7. WHO REALLY IS A DATA SCIENTIST? ANALYSIS OF REQUIREMENTS FOR DATA CENTRED ROLES JOB MARKET AND THEIR FUTURE

https://doi.org/10.18559/978-83-8211-143-9/7

Piotr Kałużny

Department of Information Systems
Poznań University of Economics and Business
piotr.kaluzny@ue.poznan.pl

Klaudia Karpińska

Comarch S.A.
klaudia.karpinska285@gmail.com

Łukasz Krawiec

Unisoft sp. z o.o.
krawieclukasz1995@gmail.com

Abstract

Data analysis and processing skills are currently required by a multitude of job offers and cover a wide variety of applications. Although mostly shaped by the development of new technologies, programming languages and libraries, they are a necessity in the world of digital economy and entrepreneurship. A multitude of reports by large consulting companies such as Deloitte predict a sharp increase in demand for data science and AI roles in the future of not only the IT sector, but also the entire economy. The following questions arise: "What skillset do these innovators that use artificial intelligence and advanced analytical skills have?" and "What skills and requirements truly make a data scientist and are they are any different to that of data analysts, data engineers or software developers and programmers?", moreover, "What is the demand for these specialists and are the university programs educating future specialists in this field or are the skills too new and need to be taught solely by business practice?" . To answer these questions, this article applies Natural Language Processing (NLP) techniques of machine learning to characterize and extract from the offers key skills important for data centred roles. The research was carried out on a preprocessed sample of 72 thousand job offers from the IT sector posted in 2019. A SVM linear classifier was applied to extract the most distinguishing technical skills and characterize the possibility of the automated classification of job postings, which resulted in about 85% precision and recall values for classifying data analyst, data scientist and data engineer roles and about 90% for classifying python developer roles.

Keywords: data scientist, data engineer, data analyst, job offers, job postings, big data, data analysis, job market, data mining, natural language processing, text mining, education.

Suggested citation:

Kałużny, P., Karpińska, K., & Krawiec, Ł. (2022). Who really is a data scientist? Analysis of requirements for data centred roles job market and their future. In H. Nowak-Mizgalska & A. Szulczewska-Remi (Eds.), *Academic entrepreneurship in theory and practic*e (pp. 107–138). Poznań: Poznań University of Economics and Business Press. https://doi.org/10.18559/978-83-8211-143-9/7

Introduction

It has been nine years since "data scientist" has been acclaimed as one of the most promising careers of the 21st century by the *Harvard Business Review* (Davenport & Patil, 2012). Despite the varying nature of the name: 'quants' (Miller & Hughes, 2017), 'data warehousing and ETL specialists', 'big data specialists', 'data scientists' or 'machine learning and AI experts', the need for people with a broad array of problem-solving and analytical skills centred around data processing and analysis has been rising steadily for the last 10 years. The McKinsey report from 2011 (Manyika et al., 2011) forecasted a need for hundreds of thousands of 'data centred' jobs in the next decade and their predictions were mostly met by the market. The value of the data analysis market, regardless of whether it is called 'AI', 'machine learning', 'big data' or 'data science' is rising and is currently facing significant skilled worker shortages all over the world (DuBois, 2021).

This may be fueled partly by the FAANG (Facebook, Amazon, Apple, Netflix, Google) tech giants' success in applying data skills and connected technologies. The largest companies in the world are now data science reliant enterprises that have built their position using large IT infrastructures and innovative analytics to achieve their position on the market. Their competition, on the other hand, needs to comply with the new market standard of using extensive data analytics.

The largest business social media platform—LinkedIn—has observed nearly a seven-fold increase in demand for "data science" roles and a ten-fold increase in machine learning jobs in 2017, compared to 2012. The growth in popularity and demand has also been present in 2020 (Seaman, 2021; Olsen, 2021) over multiple reports on most promising and rising jobs. The role of data scientist has been acclaimed "best job" by Glassdoor platform since 2016, and it fell to the third place only in 2020, due to the increase in demand for front and back end developers. This popularity, however, not only creates the need for multiple job openings in data centred areas, but also emphasizes these skills as a necessity for entrepreneurship and success in the startup and technology-oriented mindset of investors. Vision, creativity and management skills are and will remain key factors of success for rising entrepreneurs. However, in the world of AI, big data, mobile and wearable applications, fin-tech, med-tech and reg-tech, those companies need experts in applying these new technologies.

To build a knowledge-based economy, a large number of skilled specialists are needed. There is no denying that in the last few years, data processing, machine learning and AI took a large part in developing technical innovations. These skills are mostly connected with the traditional role of the data scientist—but do all companies need a data scientist or maybe do some need not one but a lot of them? To gain answers, we need to know who exactly is a data scientist and what skills he or she needs to possess. Moreover, we should also consider if there are any emerging roles with similar skillsets that may be hidden under the popularity of the name and may be crucial for innovation. What is more, in the highly competitive employee oriented job market of IT, with the shortage of data science talents, it may be hard to convince these specialists to change their well-paid positions to take up an unsure future as a startup tech leader. Still, has the pandemic changed the demand situation? Furthermore, did the shortage of data scientists come about because universities are not addressing the demand and generating these experts or is it just the result of too high demand in a short term?

The goal of this chapter is to characterize the data centred roles job market, analysing both the demand side and the supply of skilled data professionals educated in Poland, compared to the flagship data science programs. In terms of those roles, the main goal is to ascertain the current skillset sought by the market for different data centred positions. This chapter is aimed at answering the following questions:

- Is the demand for data centred roles still high in 2020/2021, how has the pandemic affected it?—For those entering into the job market or higher education, it is important to recognize whether data science is still a growing trend or the recent events have decreased the demand for the skills and possibly ended its rising popularity.
- Which skills are the most prevalent and to what extent in the data centred roles?—The concept of data science is diverse, and there is no single definition as to what contributes to the skills of a data scientist. Measuring the diversity of required skills and similarities between the data centred positions will provide answers to what is actually needed. The results of the analysis could also point out to the dynamics and changes of the current market, highlighting the new skills gaining popularity in the recent year.
- Which skills are unique to the roles identified?—The analysis of skill importance and structure may point out to the different characteristics of data centred roles influencing the entry barrier, market demand and diversity of skills required. It is important to know what differentiates e.g. a data analyst from a data engineer and how that may influence the role that the given specialist should play in the company.

The approach applied to identify the key skills and differentiate between the said roles relies on a NLP classifier applied across the collected job postings, which based on a keyword detection are assigned to one of the data centred roles. A SVM machine learning classifier is further on employed to allow for the identification of the key skills that can differentiate between the roles and to automatically classify job postings based on its contents. Our approach is in line with other researchers' studies. This will be described in the subchapter "Comparison of results with other studies", which compares the results and methods applied by multiple authors focused on similar topic, mainly on the US job market. Our proposed method of key skill extraction however adds an ontology database matching approach not used by other studies who have utilized similar NLP methods.

Additionally, we will check if key skills identified for these roles are taught in “data science” and similar academic programs in Poland and how they compare to some of the top US and UK programs. This will, of course, be a preliminary study and only the beginning to some more detailed analysis. Its aim is to confirm if today’s academic programs are in line with meeting the market demand for data centred roles in terms of the curriculums and overall analytical skills taught. It will also help in identifying the skills the alumni will need to learn on their own from different code camps, micro degrees or from business practice.

This work has the following structure: first section is the introduction, describing the problem area and the main questions motivating the research. The next section describes the market of the “data centred” roles, along with the impact that the COVID pandemic had on it, thus aiming to answer the first research question. Further on, the second section describes the issue of analysing the key skills extracted from a large sample of job postings, along with the results answering to the second and the third research question. In this section both the problem, the dataset used, the NLP machine learning approach applied and the achieved results are described. The third part aims at comparing the key skills found out in the analysis with the curriculums of various “data science” and similar academic programs. Finally, the last section presents the conclusions of the article.

7.1. Data science—demand (market)

In 2018, the World Economic Forum (WEF) published its predictions for the future workforce through to 2022. In it, WEF predicts that by 2022, 85% of all business entities will have adopted big data and analytics technologies and 96% of all enterprises will be likely to hire new permanent positions to fill these roles (World Economic Forum, 2021).

According to Grand View Research, in 2019, the global data science platform market size was valued at $3.93 billion. Moreover, the role of data scientist has become one of the most in demand jobs in both the UK and the US. Indeed, the role appeared on LinkedIn’s 2020 Emerging Job Report in both countries, featuring at number 3 in the US and number 7 in the UK, the first being AI specialist in both of the cases (LinkedIn, 2020a; LinkedIn, 2020b). In 2019 (Blake, 2021), in the UK, the demand for data scientists and data engineers tripled over the past five years, rising 231%. This is much faster than job postings overall in the UK. The annual number of job postings has more than doubled since 2014, reflecting strong growth in demand for these roles among employers.

According to the Quanthub, which started to measure the data scientist shortage in the market from 2019, there is still a 250,000 positions shortage for data scientists alone in 2020 (DuBois, 2021). In the US, in 2020, for the second time in four years, the number of jobs posted by tech companies for analysis skills, including machine learning, data science, data engineering, and visualization—surpassed traditional skills such as engineering, customer support, marketing and PR, and administration. Of note, demand for data scientists and ML and AI specialists began surging in 2016 (Ramachandran & Watson, 2021). Similarly, the Dice Tech Jobs report released in February 2020, showed that the demand for data engineers was up 50% and demand for data scientists was up 32% in 2019, compared to the prior year. "Demand for data-oriented occupations and skillsets skyrocketed in 2019," the company stated (Techhub.dice.com, 2021). In addition, these skills were needed not only in the IT sector, but also in healthcare, ecommerce and underlying logistics, the financial sector and in cyber security industries (Olsen, 2021; Motion Recruitment, 2021).

How has the pandemic affected the market? With COVID lockdowns enforced on shops and restaurants in multiple countries, online shopping and food deliveries has increased significantly. With the speed up of the digital transformation in companies caused by the increase in remote work, analytical and technical skills have become more valuable. On the other hand, companies were forced to undergo employment cuts and limit their spending on new endeavors during the pandemic. This negatively impacted the development for data analytics, but still placed more emphasis on IT infrastructure during the early stages of this crisis.

Overall, it seems that in spite of the COVID-19 pandemic and the overall decrease in job market demand, the data science job postings did not seem to be affected tremendously. Among the large companies, the demand for data centred roles still increased, as observed by interviewquery report in 2021 (Feng, 2021). The company carried out an analysis covering over 450+ tech companies, segmenting the data science roles into eight different types. Accordingly, there was an overall slight increase in the number of job offers between 2019 and 2020 in all of the roles, but also a 15% dip in interviews for the data scientist, compensated by the increase in business analyst, data analyst, and data engineering interviews. The demand among FAANG did not suffer much from the pandemic, but increased instead, which was also confirmed by Deloitte (Ramachandran & Watson, 2021). Some other reports such as from the hiring platform indeed.hiringlab.com showcased a large decrease in the demand for technical jobs during the pandemic (Konkel, 2021), although artificial intelligence and machine learning jobs have been hit the least.

Does this mean that the hype for data science has ended? It does not seem so. The opendatascience.com study confirmed a significant dip in job postings between

March and May of 2020, but also a sharp increase in July which may mean that pandemic had short lasting effects for this area of the job market (Opendatascience.com, 2020). The hiringlab results are hard to generalize, and according to other platforms, the overall data analysis area is still on rise. In Glassdoor's 50 Best Jobs in America for 2020, data scientist remains one of the top three positions in the U.S, and the recent LinkedIn study in 2021 (Seaman, 2021) emphasized that hiring in the data science domain increased 46% in 2020, compared to 2019, while the demand for AI Specialists increased 32% in the same period. In the UK, despite the challenges faced by employers in 2020 due to the COVID-19 pandemic, 2020 was the highest year to date for the number of online job vacancies related to AI and data science, with an increase of 16% from 2019 levels (Blake, 2020). Summarizing the effects the pandemic had on the data centred roles job market and answering the first research question:

- Even as the pandemic was worsening the business conditions in March 2020, job openings for roles such as: data analyst, data engineer and data architect continued to trend high for tech majors (Feng, 2021; Ramachandran & Watson, 2021) in all of 2020.
- The overall demand for data oriented jobs did not change much, but the structure of specific positions did (Opendatascience.com, 2020; Feng, 2021), emphasizing skills in the data engineering and machine learning fields and the need for more experienced data specialists.
- Small companies may be hiring less data specialists and more of the jobs seem to have bigger entry barriers for junior developers, with more specialized requirements posted. Companies are, therefore, looking for specialists who already possess extensive technical skills.
- The increase in remote work emphasized the need for analytical skills and showcased the issues of data management in companies. This was exemplified by the increased need for data engineers and the overall demand for IT specialists that surged in 2020. For now, the demand is expected to only increase, especially in cyber security, cloud computing and IoT areas (DuBois, 2021), with less emphasis on traditional engineering skills and more on data processing and analytics (Ramachandran & Watson, 2021).
- The demand for these specialists was not uniform thorough the sectors, for example, healthcare, e-commerce, logistics and financial industry increased their analytical significantly during this period. The companies which emphasized strong IT infrastructure survived better in the pandemic, where users switched more to online channels for accessing everyday services. This means more technology-savvy companies (e.g. providing e-commerce services) that might lead the transformation in the future.

7.2. Analysis of skills for data roles—online job postings study

The skillset of the new data centred roles differs significantly from traditional analysts, but that does not mean business reporting or databases are not required. It is just that nowadays the roles are more technology focused, and require programming and statistical skills, along with knowledge on how to extract, process and predict, based on diverse data sources.

We already have specialists that make analysis and inference on various data sources—statisticians. But how does this new trend of data science translate to the skills of this profession? Citing the statistics professor, Nancy Reid, in an article from 2018: "Data science may be to some point a different way of looking at data than statistics. Where statistics is cautious and deduces on data with underlying assumptions on the processes generating it, data science aims at classifying on varying and uncertain data formats, while largely emphasizing the speed and applicability in different domains. It is blend of statistical modeling and inference, data management, computing at scale, optimization, communication and visualization" (Reid, 2018).

Knowledge of statistics is important for the new data centred roles in enterprises, but so are visualization and prediction skills, including knowledge of machine learning methods and other algorithms originating from computer science. As emphasized by the paradigm of big data: the size of the data for analysis (volume) meets the requirement for real-time or short batch processing (velocity) of not only numerical, but also image and text data (variety) that may be burdened with high errors (veracity). In this environment, as stated by Davenport's article in 2012, when describing the role of a data scientist: "The traditional backgrounds of people you saw 10 to 15 years ago just don't cut it these days" (Davenport & Patil, 2012). In line with that, the focus of the data scientist it to produce a business actionable prototype using a programming language, with strong data management and analytical skills, supported by a strong foundation in math, statistics, probability and computer science (Cao, 2017).

To offer innovative services and derive insight directly from the raw sources available for the company, these experts require a variety of technical skills. The question is, which of them are absolutely necessary as of today, when the role is more mature than it was in 2012. Also, is the required skillset diverse or highly interchangeable between different roles? To answer this, the authors of this chapter have conducted an analysis of job offers in late 2020, that spanned over the whole previous year and aim to compare them with the results of similar studies.

The main goal of the study was to extract the skills that differentiate data scientist and similar roles from programmers and developers. As our dataset we have extracted over 72 thousand job offers from one of the biggest online job advertisement services in Poland. We analysed only the technical offers[1], meaning the offer was included only if it classified the job offering as an IT job. The collected offers spanned from January to December 2019 and their geographical distribution, where the expected focus on large towns is clearly visible, is shown in Figure 7.1.

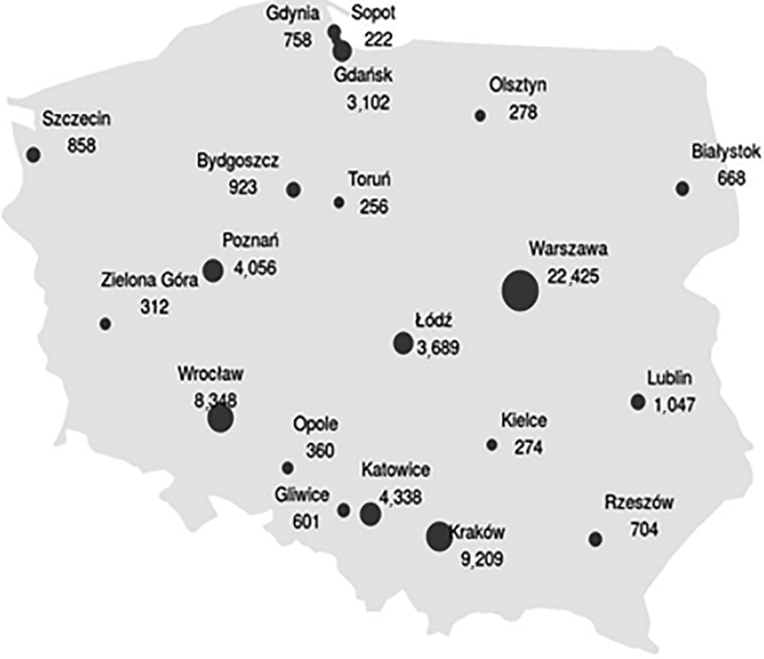


Figure 7.1. Distribution of all IT offers (duplicates removed) by city
Source: Own elaboration from authors' dataset.

We have cleaned the data to include only the offers written in Polish and English (we have used Python's "langdetect" library). The structure of extracted languages after the filtering procedure is shown in Figure 7.2. We have also excluded offers that were placed multiple times. All of the cleaning resulted in a preprocessed sample of 42,885 job offers.

Due to the size of the dataset, automatic keyword extraction and stemming and other similar natural language processing (NLP) techniques were applied as a part of our approach to the problem. The ”soft skills” extraction for now was an issue, because of the complexity of Polish language, a lot of mixed-language words and potential translation errors when comparing Polish and English offers.

[1] From "IT—administration" and "IT—Software Development" categories, the only two directly connected with IT jobs.

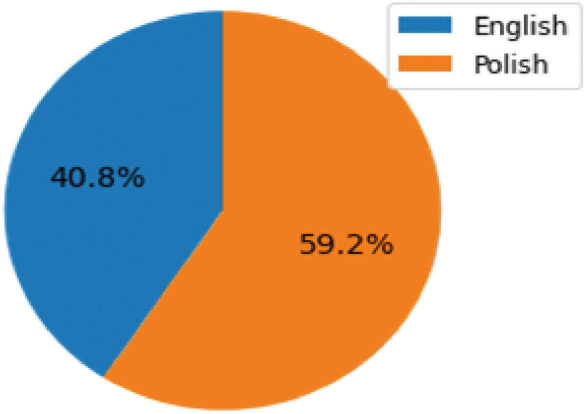


Figure 7.2. Distribution of languages in the prepared study
Source: Own elaboration from authors' dataset.

The technical skills, on the other hand, were mostly identified by a technology name: Python, SQL, Spark, Tensorflow etc., which are usually nouns, hence our approach for extracting technical skills was applicable for the offers in both languages. Due to that, and to compare the results with similar studies focused on the skillset identification, we decided to focus on technical skills. As an extension of our approach, however, extracting soft skills could be a valuable area of further work. An interesting approach for the English language has been explained in the work from 2017 (Papoutsoglou, Mittas, & Angelis, 2017). To extract the skills required by the data centred roles, the NLTK[2] and spaCy[3] Python libraries were used. Thus, the following approach was proposed:

- Part of Speech tagging was applied to extracting NNP (ang. propel nouns) as the skill names. Similarly, extracting ORG (original names) parts of speech by NLTK was tested, but it extracted significantly less names that were usable.
- Due to the fact that not all of the technical skills are single nouns, bi-gram detection was performed on the extracted nouns.
- The nouns and bi-grams representing technical skills were then compared with the Wikidata ontological database categories[4], which resulted in 24,580

[2] https://www.nltk.org/

[3] https://spacy.io/

[4] That included: programming language (Q9143), python library (Q29642950), free software (Q341), object-based language (Q899523), functional programming language (Q3839507), scripting language (Q187432), multi-paradigm programming language (Q12772052), imperative programming language (Q21562092), interpreted language (Q1993334), highlevel programming language

different technologies. **This is the main extension of this work over SOTA as it significantly increases the possibility of matching the extracted skillset with its natural language descriptions without manual labelling work.**

- Some of the ambiguous phrases such as “nice” or “possess”, which are rarely used terms in technologies, were parts of the “nice to have” and similar phrases. After careful examination, skills not matching the context were manually removed.

After the analysis of extracted skills, considering top 20 skills for all of the IT related jobs our results showcased, that:

- the most popular skill was SQL, connected with relational databases querying, and was included in 28% of all offers,
- Linux (15%) and Windows (13,5%) were commonly mentioned in posted offers,
- the most popular programming languages demanded besides SQL were: JavaScript (14%), Java (13,5%), HTML (11%), Python (9%) and C# (7,7%),
- knowledge of git versioning system was mentioned in 12% of all offers,
- analytical skills like “office” were present in 11% of the offers, and beyond this, “excel” was mentioned explicitly in 7% of the offers,
- Scrum (9%), Agile (11%) and Jira (9%) knowledge, corresponding to project management skills was also present in the job applications, and were used to identify the portion of jobs aligned with project manager roles among all of the IT offers,

(Q211496), statistical package (Q13199995), JVM language (Q56062429), procedural programming language (Q28922885), structured programming language (Q28920117), computing platform (Q241317), Web API (Q20202982), markup language (Q37045), academic discipline (Q11862829), numerical software (Q74086777), mathematical software (Q1639024), software library (Q188860), software framework (Q271680), computer science (Q21198), NoSQL database management system (Q82231), database management system (Q176165), document-oriented database (Q1235236), relational database management system (Q3932296), proprietary software (Q218616), open-source software (Q1130645), message-oriented middleware (Q1092177), programming paradigm (Q188267), artificial intelligence (Q11660), service oriented architecture (Q220644), communications protocol (Q132364), computer network protocol (Q15836568), continuous integration software (Q16947796), free and open-source software (Q506883), virtualization engine (Q7935198), web framework (Q1330336), virtual hosting (Q588365), agile software development (Q30232), computer science term (Q66747126), operating system (Q9135), software development methodology (Q1378470), protocol suite (Q67080166), Internet Standard (Q290378), distributed data store (Q339678), collaborative software (Q474157), application framework (Q756637), event-driven programming language (Q28920813), platform as a service (Q1153767), platform as a service (Q1153767), computer data processing (Q6661985), software design pattern (Q181156), architectural pattern (Q635346), search engine (Q19541), web server (Q11288), operating system shell (Q18109), certificating services provider (Q13460321), software (Q7397), modelling language (Q1941921), query language (Q845739) and generic top-level domain (Q29469).

- the most common coexisting pairs of skills were also analysed, but, mostly, the versioning systems and SQL were mentioned, along with their corresponding programming language. The only exceptions to this were the (Windows, Linux) pairing that was in 5,95% of all the offers and the (Agile, Scrum) pair. The meaning of this is twofold:
  - firstly, among the IT offers, only the knowledge of SQL and databases and the GIT versioning system can be considered basic and necessary knowledge,
  - there are no combinations of technical skills required that are strongly tied together.

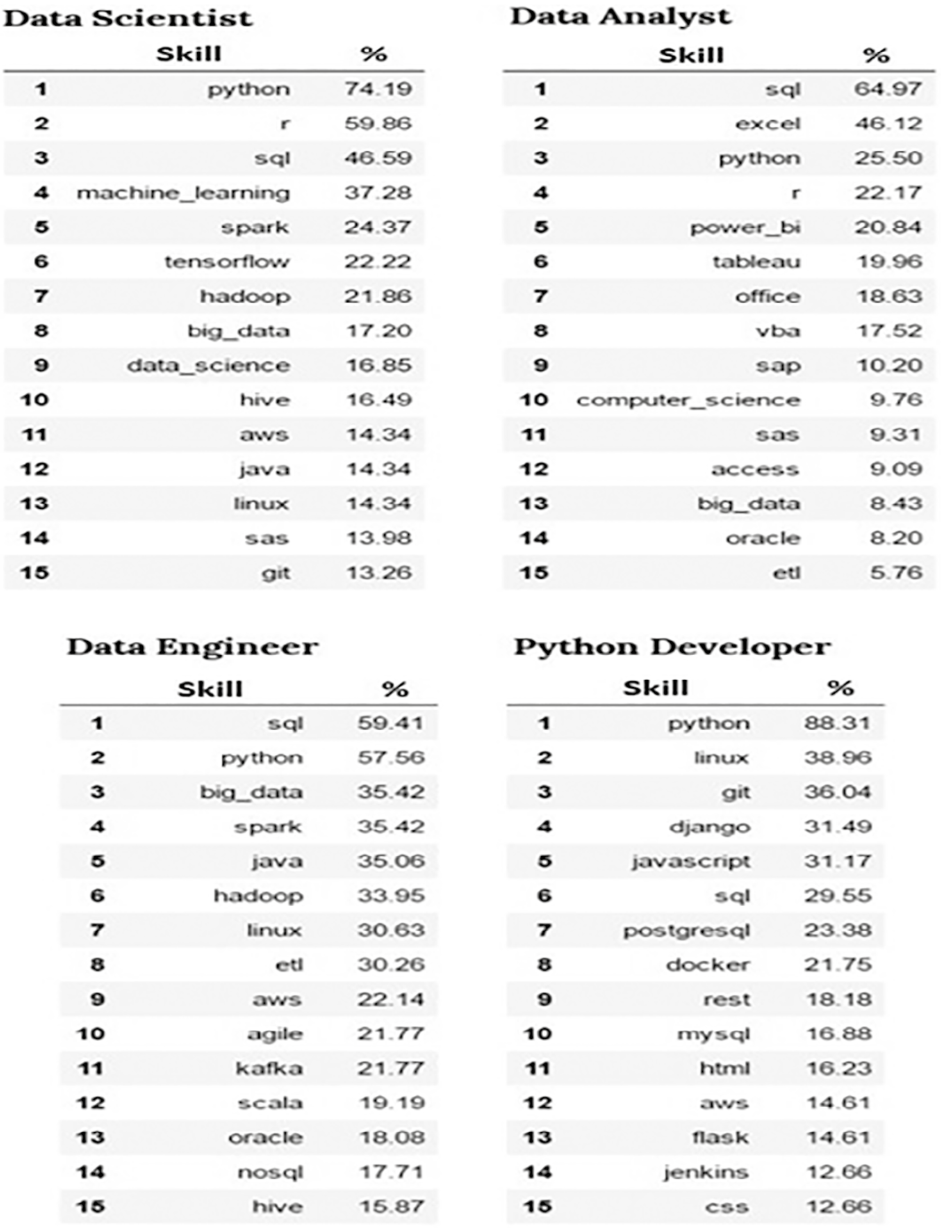
This means that there is a wide array of independent skills necessary for the data analysis jobs.

7.3. Findings—data centred roles

Further on, only offers that contained the words "data" were included, creating a subsample of all IT offers. Additionally, for comparison with the standard programmer profile, developer job offers that contained the word "Python" were included and henceforward we will refer to them as the "Python developer" role. For each of these roles, the job offers were extracted based on the inclusion of the keyword name in the "data" subsample of offers. To characterize these roles in companies, an analysis of extracted skills was performed. The top skills for each of the roles are showcased in Figure 7.3.

To measure the stability of the job's profile, a classification experiment was performed. Its goal was to showcase how stable the skill profile is, based on a sample of skills extracted for a specific role. Relying on that, TFIDF vectorizer[5] was applied on the keywords extracted. After a few experiments, top 500 words were used to represent a single job offer in the data centred area. Further on, the SVM linear classifier with $C = 1$ parameter was used to train a machine learning model that can classify the job into: data scientist, data analyst, data engineer and Python developer categories. This was just a short experiment to showcase the potential accuracy and stability of data centred profiles—the low accuracy of the classifier would mean that potentially the identified positions do not differ much in terms of required skills and there are no distinctive technologies connected with a specific occupation. Fortunately, the classifier resulted in over 87,5% average accuracy and the results on a sample of job offers can be seen in Figure 7.4. Accordingly, there

[5] Term frequency—inverse document frequency statistic implemented in the scikit-learn Python library was used.



Data Scientist

	Skill	%
1	python	74.19
2	r	59.86
3	sql	46.59
4	machine_learning	37.28
5	spark	24.37
6	tensorflow	22.22
7	hadoop	21.86
8	big_data	17.20
9	data_science	16.85
10	hive	16.49
11	aws	14.34
12	java	14.34
13	linux	14.34
14	sas	13.98
15	git	13.26

Data Analyst

	Skill	%
1	sql	64.97
2	excel	46.12
3	python	25.50
4	r	22.17
5	power_bi	20.84
6	tableau	19.96
7	office	18.63
8	vba	17.52
9	sap	10.20
10	computer_science	9.76
11	sas	9.31
12	access	9.09
13	big_data	8.43
14	oracle	8.20
15	etl	5.76

Data Engineer

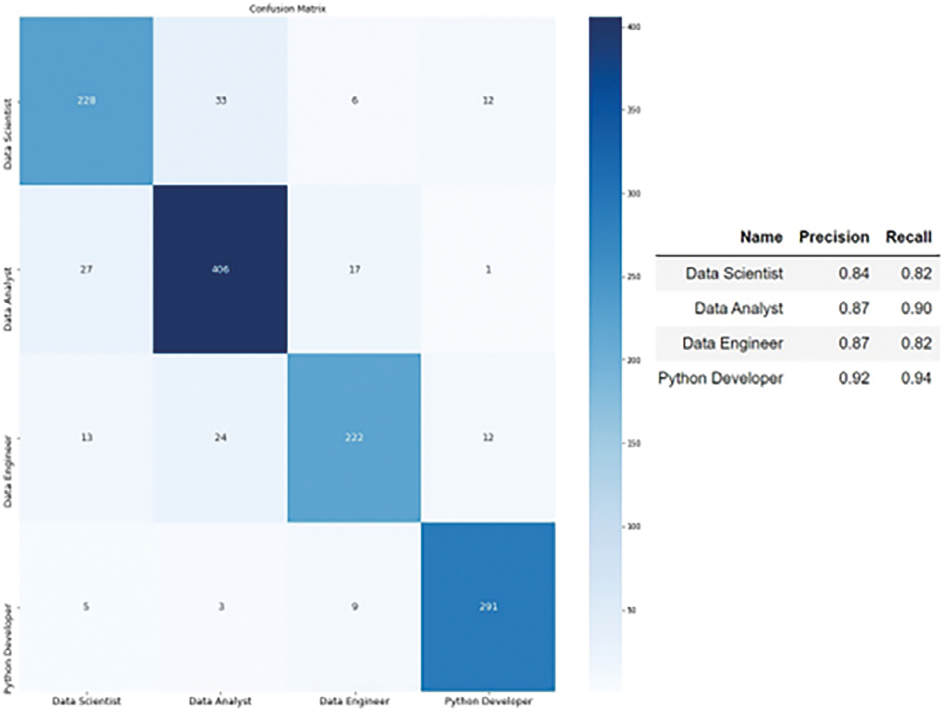
	Skill	%
1	sql	59.41
2	python	57.56
3	big_data	35.42
4	spark	35.42
5	java	35.06
6	hadoop	33.95
7	linux	30.63
8	etl	30.26
9	aws	22.14
10	agile	21.77
11	kafka	21.77
12	scala	19.19
13	oracle	18.08
14	nosql	17.71
15	hive	15.87

Python Developer

	Skill	%
1	python	88.31
2	linux	38.96
3	git	36.04
4	django	31.49
5	javascript	31.17
6	sql	29.55
7	postgresql	23.38
8	docker	21.75
9	rest	18.18
10	mysql	16.88
11	html	16.23
12	aws	14.61
13	flask	14.61
14	jenkins	12.66
15	css	12.66

Figure 7.3. Top skills extracted for data analyst, scientist, engineer and Python developer/programmer job postings

Source: Own elaboration from authors' dataset.



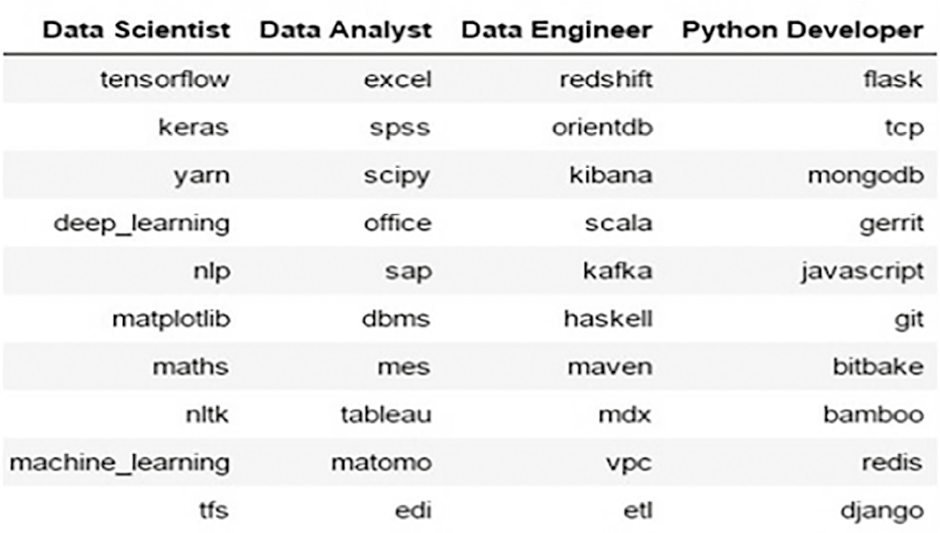
Name	Precision	Recall
Data Scientist	0.84	0.82
Data Analyst	0.87	0.90
Data Engineer	0.87	0.82
Python Developer	0.92	0.94

Figure 7.4. Classification experiment for job offers in data scientist, analyst, engineer and Python developer roles

Source: Own elaboration from authors' dataset.

were many more data analyst jobs than any other, and, actually, "data scientist" was the hardest to predict—which could mean that it required the widest variety of skills. Due to the use of a ML classifier, we could extract feature importance in terms of the unique skills that characterized each of the occupations. The resulting top 10 skills that are specific for each of the roles are presented in Figure 7.5.

As a summary of the results of both of these approaches (exemplified in Figures 7.3 and 7.5), it is easy to see that data scientist profiles include deep learning and machine learning knowledge. This state of reality partially implies that there is a strong connection of this occupation with AI and prediction. Secondly, data analysts were mostly connected with traditional analytics and office skills, but also required knowledge of python libraries and the statistics software used for data visualization, as well as other commercial analytical software. Data engineers, on the other hand, were mostly programming- and software-focused, and had the need for the knowledge of Scala, Haskell and NoSQL and big data databases, along with stream processing libraries. Python developers, in contrast, were more focused on



Data Scientist	Data Analyst	Data Engineer	Python Developer
tensorflow	excel	redshift	flask
keras	spss	orientdb	tcp
yarn	scipy	kibana	mongodb
deep_learning	office	scala	gerrit
nlp	sap	kafka	javascript
matplotlib	dbms	haskell	git
maths	mes	maven	bitbake
nltk	tableau	mdx	bamboo
machine_learning	matomo	vpc	redis
tfs	edi	etl	django

Figure 7.5. Top 10 unique skills specific for given data centred role based on feature importance of the classifier
Source: Own elaboration from authors' dataset.

providing web services with Flask and Django, along with front end and databases experience. Their skillset included a large quantity of software version management tools. Overall, the results are very interesting, but a confrontation of the results with other studies is required to build stable profiles for these roles. Summarizing, while some skills were prevalent among multiple data processing roles (SQL, and, of course, Python and R to some extent), the **analysed job profiles differ significantly** both in terms of their top skills, but also in terms of their distribution among offers, which answers both the **second** and **the third research questions**.

Data centered roles characteristics: an interesting outcome from the previous study is that some key skills for each data centred roles were identified and they do differ between the roles. This may confirm that the proposed division of the roles presented was justified. What remains to be answered, is to what extent the results correspond to similar studies and if there any dynamics which can be observed in these occupations that might have been missed. The approach proposed was unique in terms of an ontology-based approach to skill detection, which allowed a much broader keyword search for identifying the technologies used. On the other hand, the key skills identified should remain the same even in simple keyword-based approaches. Hence, to confirm the general nature of the results, before describing the role skillsets in detail, we will need to look at the results in a broader context of other analysis performed in the literature.

7.4. Comparison of results with other studies

According to the job interviews analysed by dataquery (Feng, 2021), data scientists required strong machine learning (and by the proxy of area, also AI) skills, with strong algorithmic knowledge, and ability to solve case studies, supported by statistics/AB testing and Python/SQL skills. This was also confirmed by our findings.

Similarly, the recent study by Jack Chih-Hsu Lin (Chih-Hsu Lin, 2020) on the dataset of 5,5 thousand data scientist jobs from 2020 in the US, identified similar key skills for all of the roles (as can be seen in Figure 7.6). He also pointed out that Python was more important for data scientists, while Excel and Tableau had pre-eminence for the Analysts. Spark, however, was more popular among the requirements for data engineers. What is interesting, is that some of the companies (hard to measure exactly, but about 5–10%) did seem to include skills clearly connected with other occupations—like the requirements for Tableau for data engineers or C/C++ knowledge for Analysts. This points to the fact that companies are sometimes looking for people who can actually fill multiple roles in one job offer, which may also influence the talent shortage on the market.

The last study, published in March 2021 (Shin, 2021) by Terence Shin, analysed over 15 thousand job postings from Indeed, Monster and SimplyHired platforms. Unfortunately, the approach was only limited to about 45 skills that were chosen

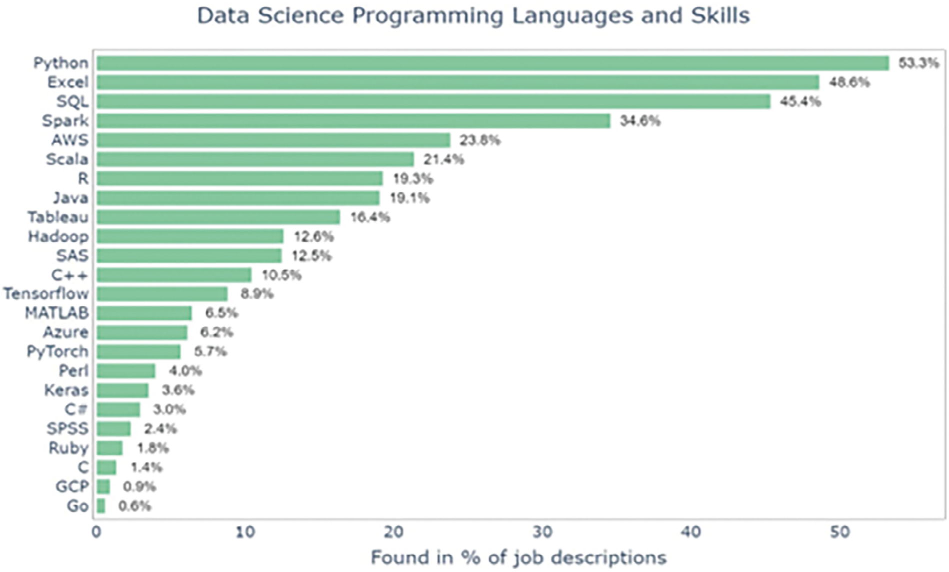


Figure 7.6. Percentage of 5,500 US job postings in 2020
Source: (Shin, 2021).

by the author. The results however, seem to be interesting due to the fact that they confirm the stable nature of the popularity of the most often mentioned technologies. The popularity of Python, SQL, R and other mentioned technologies seem to be in line with the results of our and Jack Chih-Hsu Lin's studies. Unfortunately, we are unable to confirm most of the data analyst skills, as Excel and other commercial technologies (with the exclusion of Tableau) were not included in the keywords for his experiment.

The role of data analyst has been recently mentioned in one of the largest IT studies in Poland, covering the Polish IT market in 2021 and carried out by bulldogjob.pl. It was classified as responsible for 6% of the overall demand for IT jobs and was divided into: Business Analyst, data analyst and System Analyst (bulldogjob.pl, 2021). The role was mostly connected with modelling IT software and its requirements, but the data centred role was mostly tied to business reporting, with Excel, SQL, Oracle and Python as main skills for the job. This also confirms the findings of our skills study that emphasize this job's role at providing business applicable analytical reports and its close ties to other more business oriented roles in the company—overlapping with more software and requirements analysis jobs.

As for the recent dynamics, based on the dataquery (Feng, 2021) research, data engineering specific interviews increased by 40% in 2020. The second fastest position growth within data science roles went to business and data analysts—which increased by 20%. In the data scientist role, a large emphasis on algorithms, machine learning and statistics was observed. The company profiles, however, differ between even the largest companies. For example, Amazon places a higher emphasis on machine learning and programming tasks, while Google mostly cares about statistical analysis. Moreover, a more broad classification of job offers was proposed by the authors of the research, showcasing that the demand for data scientists, data analysts, business analysts and data engineers make up about 90% of the offers for the data centred roles.[6] The overall distribution and changes in the popularity of the analysed roles can be seen in Figure 7.7. In terms of **the skills that increased in popularity in 2021**, based on the Terence Shin's (Shin, 2021) study: cloud computing technologies were more popular, especially AWS, along with deep learning libraries. What is interesting, is that the requirements for the most popular skills like SQL, Python and machine learning (scikit-learn) increased even more during this period. This may be due to the standardization of skillsets for those positions over the years, or just the fact that companies are more aware of the most popular technologies due to existing infrastructure. The most trending skills in 2021 can be seen in the bottom part of Figure 7.8.

[6] The inclusion of business analyst was an interesting choice, partially confirmed by our findings on the variety of skills for analytical roles and possible misunderstandings between data and business analysts responsibilities.

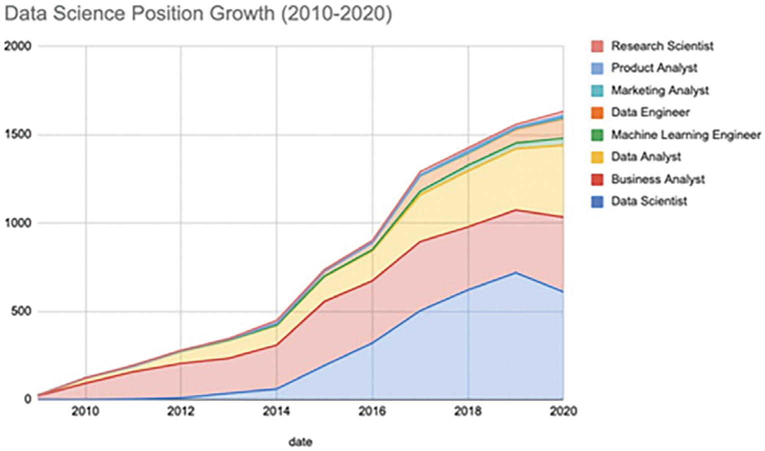


Figure 7.7. Growth of more detailed data centred roles

Source: (Feng, 2021).

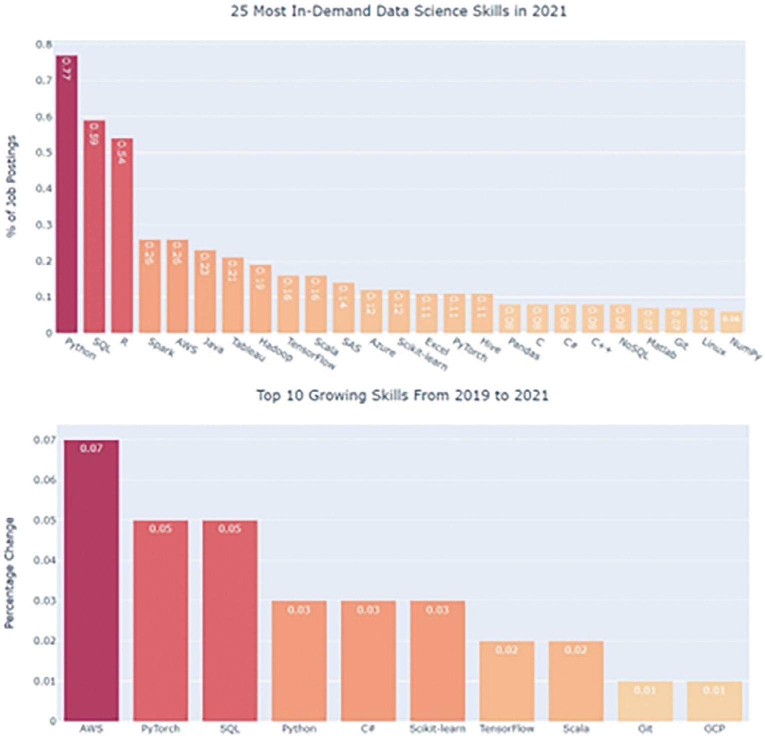


Figure 7.8. Percentage of 15,000 job postings from Indeed, Monster and SimplyHired that contained chosen skills (top) and trending skills between 2019 and 2021 (bottom)

Source: (Shin, 2021).

7.5. Summary of profiles for data centred roles

Relying on all of the previously mentioned analytical studies, we can try to summarize the responsibilities and generalize the skillsets of the data centred roles. Although multiple technologies were found, they often correspond to similar tasks and general responsibilities which these jobs are expected to perform in enterprises. Summarizing these descriptions:

- Data analysts are the most connected with traditional analysts roles (business and system analysts), and they are mostly tasked with explaining current issues and presenting them to business representatives by the use of company's data. The quanthub interpretation of their responsibilities (duBois, 2020) seems to be fitting, as they provide processed information that enables companies to make business decisions primarily working on structured data from a single source and are strongly tied to a company division. Their professional needs emphasize statistics, communications and business oriented skills to not only perform historic analysis on organized data, but also to enable data-driven decision-making on a daily basis. These positions exhibit high demand for knowledge of SQL and commercial analytical software such as Excel, Power Bi, Tableau, VBA, SAP, SAS, SPSS, Access and Oracle. They use the infrastructure, software and databases that are already in the company to provide more insight.
- Data scientists, in fulfilling their most popular role, mostly emphasize machine learning and AI skills (machine learning, Spark, Tensorflow, data science), along with strong programming (Python, R) and data processing (SQL, Spark, Hadoop, big data, Hive, AWS, SAS) knowledge. The combination of broad keywords such as machine learning, data science and big data may however, emphasize that it is either hard to specify technologies or that employees assume data scientists will choose the right tools that are good for the job. The only libraries mentioned in the top skills are tied with data processing large quantities of varying data, which showcases that data scientists should be able to perform machine learning not only on megabytes of queried data, but also thousands of gigabytes or several petabytes of varying data stores. Where data analysts provide the foundation for decision-making, data scientists should be able to extract, process and infer on the data sources, while providing predictions and self-sustaining data products for mostly internal purposes. They seem to be expected to work throughout the entire process of developing a business prototype, and be able to resolve outstanding issues, which explains the higher variability of skills required for the role and the focus on prediction methods.
- Data engineer skills are much more focused on the deployment of the analytical software prototype in an effective way and providing data for the

analysts. The technologies required describe a skillset similar to traditional database and ETL specialists (SQL, ETL, Oracle, Java[7]), but highly extended with big data technologies for data processing and storage (big data, Spark, Hadoop, NoSQL, Hive, Kafka) and cloud computing knowledge (AWS, GCP). They lean more toward service management and providing the most effective ways of data processing company data by the use of dedicated programming languages (Java, Scala, Python). They are more in line with DevOps than machine learning or git skills as they can also work in deploying the data scientists' results.

- Python developers, effectively utilizing the same technologies as previously mentioned positions, are mostly characterized by very strong programming (Python) and SQL skills, along with the ability to manage self-sustained software (Linux, git, Docker, MySQL, Jenkins), supported by front-end frameworks (Javascript, CSS, HTML) and a database backend to create a full-stack web service (REST, Flask, Django). It can be clearly seen that the skillsets exemplified closely follow the DevOps approach for software creation, emphasizing continuous deployment, testing and release management roles with a skillset that is effectively a full-stack developer knowledge.

Overall, the analysis allowed us to generalize the responsibilities of analysed roles and compare them with developers, who use the same programming language as most of the data specialists. The number of data analysts is sure to increase and the basic "data literacy" skills connected with processing and visualization will continue to surge as companies will switch from excel to more analytical software or even data analysis libraries in languages like Python and R. The data scientist and engineer seem to be far more specialized roles. The first appears to have a very wide variety of skills, but significantly emphasized machine learning techniques (with deep learning and text mining knowledge), along with traditional data processing knowledge, supported by business and analytical skills in which the other two positions specialize in. This, in turn, means that some of data scientists may be responsible for analytical or data engineering tasks or are at least required to have basic knowledge in those areas. Data engineers, on the other hand, have a more focused skillset, emphasizing cloud and big data computing, along with technologies and libraries that allow fast and efficient processing of varying data sources in line with the big data paradigm. Their role is to bring an ETL and SQL company to the big data era and enable the processing of error-free, fast and high volume data sources that are built into various technologies.

[7] Exemplified by availability of ETL processing, JDBC drivers and ORM libraries.

7.6. Data science—supply (teaching)

An additional topic that was worth noting is the supply part of the market, exemplified mostly by the Universities' "data science" and analytical or computer science programs. As a lot of the skills described previously requires extensive mathematical, statistical and computer science knowledge, so it is reasonable to assume Universities are the first step in addressing the required skills shortage and the over-increasing gap of experts in this field. As an additional argument, most of the market studies showcased that over 90% of the currently employed data scientists and engineers have a master's degree, which only solidifies the previous assumption.

The goal of this analysis is not to showcase what technologies to teach, as it would break the general goal of a University degree. However, it should be an assumption that when a data science or data analyst program is offered by a University, it should teach at least general skills connected with the roles described earlier. As such, for example, there is no difference as to which library will be used, as long as data scientists will learn machine learning, deep learning or natural Language processing skills and possibly will be able to process large and varied datasets in line with the big data paradigm in any of the technologies that allow it. These are valid academic subjects often originating from computer science and are the core skillset of the data centred roles in the market—as confirmed by our study.

To confront the skills with the academic curricula, a preliminary study of subjects taught in the main path of English and Polish master's degree programs were examined. As a model example, two programs from UC Berkley (exemplifying more computer science skills) and University College of London (more in line with economic studies) were used. These two were then compared with a few selected Polish master's programs which seemed to be connected with the area. Excel analytical skills were not included, as they are assumed to be basic skills learned during bachelor's studies or high school. In most of the programs, statistics and data visualization classes were taught, while some of them focused on data storytelling—which points out that most of the alumni would have no problems in these tasks. Overall, from the characteristics which were summarized in Table 7.1, characterizing the programs more in depth[8]:

- UC Berkeley (University of California, Berkeley, 2021) was chosen as the main example, with very clear curriculums exemplifying technologies used, and a high focus on cloud development. No classes seemed to be directly

[8] The analysis of the study programs was carried out in September 2021, and all changes in favor of updating the syllabuses to meet the described skill requirements from this date should be interpreted as a high level of awareness of said programs and treated with the utmost respect to their program committees.

Table 7.1. Short description of skills covered by chosen Master’s programs connected with the data centred roles

Name of studies	General skills	Programmer			Analyst			Data scientist					Data engineer				
	Project Management	Git	Web services rest	Programming	Data Visualization	Databases data wareh using	Statistics and statistical software	Big Data (general)	Machine Learning	Deep Learning	NLP	Big Data processing (Spark)	NoSQL and Big Databases Hadoop	Linux, Dev Ops	Cloud computing	Stream Processing	Java/ Scala
Master of Information and Data Science (UC Berkley)	X	X		X	X	X	X	X	X	X	X	X	X	X	X	partially	
Data Science (UCL)			X	X	X	X	X	X	X	elective	elective	X	X		X	X	
Data Science and Business Analytycs (UW)		X	X	X	X	X	X	X	X	X	S	X	X	X	X		
Data Science (Uwr)	X			X			X	X	X	elective	elective	elective*	elective*		elective*	elective*	elective*
Data Science (PW)	X			X	X	X	X	X	X	X	S	X		X	X	X	partially
Informatics and Economics (UEP): ISfBA track		X	elective	X	X	X	X	X	X	elective	S	X		X	elective	X	
Informatics and Economics (SGGW): “Big Data”	X		X	X		X	X	X	X	X		?					
DataAnalysis-BigData (SGH)				X	X	X	X	X	X	elective		X	X		X	X	

Source: Own elaboration.

tied to traditional REST web services, but the use of various API's were mentioned. Additionally, data visualization classes included not only Tableau, but also Javascript visualization libraries. One of the final subjects was a data science problem-solving project using real-world data and covering the issues of collaboration and project management skills.

- The data science program for UCL is more focused towards financial risk, stochastic methods and quantitative modelling (University College London, 2021). It, however, contains most of the key data analyst and data scientist skills and there are no general subject like management, economic theory or similar. All of the studies that correspond to the domain area of finance and economics are closely tied to analytics and the descriptions are phrased in a way that describes the usefulness of these subjects to analysts. UCL also has a more computer science-oriented program called "Data Science and Machine Learning".
- Comparatively in Poland, there seem to be a lot of programs tied to the economics departments (UW, UEP, SGH) compared to that of computer science or mathematics. The overall interest in this area for economics is the natural evolution of economic computer science courses (Korczak, Abramowicz, Gołuchowski, Kobyliński, & Owoc, 2014) that have a strong tradition of being connected with business information systems and data analytics. The changes are mostly due to the shift in the paradigm of data analysis services: dig data replaced data warehousing, cloud replaced web services and Tableau, along with visualization tools, took over ERP systems for enterprise analytics and business intelligence).
- Overall, UW (University of Warsaw, 2021a) seems to propose a very intriguing program fitting almost all of the criteria, with a very interesting course called "Reproducible Research", which covers a lot of the issues required by the business requirements [28], including: linux, git and cloud computing.
- The programs of UWr (University of Wrocław, 2021), SGGW (Warsaw University of Life Sciences, 2021a, 2021b) and PW (Warsaw University of Technology, 2021a) have the issue that most of the key skills concerning big data technologies, data processing etc. are stacked into very small, often elective courses or not offered at all. For example, in UWr, a lot of the skills (SQL, Scala, Hadoop, Spark, Cloud Computing, Stream processing) are taught in one course that is limited to 15 participants called "Tools and Methods in Big Data Processing". This includes all of the elective courses with an asterix (*) in the table, which the University can have graduates with no knowledge over the main skill areas of big data processing and software development. In case of PW, the same role (but fortunately in less areas) is carried out by the "Big Data Analytics" course (Warsaw University of Technology, 2021b).

- Other economic programs like SGH (SGH Warsaw School of Economics, 2021a, 2021b), UEP (Poznań University of Economics and Business, 2021) have a significant number of data centred skills taught (but slightly less than UW). It is also visible they have much more very general economics subjects than any of the other master programs. UEP did not label their studies as data science, so this is more understandable, however, overall, the subjects taught position it very close to a full data science program, as compared with other studies. The same issue of emphasis on general subjects applies for UWr and mathematics. It can be clearly seen that most of the subjects are very similar to the ones offered by other master's degree programs by the same faculty.
- The non-data analysis[9] skills take as low as 10% (UW), about 30% (SGGW, UW, UEP) to 50% (UWr) of the program, although in this case, the strong mathematical background may have been intended. The most focused program is "Data Science Masters" offered by PW. This has only two theoretical electronic courses: "Electronic Principles" and "Data Transmission", despite it being tied to the Information Science and Mathematics faculty, it does not offer that many theoretical subjects out of the scope for data centred roles, but it is still behind the Berkeley program in that regard.

Overall, in our analysis we did not include some studies offered, for example, by UAM (Adam Mickiewicz University, 2021) or PJATK (Polish-Japanese Academy of Information Technology, 2021), partially due to the fact, that to the best of authors' knowledge, there were no widely available descriptions of the syllabus content and so the list may be extended in future. The results are based only on a preliminary study of information found online, but some may argue that this might be the same source of information students will seek. Non-master postgraduate studies were also excluded from our study.

Interestingly, a lot of Polish programs analysed include a large number of general subjects that differ from University to University. Where PW includes, for example, "Principles of Electronics", clearly an aspect of computer engineering, UW, SGH and UEP propose courses in mathematical economics, marketing or finance (actuarial methods). UWr, on the other hand, offers a lot of courses focusing on the mathematical principles behind data science. The problem is not with these subjects, but that, sometimes, the hours they are assigned are disproportional to the number of courses exemplifying the key data science skills identified by us in the study. For example, the UWr program educates mathematicians and provides them with some analytical and data science skills. This showcases that Polish curriculums are significantly more tied to main scientific disciplines and less focused

[9] Meaning not teaching any of the skills identified in the analysis or general skills tied to responsibilities like data processing, visualization, data management etc.

on teaching classes aimed at developing analytical and data processing skills than are their top tier foreign counterparts.

While nearly all of the key skills were included in the study programs, sometimes a multitude of important skills are taught in 30 h of a single course. This creates an interesting conundrum, where more emphasis and time is spent on teaching the main discipline theoretical subjects than on data science in a data science master's course. This might happen partially due to adding data science (which is a very promising and popular area of education) to current programs and curriculums that are tied to the existing faculties that want to integrate subjects taught for multiple programs. This emphasis, however, is partially based on the requirement for Polish study programs to strictly follow the guidelines of assigning ECTS to underlying scientific fields of science (Economics, Computer Science, Computer Engineering etc.). Compared to the Polish programs, US and UK top data science programs are significantly more focused on analytics, and exclude many general subjects. For example, Berkeley's data science program offers more hours of training in key analytical skills and the same can be said for other programs from Carnegie Mellon University, Harvard or Colombia.

Overall, however, despite the fact that multiple the important issues in Polish curriculums are covered for some programs in only one subject, the coverage of the skills by the curricula seems to be good. However, the small number of subjects covering Hadoop and Spark, paired with the significant underrepresentation of Git, DevOps and Java or Scala skills seems to be an issue in teaching data engineering skills. Still, most of the programs teach both Python and R, with the use of R mostly for teaching statistical inference, regression and time series and panel data analysis—which is in line with global trends. The big data databases (Hadoop) and streaming environments seem to be lacking in some of the economics programs, and are sometimes overshadowed by SAS software, which is a significant over representation of commercial big data analysis software for these academic programs.

7.7. Recommendations for data science programs

As for the general remarks to the changes in data science programs compared to the other areas, based on the recommendations by (Hicks & Irizarry, 2018), the teaching needs to consider more training in computing (data management and processing), as well as in connecting—connecting the subject matter question with the appropriate dataset and analysis tools—which is not currently prioritized in statistics curricula (in terms of researching the dataset, data transformations, visualizations, cleaning and processing). In addition, emphasis should be placed on creating: in terms of searching for answers in the data, processing it by known methods and evaluating the results.

The courses should also emphasize real world problems, more projects and less memorization (which seems rarely useful in terms of the ever changing nature of the technologies and syntax used). Since, data science is more problem-solving oriented and less descriptory, this means that teaching data science emphasizes connection to solving business issues by employing the scientific method—and this emphasis should be visible in the nature of the studying process. It could include: more classes held by business professionals (Kross & Guo, 2019), or more group projects and data experiments. Overall, most of the Polish data science programs suffer from the very large number of general subjects, which significantly limits the key skills in big data, cloud computing and deep learning that they should provide. Summarizing the outcomes, a good data centred program should:

- In terms of bachelor's degrees, focus on teaching the basics of data analyst skills, which can include management of IT systems widely used in business, such as SAP, SAS, Office etc. Training in these should be expanded with developing skills in using PowerBI, Tableau or Web Analytics tools, as the work for IT professionals is more focused on analysis than on manual programming. This is supported by the requirements for skilled business professionals, in line with the findings of (Waller & Fawcett, 2013).For training in data analysis and other analytical roles, the inclusion of these skills, along with having strong statistical and mathematical backgrounds is crucial.
- The importance of teaching programming languages has not changed. Python has proven to be effective for teaching algorithms and data structures. It is not suited for studies which require low level embedded or network programming, such as computer engineering, automatics and robotics etc., but knowledge of it is very good for data centred roles, as Java and Scala programming are encouraged to be included or remain in the main-stream programs.
- Statistics, econometrics and similar courses should be accompanied by machine learning. What is more, the contents of the subjects should follow a coordinated approach for teaching the students how to solve different business problems while relying on a given source of data. Students should also learn multiple methods for analysing, classifying and predicting based on mining different types of data. Furthermore, training in NLP and graph analysis should be mandatory and non-elective to provide alumni with expert knowledge on the subject of data analysis.
- There should be significantly more focus on solving data-oriented problems and building services, and, later on, in using cloud computing or building API's, concept supported by other studies such as (Donoghue, Voytek, & Ellis, 2021). Theory should be accompanied by less textbook examples and more open datasets in which the limits of some methods should be learned. As noted by previous researchers, students must handle real, demanding data

to be prepared to be able to clarify situation wherein some assumptions of the statistical tests are not met.

- Computer networks, operating systems, databases and similar “hard” IT subjects should include how to prepare students in choosing the right IT architecture for the solution, and less on theoretical principles that are useful for engineers.
- As an addition, students should learn the development process, both from an analyst perspective (by being able to discern UML diagrams) and that of project manager (by understanding requirements analysis, task management in tools like JIRA and differences between approaches of Waterfall and Agile (including Scrum)—this is emphasized by the requirements of Agile knowledge and by the overall nature of the current IT management paradigm.
- The program should be highly focused on teaching all of the key skills, with a reasonable number of hours and ECTS assigned to develop key knowledge and data management expertise.

Conclusions

This article solidified the demand expectations for the role of data experts, identifying multiple diverse roles tied to this area (data analysts, data scientists and data engineers), along with a multitude of professions in great demand (research scientists and machine learning engineers). The need for those roles did not seem to be affected much by the pandemic: the slowing of the digital transformation of reluctant companies is countered by increases in demand for those experts among the leaders of the IT transformation. Automatic analysis of key skills relying on semantic databases and NLP knowledge, allowed building clear skill profiles for the major roles in this environment and showcased significant differences in skillsets that correspond to the responsibilities held by different data experts within commercial, business and industrial enterprises.

The research provided interesting results in profiling the multiple data centred roles present in the job market. Our approach for machine learning has brought interesting results, but it could be extended further on as a possible tool to differentiate between the roles and possibly to help in classifying new job offers for companies which are unsure what occupation of data expert they are looking for. Some of the new occupations like machine learning engineer and research scientist were not included, as they had far too few offers in Poland as of the time of writing this article, but this might change in the near future.

As a limitation of our research, in our approach, not all of the skills were included, especially not soft skills, which could potentially create some new insights into these roles—which is an interesting area of further work. However, based on the technical skills analysis, the Polish market of IT skill requirements for data centred roles do not seem to differ in any way from US and global trends, which means that the specialists are interchangeable in this global market. The demand for data centred roles seems to be similar, emphasizing that the talent shortage also appears in Poland.

Due to the diversity and changing nature of the skills, not all of the required technologies are taught broadly in the University programs, but according to the preliminary study of curriculums, most of the key skills are included in current offerings. Unfortunately, most of the Polish master programs lack focus on the main analytical skills found to be crucial in this study, and it seems that data science was just thrown into the already existing master programs, sometimes with a name change. Data specialists need to have a very wide portfolio of technical skills, which they must update every year—which means that the technologies are expected to be shifting. However, the big data paradigm of computing and connected technologies, cloud computing, machine learning, deep learning or stream processing have been present in the scientific literature and different study programs for years.

This may highlight the reason for the shortage of data scientists and engineers on the market—as the requirements of companies seem to require a multitude of different skills, which universities struggle to prepare their alumni for. Meeting the average skillset for data scientist or engineer requires extensive experience in dealing with technologies that are often hard to self-learn (e.g. cloud computing). As noticed by Song and Zhu (2016), “the biggest bottleneck in the big data era is the production of capable data scientists, and producing such capable people takes time”. Due to this, the programs offered could specialize more, there is no possibility of graduating IT experts alone, as the area has broadened so much in the past 10 years that the notion is unrealistic. Similarly, as there is no single program to produce an engineer, with the profession’s multitude of specialties and disciplines and unique key skills, there is no way to build an all-around IT expert.

On the other hand, these specialists should not be expected to be the ’one man army’ that the current entrepreneurs are looking for. It is crucial to understand that average developers are not machine learning experts and run of the mill data scientists are not developers. While a data scientist is crucial to have in the team when collaborating on developing an innovative data-driven solution, they are mostly not equipped to scale the solution, which is what data engineers (or DevOps specialists) are for. This, of course, does not mean there are no experts that will meet these criteria, but if the focus is on minimizing the gap—narrowing the responsibilities for each of the roles may contribute to higher chances of educating future data experts and decreasing the talent shortage. The description of responsibilities contained within our study may also help future entrepreneurs in specifying what experts do they need in their companies, narrowing the expectations may lead them to more effective processes of developing innovative products and services.

The profiles extracted for these roles allowed us to clearly differentiate between the responsibilities of data centred roles, but the skills demanded from a single company may still differ from the averaged profiles that we created. We may expect some unification in the future, as well as the appearance of roles which better emphasize some of the current roles, hence, narrowing the skillset (machine learning engineers or cloud architects), which may, in turn, influence the skills of the main-stream occupations.

Our analysis of the curriculums proved to be interesting in terms of assessing the supply of data skills. Still, our reasoning only provides conclusions that might help in meeting the shortage of experts in the data market. The results of this article may offer some interesting insights for some of the stakeholders in the job market:

- **For companies interested in data analytics:** With an increase in demand for data analysts, data scientists will play a more precise role as machine learning / AI specialists, applying their skills to develop prototypes of new data-driven solutions. The technologies they use may be influenced by existing company structure and market trends. Executives should take a selective approach to determining which analytics specialists are really needed and whether the goals can be accomplished with automated tools or available services (Ramachandran & Watson, 2021). The role played by a data scientist or any other data expert should be in line with the company strategy and their skills should be used to address specific needs (prediction, AI, new products) in which the company requires custom tools or wants to gain a competitive advantage. On the other hand, the majority of the daily analytical jobs are to be performed by data analysts whose held skillset must slowly evolve from PowerPoint and Excel, to the world of Python/R/SQL and Tableau dashboards or other visualizations utilizing the vast data sources transformed by data engineers.
- **For innovative entrepreneurs and startups:** The outcome of this article exemplifying key analytical skills should help in discovering bona fide data scientists among the students and alumni of University programs that meet their needs. The entrepreneurs should, however, be aware of the possibilities that big data, machine learning and connected technologies can offer. Data scientists will not be these magical experts that can solve all business, technological and organizational challenges in a startup. The skillset analysed describes them as experts with very precise knowledge on data processing and predictive analytics, with a much more goal- and product-oriented approach than that of data analysts.
- **For aspiring data scientists:** Investing time in learning either cloud computing or machine learning and deep learning technologies is the most promising

area for meeting the skill shortage in the near future. With hands-on experience with AWS/GCP/Azure, Spark/Hadoop and other big data processing and database knowledge engines, there might be a lot of data engineer jobs opening in the future, as the trend for this role is on the rise. On the other hand, it does not seem like the need for data-centred roles is going to decrease anywhere soon, but the emphasis on specialization may make it easier to find suitable and responsible job positions. Overall, the need for both specialized data experts and those who will perform daily business analytics will increase as companies are expected to build more and more complex analytical pipelines in the future.

- **General University programs remarks:** For now, Polish Universities offering "data science" and similar programs do not seem to have a clear skillset profile of the alumni for their Master programs or lack focus on key skills. Their approach seem to be more in line with teaching a good base of general skills tied to the main discipline of the studies that is only enriched by data science knowledge. However, due to the very broad nature of skills that are required by data expert roles, they might consider putting more focus on big data, cloud computing and similar key skills which are now a requirement for data scientists and engineers. This might help in addressing the shortage of skills in the global economy and teach a new generation of experts, who being equipped with the latest knowledge, might create much more opportunities in terms of innovation. Ensuring all of the 'hard skills' are covered by their curricula by adding in a reasonable number of ECTS might be a solution for the issues described—as the market does not require just some postgraduates, but it needs data scientists, engineers and analysts with a well-defined set of skills applicable in modern enterprises. In a world where big data, NLP, AI and machine learning are used widely even in academic studies, there is no reason to limit these subjects in master programs that are directly focused on data science, big data or business analytics. The possibility of undertaking this and achieving high success rates has been demonstrated by non-Polish top tier universities offering similar master studies with much more focused study programs in this domain.

The findings of this article are only the beginning for a full description of the responsibilities and skills evidenced in those occupying data-centred roles. The analysis of technical skills allowed showcasing key responsibilities and reveals that it is possible to differentiate between some of the data expert activities. However, with the appearance of more specialized roles in the upcoming years, the exact skillset may still change. Technical analysis of demand for particular skills over the years should, however, be the base for any governmental, university or private educational programs addressing the education of AI and data experts. The stability of measured skills over the years demonstrates that some key responsibilities, e.g. data processing and prediction skills, remained stable over time despite the adoption of new tools. This might provide incentive to teach those key skills, which are we do not suspect to significantly change in the upcoming three or five years. Addressing them can decrease the data expert talent shortage that the world is currently facing.

References

Adam Mickiewicz University. (2021). *Study program for data analysis and data processing master's.* Retrieved from https://wmi.amu.edu.pl/dlakandydata/studia-ii-stopnia/analiza-i-przetwarzanie-danych

Blake, A. (2020). *Understanding the UK AI labour market: 2020 executive summary.* Retrieved from https://www.gov.uk/government/publications/understanding-the-uk-ai-labour-market-2020

Blake, A. (2021). *Dynamics of data science skills—how can all sectors benefit from data science talent?*. Retrieved from https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/

Bulldogjob.pl. (2021). *Raport „badanie społeczności IT" dla stanowiska analityk IT*. Retrieved from https://bulldogjob.pl/it-report/2021/analyst

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, *50*(3), 1–42.

Chih-Hsu Lin, J. (2020). *Analysis of 5,500 data science jobs (2020): Popular skills and location.* Retrieved from https://medium.com/@chihhsulin/5-500-datascientist-jobs-report-2020-adefe1d364d3

Davenport, T. H., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(5), 70–76. Retrieved from https://sociology.berkeley.edu/sites/default/files/documents/job_market/Data%20Scientist%20--%20HBR%20 2012.pdf

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education*, *29*, 27–39.

duBois, J. (2021). *The data scientist shortage in 2020.* Retrieved from https://quanthub.com/data-scientist-shortage-2020/

Feng, J. (2021). *The 2021 data science interview report.* Retrieved from https://www.interviewquery.com/blog-data-science-interview-report/

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician,* 72(4), 382–391. https://doi.org/10.1080/00031305.2017.1356747

Konkel, A. (2021). *Tech sector feeling covid-19's economic pain*. Retrieved from https://www.hiringlab.org/2020/07/30/tech-sector-covid19-impact/

Korczak, J., Abramowicz, W., Gołuchowski, J., Kobyliński, A., & Owoc, M. (2014). Wzorcowy program studiów licencjackich kierunku informatyka ekonomiczna–koncepcja wstępna. *Informatyka Ekonomiczna*, *2*(32), 311–337.

Kross, S., & Guo, P. J. (2019). *Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges*. (Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1–14).

LinkedIn. (2020a). *2020 emerging jobs report UK.* Retrieved from https://business.linkedin.com/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-uk-cont-fact

LinkedIn. (2020b). *2020 emerging jobs report US.* Retrieved from https://business.linkedin.com/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-us

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition (Vol. 5, No. 6). and productivity.* McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_exec_summary.ashx

Miller, S., & Hughes, D. (2017). *The quant crunch: How the demand for data science skills is disrupting the job market. Burning Glass Technologies.* Retrieved from https://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market

Motion Recruitment. (2021). *The 5 tech roles in highest demand for 2021.* Retrieved from https://info.motionrecruitment.com/talent-shortage-in-2021

Olsen, G. (2021). *Top industries hiring for machine learning, data science in 2020.* Retrieved from https://www.linkedin.com/pulse/top-industries-hiring-machine-learning-data-science-2020-greg-olsen

Opendatascience.com. (2020). *Looking for data science jobs in the pandemic? Good news and not so good news.* Retrieved from https://medium.com/@ODSC/lookingfor-data-science-jobs-in-the-pandemic-good-news-and-not-sogood-news-1add9367c861

Papoutsoglou, M., Mittas, N., & Angelis, L. (2017). *Mining people analytics from stackoverflow job advertisements*. (43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 108–115). https://doi.org/10.1109/SEAA.2017.50

Poznan University of Economics and Business. (2021). *Study program for informatics and econometrics master's—information systems in business and administration specialty track.* Retrieved from https://esylabus.ue.poznan.pl/pl/21/S/2/all/IiE/IGA

Ramachandran, K., & Watson, J. (2021). *Tech looks to analytics skills to bolster its workforce—addressing the analysis talent shortage.* Retrieved from https://www2.deloitte.com/us/en/insights/industry/technology/data-analytics-skills-shortage.html

Reid, N. (2018). The role of statistics in the era of big data. *Statistics Probability Letters*, 136, 1–3. https://doi.org/10.1016/j.spl.2018.04.009

Seaman, A. (2021). *Linkedin jobs on the rise: 15 opportunities that are in demand and hiring now.* Retrieved from https://www.linkedin.com/pulse/linkedin-jobs-rise-15-opportunities-demand-hiring-now-andrew-seaman

SGH Warsaw School of Economics. (2021a). *Curricula for first and second cycle programs in English at SGH.* Retrieved from https://www.sgh.waw.pl/en/study-plan-graduate-studies

SGH Warsaw School of Economics. (2021b). *Study programs curriculums and courses*. Retrieved from https://usosweb.sgh.waw.pl/

Shin, T. (2021). *The most in-demand skills for data scientists in 2021.* Retrieved from https://towardsdatascience.com/the-most-in-demand-skillsfor-data-scientists-in-2021-4b2a808f4005

Song, I. Y., & Zhu, Y. (2016). Big data and data science: What should we teach?. *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Techhub.dice.com. (2021). *Dice tech job report Q1 2020—the fastest growing hubs, roles and skills*. Retrieved from https://techhub.dice.com/Dice-2020-TechJob-Report.html

University College London. (2021). *Data science master's studies.* Retrieved from https://www.ucl.ac.uk/prospective-students/graduate/taught-degrees/data-science-msc

University of California. (2021). *Data science master's studies.* Retrieved from https://ischoolonline.berkeley.edu/data-science/curriculum/

University of Warsaw. (2021a). *Data science master's studies.* Retrieved from https://www.wne.uw.edu.pl/pl/dla-studentow/plany-zajec-i-programy-studiow

University of Warsaw. (2021b). *Reproducible research course.* Retrieved from https://usosweb.uw.edu.pl/kontroler.php?_action=katalog2/przedmioty/pokazPrzedmiot&prz_kod=2400-DS2RR

University of Wrocław. (2021). *Data science master's studies*. Retrieved from http://datascience.uni.wroc.pl/courses.html

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84.

Warsaw University of Life Sciences. (2021a). *Data science master's studies.* Retrieved from http://www.wzim.sggw.pl/studia/planystudiow/

Warsaw University of Life Sciences. (2021b). *Study programs curriculums*. Retrieved from http://www.wzim.sggw.pl/studia/sylabusy/

Warsaw University of Technology. (2021a). *Big data analytics course for data science masters studies.* Retrieved from https://ects.coi.pw.edu.pl/menu2/detail2test/idProgram/2157/idWydzial/13/idStopien/2

Warsaw University of Technology. (2021b). *Study programs curriculums*. Retrieved from https://pages.mini.pw.edu.pl/~estatic/e.mini.pw.edu.pl/pl/programmes/2020-2021/

World Economic Forum (2020). *The future of jobs report 2020*. Retrieved from http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf