




Preparing for the future of work: a novel data-driven approach for the identification of future skills

Julia Brasse¹ · Maximilian Förster¹ · Philipp Hühn¹ · Julia Klier²  · Mathias Klier¹ · Lars Moestue¹

Accepted: 19 June 2023 / Published online: 24 July 2023
© The Author(s) 2023

Abstract

The future of work is changing rapidly as result of fast technological developments, decarbonization and social upheavals. Thus, employees need a new skillset to be successful in the future workforce. However, current approaches for the identification of future skills are either based on a small sample of expert opinions or the result of researchers interpreting the results of data-driven approaches and thus not meaningful for the stakeholders. Against this background, we propose a novel process for the identification of future skills incorporating a data-driven approach with expert interviews. This enables identifying future skills that are comprehensive and representative for a whole industry and region as well as meaningful for the stakeholders. We demonstrate the applicability and utility of our process by means of a case study, where we identify 33 future skills for the manufacturing industry in Baden-Wuerttemberg, Germany. Our work contributes to the identification of comprehensive and representative future skills (for whole industries).

Keywords Future skills · Future skill identification · Job advertisement analysis · Case Study

JEL Classification J24 · M00

✉ Julia Klier
julia.klier@informatik.uni-regensburg.de

¹ Institute of Business Analytics, University of Ulm, Helmholtzstr. 22, D-89081 Ulm, Germany

² Department of Management Information Systems, University of Regensburg, Universitätsstraße 31, D-93053 Regensburg, Germany

1 Introduction

The pace of rapid technological development, decarbonization, and social and demographic upheavals will fundamentally change the future of work (Leopold et al. 2016). Digitalization is affecting every aspect of people's lives and is happening so fast that today's period is said to be the fourth industrial revolution (Leopold et al. 2016). Indeed, organizations need to build new capabilities to respond to external challenges (Wirtky et al. 2016). This requires a majority of employees to acquire new skills (Klus and Müller 2021; Zahidi et al. 2020). Overall, companies expect that 40% of employees worldwide will need to learn new skills by 2025 (Zahidi et al. 2020). The challenges ahead require a society-wide effort and proactive adaptation by companies, governments, and individuals in dealing with this significant skill shift.

It is critical that companies actively support their workforce through up- and re-skilling, that individuals pursue lifelong learning, and that governments entail innovative education and labor-related policymaking (Heuser et al. 2022; Leopold et al. 2016). For this purpose, it is crucial to have transparent knowledge about the skills needed to cope with future challenges (Zahidi et al. 2020). Human resource management (HRM) is indispensable for organizations to be successful in the future (Wirtky et al. 2016). Especially against the background of the enormous amount of data available in this field (Buchmann et al. 2022), e-HRM in terms of "systems that support operative, relational and transformational HR functions" (Ilek et al. 2022) seems particularly promising. However, to date this potential is barely leveraged and little is known about HR planning that uses external market data such as electronic marketplaces or external recruiting websites (Wirtky et al. 2016). Against this background, we aim to identify future skills based on a data-driven approach that enables organizations and individuals to adapt to the future workforce and the transformation of society. Following the Design Science Research (DSR) methodology (Hevner et al. 2004), we propose a novel artifact to identify future skills, i.e. the future skills identification process. The future skills identification process integrates and expands existing data-driven and expert-based concepts into a transparent and replicable approach to identify future skills for a given region and industry. It includes data-driven derivation of future skills based on a large number of job advertisements to identify a comprehensive and representative set of future skills. Furthermore, it integrates stakeholders as experts to determine the level of aggregation of future skills which ensures that the results are meaningful to stakeholders. We incorporate these concepts into a process that systematically guides the identification of future skills, inspired by the well-established standard process for data mining in research and practice CRISP-DM (Martinez-Plumed et al. 2019). We demonstrate the practical applicability and utility of our approach through a case study where we identify future skills for the manufacturing industry in Baden-Wuerttemberg, Germany, based on 1.16 million job advertisements. The approach results in 33 future skills assigned to four categories, i.e., generic skills, digital skills, technological skills, and industrial skills. The results allow further in-depth analysis of the importance of the future skills for different branches. A survey with 294 companies underpins

that the results of the future skills identification process indeed represent skills with increasing importance for the manufacturing industry in Baden-Wuerttemberg. Our contribution to research and practice is threefold: First, our approach is suitable to identify a comprehensive and representative set of future skills for a given region and industry. Second, our approach incorporates existing literature streams in the context of future skills identification in a mixed-method approach, i.e., data-driven and expert-based identification of future skills, leveraging complementary strengths and providing multifaceted insights (Reis et al. 2022). Third, the future skills identified by application of our approach in a case study help governments, companies, educational institutions, and individuals to adapt to the future workforce and the transformation of society. With our case study, we are the first to leverage the potential of big data and derive future skills for an entire industry in an efficient way.

The remainder of this paper is structured as follows: In the next section, we discuss relevant literature related to future skills and their identification. In Sect. 3, we propose a novel approach to identify the future skills for a given region and industry. In Sect. 4, we demonstrate the applicability and utility of our approach through a case study. We conclude the paper with implications for theory and practice, limitations of our work, and directions for further research.

2 Related work

2.1 Future skills

Requirements for the current workforce are continuously changing, especially in areas with rapid and radical changes such as the IT sector. Knowledge about which skills are needed to prepare individuals to thrive in the face of an uncertain future not only determines the success of companies and individuals, but also the success of future economy and society (Ehlers 2020; Kotsiou et al. 2022; Zahidi et al. 2020). Literature in the information systems and business economics field has been dealing with the question of how to acquire and retain skilled employees for many years (Pflüger et al. 2018; Frei and Grund 2022). Thereby, the mismatch of required and available skills between jobs and people can lead to great individual dissatisfaction (Frei and Grund 2022; Kalleberg 2008). Against this background, literature proposes training employees with appropriate future skills to advance their career development (Pflüger et al. 2018). Overall, the identification of future skills attracts major attention among IS researchers (Prommegger et al. 2020a).

Future skills allow employees to be successful in the future workforce (Schallock et al. 2018) and enable citizens to participate in a globalized and fast-changing society (Ehlers 2020). Their identification allows people and organizations to adapt to future requirements and is thus of interest to several stakeholders (Ehlers 2020; Leopold et al. 2016; Rios et al. 2020). First, companies use information about future skills to align their recruitment and qualification strategies (Leopold et al. 2016). In addition to preparing for future challenges, aligning the recruitment and qualification strategies based on future skills is crucial for companies to retain skilled employees (Pflüger et al. 2018). Second, governments have an obligation to react to

the major changes in the labor market and accordingly adapt educational programs to enable the success of companies (Rios et al. 2020; Zahidi et al. 2020) as well as allow all citizens to participate in society (Ehlers 2020). Third, educational institutions focusing on vocational training need to adapt their educational programs to the future skills necessary in the current and prospective workforce (Rakowska and Juana-Espinosa 2021; Zahidi et al. 2019). Finally, against the background of the concept of lifelong learning, individuals knowing the skills necessary in the future can ensure their employability (Ehlers 2020) and plan their professional careers accordingly (Zahidi et al. 2020). The constant threat of the so-called skills obsolescence (Chilton et al. 2010; Joseph et al. 2011) and natural evolution, for example in IT professions, forces employees to constantly adapt their skills (Prommegger et al. 2020b). In the remainder of this section, we will present three different literature strands that have evolved for the identification of future skills: literature reviews, expert assessment, and data-driven approaches (Rios et al. 2020).

2.2 Approaches to identify future skills based on literature reviews

One research direction to identify future skills is by reviewing relevant literature (Kotsiou et al. 2022). This is based on well-established approaches for literature reviews (cf. vom Brocke et al. 2015). In this line, a wide range of different perspectives on future skills can be included, categorized, and evaluated (Kotsiou et al. 2022). This results in unambiguous future skills since the researchers deduce it from prior research (Kotsiou et al. 2022). In this research field, two types of literature reviews exist. First, literature reviews summarizing studies and prior research on certain trends and their impact on employment and the workforce across all industries to deduct which skills are needed to cope with these changes. These literature reviews determine future skills against the background of technological and societal developments such as Industry 4.0 (Ra et al. 2019; Rahmat et al. 2020) or digitalization (Kotsiou et al. 2022; Kurtzo et al. 2016). Rahmat et al. (2020), for example, find that digital literacy is one of the most important skills for graduates to remain competitive in the job market. Second, literature reviews gaining a specific and deeper insight into future skills in a single industry by reviewing both future skills studies for all industries as well as trend studies for the particular industry. For example, Lieu et al. (2018) determine future skills for engineering such as big data analytics, augmented realities, and mathematics, while Cicek et al. (2019) focus on the maritime industry identifying for instance equipment maintenance, information, and data processing as well as safety awareness as future skills. Literature reviews can provide broad and scientifically accepted knowledge on future skills (Bandara et al. 2011). A drawback of literature reviews is their past-orientation as they review already published work. Thereby, they might oversee future skills not yet established especially if they arise of emerging trends not yet discussed in prior literature (Kotsiou et al. 2022).

2.3 Approaches to identify future skills based on expert assessments

Intending to consider future skills arising from emerging trends, the strand of using expert assessments as a qualitative research method has been established. Experts are usually aware of established trends as well as new developments on the rise and related future skills (Bogner et al. 2009). Intending to gain a broad understanding of trends and their consequences for future skills requires both the background—ranging from researchers and lectures over training coaches to private and public managers—and the seniority level of the interviewed experts to vary greatly. A well-established method to perform expert assessments are individual interviews with several experts. Such individual interviews are used to identify future skills for job fields such as agricultural communication (Kurtzo et al. 2016) or against the cultural changes toward a more diverse society and thus a more diverse workforce (Wentling and Palma-Rivas 1998). An advantage of these individual interviews is that they allow in-depth questioning and inquiries. However, they are time-consuming, accordingly, the number of experts is usually rather small. A less cumbersome method of expert assessments are online surveys, which allow interviewing a larger number of persons (Evans and Mathur 2005), but without the possibility of inquiries and without control over who exactly participates in the assessment (Evans and Mathur 2005). In the context of future skills identification, they are especially used to establish the perspective of managers (Kirchherr et al. 2018; Leopold et al. 2016; Zahidi et al. 2020). However, the results of interviews are aggregated by the researchers, which can lead to a bias towards extreme answers (Aichholzer 2009). A promising approach to overcome this demerit is the use of focus groups (Tremblay et al. 2010), where experts discuss opinions and experiences. Davies et al. (2011) use this approach to determine six disruptive drivers that will reshape the future workforce and ten skills to successfully face these challenges. Prifti et al. (2017) identify future skills in eight categories (e.g., analyzing and interpreting data) in the context of Industry 4.0 using focus groups. To avoid a bias towards rhetorical adroit experts or experts with higher status, researchers stress the importance of carefully designing and moderating the discussion to avoid group pressure situations within focus groups (Aichholzer 2009). Another challenge of focus group research is that they demand substantial resources of experts and researchers in preparation and implementation (Krueger and Casey 2015).

In recent years, researchers combine different methods based on expert assessments to identify future skills. Several studies analyzing future skills in countries such as Germany (Kirchherr et al. 2018; Sczogiel et al. 2019) or Poland, Spain, and Thailand (Rakowska and Juana-Espinosa 2021) have used this procedure. For instance, a combination of individual interviews and focus groups promises stronger evidence (Johnson and Onwuegbuzie 2004), since it allows to gain deep insights from single experts within individual interviews as well as a broad overview of opinions of a larger number of experts using focus groups. Still, mixed methods that identify future skills solely based on interviews are often considered to only include

a small sample of opinions, given that a limited number of experts is involved (Bogner et al. 2009; Rios et al. 2020). Furthermore, the results of expert-based methods are heavily dependent on the selection of experts, i.e., their individual opinions and positions (Bogner et al. 2009; Rios et al. 2020).

2.4 Data-driven approaches for the identification of future skills

Driven by the large availability of data including opinion-rich resources on the internet and tremendous technological advances such as machine learning, data-driven approaches enable the identification of future skills. Thereby, the identified future skills can be representative for whole regions and industries under consideration and derived more objectively than by interviewing single experts (Rios et al. 2020). Most frequently, job advertisements are utilized as the underlying database (cf. Ang et al. 2013; Litecky et al. 2010) and have been used as a source of data by researchers for several years (cf. e.g., Todd et al. 1995). They usually consist of different parts such as information about the company, benefits of the job, information about the application process (Buchmann et al. 2022; Ganesan et al. 2018) and, most importantly for these approaches, detailed information about the demanded skills considered to be necessary for the future by employers (Rios et al. 2020). Given that nowadays almost all job advertisements are posted online (Buchmann et al. 2022), online job advertisements can provide a representative sample of skills demanded by the labor market in the future (Carnevale et al. 2016; Rios et al. 2020). However, they are based on the opinions of employers and can therefore contain biases such as gender bias (Khaouja et al. 2021). Data-driven approaches to identify future skills predominantly consist of two steps. In a first step, all skills are extracted from the job advertisements using different methods such as text mining or skill dictionaries. In a second step, the extracted skills are analyzed using various quantitative methods. A common approach to identify future skills from online job advertisements is using empirical analysis. Empirical analysis is frequently used to identify future skills for specific job positions such as early career researchers (Maer-Matei et al. 2019), marketing jobs (Pefanis Schlee and Harich 2010), college graduates (Rios et al. 2020), and librarians (Yang et al. 2016). Findings underpin the importance of both specific job-related skills as well as generic skills such as communication and problem-solving. Thereby, the basic idea is to analyze which skills appear most frequently in the job advertisements. While this method provides a good overview of frequently demanded future skills, it might overlook especially emerging future skills since such skills do not yet appear relatively often in job advertisements. Moreover, it is prone to favor skills with ambiguities (i.e., MS for Microsoft as well as mass spectrometry) over skills with multiple names (i.e., commitment and engagement).

Another commonly used approach to derive future skills from job advertisements are explorative methods which discover future skills based on analyzing similarities in skills (i.e., skills that frequently appear together in job advertisements). Thereby, this approach can overcome the weaknesses of empirical analysis (Pejic-Bach et al.

2020), especially with respect to the identification of emerging skills since it can perceive emerging but small trends (Gurcan et al. 2021). A popular approach in this context is topic modeling. Topic modeling is used to determine different topics (which can be interpreted as skills) in different documents. For instance, Latent Dirichlet Allocation (LDA) (Blei et al. 2003) has been used to identify future skills in different IS jobs (Föll et al. 2018) and data science (Gurcan and Cagiltay 2019; Handali et al. 2021; Michalczyk et al. 2021). However, LDA has a high runtime complexity (Sonntag and Roy 2011) and is therefore only applicable for small datasets (Sonntag and Roy 2011). Accordingly, existing approaches using topic modeling concentrate on the identification of future skills for single job positions and use datasets based only on single job portals thereby impeding the representativeness of the results (Rios et al. 2020). An approach far less complex in runtime than topic modeling are clustering algorithms (Firdaus and Uddin 2018). These algorithms group skills in clusters according to how frequently they appear together in job advertisements. Recently, researchers mostly make use of hierarchical clustering methods, which are used frequently and successfully for text mining problems (Inzalkar and Sharma 2015) allowing to find meaningful skill clusters (Pejic-Bach et al. 2020) and adjusting the level of aggregation (Adl et al. 2011). For instance, Litcecky et al. (2010) use this approach to determine the future skills of 20 computing professions, while Pejic-Bach et al. (2020) identify eight skill clusters for Industry 4.0 jobs. Hence, hierarchical clustering has the potential to comprehensively determine future skills without bias toward particular branches or regions, provided the underlying sample of job advertisements is representative (i.e., the sample is large enough and not skewed toward particular branches or regions). However, as with all data-driven approaches, it is not capable of interpreting the underlying generic concept of similar skills. Indeed, experienced researchers need to interpret the results of the algorithms (i.e., clusters) to obtain future skills. However, thereby the future skills may not be meaningful for the stakeholder since the level of aggregation of the future skills may differ from the stakeholders' expertise if researchers conduct the interpretation. Especially when the future skills of a whole economy are investigated since it is almost impossible for a small group of researchers to have a deep insight into the whole economy of a region (Pejic-Bach et al. 2020; Rios et al. 2020).

Against the background of major disruptions in the workforce (Föll et al. 2018; Leopold et al. 2016), companies, governments, educational institutions, and individuals alike need to react and prepare employees and themselves for the future of work (Leopold et al. 2016; Rakowska and Juana-Espinosa 2021; Rios et al. 2020). Therefore, the identification of future skills is critical to the success of companies and society as a whole (Zahidi et al. 2020). Literature provides three different strands to identify future skills, i.e., based on literature reviews, expert assessments, and data-driven approaches. In order to combine the strengths and to overcome the weaknesses of these strands, we propose a new approach that combines existing strands, resulting in future skills that are meaningful for stakeholders, comprehensive, and representative of the region and industry under consideration.

3 A novel approach to identify future skills

Following the DSR methodology (Hevner et al. 2004; Gregor and Hevner 2013), in this section, we describe our proposed artifact: a novel approach to identify the future skills of a selected region and industry. Our approach builds on prior literature in the context of data-driven and expert-based identification of future skills. By integrating and expanding existing concepts into a transparent and replicable approach, we obtain comprehensive and representative future skills that enable companies, governments, educational institutions, and individuals to adapt to the future workforce and the transformation of society.

3.1 Basic idea

When identifying future skills, three core challenges need to be addressed. First, the future skills must be comprehensive and representative for the given region and industry (Rios et al. 2020). Second, the future skills must be meaningful for the stakeholders, i.e., free of overlap and on the correct level of aggregation (Ehlers 2020; van Laar et al. 2019). Third, the future skills must be identified by means of a structured and replicable approach (Föll et al. 2018; Pejic-Bach et al. 2020). To address these challenges, our approach is based on three core concepts building on literature in the context of data-driven (Pejic-Bach et al. 2020) and expert-based (Rakowska and Juana-Espinosa 2021) identification of future skills as well as processes in the context of knowledge extraction from data (Martinez-Plumed et al. 2019): (i) data-driven derivation of future skills based on a comprehensive and representative set of job advertisements, (ii) integration of stakeholders as experts to determine the level of aggregation of future skills, and (iii) incorporation of these concepts into a process that systematically guides the identification of future skills.

First, a key challenge is to identify a comprehensive and representative set of future skills for the given region and industry. This means that future skills are not distorted by sub-regions or sub-industries (Rios et al. 2020) and that they also contain emerging future skills which are especially difficult to identify (Ehlers 2020). To address this challenge, we *derive future skills by means of a data-driven approach based on a comprehensive and representative set of job advertisements*. Against the background of large recruitment and training costs for new employees (Dash et al. 2018; Sorensen and Ladd 2020; Tan and Laswad 2018) employers are careful to describe skills they actually require (Brooks et al. 2018). Therefore, job advertisements usually contain the most desired and valued skills by employers (Tan and Laswad 2018) and thus the skills considered to be necessary and most relevant for employees to ensure the current and future success of companies (Buchmann et al. 2022; Carnevale et al. 2016; Rios et al. 2020). There is an ongoing discourse on the forecast horizon for reliably predicting future skills from online job advertisements, which in turn depends on the industries analyzed and on the type of future skills investigated (Ra et al. 2019). The possible forecast horizon therefore depends on the use case and should be determined accordingly for each application, for example by expert assessment. However, especially for a time span of up to five years (Grimes

and Grimes 2008), job advertisements are considered one of the most promising sources for all types of labor market analyses, but particularly the analysis of skills, containing information barely available in other sources (Descy et al. 2019). Thus, using job advertisements to gain insights into which skills are required to enable employees' success in the future workforce, i.e., future skills, is a proven approach supported by multiple studies (Brooks et al. 2018; Tan and Laswad 2018). In this line, the fact that nowadays almost all job advertisements are posted online (Buchmann et al. 2022) such as on job search engines, allows for the collection of a dataset that comprehensively represents the required skills in a given region and industry for a medium-term future (Buchmann et al. 2022; Rios et al. 2020). Based on a comprehensive and representative set of job advertisements, we derive future skills by means of a data-driven approach (Sonnewald et al. 2020). As data-driven approach, we choose hierarchical clustering which is an unsupervised machine learning method and thus allows for the discovery of yet unknown and potentially emerging future skills (Gurcan et al. 2021). Moreover, it allows an ex-post determination of the level of aggregation due to the underlying architecture (Adl et al. 2011; Ayad and Kamel 2008). The data-driven derivation of future skills based on a comprehensive and representative set of job advertisements ensures that a comprehensive and representative set of future skills can be identified for a given region and industry. In sum, the data-driven approach allows for extracting future skills based on a large number of job advertisements that comprehensively represent the sought skills of a given region and industry. Accordingly, it ensures to identify a comprehensive and representative set of future skills.

Second, future skills must be meaningful for the stakeholders (van Laar et al. 2019), i.e., companies, governments, educational institutions, and individuals, need actionable future skills to help them adapt to the future workforce and the transformation of society. This demands that future skills are unambiguous, i.e., free of overlap (Rios et al. 2020). Furthermore, skills in job advertisements are often on a different level of aggregation (e.g., data science vs. Python) (Gardiner et al. 2018). To be meaningful, future skills' level of aggregation must be consistent with their stakeholders' expertise of future skills (Djurnalieva and Sleeman 2018). While hierarchical clustering naturally leads to unambiguous results (Rios et al. 2020), the results can be interpreted on varying levels of aggregation (Adl et al. 2011; Ayad and Kamel 2008). *We integrate stakeholders as experts to determine the level of aggregation of future skills.* In our approach, experts are chosen to represent the relevant stakeholders of the given region and industry and asked to align the level of aggregation of the clustering algorithm's results with their own expertise. This ensures that the interpretation of the results of the data-driven derivation of future skills – described as a major challenge in literature (Ehlers 2020; van Laar et al. 2019)—is in accordance with the stakeholders' expertise. Consequently, future skills are meaningful for their stakeholders.

Finally, future skills must be identified by means of a structured and replicable approach. To yield comprehensive and representative as well as meaningful future skills, data-driven and expert-based methods need to be combined and their parts need to be specified. We propose to *incorporate these concepts into a process that systematically guides the identification of future skills.* The process is inspired by

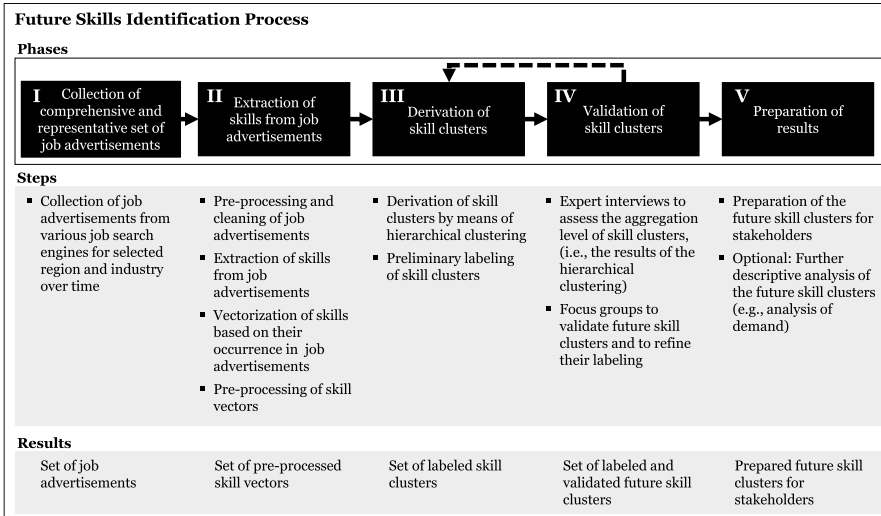


Fig. 1 Phases of the future skills identification process

the well-established process CRISP-DM, which provides a structured and replicable framework to systematically guide knowledge discovery in data (Martinez-Plumed et al. 2019). CRISP-DM is the standard process for data mining in research and practice (Martinez-Plumed et al. 2019) and serves to minimize risks through various validation steps, reveal and remedy errors, and to facilitate resource allocation (Marbán et al. 2009; Rollins 2015). It consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, which we adopt for our process. The business understanding phase focuses on understanding the objectives and requirements (Wirth and Hipp 2000), which we obtained through a literature review (cf. Section 2). We have adapted the subsequent five phases in our approach: First, the three CRISP-DM phases ‘data understanding’, ‘data preparation’, and ‘modeling’ (Wirth and Hipp 2000) to guide the core concept of data-driven derivation of future skills based on a comprehensive and representative set of job advertisements. Second, the CRISP-DM phase ‘evaluation’ to guide the core concept of the integration of stakeholders as experts to determine the level of aggregation of future skills. We interlink these core concepts as in CRISP-DM, where ‘evaluation’ can induce an iteration of ‘modeling’ (Wirth and Hipp 2000): In our process, the experts’ assessment of the aggregation level of future skills can induce an adjustment of the aggregation level of the results of the hierarchical clustering. Finally, the CRISP-DM phase ‘deployment’ (Wirth and Hipp 2000) to guide the presentation of the future skills to the stakeholders. Designing the identification of future skills as a process inspired by CRISP-DM represents a structured and replicable approach.

In the following, we describe the phases of the proposed future skills identification process (Fig. 1).

3.2 Phases of the future skills identification process

The first phase aims at the *collection of a comprehensive and representative set of job advertisements*. It is inspired by the CRISP-DM phase ‘data understanding’, which includes all activities related to data collection and familiarization with the data (Wirth and Hipp 2000). Job advertisements are considered to comprehensively represent the future needs of the labor market (Carnevale et al. 2016). Nowadays, almost all job advertisements are published online (Buchmann et al. 2022). Thus, we propose to collect a set of job advertisements from various job search engines over time. While accessing different job search engines counteracts potential bias towards companies using certain platforms more frequently, collection over time mitigates potential bias due to short-term effects such as seasonal fluctuations (Föll et al. 2018). The time span for the collection should be chosen according to the needs of the specific application context (i.e., whether short-term or long-term effects are of interest). Using shorter time spans for the collection leads to results more heavily impacted by short-term effects in the labor market. If seasonal effects are not of interest, a respective minimum time span (e.g., one year) is recommended. To operationalize the collection process, for example, job advertisements can be crawled from job search engines or platforms over a certain time span, with a filter on the respective region and industry. An advantage of this method is to provide a set of job advertisements that can be flexibly tailored to individual needs. However, the variety of different job search engines or platforms requires high effort and makes it difficult to generate a representative set of job advertisements. Another possibility is to purchase such datasets from different job search engines, platforms, or companies that specialize in crawling job advertisements. While this might be associated with relatively high costs, it allows to obtain a large set of job advertisements with a long history. As a result of this phase, a set of job advertisements is available and can be prepared and analyzed in the following.

The second phase aims at the *extraction of skills from job advertisements*. This corresponds to the ‘data preparation’ phase in CRISP-DM which includes the selection, cleaning, and formatting of relevant data to construct the final dataset for subsequent analysis (Martinez-Plumed et al. 2019; Wirth and Hipp 2000). In our case, the relevant data to be analyzed are skills. To achieve a set of skills that are ready for analysis, data preparation consists of four steps. First, the set of job advertisements must be pre-processed and cleaned (Pejic-Bach et al. 2020). It is common for the same job advertisement to be published several times, both on the same job search engine and on various job search engines (Föll et al. 2018). To ensure that certain skills are not overrepresented, duplicates of job advertisements must be removed so that each job advertisement only appears once in the dataset. Based on a dataset of unique job advertisements, in a second step, the skills are extracted. Thereby, text passages are removed from job advertisements that do not represent skills, such as company descriptions or employee benefits (Yang et al. 2016). Literature provides two suitable approaches to extract skills from job advertisements: First, skills can be automatically detected by keyword matching with a skill dictionary (Brancatelli et al. 2020) such as the ESCO skill ontology (ESCO 2020) or the O*NET taxonomy (O*NET 2022). Skill dictionaries can for instance be generated and kept up to

date by experts to guarantee that they also include recent trends (Stefanić and Šimić 2021). They can further be enriched by the use of existing future skill studies or literature reviews. This approach requires a lot of manual effort from experts but allows for the reaction to upcoming trends in an early stage (Gugnani and Misra 2020). Second, machine learning approaches can be trained to automatically detect skills in job advertisements (Buchmann et al. 2022). Thereby, machine learning algorithms are trained on a large amount of labeled data to first divide job postings into different segments (e.g., requirements or benefits) and then classify those segments. Then, the skills are extracted from the requirements segment using named entity regression (Buchmann et al. 2022). As with all supervised machine learning models, it requires some initial manual effort to label the data, but minimal effort is required afterwards to use the model. In the third step, skills are vectorized so that they can be analyzed by data-driven algorithms in the subsequent phase. To contain the information on skills' occurrence in the job advertisements, we vectorize each skill as a binary (row) vector, with each column representing one job advertisement. A column's entry is one if and only if the skill is contained in the respective job advertisement. Especially for large regions and industries, this results in high-dimensional skill vectors. Against the background that analysis of high-dimensional binary data is difficult and complex in run-time (Ray et al. 2021), in the final step of this phase, skill vectors are further pre-processed using dimensionality reduction. For this purpose, established dimensionality reduction algorithms such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), isometric mapping (Isomap), or uniform manifold approximation and projection (UMAP) can be applied (Anowar et al. 2021; Becht et al. 2018). The choice of the algorithm depends on the amount of data and noise present, as each algorithm has different advantages and disadvantages. PCA summarizes multiple dimensions into fewer dimensions by linearly combining the dimensions while preserving maximum variance (Bro and Smilde 2014). It is relatively simple and thus computationally efficient, allowing it to handle very large amounts of data (Bro and Smilde 2014). However, it is designed for continuous data, which can lead to poor results using the binary data available in the present context (Kolenikov and Angeles 2005). Thus, PCA should only be used when the amount of data leaves no other choice. Isomap, on the other hand, is a nonlinear dimensionality reduction method based on the spectral theory which aims to preserve the geodesic distance in the lower dimension (Anowar et al. 2021). It can take nonlinear structures into account with the disadvantage of high computational complexity and sensitivity to noise (Anowar et al. 2021; Yousaf et al. 2021). Therefore, it can be used when the amount of data is relatively small and of high quality. Otherwise, UMAP and t-SNE may be better choices. Both algorithms are based on the idea of preserving the similarity between points (McInnes et al. 2018). Thereby, UMAP is based on Riemannian manifolds (McInnes et al. 2018), while t-SNE is a probabilistic approach based on the t-distribution (Wattenberg et al. 2016). For most datasets, UMAP performs slightly better (Becht et al. 2020). However, UMAP requires parameter fine-tuning, which can be time-intensive (Vermeulen et al. 2021). Thus, in case time is a limiting factor, it may be worthwhile to assess whether t-SNE is capable of producing satisfactory outcomes. Altogether, these steps result in pre-processed skill vectors that were extracted from a set of unique job advertisements.

The third phase aims at the *derivation of skill clusters*. It is inspired by the CRISP-DM phase ‘modeling’ in which a selected model is applied to the dataset to generate new insights with respect to the defined goals (Wirth and Hipp 2000). Skills in job advertisements occur on different granularity levels (e.g., a concrete programming language like Python vs. a complete discipline like data science) (Gardiner et al. 2018). Together with the fact that a dataset that comprehensively represents a region and industry is likely to be large, a direct interpretation of the set of skills is not feasible. To extract meaningful knowledge from the set of skill vectors, we suggest to derive skill clusters comprising semantically similar skills. Thereby, the basic idea is that the similarity of skills translates into their co-occurrence in the same job advertisements. This is based on the fact that job advertisements are aimed at specific job profiles. We propose to apply hierarchical clustering to derive skill clusters based on the pre-processed skill vectors as the first step in this phase. Hierarchical clustering is selected due to three reasons. First, this method represents a data-driven approach that analyzes the co-occurrence of skills in job advertisements (Rios et al. 2020). Second, representing an unsupervised machine learning method, it allows for the discovery of a priori unknown skill clusters (Madhulatha 2012). Third, hierarchical clustering provides unambiguous skill clusters while their level of aggregation can be determined ex-post due to the underlying architecture with dendrograms (Adl et al. 2011; Ayad and Kamel 2008). Based on the results of the hierarchical clustering, in a second step, a preliminary level of aggregation is chosen and preliminary labels are assigned to each skill cluster. These steps result in a set of labeled skill clusters, which can be interpreted and evaluated in the next phase.

The aim of the fourth phase is the *validation of skill clusters*. It is inspired by the ‘evaluation’ phase in CRISP-DM, where the results are evaluated from a professional perspective (Wirth and Hipp 2000). If necessary, this phase can trigger a new modeling iteration (Wirth and Hipp 2000). While the previous phases serve to identify a comprehensive and representative set of skill clusters for the given region and industry, the fourth phase serves to assure that the results are meaningful to stakeholders. To this end, the stakeholders of the future skills (e.g., companies, governments, educational institutions, and individuals) are incorporated as experts in two steps. In the first step, stakeholders are questioned in individual expert interviews to assess whether the level of aggregation of the skill clusters is consistent with their expertise. Individual expert interviews are chosen since they allow in-depth inquiries on certain skill clusters (Evans and Mathur 2005). For instance, experts might find that skill clusters contain skills from diverse skill profiles or that certain skill profiles are dispersed in various skill clusters. In this case, this phase should be discontinued and instead resumed with the previous phase to choose a higher or lower level of aggregation of the results of the hierarchical clustering. If the majority of experts find the level of aggregation appropriate, it is guaranteed that the skill clusters are meaningful to stakeholders and the process can continue. In the next step, focus groups are conducted to validate that skill clusters indeed represent future skill clusters and to refine their labeling. For instance, certain skill profiles, such as human resources management, are always important and are therefore always sought in job advertisements; however, they may not represent the skills necessary

for the success of employees in the future workforce and thus be removed. Focus groups are chosen because they allow experts to share and discuss their opinions and experiences (Tremblay et al. 2010). In order to gain rich insights (Aichholzer 2009), the following guidelines are established. First, at least two focus groups are conducted to gather diverse opinions from different focus groups and thus to provide sufficient evidence for the final future skill clusters (Tremblay et al. 2010). Second, focus groups are composed of diverse labor market experts who are familiar with the examined industries (Krueger and Casey 2015; Tremblay et al. 2010). Third, moderators structure focus groups from most to least important topics, and from general to specific (Krueger and Casey 2015). In this process, all skill clusters are presented to the experts, and they discuss which of the skill clusters actually represent future skill clusters, in order to obtain evidence on all clusters. In addition, the experts are asked to refine the respective labels according to their expertise, i.e., the wording of each label. Since focus groups do not necessarily converge on a joint opinion, diverse opinions are collected from different focus groups to obtain sufficient evidence for the final future skill clusters (Tremblay et al. 2010). Fourth, the researchers conduct a note-based analysis of the focus group results and decide on the proposed exclusion of skill clusters as well as adjustments to their labels (Krueger and Casey 2015; Tremblay et al. 2010). For both, interviews and focus groups, it is crucial to select a mix of experts with a broad background so that both the diversity of stakeholders as well as the whole region and industry are represented. Moreover, it is possible to use material such as literature reviews to prepare experts for the interviews. At the end of this phase, future skill clusters validated by experts and meaningful for the stakeholders are present.

The final phase comprises the *preparation of the results*, inspired by the ‘deployment’ phase in CRISP-DM. According to CRISP-DM, the results should be prepared so that the stakeholders can utilize the knowledge gained and put it to practical use (Martinez-Plumed et al. 2019). This translates into the preparation of future skill clusters so that companies, governments, educational institutions, and individuals can adapt to the future workforce and the transformation of society. The data-driven derivation of future skills based on a large dataset of job advertisements allows to provide stakeholders with rich information. As a first step in this phase, future skill clusters are prepared to include not only the labels of the future skill clusters but also the most relevant skills of each future skill cluster, for instance, those that appear most frequently in job advertisements. Moreover, in a second step, the data-driven approach allows for further descriptive analysis of future skill clusters at the request of specific stakeholders. For instance, (development of) demand of future skill clusters (over time) can be determined for selected sub-regions or sub-industries, based on the analysis of a subset of job advertisements over time. This provides insights for stakeholders on the importance of future skill clusters for selected sub-regions or sub-industries as well as on the development of the importance of future skill clusters over time. At the end of this phase, the future skill clusters as derived in the phases one to four are prepared to inform their stakeholders.

4 Case study

Following the DSR methodology (Hevner et al. 2004; Gregor and Hevner 2013), we demonstrate the practical applicability and utility of our artifact. To this end, we assess the novel future skills identification process through a case study (Yin 1981). The future skills identification process was instantiated to identify the future skills within the manufacturing industry in Baden-Wuerttemberg, i.e., the third-largest federal state in Germany with a population of over 11 million people. In line with the literature and supported by our expert assessment, we set a forecast horizon of five years (Grimes and Grimes 2008). The study was conducted from June–September 2021 in cooperation with the Ministry of Economy, Labor and Tourism Baden-Wuerttemberg and the leading employers' as well as employees' associations of the state.

Baden-Wuerttemberg is one of the most economically strong and innovative regions in Europe (Hörisch and Wurster 2019) with manufacturing as its most important industry (Hörisch and Wurster 2019). The manufacturing industry in Baden-Wuerttemberg consists of a wide range of branches such as mechanical engineering, automotive industry, electrical engineering, metal engineering, and medical engineering. It includes both large global players such as the Mercedes-Benz Group, the Robert Bosch GmbH, and the ZF Friedrichshafen AG as well as a variety of (often globally acting) small and medium-sized enterprises (SMEs) (Schenkhofer 2022). Global challenges and trends, such as decarbonization, digitalization, and demographic change particularly affect the manufacturing industry (Suuronen et al. 2022); accordingly, also the manufacturing industry in Baden-Wuerttemberg is in the need of a major structural transformation (Abdallah et al. 2021). Thus, a fundamental skill shift of employees is required (Doyle-Kent and Kopacek 2020). Our aim was to identify the future skills of the manufacturing industry in Baden-Wuerttemberg to enable the respective companies, government, educational institutions, and individuals to adapt to the future workforce and the transformation of society.

4.1 Application of the novel approach to identify future skills

In the first phase of our process, a comprehensive and representative set of job advertisements for the given region and industry needs to be collected. To this end, job advertisements were crawled with a regional filter from the most important job search engines and job platforms in Baden-Wuerttemberg, including indeed.com, monster.de, and jobboerse.arbeitsagentur.de. To account for the large number of SMEs in Baden-Wuerttemberg, also job platforms with a focus on SMEs, such as yourfirm.de, were targeted. To mitigate bias due to seasonal effects and to counteract an excessive influence of the global Covid-19 pandemic on the results, we chose to collect the data within a rather wide time span of three years (2018–2020) (Carnevale et al. 2016; Rios et al. 2020). This choice was also supported by the experts involved in the case study, who did not expect the labor market disruptions of Covid-19 to be as severe in the next five years as they have been in 2020. The resulting

dataset representatively comprised job advertisements of the manufacturing industry in Baden-Wuerttemberg.

The second phase serves to extract skills from the job advertisements and prepare them for further analysis. In the first step, the job advertisements were pre-processed and cleaned according to standard text preprocessing frameworks (Hickman et al. 2022), including the removal of html-tags and punctuation as well as lowercasing (Pejic-Bach et al. 2020). We did not use lemmatization and stemming, which is in line with similar approaches (Pejic-Bach et al. 2020). Also, for our purposes, the disadvantage of potentially reducing different skills that belong to the same word stem did not outweigh the few advantages, such as accounting for few misspelled words. Moreover, this step served to ensure that each job advertisement only appears once in the dataset. Thus, we removed duplicates using the Levenshtein algorithm which is commonly applied in the context of job advertisements (Föll et al. 2018). This algorithm calculates the similarity of two texts as the minimum number of changes of single characters needed to transform one text into the other, normalized with respect to the length of the texts. Texts below a certain threshold, which we chose in line with literature as 0.15, were classified as duplicates and sorted out (Di Lucca et al. 2002; Peng et al. 2014). Applying the Levenshtein algorithm on our data resulted in a set of 1.16 million unique job advertisements. On this basis, in the second step, skills were extracted using keyword matching with a skill dictionary (Brancatelli et al. 2020), where all job advertisements were searched for terms occurring in the skill dictionary. To ensure the inclusion of recent trends, we used a skill dictionary that was constantly being updated by several experts (e.g., by searching through patents to detect new technological trends) (Stefanić and Šimić 2021). The skill dictionary contained over 10 million skills as well as synonyms and translations in several languages for skills. Keyword matching with this skill dictionary resulted in 6769 distinct skills occurring at least once within the set of 1.16 million unique job advertisements. This means that each skill (e.g., “Python”, “neural network”, “machine learning”) was included in the data independently at this point. To prepare these skills for analysis in the subsequent phase, in the third step, the skills were vectorized based on their occurrence in the job advertisements. This resulted in 6,769 binary skill vectors with 1.16 million dimensions (i.e., the number of unique job advertisements). This made it possible to compare skills in terms of their occurrence in job advertisements and to find those that occurred together most frequently (e.g., the skill “neural network” was often included in job advertisements that also included the skill “machine learning” because neural networks are a prominent machine learning method). As expected, for large regions and industries such as the manufacturing industry in Baden-Wuerttemberg, analysis of binary skill vectors based on their occurrence in job advertisements was unfeasible due to high dimensionality. Thus, in the final step of this phase, we pre-processed skill vectors using dimensionality reduction. To this end, we applied UMAP, which is well-established to reduce binary vectors to lower-dimensional real-valued vectors (Becht et al. 2018). UMAP requires to set three parameters: number of dimensions, number of neighbors and minimum distance. We chose these parameters in order to prepare the skill vectors for the application of a clustering algorithm (cf. next phase). Regarding the number of dimensions (i.e., the number of dimensions after the reduction), the

more complex the underlying dataset is, the more dimensions are needed to preserve its information. In our case, the initial number of dimensions was high with more than one million dimensions. However, the data was only binary, so we decided to reduce the data to two dimensions (Rugard et al. 2021). The second parameter is the number of neighbors. UMAP attempts to project the structure of high-dimensional space into low-dimensional space. This is achieved by measuring the structure in high-dimensional space for each point based on its nearest neighbors and their distance. The number of neighbors determines how many nearest neighbors are considered for each point to determine the structure. Accordingly, the number of neighbors specifies whether the local or global structure of the data is preserved. With a choice of a small number, only the local structure is preserved, and the points in the low-dimensional space are widely distributed over the plane. In contrast, with a large number, the points that are similar in the high-dimensional space are close in the low dimensional space as the global structure is preserved¹ (McInnes 2018). For clustering, the preservation of the global structure is more important (McInnes 2018). Against this background, we chose 50, which represents a high number of neighbors. The last parameter, the minimum distance allowed between two points in the dimensionally reduced space, controls how close points can be in the low-dimensional space. A higher minimum distance spreads the points more over the plane, while a lower minimum distance allows similar points to remain clustered together in the low-dimensional space¹. The latter is important for clustering (McInnes 2018), which is why we chose the rather small value of 0.05. The result of this phase were 6,769 real-valued two-dimensional skill vectors extracted from 1.16 million unique job advertisements. Here, skills such as “neural network” and “machine learning”, which often appeared together in job advertisements, were represented by points close to each other in the two-dimensional space.

Direct interpretation of 6769 skills is not feasible. Thus, in the third phase of our process and based on the set of skill vectors, we derived skill clusters by means of hierarchical clustering. The two main parameters to be chosen for hierarchical clustering algorithms are distance metric and linkage criterion (Madhulatha 2012). As distance metric, we selected the Euclidean distance, which is the natural distance between real-valued vectors (Madhulatha 2012). Moreover, the Euclidean distance was also used in UMAP to calculate distances between points in the low-dimensional space and was hence the logical choice for the clustering algorithm as well. As linkage criterion we chose the Ward criterion, which is a variance-minimizing approach, that minimizes the sum of squared differences within all clusters (Laudau et al. 2011) and fosters interpretability of the results (Ros and Guillaume 2019). The results of hierarchical clustering take the form of a dendrogram which allows determining the number of clusters ex-post. Representing the second step of this phase, we chose 100 as the preliminary number of skill clusters. Two researchers assigned labels to each of the 100 skill clusters based on the skills within the corresponding cluster. Both the number of skill clusters and the corresponding labels

¹ For a more detailed discussion of the parameters with exemplary plots we refer to the documentation of the Python implementation of UMAP: <https://umap-learn.readthedocs.io/en/latest/parameters.html>.

were preliminary and refined within two iterations (cf. next phase). Exemplary, the skills “neural network” and “machine learning” were now part of a cluster “Data Science”.

The fourth phase serves to validate future skill clusters. In the first step, it has to be validated whether the level of aggregation of skill clusters is consistent with stakeholders’ expertise. To this end, we conducted semi-structured individual interviews (Dearnley 2005) with eight experts with a broad background in the manufacturing industry in Baden-Wuerttemberg. More concretely, we chose experts with diverse positions in companies in the manufacturing industry in Baden-Wuerttemberg (3 department heads, 3 employee representatives, 1 chief human resources officer, and 1 employer’s representative) and from different branches (3 mechanical engineering, 2 automotive, 2 metal-working industry, 1 medical engineering). Each interview took 60 min and was conducted remotely via Zoom. In these interviews, experts were presented with the labels of the skill clusters and the most relevant corresponding skills with respect to occurrence in the job advertisements. Based on a pre-defined interview guide and the option to make use of in-depth inquiries, if necessary (Evans and Mathur 2005), experts were asked to assess the level of aggregation of the skill clusters. The interviews were conducted as follows. After a short introduction explaining the structure and purpose of the interviews, the experts were presented with the future skills clusters in combination with their most important sub-skills. They were asked whether the level of aggregation across the different clusters was appropriate and, if not, which clusters were particularly critical. After two iterations, which included an update of the number of skill clusters and labeling by the two researchers (cf. third phase), the experts found the level of aggregation appropriate, resulting in 57 skill clusters. To validate that skill clusters indeed represented future skill clusters and to refine their labeling, we conducted focus groups. These focus groups were designed according to the framework of Tremblay et al. (2010) for confirmatory focus groups. In line with the guidelines by Tremblay et al. (2010), we conducted four confirmatory focus groups with a total of 25 experts. Each focus group consisted of experts that represent diverse positions and branches of the manufacturing industry – similar to the individual interviews in the previous step. The focus groups were moderated by a researcher who was observed by another researcher to mitigate personal bias (Tremblay et al. 2010). Moderators adhered to a questioning routine which was developed and tested with a pilot focus group of researchers and students in advance (Tremblay et al. 2010). In the focus groups, experts were shown the labels of the skill clusters and the most relevant corresponding skills with respect to occurrence in the job advertisements. Their task was to decide which of the presented skill clusters indeed represented future skills of the manufacturing industry in Baden-Wuerttemberg. Thereby, they exemplary excluded skill clusters such as a cluster dedicated solely to Covid-19 related jobs (i.e., mainly security and administrative work for Covid-19 test centers as well as the conduction of Covid-19 tests), where the experts agreed that despite it was frequently sought in job advertisements did not represent necessary skills for the success of employees in the future workforce. For validated future skill clusters, they refined the respective labels according to their expertise. For example, they changed the label “Data Science” to “Data Science and AI” and “Electrical Drive Technologies” to “Alternative

Drive Technologies". Additionally, the participants of the focus groups assigned each of the future skill clusters to a skill category. This procedure led to 33 future skill clusters assigned to four categories, i.e., generic skills, digital skills, technological skills, and industrial skills (cf. Table 1). These categories correspond to established future skills frameworks (Kirchherr et al. 2018), whereby the category industrial skills considers the special role of the industrial sector in Baden-Wuerttemberg.

In the final phase, we prepared the results so that stakeholders can utilize the gained knowledge. To this end, in a first step, we published a report including the 33 future skill clusters (labels + most relevant corresponding skills with respect to occurrence in the job advertisements). These results were also presented to relevant stakeholders. The manufacturing industry in Baden-Wuerttemberg consists of a wide range of branches including mechanical engineering, automotive industry, electrical engineering, metal engineering, and medical engineering. Stakeholders were especially interested in the importance of the future skill clusters for their specific branch. Thus, in a second step of this phase, we analyzed the importance of the future skill clusters for different branches. To this end, for each branch, we first filtered the skills in the dataset by matching the job advertisements' publishers with lists of companies of a certain branch. This list was obtained by crawling industry registers. On that basis, to indicate the importance of future skill clusters for a certain branch, we analyzed the absolute demand for the future skill clusters, i.e., the amount of job advertisements seeking at least one skill of the corresponding future skill cluster from 2018–2020.

4.2 Results

The data-driven approach by means of hierarchical clustering of skills from over 1 million job advertisements resulted in 33 future skill clusters for the manufacturing industry in Baden-Wuerttemberg assigned to four categories (Table 1). These future skills were validated by stakeholders in expert interviews and focus groups. The generic future skill clusters comprise skills that go beyond the typical professional skills and include, among others, Leadership, Problem Solving, and Resilience. The technological future skill clusters reflect the upcoming digital transformation (e.g., Software Development) as well as key technologies such as Data Science & AI. The digital future skill clusters comprise skills that enable employees to cope with a digitized environment and actively participate in shaping it (e.g., Digital and Data literacy, Digital Collaboration & Interaction). Finally, the industrial future skill clusters include industry-specific skills such as Alternative Drive Technologies and Industrial Engineering.

In order to provide stakeholders with additional information about the future skill clusters, we conducted an in-depth analysis. Therefore, in a first step, we analyzed the demand for the future skill clusters in the manufacturing industry in Baden-Wuerttemberg based on the data (Fig. 2). The future skill clusters with the highest demand are the generic skill clusters Organizational Skills (corresponding skills comprise 9.5% of all future skills sought in job advertisements from 2018 to 2020)

Table 1 Future Skills for the Manufacturing Industry in Baden-Wuerttemberg

	Future skill cluster	Most relevant skills based on job advertisements
Technological Skills	Cybersecurity	Firewalls; Information Security; security Incident Handling; VPN
	Data Management	Data Quality Assessment & Management; Databases; Data Processing
	Data Science & AI	Big Data Analytics; Deep Learning; Machine Learning; Python
	Design	Human-Machine-Interaction; Design of Interfaces; UI/UX/Interaction design; Web-Frontend Development; Visualization
	Intelligent Hardware & Robotics	Communications Systems; Embedded Systems; Hardware-in-the-Loop
	IT-Infrastructure & Cloud Systems	Cloud Computing; Cloud Security; Cloud Services; Computer Centre- & Server-Management; System Integration; Technical Consulting
	Sensors & IoT	Data Transmission in IoT; Development of Microsystems; Sensor Integration
	Software-Based Control of Business Processes	Customer Relationship Management; Digital Material Planning & Procurement; Digital Eco-Systems & Platforms; Process Management
	Software Development	Agile Software Development; App & Web Development; Automatic Programming; Container Technologies (Docker); DevOps
	Sustainable & Resource-Friendly Technologies	Green Technologies; Recycling Economy; Environmental Management; Environmental Compliance
	Agile Methods	Agile Working; Agile Project Management; Product Ownership
	Basic IT-Skills	Office Suites; Operating Systems; Data Protection
	Digital & Data Literacy	Search, Assessment & Selection of Digital Information; Critical and Ethical Data Handling; Online Security & Digital Identity
	Digital Collaboration & Interaction. Programming Skills	Digital interaction; (Digital) Teamwork; Collegial (Digital) Collaboration Object-Orientated Programming; Web Programming

Table 1 (continued)

Industrial Skills	Future skill cluster	Most relevant skills based on job advertisements
Alternative Drive Technologies		Battery Development; E-Fuels; Powertrain Engineering; Hydrogen
Analytical Chemistry		Material Analysis; Corrective & Preventive Action
Assisted & Autonomous Driving		Data Standards & Processing; Development of Driver Assistance Systems; Functional Safety; Legal Requirements; AutoSAR; Interlinkage of Cars
Biotechnology		Biochemical Analysis; Genome Editing; Cell Cultivation
Development of Medical Devices		Imaging Techniques; Connected Medical Devices; Digitalization of Medical Devices; Intelligent Medical Instruments; Wearables
Electrical Engineering		Digital Electronic; Laser; Performance Optimization; Microtechnology
Industrial Engineering		Automatization; High-Performance Plastic; Preventive & Predictive Maintenance; Simulation & Digital Twin; Technical Drawing and Modeling
Pharmaceutical Development of Products & Processes		Biopharmaceuticals (mRNA, RNA); Quality by Design; Therapeutic Development
Communication		Active Listening; Real-Time Communication; Storytelling
Customer Orientation		Customer Experience Management; Customer Understanding
Creativity		Innovative Thinking; Open-Mindedness
Flexibility		Adaptability; Willingness to Change
Goal Orientation		Efficiency; Objective and Key Results; Structured Working
Initiative		Engagement; Enthusiasm; Decision-Making; Proactivity
Leadership		Coaching & Guidance; (Critical) Feedback; Positive Leadership; Empathy
Organization Skills		Result-Orientated & Systematic Working; Self-Management; Reliability
Problem Solving		Coordination; Problem Solving; Structuring & Conceptualization
Resilience		Ambiguity Tolerance; Perseverance; Resistance

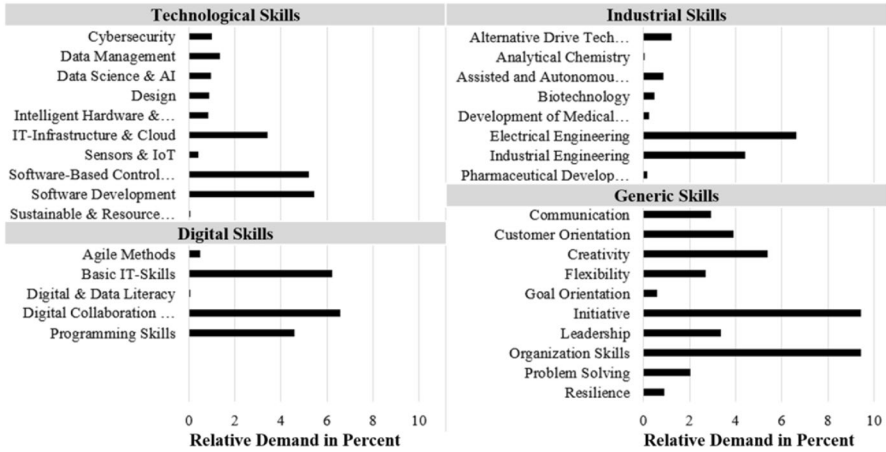


Fig. 2 Relative demand for the identified future skill clusters for the manufacturing industry in Baden-Wuerttemberg Based on job advertisements

and Initiative (9.4%) as well as the technological skill cluster Electrical Engineering (6.9%) (Fig. 2).

Moreover, we analyzed the growth of demand for the future skill clusters from 2018–2019, which indicates a growth in the importance of future skill clusters just before the global Covid-19 pandemic distracted the labor market. The future skill clusters with the highest growth in demand are Biotechnology (47.4%), Analytical Chemistry (45.2%), and Data Science & AI (42.9%) while all of them still show a rather low demand (Fig. 2).

In a second step, we analyzed specific sub-sectors of the manufacturing industry in Baden-Wuerttemberg to identify which future skill clusters are of particular interest in these specific industries. Therefore, we compared the relative demand for the future skill clusters in the whole manufacturing industry with the relative demand in specific industries (Appendix, Fig. 4). This analysis reveals interesting sub-sector-specific differences. For example, the future skill clusters Cybersecurity and Data Science & AI are particularly relevant for the automotive industry while the future skill clusters Sensors & IoT, Software Development, and Programming Skills are more frequently requested within the medical engineering industry (Appendix, Fig. 4).

In order to reflect the results of the future skills identification process with companies’ direct assessment of the status quo regarding the current workforce’s equipment with and the expected increase in the demand for the future skills, we conducted an online survey with companies in the manufacturing industry in Baden-Wuerttemberg. To ensure consistency with the results of our analysis, the survey was based on the same forecast horizon of five years as used in the case study. The survey consisted of two parts. In the first part, participants were asked to provide information on their company’s branch, their position in the company, and to provide the

number of employees in their company as well as an estimated number of employees in five years. In the second part, participants were presented with the 33 future skill clusters (Table 1) and asked how many percent of employees in their company were already equipped with skills of each future skill cluster as well as how many percent of employees should be equipped with such skills in five years. On this basis, for each future skill cluster, we estimated the current share of employees equipped with respective skills as well as the expected increase of this share in five years for the manufacturing industry in Baden-Wuerttemberg (Table 2). To derive these estimates, we first calculated the average values per branch by weighting the responses per number of employees of the company. Then, to obtain the results for the entire manufacturing industry in Baden-Wuerttemberg, we weighted the averages of the branches by the number of employees of the respective branches.

We distributed the survey via mailing lists from the federal state's leading employers' association and employees' association. Experts from 294 companies participated in the survey, representing diverse positions (53.4% human resources, 19.7% management, 12.9% works council, 4.4% production, 3.1% research and development, 6.5% other positions), company sizes (160 SMEs, 134 large companies), and industries (30.3% mechanical engineering, 25.9% metal-working industry, 22.1% automotive, 10.2% electrical engineering, 5.1% medical engineering, 6.5% other industries).

The results show that for each of the 33 future skill clusters, companies expect an increase of employees that are equipped with respective skills in the next five years (Table 2). The results are significant for each future skill cluster according to a Wilcoxon signed-rank test ($p < 0.001$). This fact underpins that the results of our future skills identification process indeed represent skills with increasing importance for the manufacturing industry in Baden-Wuerttemberg.

Finally, we compared the results of the data-driven analysis with the online-survey. Indeed, Fig. 3 shows a significant correlation between the demand of the future skill clusters from our data-driven analysis and the demand in the online survey ($r = 0.291$, $p < 0.1$). However, it also reveals a few outliers characterized by lower demand in the data-driven analysis but high demand according to the survey (e.g., "Digital & Data literacy"). These outliers underscore the benefit of incorporating expert-based concepts into our approach. By using focus groups with experts to decide which of the identified skill clusters indeed represent future skill clusters, our approach ensures the identification of important future skill clusters, even when their importance may not be completely obvious based on the data-driven analysis.

5 Discussion and conclusion

We developed a novel approach to identify future skills based on prior literature in the context of data-driven and expert-based identification of future skills (Pejic-Bach et al. 2020; Rakowska and Juana-Espinosa 2021). We applied the approach in a case study where we identified future skills for the manufacturing industry in

Table 2 Current Share of Employees Equipped with Skills (Cur.) and Expected Increase of this Share in 5 Years (Incr.) in the Manufacturing Industry in Baden-Wuerttemberg

Future skill cluster	Cur (%)	Incr (%)	Future skill cluster	Cur.(%)	Incr.(%)
<i>Technological Skills</i>					
Cybersecurity	21	42	Alternative Drive Technologies	15	79
Data Management	16	102	Analytical Chemistry	2	160
Data Science & AI	8	197	Assisted & Autonomous driving	10	80
Design	11	125	Biotechnology	1	42
Intelligent Hardware & Robotics	12	119	Development of Medical Devices	1	98
IT-Infrastructure & Cloud Systems	14	68	Electrical Engineering	19	52
Sensors & IoT	14	112	Industrial Engineering	25	22
Software-Based Control of Business Processes	17	105	Pharmaceutical Development of Products & Processes	1	255
Software Development	12	145	<i>Generic Skills</i>		
Sustainable & Resource-Friendly Technologies	15	99	Communication	46	52
<i>Digital Skills</i>					
Agile Methods	34	78	Customer Orientation	61	31
Basic IT-Skills	41	36	Creativity	47	44
Digital & Data Literacy	57	71	Flexibility	49	55
Digital Collaboration & interaction.	52	48	Goal Orientation	59	29
Programming Skills	16	103	Initiative	50	41
			Leadership	35	71
			Organization Skills	60	29
			Problem Solving	55	32
			Resilience	52	37

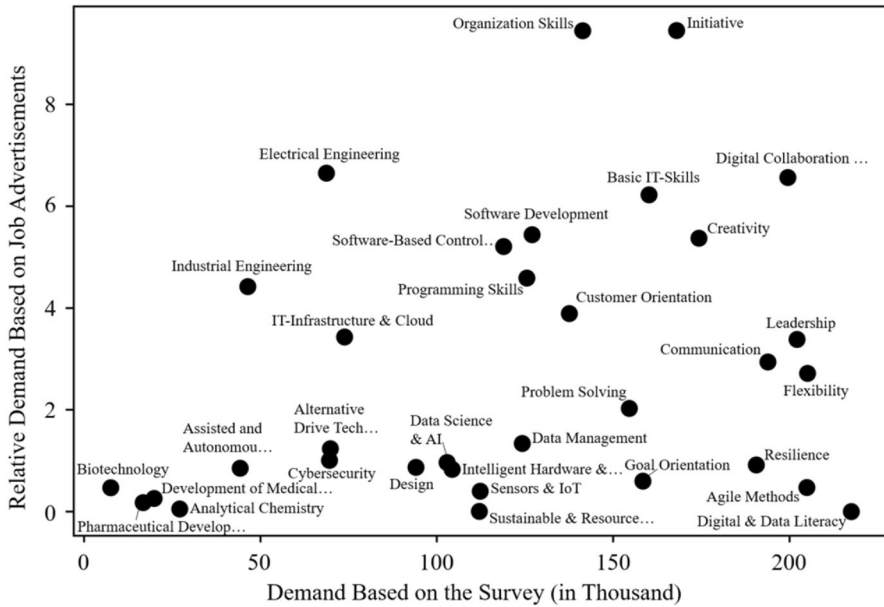


Fig. 3 Comparison of the demand for future skills based on job advertisements vs. the survey

Baden-Wuerttemberg to demonstrate its practical applicability and utility. The findings contribute to theory and practice in several ways.

5.1 Implications for theory and practice

From a theoretical point of view, our results indicate the following two implications. First, analyzing over 1 million job advertisements between 2018 and 2020, our novel approach to identify future skills by means of hierarchical clustering as an unsupervised machine learning method has proven successful, resulting in 33 future skill clusters assigned to four categories (generic skills, digital skills, technological skills, and industrial skills) for the manufacturing industry in Baden-Wuerttemberg. The data-driven approach allows for an effective analysis of a representative set of job advertisements of a given region and industry to derive future skills (Sonnewald et al. 2020). Thereby, the unsupervised machine learning method hierarchical clustering on a set of skill vectors with reduced dimensionality enables to process a large amount of data (Madhulatha 2012). While our results confirm common future skills (e.g., the generic skills like Communication, Flexibility, and Leadership) which can be found in well-known taxonomies (Kirchherr et al. 2018; Kotsiou et al. 2022; Rakowska and Juana-Espinosa 2021; Ra et al. 2019), other future skill clusters expand extant future skills taxonomies (e.g., Basic IT-Skills and all Industrial Skills). The identification of industrial skills shows, that our data-driven approach

is well-suited regarding a consideration of regional and industry-specific peculiarities. The identification of Basic IT-Skills as a future skill cluster might be surprising at first glance, as the importance of basic IT-skills has been known for a long time. Our focus groups and expert interviews revealed, that experts might overlook future skills like basic IT-skills, but confirm them if they are presented as input. The data-driven derivation of future skills in our approach serves well to generate important and otherwise overlooked input for future skills which are verified by humans in a second step. Other researchers have employed data-driven approaches to identify future skills as well. However, these studies are based on much smaller datasets and accordingly limited to the identification of future skills for single job positions. For instance, Föll et al. (2018) identify future skills for IS jobs based on 12875 job advertisements using LDA. Many existing data-driven approaches are unable to analyze large amounts of data due to a high runtime complexity (Sonntag and Roy 2011). With our case study, we are the first to leverage the potential of big data and derive future skills for an entire industry in an efficient way.

Second, we introduce a mixed-method approach (Creswell and Clark 2017; Reis et al. 2022) combining the advantages of data-driven (Rios et al. 2020) and expert-based (Rakowska and Juana-Espinosa 2021) methods for future skills identification. In our approach, data-driven derivation of future skills allows us to identify a comprehensive and representative set of future skills for a given region and industry. However, the interpretation of the results of data-driven approaches is described as a major challenge in literature (Ehlers 2020; van Laar et al. 2019). To enable governments, companies, educational institutions, and individuals to adapt to the future workforce and the transformation of society, it is paramount that future skills are meaningful to stakeholders. This means that the aggregation level of future skills is consistent with stakeholders' expertise. Against this background, our approach integrates stakeholders as experts to determine the level of aggregation of future skills. Indeed, in our case study, the stakeholders' assessment of the skill clusters resulting from the data-driven approach led to two iterations where the aggregation level was adapted. By incorporating both data-driven and expert-based concepts into a unified process, we leverage the strengths of both literature streams which allows us to identify a comprehensive set of future skills that are meaningful to stakeholders.

Beside these theoretical insights, our findings indicate four practical implications. First, governments around the world can use our novel process to accurately identify future skills regardless of industry, region, or job position under consideration. Using our novel future skills identification process for the manufacturing industry in Baden-Wuerttemberg, we were able to identify 33 future skill clusters, thereby enabling valuable insights into the change of the labor market in the manufacturing industry in Baden-Wuerttemberg in the next five years. Identifying future skills using our process enables governments to react to the shift of employees' skillsets in the future. Governments can proactively shape the conditions for an effective future of work by developing a strategic plan for the ecosystem, identifying skill needs, and using funding to initiate programs for reskilling and upskilling employees to close the skills gap (Leopold et al. 2016;

Zahidi et al. 2020) and therefore secure prosperity and social cohesion. Beyond, our approach enables fast and with little effort more in-depth analyses of either the whole industry under investigation or any of its' subgroups. For instance, analyzing (growth of) demand of the future skills in the underlying job advertisements allows identifying particularly relevant future skills for specific industries or regions. This facilitates the use of results obtained with our process, as practitioners can derive valuable insights for any subgroup of the industry and region investigated without the need to repeat the entire process. Second, education providers can use the results to align their education programs to future needs and design them accordingly (Föll et al. 2018). Third, companies can plan the indispensable realignment of their employees' skillset according to future challenges such as the surging demand for employees in the green economy or at the forefront of the data and AI economy (Leopold et al. 2016; Zahidi et al. 2020). With the identified future skills, especially SMEs can easily determine in which areas and to what extent employee qualification is necessary, thus facilitating the adaptation of vocational training and further qualification of employees. Furthermore, especially for SMEs our approach provides a convenient opportunity to compare with competitors or other industries (e.g., by analyzing the discrepancies in the demand for future skills between companies as well as industries). Fourth, our approach enables a quick and easy update of results as it is necessary especially after unforeseen disruptions of the labor market (e.g., as caused by the Covid-19 pandemic or Russia's invasion of Ukraine). Indeed, in the past the skills required in the job market changed significantly due to such disruptions (Zahidi et al. 2020). Thus, an adaptation of existing skill ontologies is necessary. In contrast to approaches solely based on literature reviews or experts, our approach can be updated with little effort. For example, in case of a labor market disruption after the completion of the future skills identification using our approach, the data-driven part of our approach can be updated by recalculating the demand for all skill clusters which have been present in the original data using more recent data. Such a recalculation enables a reassessment of the importance of skill clusters present in the original data when considering the disruptive event. This enables especially small companies with limited budget (i.e., regarding financial and working time resources) to regularly update the identified future skills and to react to changes in the labor market.

5.2 Limitations and future research

Although our research provides a substantial step towards the identification of future skills, it is subject to several limitations. First, we demonstrated the applicability and utility of our novel future skills identification approach only for one single use case. More concretely, the process was applied to identify the future skills for the manufacturing industry in Baden-Wuerttemberg, based on over 1 million job advertisements. Nevertheless, the approach is transferable to

other regions and industries. Therefore, we encourage future research to apply and evaluate the proposed process for other regions of varying sizes, to further test scalability, as well as for other industries. Second, as part of our data-driven derivation of future skills, skills are clustered based on their co-occurrence in the same job advertisements. While this constitutes a reasonable approach as job advertisements are targeted at specific job profiles (e.g., Pejic-Bach et al. 2020), other approaches can be applied to operationalize semantic similarity of skills, such as word embeddings (cf. Mikolov et al. 2013). Hence, we invite researchers to investigate alternative data-driven approaches to derive future skills based on job advertisements and compare the results. Third, the derivation of future skills as part of our process is based on the assumption that a comprehensive and representative set of job advertisements depict the future skills of a given region and industry. While job advertisements are considered to be one of the most promising sources for the analysis of skills (Descy et al. 2019), they might be criticized due to several reasons. Indeed, there is only little literature about how the skills demanded in job advertisements correlate to the actual skills needed on the job (Arcordia et al. 2020). For instance, when creating job advertisements recruiters may use outdated standard profiles or may be bound by legal requirements. This may result in job advertisements created without evidence of the actual skill demand and thus in differences between perceived and actually demanded skills. Moreover, it is uncertain whether the demand for skills strongly correlates with the on-the-job performance of the person with these skills. Therefore, we acknowledge that job advertisements may not comprise a full picture of skills that allow employees to be successful in the future workforce and enable citizens to participate in a globalized and fast-changing society (Föll et al. 2018). In addition, job advertisements as a data source may contain biases by their authors such as gender bias (Khaouja et al. 2021). In our approach, we aim at addressing these limitations by collecting a large number of job advertisements. Nonetheless, we encourage future research to explore how job advertisements can be complemented by other data sources for the derivation of future skills. Finally, although the online survey conducted to reflect the results of the future skills identification process provides first interesting and valuable insights, it does not represent a profound evaluation. Thus, we encourage future research to explore further possibilities to strengthen respective online survey-based evaluation, for example by including placebo skill clusters in the survey to analyze a potential “more of everything” effect in the responses. Beyond addressing these limitations, we invite future research to generalize our approach and explore further application domains. In particular, our approach could be used to examine and analyze more information available in job advertisements such as benefits offered to employees. Companies could make use of such an application, for example, to draw valuable conclusions for their employer branding campaigns.

Appendix

See Fig. 4.

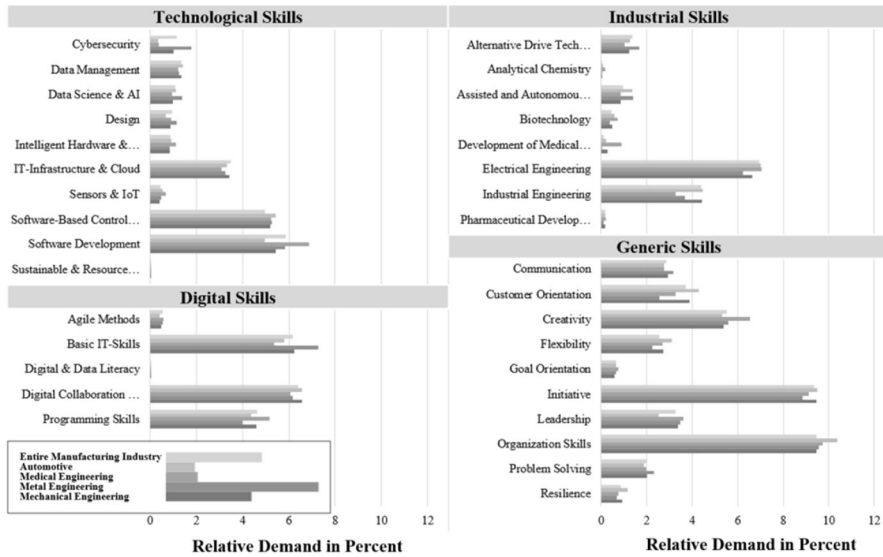


Fig. 4. Relative demand for the identified future skill clusters for sub-sectors of the manufacturing industry in Baden-Wuerttemberg

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The dataset generated during the current study is not publicly available as it contains proprietary information that the authors acquired through a license. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Abdallah YO, Shehab E, Al-Ashaab A (2021) Understanding digital transformation in the manufacturing industry: a systematic literature review and future trends. *Prod Manage Dev* 19(1):1–12

- Adl ST, Givchi A, Saraee M, Eshraghi A (2011) Disordered metabolic evaluation in renal stone recurrence: a data mining approach. *J Appl Comput Sci Math* 11:64–68
- Aichholzer G (2009) The Delphi Method: Eliciting Experts' Knowledge in Technology Foresight. In: Bogner A, Littig B, Menz W (eds) *Interviewing Experts*. Houndmills, UK, pp 252–274
- Ang S, Joseph D, Slaughter SA (2015) IT Professionals and the IT Profession. *Encycl Manag Inf Syst* 7:1–13
- Anowar F, Sadaoui S, Selim B (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput Sci Rev* 40:1–13
- Arcodia C, Novais MA, Le TH (2020) Using job advertisements to Advance Event Management Research. *Event Manage* 24(5):661–664
- Ayad HG, Kamel MS (2008) Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans Pattern Anal Mach Intell* 30(1):160–173
- Bandara W, Miskon S, Fiel E (2011) A systematic, tool-supported method for conducting literature reviews in information systems. In: *Proceedings of the 19th European Conference on Information Systems (ECIS 2011)*, Helsinki, Finland
- Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37(1):38–44
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bogner A, Littig B, Menz W (2009) Introduction: Expert Interviews—An Introduction to a New Methodological Debate. In: Bogner A, Littig B, Menz W (eds) *Interviewing Experts*. Houndmills, UK, pp 1–16
- Brancatelli C, Marguerie AC, Brodmann S (2020) Job creation and demand for skills in Kosovo: what can we learn from Job Portal Data? *World Bank Policy Research Working Paper* 9266:1–33
- Bro R, Smilde AK (2014) Principal component analysis. *Anal Methods* 6(9):2812–2831
- Brooks NG, Greer TH, Morris SA (2018) Information systems security job advertisement analysis: skills review and implications for information systems curriculum. *J Educ Bus* 93(5):213–221
- Buchmann M, Buchs H, Busch F, Clematide S, Gnehm A-S, Müller J (2022) Swiss job market monitor: a rich source of demand-side micro data of the labour market. *Eur Sociol Rev* 38(6):1–14
- Carnevale AP, Jayasundera T, Gulish A (2016) America's divided recovery: college haves and have-nots. *Georgetown University center on education and the workforce* pp 1–36
- Chilton MA, Hardgrave BC, Armstrong DJ (2010) Performance and strain levels of it workers engaged in rapidly changing environments: a person-job fit perspective. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 41(1):8–35
- Cicek K, Akyuz E, Celik M (2019) Future skills requirements analysis in Maritime Industry. *Procedia Comput Sci* 158:270–274
- Creswell JW, Clark VLP (2017) *Designing and conducting mixed methods research*. SAGE Publications, Thousand Oaks, CA
- Dash M, Faforia NG, Muthyala A (2018) A model for recruitment process costs in the indian IT industry. *J Strategic Hum Resource Manage* 7(1):1–8
- Davies A, Fidler D, Gorbis M (2011) *Future Work Skills 2020*. Institute for the Future for University of Phoenix Research Institute 540:1–14
- Dearnley C (2005) A reflection on the use of semi-structured interviews. *Nurse Res* 13(1):19–28
- Descy P, Kvetan V, Wirthmann A, Reis F (2019) Towards a shared infrastructure for online job advertisement data. *Stat J IAOS* 35(4):669–675
- Di Lucca GA, Di Penta M, Fasolino AR (2002) An approach to identify duplicated web pages. In: *Proceedings of the 26th International Computer Software and Applications Conference (COMPSAC)*, Oxford, UK
- Djumaliev J, Sleeman C (2018) An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. In: ESCoE (ed.) *Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring Across Europe*, pp 425–454
- Doyle-Kent M, Kopacek P (2020) Industry 5.0: Is the Manufacturing Industry on the Cusp of a New Revolution?. In: *Proceedings of the International Symposium for Production Research 2019*, Vienna, Austria
- Ehlers U-D (2020) *Future Skills*. Springer, Germany <https://doi.org/10.1007/978-3-658-29297-3>

- Ehlers, U.D. & Kellermann, S.A. (2019). Future Skills The future of learning and higher education. Karlsruhe. <https://www.learntechlib.org/p/208249/> Retrieved 26 June 2023
- ESCO (2020) About ESCO. <https://www.esco-projects.eu/esco/portal/howtouse/dc9a812c-8135-4f46-92f6-7364a1714ae0> Accessed 30 Sep 2022
- Evans JR, Mathur A (2005) The value of online surveys. *Internet Res* 15(2):195–219
- Firdaus S, Uddin MA (2018) A Survey on Clustering Algorithms and Complexity Analysis. *Int J Comput Sci Issues (IJCSI)* 12(2):62–85
- Föll P, Hauser M, Thiesse F (2018) Identifying the Skills Expected of IS Graduates by Industry: A Text Mining Approach. In: *ICIS 2018 Proceedings*, San Francisco, CA
- Frei I, Grund C (2022) Working-time mismatch and job satisfaction of junior academics. *J Bus Econ* 92:1125–1166
- Ganesan M, Antony SP, George EP (2018) Dimensions of job advertisement as signals for achieving job seeker's application intention. *J Manage Dev* 37(5):425–438
- Gardiner A, Aasheim C, Rutner P, Williams S (2018) Skill requirements in Big Data: a content analysis of job advertisements. *J Comput Inform Syst* 58(4):374–384
- Gregor S, Hevner AR (2013) Positioning and presenting Design Science Research for Maximum Impact. *MIS Q* 37(2):337–355
- Grimes MF, Grimes PW (2008) The academic librarian Labor Market and the role of the Master of Library Science Degree: 1975 through 2005. *J Acad Librariansh* 34(4):332–339
- Gugnani A, Misra H (2020) Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation. In: *Proceedings of the 2020 AAAI Conference on Artificial Intelligence*, A virtual conference
- Gurcan F, Cagiltay NE (2019) Big data software engineering: analysis of knowledge domains and Skill Sets using LDA-Based topic modeling. *IEEE Access* 7:82541–82552
- Gurcan F, Ozyurt O, Cagiltay NE (2021) Investigation of emerging Trends in the E-Learning field using latent Dirichlet Allocation. *Int Rev Res Open Distrib Learn* 22(2):1–18
- Handali JP, Schneider J, Dennehy D, Hoffmeister B, Conboy K, Becker J (2021) Industry Demand for Analytics: A Longitudinal Study. In: *European Conference on Information Systems 2021*, Marrakech, Morocco
- Helmcke S, Heuss R, Hieronimus S, Engel H (2021) *Net-Zero Deutschland*. McKinsey & Company, Germany
- Heuser P, Letmathe P, Schinner M (2022) Workforce planning in production with flexible or budgeted employee training and volatile demand. *J Bus Econ* 92:1093–1124
- Hevner AR, March ST, Park J, Ram S (2004) Design Science in Information Systems Research. *MIS Q* 28:1
- Hickman L, Thapa S, Tay L, Cao M, Srinivasan P (2022) Text preprocessing for text mining in Organizational Research: review and recommendations. *Organizational Res Methods* 25(1):114–146
- Hörisch F, Wurster S (2019) The policies of the First Green-Red Government in the german Federal State of Baden-Württemberg, 2011–2016. *Politische Vierteljahresschrift* 60:513–538
- Ilek T, Maier C, Weinert C (2022) Electronic Human Resource Management: A Literature Analysis of Drivers, Challenges, and Consequences. In: *Wirtschaftsinformatik 2022 Proceedings*, Nuremberg, Germany
- Inzalkar S, Sharma J (2015) A survey on text mining-techniques and application. *Int J Res Sci Eng* 24:1–14
- Johnson RB, Onwuegbuzie AJ (2004) Mixed methods research: a Research Paradigm whose time has come. *Educational Researcher* 33:7
- Joseph D, Tan ML, Ang S (2011) Is updating play or work?: the mediating role of updating orientation in linking threat of professional obsolescence to turnover/turnaway intentions. *Int J Social Organizational Dynamics IT* 1(4):37–47
- Kalleberg AL (2008) The mismatched worker: when people don't fit their Jobs. *Acad Manage Perspect* 22(1):24–40
- Khaouja I, Kassou I, Ghogho M (2021) A survey on skill identification from online job ads. *IEEE Access* 9:118134–118153
- Kirchherr J, Klier J, Lehmann-Brauns C, Winde M (2018) Future Skills Welche Kompetenzen Deutschland fehlen. In: *Stifterverband (ed) Future Skills Diskussionspapier*, McKinsey&Company, Germany
- Klus MF, Müller J (2021) „The digital leader: what one needs to master today's organisational challenges. *J Bus Econ* 91:1189–1223

- Kolenikov S, Angeles G (2005) The use of discrete data in principal component analysis for socio-economic status evaluation. University of North Carolina at Chapel Hill, "Chapel Hill, NC
- Kotsiou A, Fajardo-Tovar DD, Cowhitt T, Major L, Wegerif R (2022) A scoping review of future skills frameworks. *Ir Educational Stud* 41(1):171–186
- Krueger RA, Casey MA (2015) *Focus Groups: a practical guide for applied research*. Sage Publications, California
- Kurtzo F, Hansen MJ, Rucker KJ, Edgar LD (2016) Agricultural Communications: perspectives from the experts. *J Appl Commun* 100(1):33–45
- Landau S, Leese M, Stahl D, Everitt B (2011) *Cluster analysis*. Wiley, "Chichester West Sussex, UK.
- Leopold TA, Ratcheva VS, Zahidi S (2016) The future of jobs: employment, skills, and workforce strategy for the fourth industrial revolution. *World Economic Forum, Switzerland*
- Lieu TTB, Duc NH, Gleason NW, Hai DT, Tam ND (2018) Approaches in developing undergraduate IT Engineering Curriculum for the Fourth Industrial Revolution in Malaysia and Vietnam. *Creative Educ* 09(16):2752–2772
- Litecky C, Aken A, Ahmad A, Nelson HJ (2010) Mining for Computing Jobs. *IEEE Softw* 27(1):78–85
- Martínez-Plumed F, Contreras-Ochando L, Ferri C, Orallo H, Kull J, Lachiche M, Quintana N Ramirez, M. J, and Flach PA, (2019) CRISP-DM Twenty Years later: from data mining processes to data-science trajectories. *IEEE Trans Knowl Data Eng* 33(8):3048–3061
- Madhulatha TS (2012) An overview on clustering methods. *IOSR J Eng* 2(4):719–725
- Maer-Matei MM, Mocanu C, Zamfir A-M, Georgescu TM (2019) Skill needs for early Career Researchers—A text Mining Approach. *Sustainability* 11(10):1–17
- Marbán O, Segovia J, Menasalvas E, Fernández-Baizán C (2009) Toward data mining engineering: a software engineering approach. *Inform Syst* 34:87–107
- McInnes L (2018) UMAP Documentation: Basic UMAP Parameters <https://umap-learn.readthedocs.io/en/latest/parameters.html>. Accessed on 1 Apr 2022
- McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*
- Michalczyk S, Nadj M, Maedche A, Gröger C (2021) Demystifying Job Roles in Data Science: A Text Mining Approach. In: *European Conference on Information Systems 2021, Marrakech, Morocco*
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 26:3111–3119
- O*NET (2022) About O*NET <https://www.onetcenter.org/overview.html#data>. Accessed 30 Sep 2022
- Pefanis Schlee R, Harich KR (2010) Knowledge and skill requirements for marketing Jobs in the 21st Century. *J Mark Educ* 32(3):341–352
- Pejic-Bach M, Bertoncel T, Meško M, Krstić Ž (2020) Text mining of industry 4.0 job advertisements. *Int J Inf Manag* 50:416–431
- Peng T, Li L, Kennedy J (2014) A comparison of techniques for name matching. *GSTF J Comput* 2(1):55–62
- Pflügler C, Becker N, Wiesche M, Krcmar H (2018) Strategies for retaining key IT professionals. *MIS Q Exec* 17:4
- Prifti L, Knigge M, Kienegger H, Krcmar H (2017) A Competency Model for Industrie 4.0 Employees. In: *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, Switzerland*
- Prommegger B, Wiesche M, Krcmar H (2020a) What makes IT professionals special? A literature review on context-specific theorizing in IT workforce research. In: *Proceedings of the 2020 on Computers and People Research Conference*, pp 81–90
- Prommegger B, Intane J, Wiesche M, Krcmar H (2020b) What Attracts the New Generation? Career Decisions of Young IT Professionals. In: *Proceedings of the 28th European Conference on Information Systems, Marrakech, Morocco*
- Ra S, Shrestha U, Khatiwada S, Yoon SW, Kwon K (2019) The rise of technology and impact on skills. *Int J Train Res* 17:51
- Rahmat AM, Adnan AHM, Mohtar NM (2020) Becoming Industry 4.0 workers: Technical, methodological, social, personal and digital processing capabilities. The management of work performance, safety and entrepreneurship trend in Malaysia, pp 103–111
- Rakowska A, de Juana-Espinosa S (2021) Ready for the future? Employability skills and competencies in the twenty-first century: the view of international experts. *Hum Syst Manage* 40(5):669–684

- Ray P, Reddy SS, Banerjee T (2021) Various dimension reduction techniques for high dimensional data analysis: a review. *Artif Intell Rev* 54(5):3473–3515
- Reis L, Maier C, Weitzel T (2022) Mixed-methods in information systems research: status quo, core concepts, and future research implications. *Commun Associ Inf Syst* 51:17
- Rios JA, Ling G, Pugh R, Becker D, Bacall A (2020) Identifying critical 21st century skills for Workplace Success: a content analysis of job advertisements. *Educational Researcher* 49(2):80–89
- Rollins JB (2015) Foundational methodology for Data Science. IBM Corporation, Somers, NY
- Ros F, Guillaume S (2019) A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Syst Appl* 128:96–108
- Rugard M, Jaylet T, Taboureau O, Tromelin A, Audouze K (2021) Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. *PLoS ONE* 16(5):1–17
- Schallock B, Rybski C, Jochem R, Kohl H (2018) Learning factory for industry 4.0 to provide future skills beyond technical training. *Procedia Manuf* 23:27–32
- Schenkenhofer J (2022) Hidden champions: a review of the literature & future research avenues. *Manage Rev Q* 72(2):1–66
- Sczogiel S, Schmitt-Rüth S, Göller A, Wiliger B (2019) “Future Digital Job Skills: Die Zukunft kaufmännischer Berufe - Langversion.” *Industrie - und Handelskammer Nuernberg für Mittelfranken* (ed.), Nuernberg, Germany (in German)
- Sonnwald M, Dutkiewicz S, Hill C, Forget G (2020) Elucidating ecological complexity: unsupervised learning determines global marine eco-provinces. *Sci Adv* 6(22):1–11
- Sonntag D, Roy D (2011) Complexity of inference in latent dirichlet allocation. *Adv Neural Inf Process Syst* 24:1008–1016
- Sorensen LC, Ladd HF (2020) The hidden costs of teacher turnover. *AERA Open* 6(1):1–24
- Stefanić J, Šimić D (2021) An overview of skills foresight methods. In: *Central European Conference on Information and Intelligent Systems*, pp 283–289
- Suuronen S, Ukko J, Eskola R, Semken RS, Rantanen H (2022) A systematic literature review for digital business ecosystems in the manufacturing industry: prerequisites, challenges, and benefits. *CIRP J Manufact Sci Technol* 37:414–426
- Tan LM, Laswad F (2018) Professional skills required of accountants: what do job advertisements tell us? *Acc Educ* 27(4):403–432
- Todd PA, McKeen JD, Gallupe RB (1995) The evolution of IS job skills: a content analysis of IS job advertisements from 1970 to 1990. *MIS Q* 19:1–27
- Tremblay MC, Hevner AR, Berndt DJ (2010) Focus groups for artifact refinement and evaluation in Design Research. *Commun association Inform Syst* 26(1):27
- van Laar E, van Deursen AJAM, van Dijk JA, de Haan J (2019) Twenty-first century digital skills for the creative industries workforce: perspectives from industry experts. *First Monday* 24(1):1–16
- Vermeulen M, Smith K, Eremin K, Rayner G, Walton M (2021) Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks. *Spectrochim acta Part A Mole Biomol Spectrosc* 252:119547
- Viégas FB, Wattenberg M, Dave K (2004) Studying cooperation and conflict between authors with history flow visualizations. In: *Proceedings of the 22th International Conference on Human Factors in Computing Systems*, Vienna, Austria
- vom Brocke J, Simons A, Riemer K, Niehaves B, Plattfaut R, Cleven A (2015) Standing on the shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Commun Association Inform Syst* 37(9):205–224
- Wattenberg M, Viégas F, Johnson I (2016) How to use t-SNE effectively. *Distill* 1(10):2
- Wentling RM, Palma-Rivas N (1998) Current status and future trends of diversity initiatives in the workplace: diversity experts’ perspective. *Hum Res Dev Q* 9(3):235–253
- Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th International conference on the practical applications of knowledge discovery and data mining*, Manchester, UK
- Wirsky T, Laumer S, Eckhardt A, Weitzel T (2016) On the untapped value of e-HRM: a literature review. *Commun association Inform Syst* 38(1):2
- Yang Q, Zhang X, Du X, Bielefeld A, Liu Y (2016) Current market demand for core competencies of librarianship—a text mining study of American library association’s advertisements from 2009 through 2014. *Appl Sci* 6(2):48
- Yin RK (1981) The Case Study as a Serious Research Strategy. *Knowledge* 3(1):97–114

- Yousaf M, Khan MSS, Rehman TU, Ullah S, Jing L (2021) NRIC: a noise removal Approach for Nonlinear Isomap Method. *Neural Process Lett* 53(3):2277–2304
- Zahidi S, Geiger T, Crotti R, Brown S, Hingel G (2019) The global competitiveness report 2019. World Economic Forum, Switzerland
- Zahidi S, Ratcheva VS, Hingel G, Brown S (2020) The future of jobs report 2020. World Economic Forum, Switzerland

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.