table 1. EDISON's competency groups and definitions (adapted from Demchenko et al., 2017, p. 16).

Competency	Description
DS Analytics	Use appropriate data analytics and statistical techniques on available data to discover new relations and deliver insights into research problem or organizational processes and support decision-making.
DS Engineering	Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.
Data Management	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.
Research Methods/ Business Process Management	Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals.
Domain Knowledge	Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations.

In a representative study, activities like processing data (69%) or collecting data (64%) have the highest potential to be automated (Manyika, 2017). Although the importance of skills required to use and develop AI systems are well known, there is a lack of focus on developing these required skills among future employees (Pumplun et al., 2019). Recent developments have shown that research on AI Literacy can help to expand the areas of competency and prepare students with foundational knowledge for the AI workplace (Faruque et al., 2021).

The increasing demand for data experts and the simultaneous shortage of skilled workers make it more difficult for companies to find the "engineer of the future" (van der Aalst, 2014). With the help of automation frameworks, time-consuming tasks can be significantly reduced, and information can be delivered in time (Abbaszadegan and Grau, 2015). Furthermore, automating testing procedures can reduce costs, optimize the functionality of software modules and, in the case of the Robot Framework, shows that automated testing is on average 80.46% faster than manually performed testing (Chakravarthy and Padma, 2023). It shows that there has already been a progress in supporting human through modern tools in their daily work.

RESEARCH METHOD

To address the research questions, a literature review is conducted based on the recommendations of Webster and Watson (2002), including keyword and reverse search. The process is documented according to the procedure of vom Brocke et al. (2009). In Figure 1, the used approach for this literature analysis is shown.

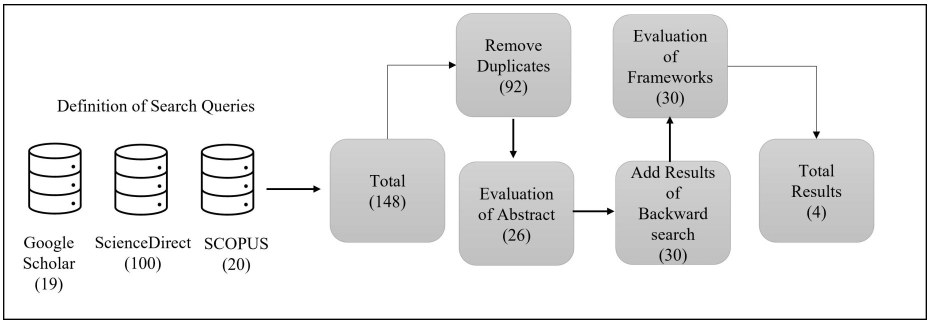

Figure 1: Literature review process.

The scientific literature databases Google Scholar, ScienceDirect and Scopus are used. The search space was narrowed down by using the five competence groups, data analytics, data engineering, data management, research methods & project management, as well as business analytics, from the CF-DS as a starting point for the search for automation frameworks with search strings e.g., "automated data analytics". The period was limited to 2017 when the CF-DS was published. This literature review process led to 148 articles. After removing duplicates, 92 relevant articles were evaluated by their abstract. Adding the 4 results from the backward search, 30 articles were evaluated by following characteristics: automation, flexibility, ease of use and interoperability. The remaining frameworks for the further qualitative content analysis are AutoPrep, AutoGluon, AutoClust and DeepEye.

For the qualitative content analysis, inductive category formation as described in Kuckartz (2022) is applied. Our aim is to find out which competencies or skills are described in the concepts of the automation frameworks. For this purpose, the category system is formed based on the structure of the CF-DS: The main categories correspond to the competence groups in Table 1, and the subcategories correspond to the more detailed descriptions or subdivisions (e.g., DSDA01, Appendix A). In addition, the category *Others* was added for possible further information. As recommended by Kuckartz (2022), to improve the quality of the results, all documents were coded by two of the authors and the results were merged. The assignment of competencies used in the CF-DS to the CRISP-DM process model was added, from which individual process steps such as data preprocessing emerge more clearly than from the actual CF-DS[1]. A total of 96 codes were identified, distributed as follows: Data Analytics with 88 codes, Data Science Engineering with one code and Data Management with four codes. No information could be identified for the other competence groups.

DATA SCIENCE COMPETENCIES OF AUTOMATION FRAMEWORKS

AutoClust is a clustering framework from the field of automated machine learning (AutoML) and uses an unsupervised learning approach. Like

[1] See EDISON CF-DS p. 41 for competence mapping

all AutoML frameworks, AutoClust aims to minimize human manual intervention during machine learning tasks. The framework focuses on two main modules: automated algorithm selection and hyperparameter tuning. (Poulakis et al., 2020)

Examination of the codes revealed that AutoClust uses analysis techniques to make *"automatic clustering algorithm selection based on meta-learning and cluster validity indices"* (Poulakis et al., 2020, p. 1220). This can be mapped to the capabilities of the DSDA01 competency of the EDISON framework. However, this does not correspond to the variety of analysis techniques such as supervised and unsupervised learning mentioned there and thus also required in the context of the entire data lifecycle. We designate thereby a capability as automated, understand this however thereby only as supporting capability of a necessary, existing human authority, which must fall back if necessary on other analysis techniques manually.

In its description, the authors of AutoClust specify two capabilities that can be assigned to statistics (DSDA02). This includes Bayesian optimization techniques and *"regression of cluster validity indices"* (Poulakis et al., 2020, p. 1220). Regression is used as an optimization criterion to vote on hyperparameters of clustering algorithms (Poulakis et al., 2020). The authors mention two experiments which they used for model validation (DSDA04). They show that AutoClust "*usually outperforms the common practice of running multiple clustering algorithms and keeping the best*" (Poulakis et al., 2020, p. 1224) and refer directly to a comparison objective they defined that *"simulate the case of a data science practitioner selecting an index from the available indices for evaluating clustering results*" (Poulakis et al., 2020, p. 1223).

AutoGluon-Tabular is an AutoML framework that trains machine learning models on an unprocessed tabular dataset. It requires only one line of Python code. (Erickson et al., 2020)

The authors of AutoGluon-Tabluar state that their framework is useful for beginners as well as experts. This gives a first hint that a competence could involve the interaction between the user and the framework. Things get more concrete when it comes to the possibilities offered to experts. Here it is stated that they have to implement the "*best ML practices only once*" (Erickson et al., 2020, p. 1). This allows the person's applied knowledge to be applied to many problems without the need for frequent manual intervention (Erickson et al., 2020). The mentioned skills like strategies of model selection, preprocessing of data or hyperparameter tuning refer to different competencies of the EDISON framework, but can be assigned to the Data Analytics group.

However, the mentioned capabilities alone are not sufficient to fully cover a competency of the group. The individual competencies can be supported by the automated skills as follows, with the core capability being DSDA04. During model validation, AutoGluon is able to automatically process raw data, identify the prediction problem (e.g., regression), and then partition the data for model training (Erickson et al., 2020). Finally, an *"optimized model* *ensemble that outperforms any of the individual trained models"* (Erickson et al., 2020, p. 2) is created. Besides model validation, which was identified as the core capability of AutoGluon, other text passages could be assigned to different competencies (DSDA01, DSDA02, DSDA03, DSENG, DSDM03, DSDM05). Most of them could rather be interpreted as skills or at least partial aspects of a skill. This includes for example different tasks of data preprocessing like "*to deal with missing discrete variables*" (Erickson et al., 2020, p. 3), where no value is estimated/inserted, but an additional strategy *unkown* is created (Erickson et al., 2020).

DeepEye aims to automate data visualization to avoid heavy user intervention compared to existing systems. In doing so, DeepEye should be able to determine whether a visualization of a data set is of interest from a human perspective and whether a data set that is not of interest becomes interesting through operations such as aggregation. Timely feedback of results is also relevant (Qin et al., 2018).

A number of passages in the article could be identified that match the following competencies: DSDA02, DSDA03, DSDA04, and DSDA06. In the broadest sense, consideration of data characteristics can be interpreted as a kind of statistical skill when the *"correlation between two columns"* (Qin et al., 2018, p. 77) needs to be determined (e.g., positive linear, log, exponential, etc.). The ability to transform data "*by using appropriate transformation operators*" (Qin et al., 2018, p. 79) can be assigned to competency DSDA03. This is interpreted as a skill of competence, which means that it cannot cover all areas of DSDA03. For model validation (DSDA04), DeepEye uses a decision tree to *"determine whether a visualization is good"* (Qin et al., 2018, p. 79). The backend then passes on a ranking of the top-k best visualizations (Qin et al., 2018). This ability can be evaluated as automated, which according to EDISON does not cover the whole competence, however depending upon employment as sufficiently and thus no further model validation would be necessary would be considered.

Most of the coded text passages refer to the core aspect of the paper, namely data visualization (DSDA06). DeepEye states the main task of visualization is storytelling by "*selecting, filtering, and transforming the data, and picking the right visualization type such as bar charts or line charts*" (Qin et al., 2018, p. 75) from a data set. Not only visualization, but also dashboard design is anchored in the front-end (Qin et al., 2018). Storytelling is also mentioned as an essential aspect, but this cannot be interpreted as a competence or ability according to EDISON. On the one hand, no concrete storytelling methods are mentioned, on the other hand, the story of the data visualization is not only dependent on the user, but also on the recipient to whom the user may want to present the results. DeepEye's description of this capability can rather be understood as a combined competency, as the user could obtain meaningful stories by "*interacting with the visualization result*" (Qin et al., 2018, p. 79). With reference to the definition of a competency used, it can be concluded here that data visualization and dashboard design skills produce observable results, but this is probably not possible with storytelling.

AutoPrep is a simple Python-based auto-preprocessing architecture designed for automated machine learning. The framework provides automated and interactive support for users to perform data preprocessing tasks (Bilal et al., 2022).

The competency interpretation of the framework focuses on data preprocessing, as its name suggests (DSDA03). AutoPrep thus specifically addresses standard tasks that arise in this area. These include identification of the data problem followed by visualization for the user (Bilal et al., 2022). It then "*recommends the most effective data cleaning and preparation method to the user after evaluating the state-of-the-art candidate techniques*" (Bilal et al., 2022, p. 107764). After this step, AutoPrep automates the preprocessing so that the dataset can then be used by any ML module (Bilal et al., 2022). This includes: "*comprise missing values, Categorial attributes encoding, features scaling, and features reduction*" (Bilal et al., 2022, p. 107765).

Again, from a theoretical point of view, the entire competency DSDA03 of the EDISON framework cannot be covered, since the competency definition is comprehensive, goes far beyond the data pre-processing part and already starts with the identification of relevant heterogeneous data. Further statements could be assigned to the competencies DSDA01, DSDA04, DSDA05 and DSDM05, which could for example represent skills in the area of the use of metrics for model validation or contribute to the assurance of data quality.

Table 2 summarizes our findings and shows a subsequent overview of the competence groups from the EDISON CF-DS to which at least one description of a theoretically suitable ability could be assigned. Competencies in the area of research methods, project management or domain knowledge could not be identified, which is why they were not included in the presentation. The foci of the automation frameworks provide the most theoretical support for the respective competencies (c). A framework was also mapped to other competencies as providing support if the description of at least one other capability emerged from the text and was interpreted as one.

These results show that based on the frameworks selected from the literature review, almost exclusively competencies from the field of data analysis could be supported in an automated way, but none can be completely replaced by an AutoFM. Despite the broad definition of a competence, the individual descriptions of the EDISON CF-DS still cover a broader spectrum of skills and knowledge, which cannot theoretically be covered by the frameworks. Here, the consideration of a combined competence could be made, whereby humans and technology contribute skills and only through this interaction is a competence covered.

Table 2. Results of matching competence groups & automation frameworks.

Competence Group / Competence	Data Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management (DSDM)
01	**AutoClust** (C) AutoPrep, AutoGluon		
02	AutoGluon, AutoClust, DeepEye		
03	**AutoPrep** (C) AutoGluon, DeepEye		AutoGluon, DeepEye
04	**AutoGluon** (C) AutoPrep, DeepEye		
05			AutoPrep, AutoGluon
06	**DeepEye** (C)		
undefined		AutoGluon	

(C) = Core function of framework

CONCLUSION

This paper provides an initial overview of how automation frameworks can demonstrate data science capabilities based on their conceptual descriptions. We show which competencies could benefit from AutoFMs and their skills. Using a systematic literature review, we identified four automation frameworks based on the competency groups from the EDISON competency framework for Data Science: AutoClust, AutoPrep, DeepEye, and AutoGluon-Tabular. In a qualitative content analysis, we identified skills that could be mapped to different DS competencies. We show that almost exclusively skills from the Data Analytics competency group were automated, but never sufficiently to cover the entire competency. The use of AutoFMs can therefore support in different areas, but still requires the human in the loop.

Scientists can use these results to research a possible construct definition of a combined competence of humans and AI, to investigate further frameworks for their capabilities and competencies, or to compare these results in case studies when using AutoFMs in practice. Practitioners can use the subsequent overview to find out which data analysis competencies can be supported by AI or AutoFMs, e.g., to close capability gaps.

Limitations of the results are on the one hand the pure interpretation of theoretical contents, since the papers of the respective frameworks served as data basis and it does not concern observations or protocols with the employment of the frameworks in practice, as e.g. with a case study. The described "capabilities" of the frameworks could lead to different results when used in practice if they do not work as described. The presentation of the subsequent is also not complete for all common AutoFMs, but is to be considered exemplary based on the results of the literature research.

APPENDIX A: EDISON COMPETENCES DSDA, DSENG AND DSDM

Data Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management (DSDM)
DSDA01: Effectively use variety of data analytics techniques, such as ML, Data Mining, Prescriptive, Predictive Analytics, for complex data analysis through the whole data lifecycle	DSENG01: Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis, and visualisation	DSDM01: Develop & implement data strategy, in particular, in a form of data management policy & Data Management Plan (DMP)
DSDA02: Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction	DSENG02: Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with the special focus on Big Data technologies for large datasets and cloud-based data analytics platforms	DSDM02: Develop and implement relevant data models, define metadata using common standards and practices, for different data sources in variety of scientific and industry domains
DSDA03: Identify, extract, and pull together available and heterogeneous data, including modern data sources such as social media data, open data, governmental data	DSENG03: Develop & prototype specialised data analysis applications, tools & supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch & streaming processing platforms, including online & cloud-based solutions for on-demand provisioned and scalable services	DSDM03: Integrate heterogeneous data from multiple source and provide them for further analysis and use
DSDA04: Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	DSENG04: Develop, deploy & operate large scale data storage & processing solutions using different distributed & cloudbased platforms for storing data (e.g. DataLakes, MongoDB, Accumulo, DynamoDB, others)	DSDM04: Maintain historical information on data handling, including reference to published data and corresponding data sources
DSDA05: Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	DSENG05: Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	DSDM05: Ensure data quality, accessibility, interoperability, Compliance to standards, & publication (data curation)
DSDA06: Visualise results of data analysis, design dashboard and use storytelling methods	DSENG06: Design, build, operate relational & nonrelational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	DSDM06: Develop and manage/ supervise policies on data protection, privacy, IPR and ethical issues in data management

REFERENCES

Abbaszadegan, A.; Grau, D. (2015): Assessing the influenece of automated data analytics on cost and schedule performance. Procedia Engineering 123, p. 3–6.

Bakhshi, H.; Downing, J. M.; Osborne, M. A.; Schneider, P. (2017): The Future of Skills: Employment in 2023. Pearson and Nesta. Futuresikills.pearson.com.

Berger, T.; Frey, C. B. (2016): Digitalization, Jobs and Convergence in Europe: Strategies for Closing the Skills Gap. Oxford Martin School. Prepared for the European Commission.

Bilal, M.; Ali, G.; Iqbal, M. W.; Anwar, M.; Malik, M.; Kadir, R. (2022): Auto-Prep: Efficient and Automated Data Preprocessing Pipeline. IEEE Access, Volume 10, p. 107764–107784.

Butler, F. C. (1978): The Concept of Competence: An Operational Definition. Educational Technology 18 (1), p. 7–18. Available online at https://www.jstor.org/stable/44418395.

Chakravarthy, N. A.; Padma, U. (2023): A Comprehensive Study of Automation Using a WebApp Tool for Robot Framework. In: Intelligent Cyber Physical Systems and Internet of Things, ICoICI 2022.

Chang, W. L.; Grady, N. (2019): NIST Big Data Interoperability Framework: Volume 1, Definitions. Available online at https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions.

Demchenko, Y./Belloum, A./Wiktorski, T. (2017). Data Science Competence Framework (CF-DS). EDISON Project. Available online at https://EDISON-project.eu/data-science-competence-framework-cf-ds/ (accessed 4/16/2023).

e-CF (2014): European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. Available online at https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/european-e-competence-framework-e-cf (accessed 4/16/2023).

Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li M.; Smola, A. (2020): AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv:2003.06505.

Faruque, F.; Watkins, R.; Medsker, L. (2021): Competency Model Approach to AI Literacy: Research-based Path from Initial Framework to Model. arXiv preprint arXiv:2108.05809.

Gale, L. E.; Pol, G. (1975): Competence: A Definition and Conceptual Scheme. Educational Technology 15 (6), p. 19–25. Available online at https://www.jstor.org/stable/44417993.

Gogolin, I.; Krüger, H.; Prenzel, M. (2008): Kompetenzdiagnostik. Wiesbaden, VS Verlag für Sozialwissenschaften.

Hinz, J.-R., Heinen, M. (2021): EY Jobstudie 2021: Digitalisierung im Arbeitsleben (accessed 4/16/2023).

Jaiswal, A.; Arun, C. J.; Varma, A. (2021): Rebooting employees: upskilling for artificial intelligence in multinational corporations. The International Journal of Human Resource Management.

Kuckartz, U. (2018) Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung. Beltz Verlagsgruppe, Weinheim.

Manyika, J.; Chui, M.; Miremadi, M.; Bughin, J.; George, K.; Willmott, P.; Dewhurst, M. (2017): A Future that Works: Automation, Employment, and Productivity. McKinsey & Company, McKinsey Global Institute.

Necula, S.-C. (2023): AI Impact on the Labour Force-Searching for the Analytical Skills of the Future Software Engineers. arXiv:2302.13229.

Poulakis, Y.; Doulkeridis, C.; Kyriazis, D. (2020): AutoClust: A Framework for Automated Clustering Based on Cluster Validity Indices. 2020 IEEE International Conference on Data Mining (ICDM).

Pumplun, L.; Tauchert, C.; Heidt, M. (2019): A New Organizational Chassis for Artificial Intelligence – Exploring Organizational Readiness Factors. ECIS 2019 Proceedings, p. 19.

Qin, X.; Luo, Y.; Tang, N.; Li, G. (2018): DeepEye: An automatic big data visualization framework. Big Data Mining and Analytics. Volume 1, Issue 1, p. 75–82.

Sturm, T. (2023): Exploring Human and Artificial Intelligence Collaboration and Its Impact on Organizational Performance: A Multi-Level Analysis. Dissertation. Technische Universität Darmstadt.

Van der Aalst, W. (2014): Data Scientist: The Engineer of the Future. Enterprise Interoperability VI., p. 13–26.

Vom Brocke, J.; Simons, A.; Niehaves, B.; Riemer, K.; Plattfaut, R.; Cleven, A. (2009): Reconstructing the giant: on the importance of rigour in documenting the literature search process. European Conference on Information Systems (ECIS) 2009.

Webster, J.; Watson R. T. (2002): Analyzing the past to prepare for the future: Writing a literature review. MIS Quaterly.