A systematic review of Data Analytics Job Requirements and Online Courses

Mohamad Almgerbi[a], Andrea De Mauro[b], Adham Kahlawi[a], and Valentina Poggioni[c]

[a] University of Florence, Florence, Italy; [b] University of Rome Tor Vergata, Rome, Italy; [c] University of Perugia, Perugia, Italy

Abstract

Data analytics' growing importance in modern business has left many organizations unprepared in terms of human talent. This study sheds light on the intersection between the analytics job skills currently in demand and the offer of massive online open courses for developing them. We have scraped from the web the description of more than 14,000 job posts and 3,600 Data Analytics online courses to systematically capture the need for data skills and available learning opportunities. By using an original combination of topic modeling and text mining algorithms, we provide a systematic mapping of educational offers with business needs, quantifying their presence and identifying gaps. Our study enables both educational providers to improve their offering on Data Analytics and Human Resources professionals to identify skill development opportunities. Additionally, we introduce a general methodology able to produce systematic mappings of job skills and learning opportunities in any domain.

Keywords: Data Analytics, Data Science, Big Data, Human Resources Management, Topic Modeling, Online Education

This is the **preprint** version of the article published on September 27th, 2021 in Journal of Computer Information Systems © Taylor & Francis. To cite this article: Mohamad Almgerbi, Andrea De Mauro, Adham Kahlawi & Valentina Poggioni (2021): A Systematic Review of Data Analytics Job Requirements and Online-Courses, Journal of Computer Information Systems, DOI: 10.1080/08874417.2021.1971579

1. Introduction

The acceleration of digital technologies development witnessed over the last decades has made a large amount of data available across all fields (i.e. Big Data), forcing companies to go beyond traditional methods for best leveraging it [1]. Therefore, data professionals have experienced the need to develop a broader skill set including competencies such as large database management, statistical modeling, cloud computing, computer programming in order to positively impact company performance [2]. However, the development of such skills required for Big Data Analytics is not keeping pace with the rapid increase of business needs [3,4] and generates the requirement for a more structured and continuous skill development process.

Data Science is an emerging multi-disciplinary domain that combines multiple fields such as statistics, business domain knowledge and computer science to equip data professionals to proficiently leverage Big Data. Despite Data Scientists have been defined the sexiest professionals [5] and as legendary as unicorns [6], the education and training of Data Science professionals currently lack a commonly accepted, harmonized model able to represent the multi-disciplinary knowledge required from the Data Science practitioners in modern, data-driven research and the digital economy [7,8]. Additionally, Data Science covers only part of the broader set of skills required by data professionals in companies [6,9].

Several studies have attempted to fill this gap by proposing standard data analytics skills training programs [6,10–13]. The Association for Computing Machinery (ACM) also intercepted this need and founded the ACM Data Science Task Force to articulate the role of computing discipline-specific contributions to this emerging field [14].

In our modern data-driven society, characterized by a fast technology change and strong skills demand, the education required by data professionals should be customizable and delivered in multiple forms. As noticed by Belloum et al. [15], the job market is looking for more senior and mid-career staff rather than fresh graduates and this suggests that short training education should be offered along with traditional education formats like BA, MSs, and PhDs. Candidates suited for senior or mid-career positions should have graduated at least five years before the advent of profiles related to Data Analytics and they make up a wide group of customers for shorter and more flexible professional courses. In this perspective, Massive Online Open Courses (MOOCs) lend themselves very well also due to their flexibility: in fact, MOOCs offer a prospect of education beyond the confines of individual universities and organizations, with the possibility of free participation for large numbers of learners from any geographical location and without the need to satisfy formal entry requirements. For these reasons, in the last decade MOOCs have been able to capture both the learners and data science market needs as well as in several other emerging fields.

MOOCs are an online phenomenon that offers the facilitation of obtaining knowledge and skill in different fields of study. They drew attention to their ability to transmit the course content via the recorded videos to a large group of users; thus, they can be considered as one of the most effective methods of distance learning [16,17]. MOOCs platforms like Coursera and Udacity are an example of breaking the college credit monopoly which is based on elements of traditional education (faculty, curriculum, credentials) [18,19]. Over the years, these platforms have been evolving and adapting to the needs and problems by certain rigid restrictions in the design, management, and outcome of MOOC courses. Furthermore, the spreading of MOOCs through social media has proven to motivate prospective students and to make the learning experience more impactful [20]. The university community considers MOOCs as a successful experience and the number of universities offering them is rapidly increasing. Indeed, the MOOCs come to be considered as an indicator of universities' educational technology [21]. MOOCs cannot be considered as a substitute for traditional education but have to be considered as an effective method that can be integrated with it [22].

With the intent of shedding light on the intersection between the analytics skills required by employers and the ones whose development is currently offered by MOOCs, we decided to pursue the following research questions:

- RQ1: What are the homogeneous skills which can be currently learnt through MOOCs available online today?

- RQ2: What are the relevant skills which employers demand when recruiting for Data Analytics jobs?
- RQ3: Are MOOCs able to provide a comprehensive development path for today's Data Analytics professionals? Is there any gap between offer and demand?

To answer these questions, in this study we analyzed both the labour market requirements for Data Analytics professionals, such as Business Analyst, Data Scientist, Data Developer and Systems Engineer, and the related educational offering provided by MOOC courses with the objective of assessing the efficiency of these courses in covering the needs of the labour market.

Our broader aim has been to develop a general methodology for systematically bridging job requirements and learning objectives. Such methodology can be reused on any professional domain other than Big Data Analytics and support both companies and educational providers in the systematic identification of gaps and opportunities.

This article is organized as follows: Section 2 addresses the related work; Section 3 explains the methodology used to answer the research question; Section 4 presents the results of the analysis; finally, Section 5 includes conclusions and opportunities for future works.

2. Related works

Several previous studies proposed approaches to study and evaluate the gap between offer and demand in professions related to Information Technology (IT). The disconnect between the skills required by industries and those provided in academic courses has a long history and it is one of the main criticisms made against the academy, especially in the IT world [23–28]. This is mainly due to the extreme speed at which new technologies and new demands appear in the IT-related fields, which have been always difficult to intercept in the design of academic courses. This is even more evident in the case of modern disciplines like Data Science and Data Analytics.

In [25] and more recently in [26] authors proposed studies to determine whether the importance of various skills for entry-level IT workers is perceived differently by faculties in academia than by IT managers. Moreover, in [26] authors identified areas of asymmetry between curricula and industry expectations. The methodology used in these analyses is very different from the one proposed in this paper because they were conducted by submitting surveys to people and analysing the collected data employing standard statistical tests. Clearly, these analyses are very limited in the amount of utilised data and are not scalable.

Similar studies have been conducted in new and more specialised fields related to IT like, for example, Big Data professions [9,29,30] or Supply Chain Management [31] as well as Data Science [15], Business Intelligence and Data Analytics [32–34].

In [29] the authors propose a semi-automatic procedure to extract information from job offers related to Big Data Software Engineering (BDSE). They leverage the LDA topic modeling technique and create a map comprising the essential knowledge domains, skills, and tools for big data software engineering. In particular, the authors analysed the job offers and identified 48 trending topics describing the knowledge domains and skill sets typical of BDSE; then they mapped the 48 topics into 10 core competency areas; finally, they revealed the tools and technologies required for BDSE by analyzing the data set through a keyword indexing technique. The authors suggest to improve the communication level of the industry and software engineering educational programs by using their results in the definition such programs, but they didn't provide any suggestion on how to implement this connection.

In [9] authors adopt LDA topic modelling to retrieve 9 homogeneous skill sets which appear in job offers related to Big Data. The identified skills are then mapped with four human-annotated job families, namely Business Analysts, Data Scientists, Developers and Systems Engineers. Finally, each job family is characterized by the appropriate level of mastery required within each Big Data skill set.

Similarly, [33,34] focused on Business Intelligence and Data Analytics positions. They analysed datasets of job posts extracted by well-known job recruitment web sites, but using different methodologies. [33] applied LDA for topic modelling, found 75 topics, organized them in a taxonomy and grouped them in 5 main areas that, according to the authors, should be considered the main areas where the competencies required by employers are located. [34] applied instead a content analysis method on the job descriptions based on a set of predefined skill categories and returned the distribution of the job posts within such categories. They then focused the results on the differences between four different profiles, namely Business Analyst, Business Intelligence Analyst, Data Analyst and Data Scientist.

A quite different approach has been used in [30]. In this work, the authors produced a conceptual model of knowledge, skills, and abilities required for Big Data practitioners. The model highlighted the large complexity of the Big Data discipline in terms of job skills requirements. Moreover, it showed also how the skills required are linked to the Data Science professions. The main limitation of this work resides in the proposed methodology that is essentially based on a manual annotation of input data. The authors themselves admitted that since Big Data is a fast-evolving discipline, their analysis is just a "snapshot in time" and should be updated over time. Clearly, also this approach is not reproducible systematically and on a large scale.

We found that the works most related to ours are [15,32]. In [15] authors introduce a service designed to extract relevant information from Data Science–related documents to help the comparison of demand and offer in the field of Data Science education and Human Resources (HR) management. The paper describes the software and its use but fails to provide any detail on the methodology. Moreover, a significant part of the work (all the Data Science education analysis) was performed manually and this makes the approach non-scalable to larger amounts of data and its results not reproducible. [32] provided a connection between the concepts required by employers and those offered in courses. The former are extracted by crawling job recruitment platforms and extracting knowledge from the job posts, while the latter are deduced by a manual knowledge extraction process applied to programs offered at universities and colleges in Canada (61 institutions). The authors propose the identification of semantic concepts in the texts, then the concepts are grouped into themes which are visualized in concept maps. Despite several limitations, this is the only work presenting two different datasets, one for jobs and one for courses.

In conclusion, we can say that several other works analyzed the gap between Academia and Job Market, especially in the case of new disciplines like Data

Science, Data Analytics or Business Intelligence. Most of them proposed to investigate the Job Market through the analysis of a dataset of job posts crawled by famous websites and employment platforms. All these works proposed to analyze job posts through text mining tools. The methodologies used are different but similar in the workflow: after the crawling phase, a preprocessing is implemented to obtain more reliable data (tokenization, stop words removal and in some cases stemming) and a semantic content analysis is performed on these cleaned data (LDA in most cases). They in general differ for the geographical scope, (Canada in [32]), as well as for different focus (Business Intelligence in [33], a comparison among Business Analyst, Business Intelligence Analyst, Data Analyst and Data Scientist in [34]. Results are in general interesting and organized also through taxonomies [33]. Nevertheless, all these works in general failed in producing an automatic (or semi-automatic) reproducible and scalable method able to create a correspondence between the skills required by the job market with the skills learnt in courses and Academia.

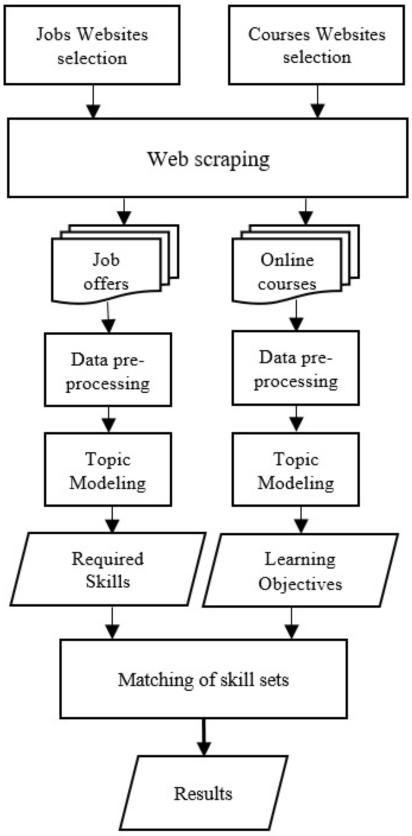

Figure 1: Flowchart of methodology

3. Methodology

In order to reach the objective of this study, we have designed a 5-steps methodology, which is depicted in Figure 1. Firstly, we have identified a set of job websites and a set of MOOCs websites. Secondly, we have extracted a large number of job offers and MOOCs using Web Scraping techniques. Thirdly, we have preprocessed the extracted data. Fourthly, we have identified relevant skill sets by applying a topic modeling algorithm on the content of both job offers and MOOCs, so to categorize them as a number of required skills and provided skills. Lastly, we have matched the skills provided by the online courses with the ones required by employers to determine the compatibility of the acquired skills from the courses with the requirements of the labour market.

3.1 *Identification of websites*

The objective of the first step was to identify a set of websites that contain data related to our study. Initially, we identified the most prominent platforms available on the World Wide Web which offered job posts and online courses. We selected websites by considering the number of posts, the geographic scope and web scraping feasibility. Tables 1 and 2 show the results of our assessment of the job post websites and the online course websites, respectively. The last column in each table reports our qualitative indication of web scraping feasibility. We individually tested each website to ensure that its data could be scraped, then we tested the quality of the extracted data. In order to ensure the robustness of our data sources, we have selected the first two websites from each table for which nearly all posts could be retrieved.

Table 1 Job Websites selection matrix.

Website	# of relevant posts	Geographic scope	Web scraping feasibility
Simply Hired	1.000+	Global	●●●
CareerBuilder	1.000+	Global	●●●
Monster	1.000+	Global	●●
Robert Half	1.000+	Global	●
ZipRecruiter	1.000+	Global	●
Google for Jobs	1.000+	Global	●

Table 2 Course Websites selection matrix.

Website	# of relevant posts	Geographic scope	Web scraping feasibility
Coursera	1.000+	Global	●●●
Udacity	100+	Global	●●●

Lynda	1.000+	Global	●●
Codemy	1.000+	Global	●●
edX	1.000+	Global	●
Udemy	1.000+	Global	●

Notes: The number of dots in the last column indicate how feasible it was to scrape from each web site (● = it is not possible to download quality results; ●● = some elements of text can be downloaded with varying quality; ●●● = nearly all posts can be successfully retrieved).

3.2 *Data collection - Web scraping*

Web scraping provides tools for extracting and using relevant data and information from the Internet [35]. Multiple studies adopted web scraping techniques to build a corpus of documents for subsequent topic modelling [9,15,29,30,32–34,36]. To extract a large number of data points related to online courses and online job offers, we built a custom web scraper software using Python able to extract titles and descriptions of every job position and course available in the selected websites. In order to extract only the data related to this study, we used six keywords which collectively cover the field of Big Data Analytics, namely: "big data", "data science", "business intelligence", "data mining", "machine learning" and "data analytics". These keywords were used to extract the data for both job offers as well as online courses. Using the web scraper, we were able to extract 14,495 online job ads published in a six-month period from March 2019 to August 2019 and 3,636 online courses that were published in the same period, we deleted the duplicate job ads and courses, then we removed those which are not related to our study. Through these progressive cleaning steps of the data set, which are summarized in Table 3, we finally obtained a dataset containing 9,067 online job ads and 764 online courses. To the best of our knowledge, this is the biggest dataset containing both job offers and courses descriptions. Other works proposed similar dataset containing only data extracted from job posts [9,15,29,30,32–34] but these datasets contain data from 3,000 job posts at most. The only one which collected a dataset also for courses is [32] but the dataset is very limited both in dimensions and geographical scope.

Table 3 Dataset size.

Preprocessing Steps	Job Posts	Online Courses
Initial dataset	14,495	3,636
Duplicate data	4,682	1,404
Data not in English	487	817
Data unrelated to our study	259	651

Final dataset	**9,067**	**764**

3.3 *Data preprocessing*

Usually, the data extracted from the internet is unstructured and needs to be preprocessed before being utilized for modelling. In this study, the preprocessing phase applied to the extracted data consisted of several sequential steps. Initially, the description of each job and course was divided into a group of text strings using Python nltk.tokenize package. We have observed that many sentences were repeated in job posts, jeopardizing the ability of the topic modelling algorithm to infer sets of words related to meaningful concepts. We noticed that such repetitions were often providing general information about the company or organization offering the job. In order to avoid any negative impact on the results of the LDA results, these duplicate sentences were removed from the database. Given the dynamic length of such sentences, we developed an original extension of the standard stop-words preprocessing step which we called stop-N-gram. N-grams consist of a set of co-occurring words within a given window [37]. In order to identify stop-N-grams we have first spotted the duplicate sentences and texts that were repeated for more than 10 times in the corpus, and created a list of such sentences (stop sentences). Then we have progressively increased the window of words in each stop sentence, creating larger N-grams until we found all long repetitions in the corpus and removed them from the original text accordingly.

After the elimination of stop-N-grams we proceeded with the more traditional pre-processing steps: we built a term-document matrix, within which each string in the corpus was represented by a word vector. Then, punctuation, web links, HTML tags, and meaningless characters were deleted. Stop words and other words not related to our study were then removed from the texts. Finally, we have created the dictionary and corpus needed for Topic Modeling.

3.4 *Topic Modeling*

LDA is a probabilistic and generative model based on an unsupervised approach, frequently used in text mining [38,39]. The main aim of this algorithm is to extract the semantic content of unstructured documents by analyzing the latent semantic structures within the documents [40]. It can be effectively applied to huge collections of documents in a given text corpus to discover semantic patterns and it has already been employed in other studies that performed the text analysis of online job advertisements in industries [9,29,41]. In this study, we have applied LDA twice considering as input first the corpus of job offers descriptions and then the corpus of online course descriptions. In LDA each document (in our case, a job offers or a course description) is represented as a finite mixture over an underlying set of topics and each topic is represented as an infinite mixture over an underlying set of topic probabilities. The topics are created considering the frequency of the terms in the documents of the corpus. LDA algorithm identifies a predetermined number of topics in the corpus, where each topic is described as a set of tuples composed of terms and frequencies, indicating the occurrence of each term in each topic. The frequency of the terms in each topic is unique and determines its conceptual meaning. An important parameter to be set a priori for LDA is the number of topics, k. In this study, different values of k ranging from 4 to 30 were tried and resulted in alternative topic models, which were subsequently assessed by their level of sense-making through expert judgement, as in [9]. As we implemented the LDA model on job offers we found that the best model was obtained with k=7 as seen in Table 4. When we have applied LDA on the data related to online courses, the ideal number of topics was k=6 resulting in the topics reported in Table 5. In the human assessment of the ideal number of topics, both the semantic consistency of the discovered topics and the distribution of the descriptive keywords in these topics were taken into consideration. Once we obtained the topic models, we proceeded to manually assign topic names to the discovered topics, in consistency with the descriptive keywords. The topic names were assigned by considering the general meaning of all keywords. Human experts were aided also by word clouds, i.e. a visual representation of text data, to visualize the conceptual content within each job and course topic as inferred by LDA.

3.5 *Matching of skill sets*

Latent Semantic Indexing (LSI) [42–44] was used to perform the process of mapping between the two topic models obtained through LDA. LSI has proven to be the similarity index which is most widely applied when assessing conceptual distance across textual documents [45,46]. In fact, LSI is based on the assumption that similar words manifest in similar text fragments. The basic principles of LSI are [47]:

- The average of word presence in all documents gives the meaning of this word;
- The multi-word constructs obtain their meaning by the words shaped within these constructs;
- The examination of word co-occurrence with every single word rivers the latent associations between these words.

For this study, the input to LSI was a synthetic document created by collating the thirty most influential words in each topic, adjusted by their importance. The adjustment was obtained by repeating each word by a number of times proportional to the within-topic weight, as inferred through LDA. More precisely, each word was repeated T times, where $T = 100 * w$, where w is the within-topic weight of each word.

Before applying LSI we executed word tokenization to the synthetic documents [48] and proceeded with the value decomposition (SVD) of the underlying Term-Document matrix. Afterwards, the documents were expressed in the semantic space and vectors representing texts in the semantic space were created. Finally, the value of the angle between the vectors was calculated and used as an indication of the degree of similarity between topics belonging to each topic model.

4. Results and Discussion

By applying LDA algorithm on each corpus of documents (namely, job offers and online courses) we have obtained two independent topic models composed of 7 and 6 topics, respectively. By analyzing the most frequent keywords identified within each topic distribution (top words) and by reading the documents carrying a noteworthy presence of each topic, we were able to infer the essential elements of each topic. In this section we provide a synthetic description of each topic as identified in LDA for each of the topic models, reporting their most salient features.

4.1 *Skill sets demand*

We have identified 7 homogenous skill sets within the corpus of online job offers, which are summarized in Table 4, together with a selection of relevant keywords by topic and the relative presence of each topic out of the total corpus.

Skill set 1: Marketing Analyst. This skill set includes the ability to gather, analyze, and synthesize marketing intelligence from data and to turn it into strategic insights for brands. Our model found this particular skill set worthy of being listed as a separate topic because of the specificity of the marketing-focused data types (such as sales trends and price points, consumer behavior logs and comments) and analytical methods (Social Network Analysis, SNA, propensity modeling, competitiveness assessments) involved. To notice, this skill set is the least covered one in the analyzed job posts (4.7% of relative presence) and this is a sign of how specific, yet relevant it appears to be in the current job market.

Skill set 2: Business Intelligence Analyst. Job posts collectively describe this skill set as the ability to determine and describe business questions and create reports and visualizations to answer them accordingly. Roles requiring this competency are required to interact with either business intelligence application for creating interactive dashboards that enable wide access to data for the organization, or directly sourcing raw data by means of structured database queries.

Skill set 3: Project Manager. The skill set refers to the capacity to provide overall direction and coordination in the planning, execution and monitoring of data transformation projects. We found that jobs calling out the need for this skill set include either: a) the responsibility of driving direct business impact with data by executing commercial or marketing activities as a response of data prescriptions, or b) the accountability for process transformation based on data-based tools, or c) the design and deployment of data infrastructure, platforms or applications. These three cases cover a broad set of data-related roles in firms, making project management a foundational skill set for all data professionals. This is corroborated by the fact that the skill set is the most frequently present in the analyzed corpus of job posts (27.0% of relative presence).

Skill set 4: Software Developer. Many job posts in our data set displayed coding and software development as an explicit requirement. Being fluent in one or more programming languages is, indeed, necessary to either customize the behaviour of already-implemented algorithms or to develop and integrate data applications. We have found that many job posts related to data required knowledge and experience in specific areas related to programming, such as the development of Application Programming Interfaces (API) for the interaction with cloud-based services and the usage of agile software engineering methodologies.

Skill set 5: Data Science. This skill set enables the usage of advanced analytical methods for the transformation of data into insights. Job posts including this skill set also refer to the responsibility of identifying patterns, designing and implementing data models and statistical methods and integrating research and best practices into problem avoidance and continuous improvement. Our model recognizes this skill set in job posts calling out for scripting or programming languages particularly popular in data science, such as Python, R and Scala, and often reference specific areas of analytical expertise, e.g. optimization and prediction.

Skill set 6: Data Engineer. The skill set focuses on building and maintaining the full technology stack which enables the utilization of Big Data in a firm. Job posts described this skill set as the ability to define and maintain the corporate data architecture, support the process of extracting, transforming and loading data (ETL) into the analytics data store and also manage cloud resources accordingly. Indeed, we have found that many job posts mention specific cloud platforms, such as Microsoft Azure and Amazon Web Services (AWS), and software frameworks for distributed computing, like Hadoop and Spark.

Skill set 7: Machine Learning. The last skill set we identified focuses on the specific ability to develop, configure, deploy and interpret results of artificial intelligence and, more specifically, machine learning models. Many of the job posts include a direct reference to deep learning, i.e. the broad exploitation of neural networks models and algorithms. Although neural networks are only a specific set of models within the wide range of machine learning algorithms, we found that deep learning is often explicitly mentioned as a self-standing skill set because of both its complexity and current popularity.

Table 4 Skill Sets identified in job posts with most relevant frequent keywords and relative presence in the corpus.

Skill set	***Selection of top keywords***	***Relative presence***
Marketing Analyst	Sales, Marketing, Products, Price, Strategy, Social media, Competition	4.7%

Business Intelligence Analyst	Business intelligence, SQL, Dashboard, Tableau, Database, Finance	14.2%
Project Manager	Project, Program, Lead, Plan, Process, Risk, Team	27.0%
Software Developer	Coding, System, Programming, Web, Cloud, Agile, API	21.3%
Data Analytics	Statistics, Model, Data Science, Mathematics, Python, Prediction, Optimization	16.2%
Data Engineer	Big data, Database, ETL, Warehouse, Hadoop, AWS, Spark	10.4%
Machine Learning Expert	Machine learning, Deep learning, Model, Artificial intelligence, Tensorflow, Python	6.2%

4.2 *Learning Objectives offer*

Through the application of LDA we have identified six topics within the descriptions of online courses scraped from the web: each topic represents a homogeneous group of Learning objectives offered by such courses. You will find in Table 5 a summary of these topics, the most relevant keywords and the relative presence for each.

Learning objective 1: Machine Learning. Many online courses promise to provide practical and theoretical expertise on artificial intelligence in general and machine learning in particular. They cover both the fundamental concepts of statistical learning (such as validation, under and overfitting, classes of learning algorithms) and concrete know-how on tools and libraries dedicated to machine learning. Some of the courses cover specific classes of algorithms or applications, like deep learning or text mining.

Learning objective 2: Application Development. This topic refers to the objective of enabling students to develop data-centred software solutions. Courses in this area cover basics of software development applied mainly to distributed computing, cloud environments, APIs and web applications, as well as introductions to specific development frameworks (including proprietary cloud solutions), languages or libraries which are popular in the domain of Big Data.

Learning objective 3: Tools for Data Analytics. This development objective aims at growing students' agility in using the application toolkit for data analytics. Courses insisting on this development objective are normally focusing on one specific software solution and include tutorials and realistic examples for making participants autonomous users of the tool. We have found that this topic is the second most present in the corpus analyzed, highlighting the importance of hand-on analytical expertise in today's business world.

Learning objective 4: Statistical Modeling. This topic covers the objective of providing the fundamentals of statistics and statistical inference to support data science needs. Generally, courses displaying this topic as prevalent guide students in the process of drawing conclusions about populations or scientific truths from data through modelling of existing data or strategies for designing data-generating experiments. Furthermore, many courses covering this objective will include data visualization techniques and tools as they appear to be indissoluble supporting skills that enable an effective delivery of insights derived from statistical models. We found that this topic had the highest coverage in the courses within our corpus, confirming that statistical modelling is one of the most sought-after learning objectives for data talents.

Learning objective 5: Marketing Analytics. This learning objective focuses on providing hybrid expertise for marketers and data professionals with the objective of letting them apply data analytics to real-world challenges. Courses in this area tend to cover the practical utilization of techniques often associated with consumer understanding and brand-building activities, such as text mining, social network analysis, sentiment analysis, real-time bidding, online campaign optimization, modelling and forecasting of marketing investments and financial impact.

Learning objective 6: Coding. This topic covers the fundamental development needs related to computer programming. We noticed that courses connected with this topic are generally providing expertise in one or more programming languages and cover syntax, semantics and libraries used within the data domain.

Table 5 Skill Sets identified in MOOCs with most relevant frequent keywords and relative presence in the corpus.

Skill set	***Selection of top keywords***	***Relative presence***
Machine Learning	Machine learning, Deep learning, Artificial intelligence, Mathematics, Supervised, Unsupervised, keras	17.1%
Application Development	JavaScript, CSS, html, Platform, API, Interface, PHP	10.2%
Tools for Data Analytics	Data analytics, Business analyst, Alteryx, Cluster, Classification, Financial, SQL	22.5%
Statistical Modeling	Statistics, Data analytics, Linear regression, Distribution, Data viz, Python, MongodB	23.3%
Marketing Analytics	Marketing analytics, Client, Customer, Trade, Investment, Forecast, Risk, Code	12.1%

Coding	Programing, Python, Algorithm, Java, Database, Cybersecurity, SQL	14.8%

4.3 *Matching*

By applying LSI to the thirty most influential words of each topic within each of the two hierarchies (Job skills and Learning objectives) we have obtained a similarity matrix which is graphically represented in the alluvial diagram reported in Figure 2. It is important to highlight that the height of each stripe connecting the topics is proportional to the degree of similarity found by each pair-wise of topics. Hence, the thinner stripes represent weak connections and, as a consequence, thinner topics constitute the ones that have little to no match in the other hierarchy. For example, Application Development is a learning objective for which it appears there is a relative low match vs. skill set most in demand. Conversely, Marketing Analytics is a learning objective which displays the broadest fit with skill sets required within data analytics jobs. We noticed that, out of the job skills identified in the study, Data engineering is the one showing the weakest match with learning objectives (lowest cumulative scores of similarity, depicted with the relative narrow stripe you can find in the picture), highlighting the need for a more robust MOOC offering in this space.

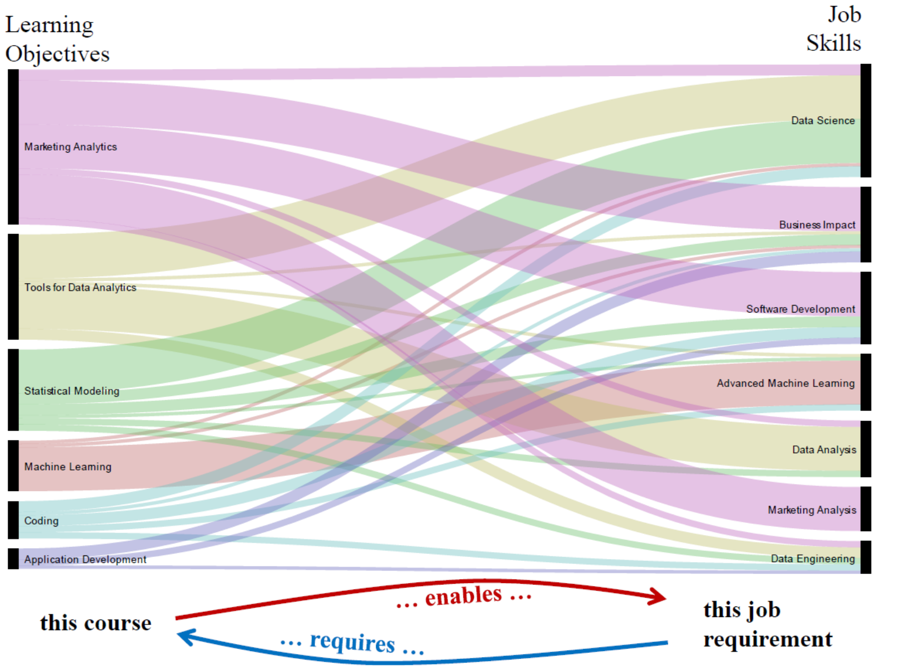

Figure 2 alluvial diagram of the similarity matrix

5. Conclusions

With the present study, we have shed light over the intersection between those analytics skills that are in demand within the job marketplace and the offering of online courses which promise to develop such skills. By using an original combination of web scraping techniques and topic modelling algorithms, we have provided a systematic mapping of educational offers and business needs, highlighting current gaps, using more than 9,000 job descriptions posted online and more than 760 course descriptions posted online.

Through the application of LDA algorithm on each corpus of documents (Job offers and Online courses) we have obtained two independent topic models composed of 7 and 6 topics, respectively, each representing a homogeneous group of skills required within job posts and learning objectives included in course descriptions. Then, thanks to the application of LSI algorithm, we obtained a pair-wise similarity value for each item of the two hierarchies (Job skills and Learning objectives), which we used to identify gaps between the two.

As a response to RQ1, we have identified 6 homogeneous Learning Objectives which can be currently reached through MOOCs available online today, which are: Machine Learning, Application Development, Tools for Data Analytics, Statistical Modeling, Marketing Analytics, Coding. Similarly. As per RQ2, we have identified the 7 most relevant skills demanded by recruiters looking for Data Analytics talent, which are: Marketing Analysis, Business Intelligence, Project Management, Software Development, Data Analytics, Data Engineering, Machine Learning Expertise. Lastly, as a response to RQ3, we matched jobs skills and learning objectives and concluded that MOOCs are able to cover for the large majority of the development needs required by today's Data Analytics professionals, although we found still opportunities of unmatched demand, especially for Data Engineering.

Our results can support education providers, HR professionals and researchers in multiple ways. Educators can build or enrich their educational curricula by closely matching current job demands. In fact, the methodology we presented in this work provides a structured approach to systematically reduce the gap between skills demand and offer, enabling a continuous validation of educational programs. HR professionals can use our mapping in two ways: first, they can identify training opportunities which best fit with their workforce requirements, designing an evidence-based framework for the development of data skills. Secondly, they can design clear job requisitions by leveraging the homogeneous sets of skills identified in this study. Additionally, academic researchers can reuse our original methodology for producing systematic mappings of job skills and educational requirements in virtually any industry domain other than Big Data Analytics, extending the scope of this paper.

We recognize a number of limitations in the current study that offer the opportunity for further research. Firstly, the feasibility of web scraping forced us to select a limited number of job post and MOOC websites to be used as a source for the research: despite the large number of documents retrieved, we might have obtained a partial view of the total universe. Secondly, although we adopted a replicable combination of methodologies like LDA and LSI, the choice of the model parameters generating the most comprehensible output has been left to human judgement, making it possibly prone to subjective biases. Lastly, in this study we decided to consider only online MOOC courses, due to the large availability of quality descriptions of each course. However, future studies might extend the scope and cover both online and traditional off-line university offering, in a way to obtain a more comprehensive view of gaps between jobs and learning opportunities. By doing so, it would be also possible to assess local-relevant patterns, i.e. the sufficiency of universities located in a geography to respond to the specific job requirements in demand of the local industry.

References

[1] N. Golchha, Big data-the information revolution, Int. J. Appl. Res. 1 (2015) 791–794.

[2] O. Kwon, N. Lee, B. Shin, Data quality management, data usage experience and acquisition intention of big data analytics, Int. J. Inf. Manage. 34 (2014) 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002.

[3] Big Data Analytics, (n.d.).

[4] A. Luntovskyy, L. Globa, Big Data: Sources and Best Practices for Analytics, in: 2019 Int. Conf. Inf. Telecommun. Technol. Radio Electron., 2019: pp. 1–6. https://doi.org/10.1109/UkrMiCo47782.2019.9165334.

[5] T. Patil, T. Davenport, Data scientist: The sexiest job of the 21st century, Harv. Bus. Rev. 90 (2012) 70–76.

[6] T. Davenport, Beyond unicorns: Educating, classifying, and certifying business data scientists, Harvard Data Sci. Rev. (2020).

[7] A. Manieri, S. Brewer, R. Riestra, Y. Demchenko, M. Hemmje, T. Wiktorski, T. Ferrari, J. Frey, Data Science Professional Uncovered: How the EDISON Project will Contribute to a Widely Accepted Profile for Data Scientists, in: 2015 IEEE 7th Int. Conf. Cloud Comput. Technol. Sci., 2015: pp. 588–593. https://doi.org/10.1109/CloudCom.2015.57.

[8] L. Cao, Data Science: Profession and Education., IEEE Intell. Syst. 34 (2019) 35–44. https://doi.org/10.1109/MIS.2019.2936705.

[9] A. De Mauro, M. Garco, M. Grimaldi, P. Ritala, Human resources for Big Data professions: A systematic classification of job roles and required skill sets, Inf. Process. Manag. 54 (2018) 807–817. https://doi.org/10.1016/j.ipm.2017.05.004.

[10] Y. Demchenko, A. Belloum, C. de Laat, C. Loomis, T. Wiktorski, E. Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, in: 2017 IEEE Int. Conf. Cloud Comput. Technol. Sci., 2017: pp. 363–368. https://doi.org/10.1109/CloudCom.2017.59.

[11] A.A. Dyumin, S. V Andrianova, MOOCs and Vendor Trainings in Academic Curriculum: Yet Another Step towards Global University, in: 2016 Int. Conf. Eng. Telecommun., 2016: pp. 39–44. https://doi.org/10.1109/ENT.2016.017.

[12] U. Fayyad, H. Hamutcu, Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards,

Harvard Data Sci. Rev. (2020). https://doi.org/10.1162/99608f92.1a99e67a.

[13] I.-Y. Song, Y. Zhu, Big data and data science: what should we teach?, Expert Syst. 33 (2016) 364–373. https://doi.org/10.1111/exsy.12130.

[14] A. Danyluk, P. Leidig, L. Cassel, C. Servin, ACM Task Force on Data Science Education: Draft Report and Opportunity for Feedback, in: Proc. 50th ACM Tech. Symp. Comput. Sci. Educ., Association for Computing Machinery, 2019: pp. 496–497. https://doi.org/10.1145/3287324.3287522.

[15] A.S.Z. Belloum, S. Koulouzis, T. Wiktorski, A. Manieri, Bridging the demand and the offer in data science, Concurr. Comput. Pract. Exp. 31 (2019). https://doi.org/10.1002/cpe.5200.

[16] S.M. North, R. Richardson, M.M. North, To Adapt MOOCs, or Not? That Is No Longer the Question., Univers. J. Educ. Res. 2 (2014) 69–72. https://doi.org/10.13189/UJER.2014.020108.

[17] S. Gao, Y. Li, H. Guo, Understanding the Value of MOOCs from the Perspectives of Students: A Value-Focused Thinking Approach, in: S.A. Al-Sharhan, A.C. Simintiras, Y.K. Dwivedi, M. Janssen, M. Mäntymäki, L. Tahat, I. Moughrabi, T.M. Ali, N.P. Rana (Eds.), Challenges Oppor. Digit. Era, Springer International Publishing, Cham, 2018: pp. 129–140. https://doi.org/10.1007/978-3-030-02131-3_13.

[18] J.G. Mazoue, The MOOC model: Challenging traditional education, Educ. Rev. Online. (2014).

[19] J.L. Martín Núñez, E. Tovar Caro, J.R. Hilera González, From Higher Education to Open Education: Challenges in the Transformation of an Online Traditional Course, IEEE Trans. Educ. 60 (2017) 134–142. https://doi.org/10.1109/TE.2016.2607693.

[20] O.B. Gené, M.M. Núñez, Á.F. Blanco, Gamification in MOOC: challenges, opportunities and proposals for advancing MOOC model, in: Proc. Second Int. Conf. Technol. Ecosyst. Enhancing Multicult., 2014: pp. 215–220. https://doi.org/10.1145/2669711.2669902.

[21] F.J. García-Peñalvo, Á. Fidalgo-Blanco, M.L. Sein-Echaluce, An adaptive hybrid MOOC model: Disrupting the MOOC concept in higher education, Telemat. Informatics. 35 (2018) 1018–1030. https://doi.org/10.1016/j.tele.2017.09.012.

[22] S.D. Krause, C. Lowe, Invasion of the MOOCs: The promises and perils of massive open online courses, San Fr. Parlor Press. (2014) 223–228.

[23] E.R. Doke, S.R. Williams, Knowledge and Skill Requirements for Information Systems Professionals: An Exploratory Study, J. Inf. Syst. Educ. 10 (2000) 10.

[24] Y. Kim, J.F. Hsu, M. Stern, An Update on the IS/IT Skills Gap, J. Inf. Syst. Educ. 17 (2006) 395–402.

[25] C.L. Aasheim, L. Li, S. Williams, Knowledge and skill requirements for entry-level information technology workers: A comparison of industry

and academia, J. Inf. Syst. Educ. 20 (2019) 10.

[26] Y.G. Sahin, U. Celikkan, Information Technology Asymmetry and Gaps Between Higher Education Institutions and Industry, J. Inf. Technol. Educ. Res. 19 (2020) 339–365. https://doi.org/10.28945/4553.

[27] N. Shmatko, G. Volkova, Bridging the Skill Gap in Robotics: Global and National Environment, SAGE Open. 10 (2020) 2158244020958736. https://doi.org/10.1177/2158244020958736.

[28] A.K. Jha, M.A.N. Agi, E.W.T. Ngai, A note on big data analytics capability development in supply chain, Decis. Support Syst. 138 (2020) 113382. https://doi.org/10.1016/j.dss.2020.113382.

[29] F. Gurcan, N.E. Cagiltay, Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling, IEEE Access. 7 (2019) 82541–82552. https://doi.org/10.1109/ACCESS.2019.2924075.

[30] A. Gardiner, C. Aasheim, P. Rutner, S. Williams, Skill Requirements in Big Data: A Content Analysis of Job Advertisements, J. Comput. Inf. Syst. 58 (2018) 374–384. https://doi.org/10.1080/08874417.2017.1289354.

[31] C. Wagner, F. Sancho-Esper, C. Rodriguez-Sanchez, Skill and knowledge requirements of entry-level logistics and supply chain management professionals: A comparative study of Ireland and Spain, J. Educ. Bus. 95 (2020) 23–36. https://doi.org/10.1080/08832323.2019.1596870.

[32] A. Persaud, Key competencies for big data analytics professions: a multimethod study, Inf. Technol. People. ahead-of-p (2020). https://doi.org/10.1108/ITP-06-2019-0290.

[33] F. Gurcan, S. Sevik, Business Intelligence and Analytics: An Understanding of the Industry Needs for Domain-Specific Competencies, in: 2019 1st Int. Informatics Softw. Eng. Conf., 2019: pp. 1–5. https://doi.org/10.1109/UBMYK48245.2019.8965457.

[34] A. Verma, K.M. Yurov, P.L. Lane, Y. V Yurova, An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements, J. Educ. Bus. 94 (2019) 243–250. https://doi.org/10.1080/08832323.2018.1520685.

[35] A. V Saurkar, K.G. Pathare, S.A. Gode, An Overview On Web Scraping Techniques And Tools, Int. J. Futur. Revolut. Comput. Sci. Commun. Eng. 4 (2018) 363–367.

[36] Y. Jung, Y. Suh, Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews, Decis. Support Syst. 123 (2019) 113074. https://doi.org/10.1016/j.dss.2019.113074.

[37] G. Sidorov, H. Gómez-Adorno, I. Markov, D. Pinto, N. Loya, Computing text similarity using tree edit distance, in: 2015Annual Conf. North Am. Fuzzy Inf. Process. Soc. Held Jointly with 2015 5th World Conf. Soft Comput., 2015: pp. 1–4. https://doi.org/10.1109/NAFIPS-

WConSC.2015.7284129.

[38] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[39] D. Sharma, B. Kumar, S. Chand, A Trend Analysis of Machine Learning Research with Topic Models and Mann-Kendall Test, Int. J. Intell. Syst. Appl. 11 (2019) 70–82. https://doi.org/10.5815/IJISA.2019.02.08.

[40] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (2004) 5228–5235.

[41] F. Gurcan, C. Kose, Analysis of software engineering industry needs and trends: Implications for education, Int. J. Eng. Educ. 33 (2017) 1361–1368.

[42] B. Rosario, Latent semantic indexing: An overview, Techn. Rep. INFOSYS. 240 (2000) 1–16.

[43] A. Mirzal, Clustering and latent semantic indexing aspects of the singular value decomposition, Int. J. Inf. Decis. Sci. 8 (2016) 53–72. https://doi.org/10.1504/IJIDS.2016.075790.

[44] N.A. Rahman, Z. Mabni, N. Omar, H.F.M. Hanum, N.N.A.T.M. Rahim, A Parallel Latent Semantic Indexing (LSI) Algorithm for Malay Hadith Translated Document Retrieval, in: M.W. Berry, A. Mohamed, B.W. Yap (Eds.), Soft Comput. Data Sci., Springer Singapore, Singapore, 2015: pp. 154–163. https://doi.org/10.1007/978-981-287-936-3_15.

[45] N.N. Amirah, T.M. Rahim, Z. Mabni, H.M. Hanum, N.A. Rahman, A Malay Hadith translated document retrieval using parallel Latent Semantic Indexing (LSI), in: 2016 Third Int. Conf. Inf. Retr. Knowl. Manag., 2016: pp. 118–123. https://doi.org/10.1109/INFRKM.2016.7806346.

[46] A. Kahlawi, C. Martelli, L. Buzzigoli, L. Grassini, A similarity matrix approach to empower ESCO interfaces for testing , debugging and in support of users ’ experience, in: A. Pollice, N. Salvati, F.S. Spagnolo (Eds.), Riun. Sci. Della Soc. Ital. Di Stat. -SIS, Pearson, Pisa, 2020: pp. 904–909.

[47] A. Rozeva, S. Zerkova, Assessing semantic similarity of texts--methods and algorithms, in: AIP Conf. Proc., 2017: p. 60012. https://doi.org/10.1063/1.5014006.

[48] D. Suna, S. Zhao, Z. Zhanga, X. Shia, A MATCH METHOD BASED ON LATENT SEMANTIC ANALYSIS FOR EARTHQUAKE HAZARD EMERGENCY PLAN, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 42 (2017).