Project acronym:	**EDSA**
Project full name:	**European Data Science Academy**
Grant agreement no:	**643937**

D1.4 Study Evaluation Report 2

Deliverable Editor:	**Leonard Mack (Open Data Institute)**
	Dr David Tarrant (Open Data institute)
Other contributors:	**Aba-Sah Dadzie (Open University) – Technical Report**
Deliverable Reviewers:	**Simon Scerri (Fraunhofer)/R.Brochenin (Tu/e)**
Deliverable due date:	**31/07/2016**
Submission date:	**29/07/2016**
Distribution level:	**Public**
Version:	**1.0**

This document is part of a research project funded
by the Horizon 2020 Framework Programme of the European Union

Change Log

Version	Date	Amended by	Changes
0.1	9/5/2016	Leonard Mack	Created document and began analysis
0.2	27/5/2016	Leonard Mack	Updated introduction and results
0.3	6/6/2016	Leonard Mack	Updated recommendations and exec summary
0.4	28/06/2016	Leonard Mack	Updated recommendations and wrote conclusions
0.5	1/7/2016	Leonard Mack	Updated future studies and exec summary
0.6	8/07/2016	Leonard Mack	Incorporated technical report and finalised document for consortium reviewers
0.7	28/7/2016	Elena Simperl	Scientific Review
1.0	29/7/2016	Aneta Tumilowicz	Final QA

Table of Contents

Change Log 2
Table of Contents 3
List of Tables 4
List of Figures 5
1. Executive Summary 6
2. Introduction 10
2.1 General background 10
2.2 Purpose and motivation of the study 13
3. Methodology 16
3.1 Study design and methodology updates 17
3.1.1 Increasing study reach 17
3.1.2 Improving the question design 18
3.1.3 Consolidating the key areas of data science 18
3.2 Summary of data collection 20
3.2.1 Interviews 21
3.2.2 Survey 22
3.2.3 Focus groups 25
3.2.4 Desk research on data science courses 26
3.2.5 Online job postings 29
3.3 Limitations of the study design 32
3.3.1 Sample boundaries 32
3.3.2 Representativeness and validity 33
3.3.3 Voluntary participation and sample bias 33
3.3.4 Data collection and use 34
3.3.5 Scraping online data and data licences 35
3.4 Study reach and key performance indicators 36
4. Results and analysis 39
4.1 Demand analysis survey and interviews 39
4.1.1 Sample size and coverage 39
4.1.2 Survey results and analysis 46
4.1.3 Interview results and analysis 56
4.2 Desk research on data science courses 65
4.2.1 Supply of training across Europe 65
4.2.2 Classification of training courses 68
4.2.3 Training languages 70
4.3 Job posting analysis 70

5. Discussion and recommendations 76
5.1 Holistic data science training 77
5.2 Technical and analytical data science training 79
5.3 Building soft skills 79
5.4 Providing data literacy training for non-data scientists 80
5.5 Exploring options for blended-learning training 81
5.6 Establishing a data science skills framework 82
5.7 Providing navigation and guidance 83
6. Conclusions 85
6.1 Conclusions 85
6.2 Future work 87
6.2.1 Deepen sectoral and country research 87
6.2.2 Increase total sample size 88
6.2.3 Explore the benefits of non-English data science training 88
6.2.4 Conduct further analysis based on learning analytics 89
7. Appendices 90

List of Tables

Table 1: Recommendations for EDSA curriculum development ---------- 9
Table 2: Purposes of D1.4 ---------- 14
Table 3: Primary data collection modes ---------- 15
Table 4: Summary overview of the subcontractor's KPI compliance ---------- 17
Table 5: Original and consolidated data science skills categories ---------- 19
Table 6: Summary of content analysis in the data science course survey ---------- 27
Table 7: KPI's compliance at M18 against target set at M6 ---------- 36
Table 8: Sample split by country and mode (descending by total count) ---------- 39
Table 9: Regional distribution after aggregating country data ---------- 41
Table 10: Organisational coverage across regions ---------- 42
Table 11: Sectoral coverage of sample ---------- 43
Table 12: Split of roles by region and organisation size ---------- 44
Table 13: Split of roles by industry sectors (listed by share of managers responses in total sector responses) ---------- 45
Table 14: Course provision of country ---------- 66
Table 15: Count of classified data science training offers ---------- 68
Table 16: Course languages ---------- 70
Table 17: Top 20 skills by frequency of mention for data set containing ca 316K job postings across Europe ---------- 73

Table 18: Comparison of selected skills across locations by frequency: Comparison of selected skills across location, showing frequency of mention in postings for the top 6 countries, and the percentage this is of the total number of postings for each country----------74

Table 19: Recommendations for EDSA curriculum development----------77

List of Figures

Figure 1: *Growth of job starters in analytics and data between 1990 and 2010* ----------12

Figure 2: *Overview of the demand analysis research process*----------16

Figure 3: *Screenshot of the online survey available through the EDSA* ----------23

Figure 4: *Knowledge framework for the acquisition and processing of online job posting data* --------30

Figure 5: *Skills that a data scientist should have* ----------48

Figure 6: *Self-assessment of own skills by data scientists (N = 355)*----------49

Figure 7: *Assessment of team's skills by manager (N = 278)*----------50

Figure 8: *Technologies, tools and languages to be included in data science training* ----------52

Figure 9: *Preferred training methods of data science professionals*----------55

Figure 10: *Job posting network graph* ----------71

Figure 11: *Spring based layout of skill co-occurrence* ----------72

1. Executive Summary

The digital economy breeds exponentially growing data - according to some estimates 2.5 exabytes per day[1]. To understand and innovate using this rich resource is the 21st century's quintessential challenge.

While organisations across Europe have identified opportunities to become data-driven, a heterogeneous group of professionals has emerged to meet industry's far reaching expectations: Classically, this group is referred to as data scientists, even though the wider profession also includes related roles such as data analysts, data engineers, and statisticians.

Employment opportunities are broad: Data from LinkedIn shows an exponentially growing share of job starters working in analytics and data science which grew by more than 1,000% between 1990 and 2010[2]. A study conducted by e-skills UK and SAS also predicted that, in the UK alone, the number of big data specialists working in large firms will increase by more than 240% between 2012 and 2017[3]. For the same period data-driven businesses are expected to contribute 58,000 new jobs to the labour market[4].

At the same time, however, the supply of new data scientists is not on a par with market demand. More efficient and effective training for data science professionals is needed to fill this new skills gap. But at the same time, the skills which data scientists apply and need vary widely by sectors, organisational background and team requirements.

Against this background, this study explores which data science skills European industries need and how the skills gap can be closed through better training.

We conducted an in-depth analysis of the data science skills and training demand across different industries in Europe, using a mixed methods approach of primary and secondary data collection. In practice this meant one of the largest studies of its kind, made up of:

- 108 face-to-face and telephone interviews with data science practitioners and managers.
- 4 focus groups.
- 584 telephone and online surveys completed by practitioners working in data science and data science team managers.
- A comprehensive desk survey of 456 data science courses across Europe.
- In depth interviews with to 19 high level managers and learning professionals on how they approach data science skills development in their organisations.

Topically, we focused on five key domains:

1. Demand for data science skills

[1] http://www.bbc.co.uk/news/business-26383058
[2] Patil (2011): Building Data Science Teams
[3] E-skills uk / SAS (2013)
[4] SAS (2012): Data equity

2. Current level of data science skills
3. Required data science tools
4. Skills acquisition strategies
5. Preferred training methods and delivery modes

1. Demand for data science skills

83 percent of our survey respondents said that data collection and analysis skills are essential skills, followed by data interpretation and visualisation skills. Perhaps the most striking quantitative result is that both advanced computing and open source skills are less in demand - which might come as a surprise given how intensely modern data science can rely on these domains.

When considered alongside our interviews, this seems to reflect on a dominant strategic expectation that data scientists act as data-driven transformers of organisations. This however requires strong presentation and communication skills in order to influence senior management and other functional departments to make the right decisions based on data.

Consequently, when we asked which additional skills data scientists should have, communication and presentation skills topped our list; other frequently demanded expertise is in teamwork, social skills and data management. Data scientists are often hired with high expectations regarding their abilities to transform business tactics and strategies; thus soft-skills such as these are seen as desirable and need greater focus in data science training.

2. Current level of data science skills

When it comes to skills that data scientists seem confident in themselves, data interpretation and analysis expertise is not among the top skills. Respondents rate their skills higher in domains such as advanced computing, machine learning and business intelligence: Interestingly, these are the same skills which they also ranked as less important to have.

In general, managers are slightly more optimistic about their team's skills. While the results for very good and good skills are roughly the same, managers do not rate their team's skills in machine learning highly, whereas they are very confident in their teams data collection and analysis skills.

3. Required data science tools

Data science is deeply dependent on the emerging provision of digital technologies, tools and languages. More than a third of those surveyed want to see training on general purpose programming languages such as R and Python. Java should be covered according to 16.5 percent of respondents. Almost the same share of participants would also like to see SQL included. The inclusion of special-purpose programming languages for data held in relational databases seems to reflect the general importance of relational database management for the work of data scientists. MatLab was the only proprietary programming language mentioned, and that by only 4 percent of users.

Interestingly this seems to reflect the great importance which data science professionals currently put on open source, highly flexible and customisable analytics solutions. This is also contradictory to their assessment of the importance of the skill, showing how the field takes this skill for granted.

4. Skills acquisition strategies

From our interviews with data science professionals we learned that self-driven, ad-hoc learning is arguably the most important approach to continuously acquire data science skills. Participants indicated that data science professionals need to be continuously learning and adapting. Data scientists, their managers and learning professionals all highlighted the great importance of informal, self-guided learning. Some even said that the majority of skills development in data science is achieved through these means.

Asynchronous training modes, such as through MOOCs and other online trainings, appear most suitable for this approach. At the same time, however, our study participants also noted that these means are less effective than face-to-face trainings. Additionally, domain specific skills are often acquired on the job. Trainings which help to acquire better domain specific skills, e.g. through sector specific examples and assignments, are scarce instead.

5. Preferred training methods and delivery modes

Particularly relevant for EDSA's curriculum development are the respondents' preferred training methods. No clear champion emerges here, with participants rating all delivery methods highly, and face-to-face trainings topping the training wish list. Additionally, sector specific trainings and assignments are high in demand as well. In our interviews, respondents frequently proposed to supplement general training contents with sector specific exercises.

The least requested feature is training with materials in users' native language; non-English was only rated essential by 18 percent of responses. Training for non-open, non-free software is the second area with limited demand. 42 percent of respondents think it is irrelevant. At the same time however, 45 percent think that training with proprietary tools is at least desirable, again reflecting the expectation that data scientists already have skills in this area.

With regards to the existing supply of trainings, we find that the size of the opportunity has led to significant growth in data science training offers in recent years. Through our desk research on the provision of data science training across Europe, we were able to identify 456 academic and professional development courses from European providers, offered in 23 European member states. With a strong supply particularly of Master degrees and professional short courses, which frequently cover tool-specific contents, European data scientists and those who want to be trained as such do not lack training options. Rather, they and their managers voiced concerns that they are finding it hard to navigate the market and identify good training offers.

Recommendations for EDSA curriculum development

Resulting from this analysis, we recommend holistically developing EDSA's curriculum and more general project work in five directions:

Table 1: Recommendations for EDSA curriculum development

Title	Summary description
1. Holistic training approach	Refine EDSA's training approach and curriculum cycle to strengthen data science skills for data science teams and data literacy across various units of each organisation.
2. Open source based training	Continue current technical and analytical training based on open source technologies; apply cross-tool focus to deliver overarching training.
3. Soft skills training	Integrate soft skill training to increase performance and organisational impact of data scientists / data science teams.
4. Basic data literacy training	Develop basic data literacy training for non-data scientists to improve basic skills across organisations and facilitate uptake of data-driven decision making and operations.
5. Blended training	Develop blended training approaches including sector-specific exercises and examples to increase effectiveness of training delivery.
6. Data science skills framework	Implement a data science skills framework to structure skills requirements, assess skills of data scientists, and identify individual skills needs.
7. Navigation and guidance	Develop quality assessment of third party courses; provide navigation support to identify relevant trainings from EDSA and third parties.

While implying a complex set of actions, we believe these recommendations can help to substantially progress EDSA, raising its project impact and sustainability chances in the long term.

2. Introduction

2.1 General background

The quintessential feature of the digital economy in the 21st century is the exponentially increasing production and consumption of data. Estimates of the world's existing data regularly produce numbers that are hard to understand for most humans: In 2012, IBM calculated that 2.5 exabytes - that's 2.5 million terabytes - of data was generated every day[5]. One year later, the Internet was estimated to host 4 zettabytes of data, one zettabyte accounting for one billion terabytes[6]. The same year, Kenneth Cukier and Viktor Mayer-Schoenberger calculated that, mainly through the Internet and digital devices, the world's data had grown so much, that there was enough to give every person alive ca. 1,200 exabytes of data. Saving all this information on CDs and stacking them up would be sufficient to build five separate CD piles, each one reaching to the moon[7].

On a daily basis most data is not processed by individuals, but the digital and digitalised economy. eBay and Google are the largest datavores, both processing around 100 petabytes[8] per day. In Europe, Spotify makes use of around 2.2 terabytes of compressed data per day[9]. Even organisations that do not classify as Internet giants use extensive amounts of data in their domains. For example, the Wellcome Trust's Sanger Institute, a British genomics and genetics research institute, uses around 1.7 terabytes per day for DNA sequencing[10].

The increasing use of large datasets, sometimes known as big data, as a tradeable and exploitable commodity, is a distinctive feature of the digital age. From an economic perspective, three foundational factors shape this development. First, raw data produced from digital transactions (e.g. in eCommerce), by IoT and mobile devices, as well as social media is becoming increasingly available. Second, there are now tools with the capacity to handle large and big data volumes at a low cost and in a reliable fashion. A variety of (open source) software tools have made complex data management and processing easier; Hadoop is probably the most prominent example[11]. The third factor is, and remains, arguably the most decisive for an organisation's ability to make use of the data abundance: human labour.

Humans steer in-depth data analysis, distill the most important findings, and make them useable for decision making and business development activities. None of these activities can be covered by artificial intelligence, yet. Instead, a new profession is supposed to master this challenge, usually referred to as data scientists. These workers' capabilities to cut through data, uncover hidden patterns, and provide the groundwork for more effective, evidence-driven business transformation has led to much advance praise, to the point where data scientist has been declared the sexiest job of the 21st century[12].

[5] http://www.bbc.co.uk/news/business-26383058
[6] https://vsatglobalseriesblog.wordpress.com/2013/06/21/in-2013-the-amount-of-data-generated-worldwide-will-reach-four-zettabytes/
[7] https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data
[8] 1 petabyte equals 1000 terabytes.
[9] https://followthedata.wordpress.com/2014/06/24/data-size-estimates/
[10] http://www.slideshare.net/insideHPC/cutts
[11] http://hadoop.apache.org/
[12] http://128.255.244.58/strategic/articles/data_scientist-the_sexiest_job_of_the_21st_century.pdf

Perhaps it is a sexy job partly due to the role of a data scientist still being undefined. The title might create the impression of a relatively uniform group of people, but in reality, it serves as the lowest common denominator for a highly disparate profession. Data scientists come from a variety of interdisciplinary backgrounds, often combining knowledge in computer science, machine learning, maths, statistics, and the business domain they work in. The name given to the role in which data science is used seems to depend strongly on sector-specific business environments, as earlier research by McKinsey has shown[13]. Data engineers, data analysts, data architects, and business intelligence analysts are just some of the roles usually working in the wider data science space.

While job profiles vary and are still in the process of being defined, organisations across Europe agree that they need this new breed of workers. Accordingly, data scientists are in high demand. Countries with well-established digital economy sectors and companies in particular want to see more data science professionals joining the workforce. Nevertheless, this is only partly a reflection of the fact that organisations have to deal with increasingly large and complex data. This demand reflects the competitive advantage which individual companies and whole economies expect from becoming data driven.

Projections on the demand for data science professionals follow a strong upward trend: Data from LinkedIn shows that the share of job starters working in analytics and data science domains grew by more than 1,000% between 1990 and 2010 (see figure 1)[14]. This growth has also taken an exponential curve.

[13] http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[14] Patil (2011): Building Data Science Teams, http://www.oreilly.com/data/free/building-data-science-teams.csp (Accessed July 06, 2016).

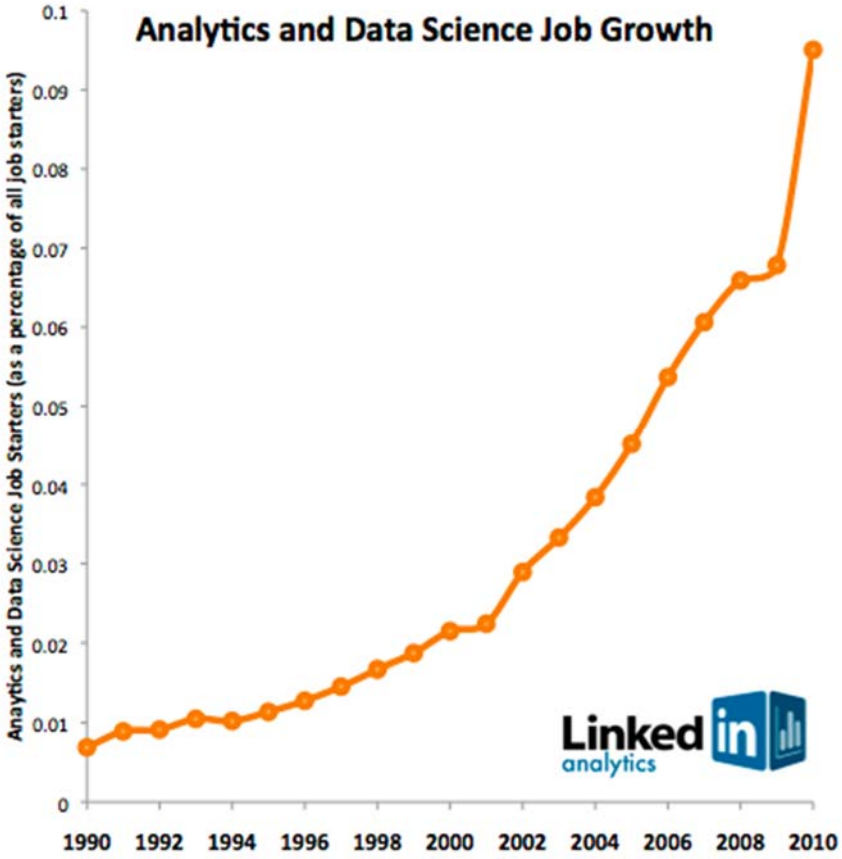


Figure 1: *Growth of job starters in analytics and data between 1990 and 2010*

Source: LinkedIn, reproduced in "Building data science teams"[15]

A study conducted by e-skills UK and SAS in 2013 predicts that, in the UK alone, the number of big data specialists working in large firms will increase by more than 240% between 2012 and 2017[16]. For the same period, SAS also expects data-driven businesses will contribute 58,000 new jobs to the UK's labour market[17].

Certainly, some of the roles covered by these predictions are not data science jobs as they include business support or sales staff working in data science companies. Nonetheless data science professionals are urgently needed to progress an ever more digital economy.

Meeting this demand is a strategical and complex objective for the EU as a whole. Accordingly, the European Commission made upskilling the EU's workforce a priority in its major initiatives on

[15] Patil (2011): Building Data Science Teams, http://www.oreilly.com/data/free/building-data-science-teams.csp (Accessed July 06, 2016).
[16] E-skills uk / SAS (2013)
[17] SAS (2012): Data equity

digitalisation[18]. Some argue that this problem is particularly challenging for the EU - after all, it lacks a Silicon Valley-like digital innovation hub and so far has not produced iconic digital giants such as Google, Amazon, or Facebook[19]. Despite such arguments, European universities and professional training providers have introduced a variety of data science degrees and courses over the past few years[20]. Additionally, numerous online learning platforms, most of them not from Europe, now offer data science training. As asynchronous trainings, the courses offer full time workers a degree of flexibility by allowing them to complete course when they can. Where a modular structure is in place, they also enable students to personalise training in line with individual needs and preferences.

The European Data Science Academy (EDSA) project's goal is to take this state-of-the-art concept further, by integrating the principles of modularity and adaptability into a rigorously assessed, demand-driven learning offer. Particularly targeted at the needs of European data scientists, EDSA aims to set a new standard in delivering multi-platform training for the next generation of European data scientists. Assessing the demand in main industry sectors is key to this objective.

The skills that data scientists apply and need vary widely by sectors and individual companies. Bearing this complexity in mind, it is both surprising and unsurprising that, until now, no comprehensive review of skills and training needs for (European) data scientists exists. Unsurprising, because the task is complex and because focusing on somewhat similar, generalist, and high level skills might appear sufficient for an emerging sector. But more critically it is surprising, because not understanding actual demand may lead to ineffective, poorly focused training. This is a severe deficit, risking producing data scientists that are ill-equipped to tackle companies' true challenges[21]. This study therefore sets out to uncover industry demand and training practice for data scientists across Europe.

2.2 Purpose and motivation of the study

One of the core objectives of EDSA is to progress Europe's competitiveness in data science. This will be achieved through the provision of high quality, industry-fit training. Uncovering industry trends and skills needs is essential to this task, as is integrating the discovered demand into EDSA's modular training framework. In this report, we seek to answer three consequential research questions:

1. What is the current demand for data skills in different European industry sectors?
2. What training should be offered in order to accommodate this demand?
3. Which options exist for EDSA to develop a sustainable offer for high-impact data science training?

To address these questions, we conducted an in-depth analysis of the data science skills and training demand across different industries in Europe. Methodologically, this report builds on the previous deliverables D1.1 and D1.2, which have prepared the study design and reported on a pilot study.

Mapped against the project's ambitions, we seek to address and promote three high impact purposes through this report, described in table 2.

[18] Most importantly, references to the need for more data scientists and related skills are for example included in the EU's Data Value Chain Strategy, the Digital Single Market Initiative, or the Digital Agenda for Europe.
[19] https://www.questia.com/library/journal/1G1-414426911/big-demand-for-big-data-scientists-in-europe
[20] See section 3.2 for the results of our survey of data science trainings in Europe.
[21] http://www.computing.co.uk/ctg/news/2433095/a-lot-of-companies-will-stop-hiring-data-scientists-when-they-realise-that-the-majority-bring-no-value-says-data-scientist

Table 2: *Purposes of D1.4*

1	**Developing the EDSA curriculum**
	The demand analysis report will guide development of the EDSA curriculum. This analysis consists of an in-depth, cross-country, cross-industry survey of the data science skills and training demand across Europe. This is essential to identify approaches for EDSA's ambition to provide both generalist and domain specific training. For this purpose, the report contains EDSA-specific recommendations for the ongoing curriculum development (see section 4).
2	**Surveying and analysing stakeholder needs**
	This report provides data and insights on current skills and training demand. This is essential for EDSA's ambition to serve as a central European access point to data science training. The report also serves as a resource and evidence base to better coordinate European initiatives in data science training along common strategical lines.
3	**Progressing research**
	Finally, this report will also help to progress research into the profession of data scientists. Until now, the specific skills needs of data scientists in different industry sectors are rarely discussed. In practice, this means that often data scientists are hired as "generalists" with rather limited domain knowledge. However, our research shows that domain-specific data science skills are high in demand - how this can be met is however not well understood.

As a basis of this report, we have conducted a series of data collections throughout the past 18 months. The essential findings of this work are evaluated and presented in this report. Our core objective was to ensure a rich, multi-dimensional view into Europe's current data science skills landscape. To achieve this goal, we structured our sample along four dimensions:

1. **Geographical spread**: Data was collected from all 28 European member states, grouped into four UN-defined regions.
2. **Roles**: We targeted data science professionals and managers of data science teams. Data was collected from high level managers as well as ground-level practitioners, ensuring a deep understanding of the required skills of a data scientist. As will be explained in section 2.3.1 of this report, we relied on a broad definition of data scientists, including various self-identified professionals working in the data science domain. Accordingly, while we use the term "data scientist" in this report, it should be noted that this is not exclusively limited to workers carrying this title. Rather it covers a broad range of profiles working in this domain.
3. **Sector**: We gathered data from organisations across 19 Eurostat-classified industrial sectors. This data helps evaluate skills demand and adoption in different industries.
4. **Organisation size**: We targeted a variety of organisation size; from micro and SME to large multinational.

Along these lines, we collected data through quantitative telephone and online surveys and qualitative interviews, from 692 respondents in total (table 3). For our study, the additional collection of qualitative data is a differentiator which allows a triangulation of results. To further back up our findings during through the course of the study, we also conducted 19 interviews with high level managers and learning professionals on their organisation's approach to developing data science skills; furthermore, we conducted four focus groups to discuss the implications of organisational and team development needs. Together, this multi-dimensional view helps us to assess the validity of quantitative results and provides additional depth in comparison to other studies which are frequently based on only one study method, most often quantitative measures.

Table 3: *Primary data collection modes*

Data collection mode	**Short description**	**Responses**
Phone and online survey	The survey was designed by the EDSA consortium. It was also made available via the project website, returning 84 responses. Additionally, it was implemented and conducted by a subcontracted research firm, to collect standardised, quantitative data on the training needs of organisations across the EU. Depending on participants' preferences, the survey was conducted online or via telephone.	**584** completed surveys
Interviews	Interviews were designed by the EDSA consortium and carried out by consortium partners, affiliates and the subcontracted research firm. Interviews helped to collect qualitative, in-depth data on the training needs of organisations across the EU.	**108** completed interviews
Supplementary qualitative data collection		
Interviews with high level managers and learning professionals	Interviews were designed to explore organisational needs and approaches to data science skills development. In the final data collection stage, these interviews helped to discuss and validate some of the core findings of our previously collected data. All interviews were conducted by the Open Data Institute (ODI), either in person, by phone or online.	**19** completed interviews
Focus groups	Four focus groups were conducted, helping to discuss and test core concepts presented in our surveys and interview questionnaires, and to gather insights into organisation skill and training challenges. Focus groups were conducted by the ODI as well as Fraunhofer with a variety of participants from both public and private sector organisations.	**4** completed focus groups

In our pilot report D1.2, we found early indications on a variety of different trends, that we explored further in this report. For example, differing opinions between data scientists and managers about what constitutes effective training. We also explored a general perception that data science skills need to be expanded but that effective training is hard to find, and that domain specific knowledge is hard to acquire.

We have integrated two further studies into this analysis. While they are not situated at the core of this report, they nevertheless add valuable contextual data on ongoing industry trends. Section 4.3 reports on the data collected through the EDSA's online demand dashboard. This data allows us to take a closer look at skills required from job postings for data science positions. Section 4.2 also summarises our findings for a cross-country survey of the data science courses offered by universities and professional training suppliers in Europe. Given the urge for more data scientists, the number of courses offered by such suppliers has been rapidly expanding over the last 2-3 years. Hence, we will also be able to reflect our findings on demand against a snapshot of Europe's current supply in data science training.

The remainder of this report is organised as a classical study design. In the next section, we report on our methodological approach, using a mixed methods research design. We then present the results and analyse them to identify significant patterns. These patterns guide us through the interpretation of results and the development of recommendations for the EDSA's curriculum. In our conclusions, we collect suggestions for further research.

3. Methodology

Figure 2 lays out the overall research process for this study. Based on the problem definition and research design outlined in D1.1, we approached the data collection stage. As laid out in section 2.2, we deployed a mixed methods approach, guided by the collection of qualitative data.

To ensure that we delivered best practice and built upon lessons learnt during the data collection, we followed an iterative implementation approach. Practically, this means that we included evaluation and redesign phases at appropriate points within the overall study timeline. Major adaptations of the study methodology have been reported in D1.2, following a pilot study, and in the interim demand analysis report which was produced for the project consortium at M12. Adaptations of the methodology after this point are recorded below.

Throughout the data collection phase, we continuously focused on the qualitative and quantitative acquisition of both self-collected primary data as well as already existing secondary data. This was essential to our work and aims, since only a holistic approach would enable us to retrieve refined insights into Europe's current data science landscape. Many other studies only use a quantitative or qualitative approach. Hence, the integration of both quantitative and qualitative data as part of a mixed methods design is a key strength and differentiator for this study.

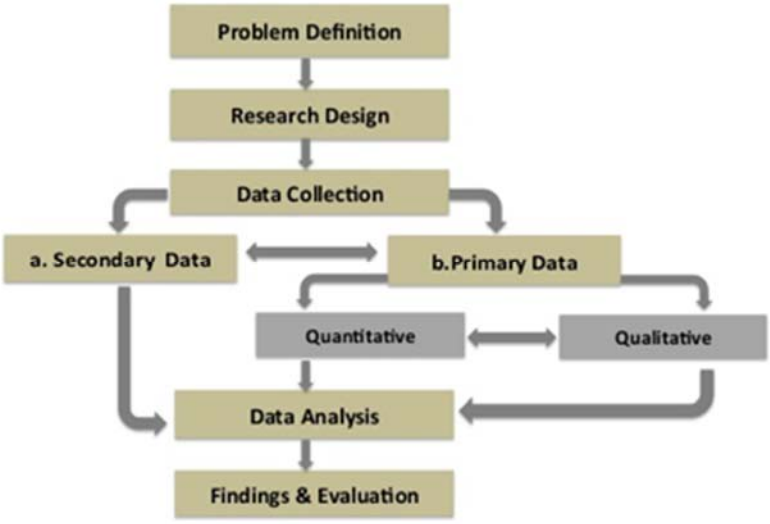


Figure 2: *Overview of the demand analysis research process*

Source: Deliverable D1.1 - Study design document[22]

[22] http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-D11-Final-v1.1-forwebsite.pdf

3.1 Study design and methodology updates

Following our study pilot (M1-M6), we conducted an extended evaluation to improve our data collection, based on initial feedback and practical lessons learnt during the first data collection. Our consultation with project partners led to some adjustments, intended to enable a more effective data collection. In sum, we concentrated our revisions on three different domains:

3.1.1 Increasing study reach

To facilitate the collection of primary quantitative and qualitative data, we briefed and received quotes for the required services from multiple research organisations, before subcontracting a company specialising in targeted multilingual telephone and online data collection. From our total sample, this research firm carried out 54 telephone interviews and 500 surveys.

In order to maximise the range of responses from different backgrounds, we provided the company with a set of coverage criteria for countries, sectors, roles and organisation sizes. To balance the amounts of qualitative insights and quantitative data, all criteria were laid out as target numbers for interviews and percentages of overall responses. Table 4 provides a summary to what extent The research company reached these metrics[23]

Table 4: *Summary overview of the subcontractor's KPI compliance*

Target description	Description of subcontractor's target compliance
2 interviews per EU member state	Compliance was achieved for all EU member states except Slovenia, where only one interview was conducted.
2 interviews per Eurostat business sector classifier[24]	The target was met for 12 of the 19 sectors. However in the remaining seven sectors, coverage was more difficult to achieve. These included mostly more traditional industry sectors with often comparatively lower ICT usage, i.e. agriculture (sector A), mining and quarrying (sector B), water supply and waste management (sector E), construction (sector F), transportation and storage (sector H), real estate (sector L), and arts and entertainment (sector R).
Each business sector accounting for 3-15 percent of sample	The target was met for 13 sectors. However, respondent shares from the ICT sector (sector J) are slightly above target, providing 16 percent of responses. Five other sectors provided less than 3 percent of responses, including agriculture (sector A), mining

[23] More detailed information on how the subcontractor performed against the KPIs is provided in Appendix 3.
[24] http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC

	and quarrying (sector B), water supply and waste management (sector E), construction (sector F), and real estate (sector L).
Each UN-defined European accounting for at least 15 percent of sample	Full compliance with the target was achieved.
55-65 percent of respondents should be managers; 35-45 percent of respondent should be non-managerial data scientists	The target was not met with respondent shares even reversing. 55 percent of respondents were data scientists and only 45 percent managers. After discussions, The research company found it substantially easier to retrieve responses from non-managers as they seemed to be more willing to discuss their insights, and also seemed more interested in the research
At least 30 percent of the sample should be SMEs; at least 40 percent of the sample should be large companies	Full compliance with the target was achieved.

3.1.2 Improving the question design

Apart from minor wording adjustments to improve the clarity of questions in the study's multilingual setting, the essence of qualitative questions remained untouched throughout the data collection. This was a crucial requirement to guarantee the comparability and compatibility of data collected during different stages of the project. Nevertheless, in the final version of our survey, we decided to add an additional question to capture more details on currently important tools, technologies and techniques in data science. Including this question allowed us to collect highly relevant information guiding the technical curriculum development in work package 2.

3.1.3 Consolidating the key areas of data science

As described in D1.1, we derived seven of our eight initial key areas of data science for this demand analysis from Drew Conway's Data Science Venn diagram[25]. Generally, these domains can also be seen as skills:

- Math and statistics
- Machine learning
- Domain expertise
- Data skills
- Advanced computing

[25] A version of the diagram is displayed in this paper of Drew Conway for the IQT Quarterly: http://static1.squarespace.com/static/5150aec6e4b0e340ec52710a/t/51525211e4b0e9fad0b56f9c/1364349457311/IQT-Quarterly_Spring-2011_Conway.pdf

- Visualisation
- Scientific method

To this list, we added "open culture" as an eighth category. Open source technologies and open access activities in data science are important drivers for the emerging data science environment. Some of the most influential tools, technologies and programming languages are open source based, such as SQL, R, Python and Hadoop[26]. Their liberal licensing regimes facilitate the customisation and replication of tools. Additionally, open data is an important, easily accessible resource for emerging data scientists to experiment with and learn new technologies with real world data, and is often used in combination with other data to generate insights in data science practice. Together, these factors led us to include awareness of "open culture" as an additional, important characteristic for data scientists.

After our pilot study, we revisited these eight original skills areas in order to align them better with other project activities, specifically the development of the curriculum. In M6, a first cycle of the curriculum had been implemented (D2.1 - Data science curricula 1[27]). The structure of the curriculum was derived from an initial market analysis of the European data science education market, conducted by the University of Southampton in the early stages of the project.

To cross-validate the skill categories, the ODI's demand analysis pilot and the University of Southampton's market research were based on two different skills categorisations. To arrive at a robust set of categories we analysed initial feedback through both the pilot and market research. This resulted in a consolidated set of skills categories, displayed in table 5.

Table 5: *Original and consolidated data science skills categories*

Initial skills categories for demand analysis pilot	**Consolidated, final skills categories**
Math and statistics	Maths and statistics
Machine learning	Machine learning and prediction
Domain expertise	Business intelligence and domain expertise
Data skills	Data collection and analysis
Advanced computing	Advanced computing and programming
Visualisation	Interpretation and visualisation
Scientific method	- *Not matched in final skills*
Open culture	Open source tools and concepts
- *Not included in final skills*	Big data

[26] http://www.techrepublic.com/blog/big-data-analytics/data-scientists-can-find-big-money-in-open-source/
[27] http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-P-D21-FINAL.pdf

As can be seen from a comparison with the initial categories, our adjustments were minimal with most areas remaining largely in line with Drew Conway's categorisation. The major advantage of the revised categories is their consistency, which allows us to test demand for the curriculum topics. It also ensures that future versions of the curriculum can be adjusted based on data from the demand analysis as well as ongoing feedback from industry and course participants.

While collecting evidence on demand according to predefined categories is essential to assess demand for the EDSA's curriculum, we also required participants' opinions on skills areas that are not covered in our framework. In a highly dynamic, emerging environment such as data science, new and emerging skills need to be monitored. To address this, we added an open ended question on 'Other sector specific skills' to our survey and interview questionnaire.

3.2 Summary of data collection

To be comprehensive, our data collection focused on the retrieval of rich primary and secondary data from a variety of different sources.

Primary data was collected through a variety of different modes, both qualitative and quantitative. We drew extensive qualitative data from semi structured telephone and face-to-face interviews with data science practitioners, managers and learning professionals (conducted in person, via telephone and VoIP online services) as well as four focus groups. To triangulate the resulting findings, we additionally collected quantitative data through a survey which was implemented as a self-administered online survey and as a guided telephone survey.

In total, we collected 108 interviews as well as 584 survey responses from non-managerial data science professionals and team managers. Furthermore, we conducted 19 interviews with senior managers and learning professionals, as well as four focus groups.

Secondary data was collected through manual desk research and retrieved from online services. The online services data initially included LinkedIn data to analyse trends. Later, we expanded this to job posting portals such as Adzuna and Trovit[28]. In our desk research, we conducted a comprehensive survey of data science training courses from European universities and professional training suppliers, covering 456 courses from across the EU. Taken together, the retrieval of this secondary data allowed us to take a snapshot of both the demand for data scientists on the European labour market as well as existing courses to train more of them. To better understand the real world developments in data science in Europe, this analysis provided us with rich contextual insights.

[28] In addition, we considered and explored options of using data from other sources such as Monster, Indeed, Jobs.ac.uk and Data Science Central. However, these were not usable for our purposes for various reasons, including non-accessibility of data and limited country coverage.

3.2.1 Interviews

Semi structured interviews formed a crucial part in the acquisition of qualitative primary data for this demand analysis. Speaking directly to data scientists and their managers allowed us to explore in detail the needs and key challenges in different European regions and industries. Additionally, following the main series of interviews, we conducted a smaller series of interviews with senior managers and learning professionals. The objective of this second series of interviews was to explore some initial trends discovered in our data as well as to explore in more depth the approaches which organisations take on data science training.

Implementation and progress

The interview questionnaire was designed and outlined at the start of the study in D1.1 and then, as discussed in sections 3.1.2 and 3.1.3, revised following the pilot study (D1.2). To match the design of this first interview questionnaire, the questionnaire for our supplementary interviews largely followed this design. Hence, in both cases, interview questions spanned across three key themes, aligning with the main focus areas of the project[29]:

1. **Practice of data science in Europe**, providing context to the project on the current state of data science across Europe. This included in particular questions on the impact of data science on organisations in Europe and the key challenges in finding skilled people for data science roles.
2. **Current provision of data science skills**, collecting information on which skills currently exist on the labour market and which training courses participants had attended. We also asked whether participants had taken alternative approaches to training (e.g. mentoring and coaching) and which key challenges they identified in finding training.
3. **Preferred training methods**, gathering information on which delivery methods are currently used, which are preferred and which would be most effective.

All questions were tested during the pilot stage of the project up to month 6 through a series of interviews conducted by the project consortium. In our pilot report D1.2[30], we found early indications on a variety of different trends affecting the demand for data science trainings.

Notably, there were differing opinions between data scientists and managers about what constitutes effective training. We also uncovered a general perception that data science skills need to be expanded but that effective training is hard to find, and that domain specific knowledge is hard to acquire.

Building on this basis, we extended our geographical and sectoral reach in the main data collection phase. As part of the large scale role out of the study, interviews were conducted by:

1. members of the project consortium as well as data science experts enrolled as 'EDSA Ambassadors', including representatives from France, Malta and Bulgaria (52 interviews);
2. the subcontracted research company specialised in market research (56 interviews).

To ensure the consistency of data collection processes, guideline instructions were given to both interviewers and interviewees. Furthermore, to guarantee a better accessibility of the interview process to non-native English speakers, interviews were translated and conducted in 20 European languages.

[29] Appendix 3 documents the full questionnaire design.

[30] http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-P-D12-FINAL.pdf

Throughout the 12 months between March 2015 and March 2016, we collected qualitative data from 108 participants.

Analysis methodology

Our analysis of the qualitative data gathered during interviews was guided by an inductive, grounded theory approach[31]. Data science is an emerging profession, consequently theories about the dynamics in this domain rarely exist. From a research perspective, a core motivation of this report was therefore to identify and formulate hypotheses on the demand for data scientists and data-science skills across Europe. Hence, rather than pre-defining analytical categories for the interviews that might eventually be ill-fit for our research purposes and skew our perspectives on this emerging phenomenon, we decided to centre on an unbiased collection of primary data. By speaking extensively to industry experts we were able to draw a detailed picture of the current data science skills demand in Europe. Building on the rich base of 108 interviews, we then conducted an in-depth, inductive data analysis to identify common themes and patterns in the responses we gathered. Eventually, these observations would then allow us to establish evidence-based categories for the data science demand in Europe.

During the pilot phase of qualitative data collection, we analysed and coded results manually. Getting a sound contextual understanding for one's qualitative data is an important aspect in implementing grounded theory, hence while being time intensive - and thus arguably not efficient - taking this deep dive into our data ahead of the main analysis was important to improve our understanding of the data.

However, as the size of our sample grew, this manual approach was no longer feasible. Instead, to facilitate our analysis, we required some automatic, preliminary structuring of the raw data that was contained in our verbatim and intelligent interview transcripts. Interview transcripts were then analysed using Thomson Reuters' OpenCalais, a Natural Language Processing web service[32], to extract topics, entities, facts, and relationships described by participants. In addition to the processed information, OpenCalais returns semantic metadata in RDF format. As a second step, we used this metadata to inform a second round of manual coding to aggregate domain specific high level themes which the OpenCalais algorithm was no longer able to capture.

3.2.2 Survey

The survey was developed to enable the collection of larger, more standardised volumes of data for the study. Furthermore, it provided a method of data collection that required a shorter engagement by respondents, thereby facilitating a less time intensive participation in the study. To align the collection of data with the interview questionnaire, the survey covered complementary areas and questions to those of the interviews. Based on these design considerations, the focus of the survey was to capture quantitative responses. In addition to the results presented in this report, survey data is where possible, available for exploration on the project survey results dashboard.

[31] Martin, Patricia Y. and Turner, Barry A. (1986): Grounded Theory and Organizational Research. In The Journal of Applied Behavioural Science, Vol.22 No.2, pp. 141-157.

[32] Thomson Reuters' OpenCalais - https://permid.org/onecalaisViewer

Implementation and Progress

To use synergies between different modes of data collection, the survey was implemented both as an online and phone survey. In the latter case, participants of the qualitative interviews were also given the option to answer the quantitative survey questions via phone.

The survey was first made available through the project website during the demand analysis pilot stage (see figure 3). With regard to user interaction, we placed value on an engaging and interactive implementation of the survey tool in order to attract more respondents and improve survey completion rates. Therefore, the survey included interactive features such as a clickable map to select the respondent's country.

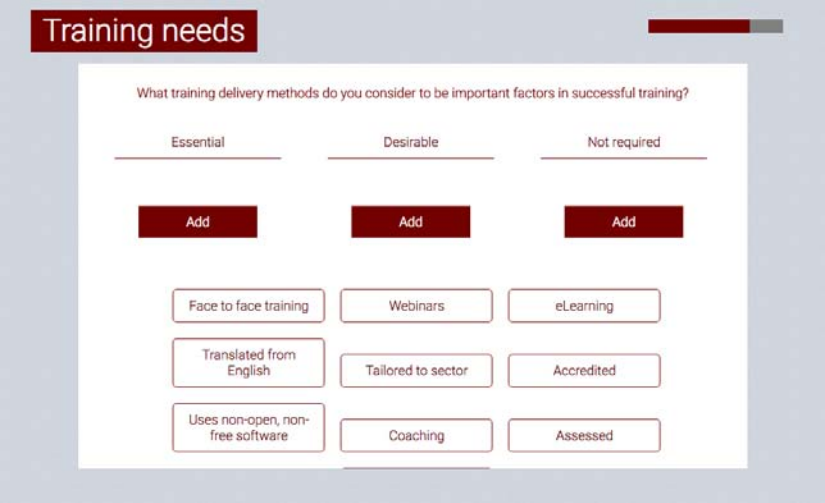


Figure 3: *Screenshot of the online survey available through the EDSA*

After the evaluation and revisions of the pilot study, we promoted the online survey to a wider audience through the EDSA's social media channels. This also involved targeted Twitter requests for participation in specific sectors and countries. We also distributed the survey through our Industry Advisory Board and circulated information at various events with EDSA representation.

To increase the accessibility of the survey across Europe, consortium partners and others in the project network provided translations of the survey into 7 languages: English, German, French, Greek, Polish, Slovenian and Swedish.

Additionally, we contracted the research firm to collect further data from a wider reaching sample. In implementing the survey, the firm offered both an online and telephone survey to study participants. In the first case, the firm deployed an adapted version of the survey, implemented through the company's own online survey system. This did not include the interactive features available through EDSA's own system. In order to facilitate participation across the EU, the research firm also translated the survey in all primary languages of EU member states. In the case of telephone surveys, users were also given the option to participate in their native language (if participants were non-native English speakers).

In total, we collected 584 valid survey responses between February 2015 and April 2016; 500 responses were collected by the research firm, another 84 through the consortium's bespoke system. An additional 48 invalid responses, collected through the consortium's survey system, have been excluded from the analysis. Invalid responses included mainly duplicates and cases where participants did not specify their role as "data scientist" or "manager", thus limiting the utility of this data for further analysis.

The survey[33] is currently still available online through the project website and will continue to feed the survey results dashboard with any new responses. However, responses after April 2016 are not covered in this report.

With regards to the continuous presentation of the survey findings, the results dashboard[34] was separated to guarantee a consistent user experience while development work progressed on the other parts of the dashboard. In addition to this report, the survey dashboards allow users to explore results through an interactive tool.

To lead as a best practice example in making research data available to external users, we are publishing the quantitative demand analysis data in CSV format under a Creative Commons Attribution 4.0 licence[35]. Not included in both the dashboard view and the downloadable data are 118 responses for which informed consent for redistribution of individual-level data had not been provided[36]. Despite this exclusion, these responses are still covered in this report.

Analysis methodology

To identify industry and regional trends, we extracted statistics for all valid survey data. In this process, we relied mostly on the production of descriptive summary statistics, specifically total counts and proportions. Given the very high dimensionality of data which reports over 28 different countries, sub-samples of different high level categories (e.g. industry sectors, country, or company size) were too small to conduct more advanced regression analyses with robust results.

Some selected data required re-coding before being usable in the quantitative analysis, specifically in three cases:

1) Open ended questions
   In some cases, both the online and phone survey included open-ended questions, e.g. on important technologies, tools, and languages that data scientists require; as well as relevant skills which are currently not covered by the EDSA's curriculum. In both cases, we used OpenCalais' semantic text analysis tools to extract topics from recorded responses. These were then grouped into larger categories, facilitating an aggregated, quantitative evaluation of results.

2) Re-coding of scale-based answers from phone surveys
   The data for some scale-based questions collected through phone surveys was not always reported according to the distinct, pre-defined categories of the online survey. This applied specifically to questions which asked participants to rank how needed certain skills are for data scientists[37] or how they would rank their own or their team's skills in certain domains[38]. Phone participants would not always settle for one category, leading the subcontractor to report both

[33] http://davetaz.github.io/EDSA/survey/
[34] http://edsa-project.eu/resources/dashboard/
[35] http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis-/
[36] A further discussion of changes regarding the informed retrieval of informed consent and data uses is included in section 3.3.4 below.
[37] Ranking on a scale ranging from "not required", "desirable" to "essential".
[38] Ranking on a scale from 1 to 5, where 5 is the maximum.

mentioned categories in those cases. When coding these answers, we took different approaches depending on the underlying scale: For questions using a scale ranging from 1 to 5, where respondents mentioned two values we coded the answer for counting and averaging purposes as equivalent to 0.5 to each value mentioned by the respondent[39]. In some other cases, the scale ranking used the categories "essential", "desirable" and "not required", with respondents answering questions as "essential/desirable" or "desirable/not required". In those cases, we coded answers as "desirable" - based on the rationale that the other options are more extreme, and likely would reflect participants' opinions less accurately.

3.2.3 Focus groups

To accompany the qualitative data collection, the consortium partners conducted four focus groups. We designed the focus groups as problem-focused workshops. Through this approach we not only acquired further insights for the project, but also supported participants and participating organisations in addressing key data science challenges and discovering data science training needs.

Implementation and progress
In total, we conducted four focus groups in the UK, Germany and Sweden, with participants coming from various industry backgrounds in public sector services, health, media, entertainment, and manufacturing. The decisive factors for this selection were to ensure a wide coverage across multiple sectors as well as to strengthen engagement with data science stakeholders from the public and private sector.

The first focus group was conducted during the pilot stage of the project to test and evaluate our design for subsequent groups. Following an interactive, discourse oriented approach, the group set out to co-define the skills of a data scientist, assess each participant's capabilities within this framework and develop a plan of action to increase data science capabilities in their respective organisations. After some refinements regarding the contents and moderator guidance notes, this general design was replicated in three further focus groups.

To deliver on our value proposition for focus group participants, we did not follow a standardised focus group design. Instead, we adapted a format and style to suit each group. Before implementing the session, representatives of the consortium and participating organisations discussed organisational needs to define each workshop's scope and targets. As a consequence, focus groups differed in their size ranging from 8 to 40 participants.

Three focus groups were delivered by the ODI to participants from single organisations. The fourth focus group was delivered by the Fraunhofer Society's Big Data Alliance and involved primarily HR professionals from German companies. Compared to the other focus groups, which consisted mainly of participants from data science and analytics teams from one organisation, the fourth focus group undertook a more strategical, high-level exploration of organisational development needs in the context of data science and overall personal development.

[39] This means that if a respondent ranked his/her skills in "interpretation and visualisation" between 4 and 5, we add a count of 0.5 to both categories.

Despite significant time investments to customise each group design in collaboration with participating organisations, we found that the focus groups were an excellent supplementary method to collect further qualitative data. The problem-based, discourse oriented approach in particular helped catalyse discussion among participants, delivering in-depth data on organisations' current and future skills needs in data science. The rich interaction with participants also serves as a good guidance to scope out potential directions for the EDSA's future.

Analysis methodology
Similar to the analysis of qualitative interview data, we took an inductive, grounded theory-led approach to analyse the outcome of the focus groups. We prepared summary notes rather than full transcripts of the discussions for several reasons. The interactive design of the sessions, including hands-on exercises were difficult to document in full detail. In addition, the discussions of single-organisation focus groups involved reflections on confidential or sensitive intra-organisation issues. To encourage a free flowing discussion, we refrained from producing word-by-word transcripts in this setting.

To supplement our qualitative findings, we therefore took a case study approach, treating each focus group as one case. By capturing, analysing and categorising the main results of the focus groups, we were able to match patterns against the findings of our qualitative interviews. The results, which are reported in the private appendix C, lend more details to some core issues on data science skills needed in Europe. Throughout the main body of this report, core insights from the focus groups are used to back up survey findings.

3.2.4 Desk research on data science courses

An important element to complement our research into demand patterns was to also understand the current supply of data science training in Europe: What is the current data science training landscape in Europe? What courses exist from universities and professional training suppliers? During recent years, this space has been developing rapidly, so we needed a detailed snapshot of the current training landscape. For this, we researched a comprehensive database of data science training offered by European higher education institutions and professional development suppliers.

Implementation and progress
To develop this image of Europe's current data science training landscape, we conducted, manual desk research in May 2016. In creating our analysis approach and template, we were guided by initial course research conducted by the University of Southampton in the first six months of the project. Spanning all 28 EU member states, we surveyed 456 courses, out of which 221 were from higher education institutions and 235 from professional development suppliers.

The interdisciplinary nature of data science required us to draw distinct lines to define which courses would be included in our research. While there is an increasing number of suppliers that offer degrees and courses labelled as "data science", these are frequently only composites of already existing seminars and modules. Hence, in restricting our research to offers that are exclusively labelled as data science degrees or courses, we would have missed a wealth of other relevant offers. This led us to also include in our sample courses that are not labelled as data science, but cover disciplines closely related to it.

Based on this rationale, we defined a narrow list of search terms to identify relevant courses, modules, and degrees. The following terms were then used in a Google incognito search to uncover relevant offers:

- Data Science
- Big Data
- Data Analytics
- Business Analytics
- Machine Learning
- Distributed Computing
- Advanced Computing

Although this list proved to be useful to capture self-identifying data science courses, it does not cover those emerging from other, related fields. For example, some Computer Science, Informatics, Digital and Software Engineering, and/or Analytical Mathematics Masters programmes contain core and optional modules that would allow students to emerge with a data science focus. However, since we chose to explicitly center on courses that had been intentionally designed with a data science focus, we considered this more selective approach to be appropriate.

Furthermore, we conducted our research only using English search terms, potentially missing courses where the same terms were used in a country's native language. However, our analysis showed this to only be the case for very few instances as an overwhelming majority of providers seem to use English either as a general training language or at least to tag their courses with the above listed English terms. A search for the equivalent terms in other languages, e.g. French and German, did not return improved results but rather overlapping ones, when compared with the standard search in English. We are therefore confident that our list represents a broad selection of the data-science training identifiable through our selection of search terms.

Analysis methodology

We identified training courses from various higher education and professional development suppliers on a country basis. In this process, we conducted a structured content analysis, to collect in-depth details on a variety of course characteristics. Table 6 below includes a full list of our survey characteristics.

Table 6: *Summary of content analysis in the data science course survey*

Content item	**Description**
Type of course	Format the course follows, e.g. undergraduate, masters, short course
Higher Education/Professional Development	Higher Education courses are offered by an academic institution. Professional development courses are offered by an organisation and aimed at professionals
Course Title	Title of course in English
Course Title (Home Language)	Title of course in original language (not applicable for courses published in English)
Course link	URL where the course is published
Course Provider	Academic body, organisation or company through which the course is run / accreditation is provided

Country	Country in which the course is delivered
Primary location (City, Online, Blended)	City in which the course is delivered
Language	Language in which the course is delivered
Length of course	Course duration
Qualification or accreditation	Qualification or accreditation that participants receive on successful completion of the course, where applicable
Study mode (FT, PT, Flexible)	Full time, part time or flexible study
Department(s), Faculty	The department and faculty the course sits within, where applicable
Core modules/units/course content	Titles of mandatory modules or units or an overview of course content in English
Core modules/units/course content (Home Language)	Titles of mandatory modules or units or an overview of course content in original language (not applicable for courses published in English)
Optional modules/units	Titles of optional modules or units
Techniques, tools, programming languages	Techniques, tools and programming languages that are taught as part of the course, where applicable
Placement	Details of any placement within another university or organisation participants can/are required to take to complete the course, where applicable
Total cost EU (starting 2016)	Total cost of course to participants from within the European Union, in currency published
Cost non-EU (starting 2016)	Total cost of course to participants from outside the European Union, in currency published
Subsidised Learning	Details of any discounts, grants or scholarships available for the course
Target audience	Type/level of participant course is aimed towards
Listed, searchable?	Third-party websites where the course is listed, where applicable
search term, tag	Search term under which the course is tagged on third-party website
Entry requirements?	Requirements of participants before undertaking the course
Frequency of course	How often the course is delivered
Start date/time	Start time and date of next course instance
Year established	Year course was first delivered
Notes	Additional information about the course which does not fit within another field

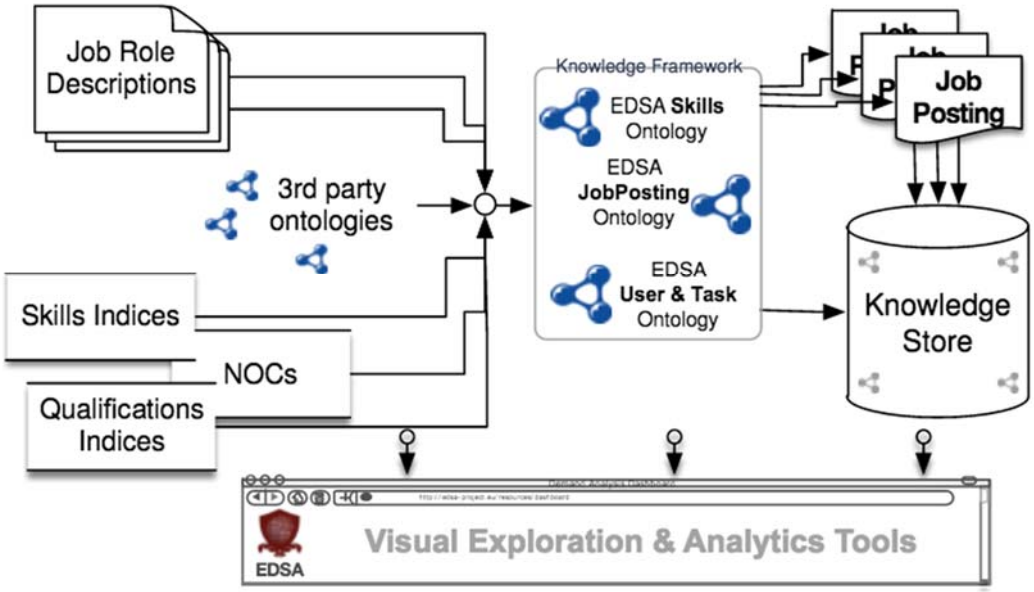
Industry partners	Lists organisations professionally associated with the course

The resulting dataset, which has been released under a Creative Commons -Attribution 4.0 license (CC BY 4.0), is accessible online[40]. External users such as potential or current data science trainees and students can use this data to identify relevant training offers. For the EDSA, this work is also highly relevant as it allows us to further explore supply analyses, ensuring that our market and competitor is comprehensive.

At this stage, we have analysed the courses data primarily to explore the existing training supply and to retrieve a rich contextual view of the training markets current evolution stage. The high-level results of this analysis, are reported in section 3.2 of this report.

3.2.5 Online job postings

To get a real life impression of the current data science job demand, we collected extensive job posting data from a variety of online sources. Offering users a self-explorative view, we included this data into the web-based 'Demand Analysis Dashboard', which is targeted at three key user types: the 'policy- or decision-maker', the 'trainee or job seeker' and the 'expert or practitioner'. Figure 4 provides an overview of the (knowledge) framework built around the user-centred methodology. This provides target users with intuitive tools to browse data and conduct exploratory analyses. These functions can also support key tasks as our research continues within the EDSA project. D1.2[41] reported findings from our initial exploratory analysis, which fed into the user and system requirements specification and initial design phases.



[40] https://theodi.github.io/data-science-courses-in-europe-2016/

[41] http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-P-D12-FINAL.pdf

Figure 4: *Knowledge framework for the acquisition and processing of online job posting data*

Our evaluation of the first working prototype revealed that users without the background knowledge of the consortium's work faced challenges in obtaining a good understanding of the data overviews. The results have fed into the redesign of the user interface (UI) to focus on the user task. This was implemented to ensure that the analysis of the demand data, the dashboard UI as well as its underlying functionality support both project partners and target end users in finding information about the demand for data science skills. This is critical if we are to correctly identify core skills and inter-relations between skills overall, as well as variation in demand with context - over *time*, across *geographical location* (and corresponding working language) and by *domain or industrial sector*. Based on our outcomes we aim to support identification of skill gaps, to feed into the design of skill- and context-driven learning material to work toward closing gap between job capacities and skills available.

In this section of the report, we focus on the high level design and implementation of the research methodology for this sub-study. The technical report attached at the end of this document provides more detailed information on the design as well as technical documentation on how the research was implemented.

Implementation and progress

EDSA's work on assessing data science skill and job demand was generally split into parts:

1. analytical research by project partners to discover demand and current capacities in the workforce, employing the demand data collected for this exercise in combination with relevant third-party resources;
2. the design and construction of user-centred tools for presentation of the input data and results of the analysis in the demand dashboard.

The explorative analysis is carried out independently by project partners following one or more paths:

1. (to support) tasks identified during user and system specification for one of the three key user perspectives, i.e. the 'policy- or decision-maker', the 'trainee or job seeker' and the 'expert or practitioner'.
2. statistical analysis of summary and usability evaluation data, to underpin the eventual release of demand benchmarks.
3. Visual analysis corresponding to the data type for one or a combination of the three key measures or indicators: (a) temporal, (b) geographical location (spatial) and (c) non-spatial skill and skill set lists. Location-based analysis, for instance, employs map-based visualisations; temporal focuses on a variety of timelines and spatio-temporal employs coupled views or multi-dimensional analysis. Network analysis and high-dimensional techniques such as matrices and parallel coordinates are used for skill frequency and correlation analysis.

An evaluation through the project consortium as well as with other data science and computer science experts returned feedback on the UI design and functionality for browsing the spatio-temporal data and skill co-occurrence, as well as keyword and location search. Comments were collected after interactive demonstration of the tools at policy and scientific fora and meetings. With smaller groups and individuals, demos were accompanied with hands-on interaction by study participants. Guided by the

task-based questionnaire[42] to be used in the next stage of usability evaluation, a think-aloud methodology was used to capture comments on the tools and the analytical support provided. Finally, a pilot of the formal usability evaluation was carried out with two practitioners. Evaluation results are being fed into further analysis and development and planning for formal evaluation. This will also assess whether the independently built tools have been successfully coupled and enable more in-depth analysis from multiple perspectives.

Analysis methodology

Working from the exploratory study described in D1.2 and by extending relevant third party vocabularies and knowledge bases, we have built and continue to refine a framework, which is formally defined in the *Skills and Recruitment Ontology* (SARO)[43]. This ontology guides the collection and processing of the demand and related data, and subsequent exploratory and more in-depth analysis. Target data sources include online advertising and recruitment portals such as Adzuna, Monster, LinkedIn, Indeed and Trovit, and other industry and domain-specific sites such as Jobs.ac.uk and Data Science Central. The data acquisition process followed two approaches – the use of:

1. custom web crawlers to collect job postings from relevant websites,
2. content providers' APIs (Application Programming Interfaces) for data collection.

Filters for identifying relevant postings are based first on (morphological) variants of the job title "Data Scientist" and the skill sets defined in D1.1, following Conway's Data Science Venn diagram. We then expanded the filter to include other technical and *soft* skills listed in matching postings from the first data collection round. A second filter was then applied to match other attributes of job postings defined in our knowledge framework (the SARO ontology). These included the date the advert was posted, job title and location - all of which were required for the analysis of demand along the three key indicators:

1. Time;
2. geographic location;
3. capacity vs. capability - captured as essential and desirable skills or skill sets.

Data on job titles along with hiring organisations will feed into future analysis of distribution by domain and industry sector.

Each posting that passes through both filters was then annotated to highlight all attributes defined in the framework described in the posting. Matching skills terms and frequency of mention are extracted from the job title and role description. Finally each posting is enriched with more detailed and precise information on geographical location[44], before being added to the RDF store as illustrated in figure 2.6. Data provenance is recorded using the source URL and any other information required for data use and by redistribution licenses.

[42] Task list and questionnaires available at: http://bit.ly/29c66tx
[43] SARO is hosted at: http://eis.iai.uni-bonn.de/vocab/saro/index.html
[44] Location data extracted from the posting is matched to the GeoNames database (http://www.geonames.org) to extract detail such as latitude, longitude and the formal place name, country name and code where necessary. Other information, such as on population for comparison of demand per capita, feeds into post-processing analysis.

As of June 2016 the demand data store contains approximately 300,000 processed postings. These cover 17 EU and EEA countries with between 5,000 and up to 90,000 postings per country. Additionally, we list 16 countries with less than 50 posting each. The data spans historical data from 2013 up to new positions advertised in June 2016.

The results obtained from the different approaches were triangulated to verify both the methodology followed and the results obtained. To validate our approach, we restricted our analysis first to the set of skills and skill sets defined in D1.1. The smaller set of 40 skills, grouped into 7 skill sets, allowed us also to manually inspect both input data and results across multiple approaches, and repeat tests with different tools to confirm initial findings.

We also evaluated factors that impact scalability of our approach; to be effective our tools must scale with data size and complexity, to allow us to refine the picture of demand as we continue to collect historical and emerging data on demand across the EU. At just over 300,000 postings the dataset already far exceeds manual inspection beyond random samples post aggregation. Early validation allows us a measure of confidence in the findings obtained considering the complete data set, which, due to size and dimensionality, now relies on semi- to fully automated analysis.

3.3 Limitations of the study design

While our research design materialised as a feasible and useful approach to assess demand trends in a highly complex and dynamic field, our study still faces limitations. In the following section, we want to explain these. Notably, these are also connected to our conclusions' discussion on the directions for potential future work.

3.3.1 Sample boundaries

As discussed in our introduction, data science is an emerging occupation area with often blurred professional boundaries. While an increasing number of organisations seek to employ team members to work on data science tasks, their role profiles and job titles display great variety. In practice, it is therefore not only a narrow group of people with the title "data scientist" who carry out data science tasks. Additionally, data analysts, data engineers and architects, business intelligence analysts and many more often work in the wider data science domain as well.

This messy landscape meant a challenge in defining our sample boundaries: We could have limited participants for our interviews and survey to those who carry, in some form, the words "data scientist" in their job titles. With companies increasingly hiring or already employing "data scientists", this would likely have resulted in a more selective, quite distinct sample. However, as indicated, we would have missed many professionals who work on data science tasks, but who are not formally identified as "data scientist" by their job title. Our sample would then have been based on the likely erroneous assumption that companies label jobs according to a uniform, schematic understanding of their underlying profiles. In reality, this is, of course, often not the case. Rather, individual organisational needs, legacies, and politics tend to influence how a role is labeled in an organisation.

We therefore applied a more informal approach, asking participants to identify themselves during the survey and interviews as primarily 'data scientist' or 'Manager of data scientists'. This enabled a wider range of data collection, based on skills, rather than job titles. It also allows us to collect data not only from professionals, but from those working in the area that wish to train further in the field.

3.3.2 Representativeness and validity

While having the benefit of giving space to a more flexible survey implementation, this approach, particularly in combination with self-administered data collection approaches (e.g. online surveys), bears a major problem, namely there is little control with regard to the representativeness of the sample and reduced control regarding the validity of survey responses.

With this being an explorative study, our intention was not, however, to produce a fully representative study. Instead, our primary intention was to gather in-depth information on the data-science skills demand in different industries across Europe to guide the EDSA's curriculum development. This goal stood above the construction of a rigorous representative sample. On the one hand, the complexity of the data science labour market would make such a study far more complex, for example necessitating an extensive industry analysis to underpin the sample construction. Conversely, a much larger sample would have been necessary to be able to make statistically robust claims on the data-science skills demand in individual countries' industry sectors. Both issues were beyond the scope of this study. We instead laid a deliberate focus on a broad, explorative approach that was sufficient to indicate major demand trends from which we can derive insights for our curriculum development. We consider this study as a starting point to look across Europe's data science landscape. In this context, our quantitative data shows us a snapshot of the current landscape, and the qualitative data provides the basis for our interpretation of this picture in this report.

Furthermore, we took a proactive approach to ensure that we reached our intended audience for the survey. In particular, we targeted links to our online survey to selected data scientists and data science forums, for example, through Twitter direct messages and other social media channels. Additionally, we built an ambassador network of data-science experts with supporters in Malta, Bulgaria and France. The EDSA ambassadors helped us to disseminate the online survey through their own expert contacts and conducted interviews with practitioners. Additionally, our subcontractor applied different methods to increase our targeted sample, including continued desk research throughout the length of the study to identify appropriate survey contacts, and snowball sampling with respondents who participated in the study identifying colleagues or others in their network.

3.3.3 Voluntary participation and sample bias

Our goal was to cover a wide range of countries and sectors across the EU, based on a broad sample of data-science professionals. To collect data, we relied on the voluntary participation of respondents. This led us to take several measures aimed at rationalising, facilitating and enriching the collection of data:

- We designed our survey to be engaging and short, so that it can be completed in 10 minutes.
- To streamline data collection, we aligned the phone and online survey with the qualitative interview questionnaire. The core motivation for this was to facilitate the collection of larger amounts of data and to prepare for the triangulation of different results from both data sources.

- We actively used the EDSA consortium partners' European network in the data science domain to reach out to stakeholders and win them as study participants.
- After the pilot study, we subcontracted the international survey firm to expand data collection through both surveys and interviews.

Through the combined efforts in keeping participation barriers low and engaging actively with relevant audiences, we were able to attract 692 valid survey and interview responses[45]. While this represents in total a large sample for an explorative study, we however noticed some skews to the data:

- The subcontracted research company reported us that data-science practitioners were more willing to participate in the study than managers. This might follow from a greater immediate interest in the practical outputs of the project. Arguably because of greater time constraints, managers were generally harder to reach out to. As discussed in section 4.1.1, ca. 58 percent of our participants identified themselves as data scientists, thus almost inverting our initial KPIs. During the evaluation of results, we thus need to consider that a majority of responses have come from data scientists. Nevertheless, this deviation seems to have no further impact on the quality and contents of our results.
- In some sectors and countries, we also found it hard to acquire a large group of participants. From a sectoral perspective, this was specifically the case for sectors which are traditionally not strongly related to ICT or statistics, such as the Tourism, Water/Waste Management and Agriculture sectors. Looking at geographic coverage, participants from Eastern European countries were particularly hard to reach. Instead, countries from Northern and Western Europe, including the UK and Germany, are well represented. While this might reflect a wider deployment of data science in businesses of these economies, our total results thus still need to be interpreted with caution given the imbalance in respondents. To balance our limited coverage in some individual countries, we decided to aggregate data on a regional basis. A further discussion of this is provided in section 4.1.1 of this report.

Given the specific, explorative scope of this study, we do not identify these deficits as major problems. Rather, we understand them as points of departure for future research, specifically into the role of data science in industry sectors which traditionally make less use of ICTs and advanced statistics. Additionally, we would encourage further research into how data science is being used in Eastern European countries, such as the Baltics.

3.3.4 Data collection and use

As a consortium we aim to license the products of our work openly whenever we can[46]. Because it contains personal data, we however cannot publish all data from our study. This is particularly the case for all qualitative interview data.

[45] Further details of the sample composition are discussed in section 4.1.1.

[46] For an in-depth discussion of how we applied this intention to different data sources developed throughout the project, please refer to the updated data management plan in D5.6.

With some exceptions, anonymised survey data is available. Out of a total of 584 survey responses, we cannot include 113 responses that were collected by our subcontractor at the start of the study. This difference results from a change in the intended use of the data which occurred during the initial phase of the data collection. At the start of the project, we planned to only release summary statistics of the quantitative data through our skills dashboard. Accordingly, participants were first informed and also agreed to their data being made accessible in an anonymous, aggregated form. Following discussions that emerged during the evaluation of the pilot study, we decided to also release pseudo-anonymised (not directly identifiable) raw data. This data would provide access to responses on an individual basis, thus adding much greater detail and utility to potential reusers of the data, but the data is only anonymised and not aggregated. Additionally, we realised that, given the small subsamples for some countries and sectors, some of the data displayed in the dashboard would not be suitably anonymised despite a properly conducted data aggregation. In M9, we therefore changed the wording of the informed consent section of the survey, stating that data could later be made publicly available in anonymised form, using an open licence.

Due to the discrepancies between both formulations and in order to reduce risks of releasing personal data without permission, we decided to not use the initial participants' data on the dashboard nor to publish them as open data. Hence, the data collected from early study participants has still been included in the analysis in D1.4, but is not available in the downloadable link and is also not included in the dashboard view.

3.3.5 Scraping online data and data licences

Our analysis of job posting data relied on the use of third party online data from various services. Naturally, these apply a variety of licenses in order to regulate access to and reuse conditions of their data. Allowing data reuse in line with EDSA's preference for open licenses, which grant unrestricted reuse rights to third parties, was therefore challenging in some instances. While we assessed the terms and conditions of third party sites, such as LinkedIn, Adzuna API, and Learning Locker, only data from LinkedIn could be reused in line with the project's requirements.

Our research however also faced additional challenges, when LinkedIn changed its terms of service, which now no longer allow external reuses of LinkedIn data. Additionally, the company has also installed various technical measures to impede web scraping from their websites. Assuming that LinkedIn's terms regarding data access and reuse would not change substantially, we had not kept a record of LinkedIn's license at the time. Weighing the risks of legal repercussions versus the benefits of continuing to reproduce now protected LinkedIn data as open data, we decided to remove this data from our own open data store.

A second challenge emerged from the use of the Trovit data to populate the jobs dashboard. As we found, Trovit's license did not allow external uses of the data. However, after careful consideration, we decided as a Consortium that the text and data mining exception for research purposes under UK law[47] allowed us to use the data as long as it was not accessible by others.

[47] http://www.legislation.gov.uk/ukdsi/2014/9780111112755 p6 (accessed 30/06/2016)

With these issues in mind, we need to highlight that the scope of web-scraped third party data which we were able to reuse is not as comprehensive as we imagined. Legal and technical obstacles limited our work and the implementation of our originally envisioned research design substantially. From our perspective, the European Court of Justice's preliminary ruling[48] to specify the scope of the EU Database Directive in the case of screen scraping may have already lead owners of publicly accessible databases to protect their contents. In our case, Trovit and LinkedIn are examples of how this practice can impede the execution of more comprehensive and arguably better research designs. Ironically, in such cases, research is held back by legal uncertainties or prohibitive limitations on data reuse – despite the fact that the same data which could facilitate this research remains publicly accessible.

Hence, it should be clear that restrictions on data use frequently prevent individuals from maximising the value of data. Instead, open data, which can be used and shared for any purpose, eventually benefits original producers through increased coverage and traffic. If a company does not want anyone to benefit financially from their work, a non-commercial licence such as CC-BY-NC 4.0[49] would still enable others to use the data and link back to the source.

3.4 Study reach and key performance indicators

We restructured and developed our KPIs to expand the scope of the study at M6. Table 7 displays our achieved coverage at the end of month 18.

Table 7: *KPI's compliance at M18 against target set at M6*

KPI	**Target (M18)**	**Actual (M18)**
Size of network (qualitative analysis)[50]	168 – 6 sectors per member state	**108 interviews** **Average number of interviews per country: ca. 3.8** **Average number of sectors covered per country: 3.1**
Number of focus groups	4	**4**
Number of sectors	17	**19 (out of 21)**
% of EU business registry sectors	80%	**90%**

[48] http://curia.europa.eu/juris/document/document.jsf?docid=161388&doclang=EN
[49] https://creativecommons.org/licenses/by-nc/4.0/
[50] For a detailed report on the achieved coverage of the network KPI, please see Appendix 1.

Importance of sectors	80%	**99.8%**
Number of EU states	ALL	**ALL**
% split of Corporate / SMEs	60%/40%	**57%/43% (14 unknown)**
% split of Managers / Data scientists	60%/40%	**42%/58%**

While a more detailed discussion of the sample composition is provided in section 3.1, we want to briefly reflect the coverage of our results against our KPIs in this section. With 692 interviews and surveys in total, as well as four focus groups, we reached a good coverage across all EU member states. From a sectoral perspective, we also exceeded coverage of Eurostat-defined business sectors[51]: Instead of 17, we reached 19 of 21 sectors, accounting for ca. 90 percent of all sectors. Together, the surveyed sectors also produce around 99.8 percent of value added in the EU's non-financial economy[52], surpassing our original benchmark of 80 percent coverage. We also nearly met our goal to achieve a split of 60 to 40 between corporates and SMEs in the sample.

Our initially targeted split of 60 to 40 between managers and data scientists inverted through the course of the data collection. As discussed in detail in section 4.1.1, we found it substantially easier to acquire data scientists than their managers as study participants during the course of the project. Naturally, less managers than data scientists exist in organisations, thus representing a smaller population. In sum, we are nevertheless not under the impression that the lower representation of managers in our sample has affected the quality or characteristics of responses we received.

We were not able to fulfill the targets for our size of network KPI. To cover at least 6 sectors in each EU member state, this would have required us to conduct at least 168 interviews. Given the emerging nature of our sample population, we had to make significant investments into identifying potential participants and then acquiring them as study participants. This necessary approach however did not allow us to achieve the minimum network coverage through interviews alone. As seen from the detailed country statistics displayed in Appendix 1, we were able to interview practitioners from at least 6 sectors in only three larger EU member states (Germany, Spain, UK).

While speaking to practitioners from countries with a lower representation, we also found that the general assumption underlying our size of network KPI might have been faulty. Reaching at least 6 sectors using data science in each country implies an already relatively broad application of data science in the different economic sectors of a country. Given these conditions, we found that we had overestimated the number of data scientists that we would be able to identify and interview in countries where this was not the case. To respond to this problem and tap more into regional communities, we

[51] We defined our sectoral coverage based on Eurostat's metadata for the statistical classification of economic activities in the European Community (Rev. 2 (2008)): http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC

[52] The according Eurostat business statistics are accessible here: http://ec.europa.eu/eurostat/statistics-explained/images/4/4e/Key_indicators%2C_EU-27%2C_2010.png

adapted our sample acquisition strategy to a snowball sampling approach. This allowed us to explore and reflect better on country and regional clusters, instead of a generic sectoral coverage.

While we notice that the lower number of conducted interviews formally impacts the achieved sample coverage, we did not find that it had reduced the quality of gathered evidence. After conducting the majority of interviews we noted that additional interviews did mostly not return new qualitative arguments. Hence, the marginal utility of conducting more interviews with the same questionnaire would have likely decreased further. Accordingly, we decided to put additional efforts into the collection of quantitative data through surveys. While it appeared that the depth of qualitative arguments collected through interviews had reached a saturation point after ca. 100 interviews, surveys helped us to substantially expand the geographical coverage of our sample.

4. Results and analysis

This section reports on the results of primary and secondary data collections conducted from February 2015 to March 2016. We start with a comprehensive summary of quantitative and qualitative data collected through surveys and interviews. We then describe the results of four adjunct focus groups and interviews on organisational training approaches. Lastly, this section will report on the main findings of the skills dashboard and survey of data science training courses.

4.1 Demand analysis survey and interviews

4.1.1 Sample size and coverage

Geographical coverage

The total sample for the main study is based on 692 responses, consisting of 584 responses to our survey as well as 108 interviews. The total split of responses by 32 countries and data collection mode is displayed in table 8. As EDSA is an EU-funded project, it should be highlighted that the sample covered all current EU member states as well as one response each from Serbia, Iceland, Norway, and Switzerland. Since our data collection and this report is however focused on EU member states, we have excluded these four countries from further analysis.

Table 8: *Sample split by country and mode (descending by total count)*

Rank (according to number of responses)	Country	Interviews	Surveys	Total
1	United Kingdom	12	43	55
2	Slovenia	3	40	43
3	Germany	9	29	38
4	France	5	31	36
5	Netherlands	3	31	34
6	Denmark	2	29	31
7	Spain	9	21	30
8	Sweden	5	24	29
9	Belgium	3	21	24
10	Poland	3	21	24
11	Bulgaria	9	14	23
12	Ireland	3	20	23
13	Austria	2	20	22
14	Italy	3	19	22
15	Malta	7	15	22

16	Greece	3	18	21
17	Romania	2	19	21
18	Latvia	2	18	20
19	Portugal	2	18	20
20	Republic of Cyprus	2	17	19
21	Croatia	2	16	18
22	Finland	2	16	18
23	Luxembourg	2	16	18
24	Slovakia	2	16	18
25	Lithuania	2	15	17
26	Estonia	3	11	14
27	Czech Republic	2	11	13
28	Hungary	2	11	13
- No rank applicable -	(unknown)		2	2
29	Iceland	1		1
30	Norway		1	1
31	Serbia		1	1
32	Switzerland	1		1
	Total	108	584	692

While the sample size appears rather large, given the specificity of the subject, and relative newness of data science as a business practice, there are imbalances in its distribution which make it difficult to assess specific trends for each individual country. Generally, Western and Northern Europe received a higher number of responses. The highest number of responses came from UK participants, accounting for 55 responses in total. Of this number, 43 responses were from surveys and 12 from interviews. Slovenia provided the second highest return (43 responses), which however only contains 3 interviews, thus limiting the available qualitative information in this case. The remaining top 10 of participating countries consists almost exclusively of Western and Northern European countries. Poland ranks 10th with 24 responses (interviews: 3; surveys: 21) followed by Bulgaria (23 responses in total; surveys: 13; interviews: 9).

On the lower end of the table, a number of Eastern and Southern European countries appear. Comparatively low response yields for Hungary and the Czech Republic (13 responses each), Estonia (14 responses), and Lithuania (17 responses) make it difficult to conduct meaningful assessments on a per country basis. This is also the case for qualitative data gathered from interviews: 21 countries provided only 2-3 interview responses. Identifying and acquiring suitable interview partners was more difficult in Eastern and Southern Europe. Overall, the qualitative data is not sufficient to conduct an in depth analysis with sufficiently robust results at a country level.

As a consequence of this sparsity of data, we opted to reduce the geographic dimensions of the sample by grouping countries according to four UN-defined regions[53]. In total, this aggregation of data helped us to achieve substantially higher numbers of responses per region, thus allowing a more robust data evaluation than on a per country basis (see Table 9). While we can still see differences between the total numbers of responses from different countries, these imbalances are less problematic given the higher total number of responses for each regional category. This is particularly the case for the number of available interviews, now reaching a threshold of at least 20 available interviews per region.

After aggregation, northern European countries account for 207 responses (176 surveys; 31 interviews), out of which more than a quarter come from the UK. Southern Europe contributes 195 responses (164 surveys; 31 interviews), out of which more than a third were received from Spain or Slovenia. Western European countries account for another 172 responses (148 surveys; 24 interviews), out of which ca. two thirds of responses come from Germany, France and the Netherlands. Eastern European countries still contribute the lowest total number of responses (112 in total), thus aligning with the limitations identified above. However out of 112 responses in total, 92 survey and 20 interview responses still allow a well-grounded cross-region evaluation beyond anecdotal evidence.

Table 9: *Regional distribution after aggregating country data*

European region	**Survey**	**Interview**	**Grand Total**
Northern Europe	176	31	207
Southern Europe	164	31	195
Western Europe	148	24	172
Eastern Europe	92	20	112
Blank (region data not recorded or not applicable)	4	2	6

Organisational and sectoral coverage

Across all regions our sample covers responses from organisations with very different sizes, reaching from large corporates with several thousand employees to sole trader businesses. While the response totals displayed in table 10 reflect a widely range coverage of different organisations, it should be noted that the majority of them come from large organisations with more than 250 employees. In total, 387 responses (i.e. ca. 57%) were collected from large organisations. A similar pattern is also replicated in the regional splits. In each of the four regions, large companies represent between 52 percent (Southern Europe) and 65 percent (Western Europe). Additionally, responses from small and medium enterprises

[53] A full list of country groupings by regions is available in Appendix 4.

with 10 to 250 employees account for another 36 percent of all responses. Their regional contribution ranks between 28 percent (Western Europe) and 42 percent (Southern Europe).

Together, large companies and SMEs represent more than 90 percent of the sample in each region. While this implies an imbalance towards larger organisations, this result can also be seen as an implicit result of our ambition to cover a wide range of industrial sectors. Data scientists in micro and sole trader businesses tend to often work in the ICT and consulting/professional services sectors. However, in other sectors that are traditionally less IT-heavy sectors, data scientists often tend to be embedded in larger businesses that have at least SME size.

Table 10: *Organisational coverage across regions*

Organisation type	Eastern Europe	Northern Europe	Southern Europe	Western Europe	(blank)	Total
Large (>250 employees)	69	112	100	105	1	387
SME (10 - 250 employees)	36	75	81	45	4	241
Micro (1 - 9 employees)	5	14	7	9	1	36
Individual (self employed / sole trader business)	2	4	5	3	0	14
Blank (organisation size not specified)	0	2	2	10	0	14

Relating to this implication, our study achieved a solid sectoral coverage across 19 Eurostat-defined sectors. Data science as a domain is closely related to and emerges from the information and communication sector, thus it is unsurprising that the highest total share of responses (19%) comes from this sector. Apart from Northern Europe, where responses from the education sector (e.g. from university research groups) were strongest, the ICT sector provided most responses in all regions as well. The second highest number of responses both in total and across all regions came from the "Professional, Scientific and Technical Activities" sector, accounting for ca. 17 percent of all responses.

Apart from this general sectoral profile, regional sectoral focusses seem to vary to some extent. A number of respondents from Northern European countries come from a background in the "Electricity, Gas, Steam and Air Conditioning Supply"-sector (21 responses) as well as the "Financial and Insurance Activities"-sector (18 responses). In both sectors, this region also contributed the majority of responses. Participants from Southern Europe instead came frequently from the education sector (18 responses), the "Public Administration and Defence; Compulsory Social Security" sector (16 responses) and the "Administrative and Support Service Activities"-sector (17 responses). In the latter two cases, Southern European participants also contributed more than a third of all responses. Western European participants instead came frequently from the "Human Health and Social Work Activities"-sector (19 responses) as well as the manufacturing sector (16 responses), providing 50 and 42 percent of responses respectively. Given the overall lower response rates from Eastern European countries, their

sectoral contribution remains low for most other sectors, only in the "Arts, Entertainment and Recreation"-sector does Eastern Europe make a more noticeable contribution with 38 percent of all responses.

Lastly, it should be noted that the coverage for some sectors at the lower end of the table remained particularly low. For the "Real Estate Activities"-, "Water Supply; Sewerage, Waste Management and Remediation Activities"-, "Agriculture, Forestry and Fishing"-, and "Mining And Quarrying"-sector, we received less than ten responses across all regions.

Table 11: *Sectoral coverage of sample*

Eurostat sector title	**Eastern Europe**	**Northern Europe**	**Southern Europe**	**Western Europe**	**Not specified**	**Total**
Information And Communication	20	27	40	42	1	130
Professional, Scientific And Technical Activities	17	28	34	36	0	115
Education	3	29	18	5	1	56
Public Administration And Defence; Compulsory Social Security	7	13	16	9	1	46
Administrative And Support Service Activities	9	10	17	5	1	42
Electricity, Gas, Steam And Air Conditioning Supply	8	21	3	5	2	39
Financial And Insurance Activities	5	18	4	11	0	38
Human Health And Social Work Activities	3	6	10	19	0	38
Manufacturing	7	9	6	16	0	38
Wholesale And Retail Trade; Repair Of Motor Vehicles And Motorcycles	8	5	9	10	0	32
Arts, Entertainment And Recreation	8	3	7	3	-	21
Other Service Activities	4	6	9	-	-	19
Transportation And Storage	4	8	3	4	-	19
Accommodation And Food Service Activities	2	6	7	3	-	18
Construction	2	5	6	1	-	14

Real Estate Activities	3	3	2	1	-	9
Water Supply; Sewerage, Waste Management And Remediation Activities	2	2	3	-	-	7
Agriculture, Forestry And Fishing	-	6	-	-	-	6
Mining And Quarrying	-	1	1	1	-	3
(blank)	-	1	-	1	-	2

Note: For each region, the sectors with the three highest response numbers are coloured in grey; the sectors with the three lowest response numbers are coloured in black.

Split of role profiles

With regards to the split of the role profile, we initially targeted a division of 60 to 40 percent between data scientists and their managers. In our total results, this distribution however inverted with 58 percent of participants being data scientists and only 42 percent being managers of such teams.

As can be seen from table 12, data scientists provided the majority of responses across almost all regions and organisation sizes. Managers only account for the majority of responses in the cases of Eastern European SMEs and Micro-sized companies as well as Western European large companies.

Table 12: *Split of roles by region and organisation size*

(DS = data scientist; M = manager; N=692)

	Eastern Europe		**Northern Europe**		**Southern Europe**		**Western Europe**		**(blank - no country specified)**	
	DS	**M**	**DS**	**M**	**DS**	**M**	**DS**	**M**	**DS**	**M**
Large	43	26	68	44	61	39	48	57	1	
SME	13	23	49	26	47	34	24	21	3	1
Micro	2	3	8	6	5	2	6	3	1	
Individual	2		4		3	2	1	2		
(blank - no organisation size specified)			2	2	2		9	1		
Total	60	52	131	76	118	77	88	84	5	1

In this context, table 13 shows that managerial responses remained only in two sectors above the initially targeted 60 percent threshold. In three further sectors, managers gave at least 50 percent of responses. Except for the agriculture sector, response numbers in these three sectors ("Wholesale And Retail Trade; Repair Of Motor Vehicles And Motorcycles"; "Education"; "Public Administration And Defence; Compulsory Social Security") were also substantially higher than in the previous two sectors. While the latter only provided a maximum of 7 participants, the former sectors returned between 32 to 56 responses.

This however does not reflect a general trend. From the two sectors with the most responses, both reflect very similar patterns with ca. 60 percent of responses coming from data scientists and only around 40 percent coming from managers. ICT companies and organisations in the professional services, science, and technology domain alone account for more than 35 percent of responses. This distribution of responses might also reflect on the more general split of data scientists versus managers in industries that are supposedly more advanced in the data science domain. Naturally, with more data scientists than managers in companies, the chances of capturing the former in a random sample are also higher than for the latter.

Table 13: *Split of roles by industry sectors (listed by share of managers responses in total sector responses)*

Sector	**Data Scientists**	**Managers**	**Total number of responses for sector**
Mining And Quarrying	0	3	3
Water Supply; Sewerage, Waste Management And Remediation Activities	2	5	7
Wholesale And Retail Trade; Repair Of Motor Vehicles And Motorcycles	13	19	32
Education	24	32	56
Public Administration And Defence; Compulsory Social Security	21	25	46
Agriculture, Forestry And Fishing	3	3	6
(Blank - no response)	1	1	2
Administrative And Support Service Activities	22	20	42
Human Health And Social Work Activities	21	17	38
Construction	8	6	14
Financial And Insurance Activities	22	16	38
Professional, Scientific And Technical Activities	68	47	115
Accommodation And Food Service Activities	11	7	18

Information And Communication	80	50	130
Arts, Entertainment And Recreation	13	8	21
Transportation And Storage	13	6	19
Manufacturing	28	10	38
Electricity, Gas, Steam And Air Conditioning Supply	30	9	39
Real Estate Activities	7	2	9
Other Service Activities	15	4	19
Total	402	290	692

4.1.2 Survey results and analysis

In this section, we present and provide an initial analysis of the results of our quantitative survey. Data has been collected mainly through an online survey, but also in a structured process during telephone interviews. The primary goal of this quantitative analysis is to explore currently existing and needed skills for data scientists. Accordingly, our questions mainly focused on the self-perceived strengths of and demand for data science skills among the study participants.

Skills needed for a data scientist

We first asked all respondents how important a range of skills in different technical, analytical, and business domains are for data scientists, specifically whether they were essential, desirable, or not required. Generally, as reflected in figure 5, all eight skills domains seem to be in strong demand. Even the skills domain with the lowest perceived demand ("Open Source Tools and Concepts"), was seen as essential or desirable by almost four fifths of respondents.

Unsurprisingly, 83 percent of respondents said that data collection and analysis skills are essential; a further 16 percent ranked them as desirable. The fact that only 1 percent of respondents saw these skills as not needed, seems to reflect on the fact that the ability to retrieve and analyse data is a genuine task for any data scientist[54].

The second highest rated skill set is in "interpretation and visualisation": 59 percent saw this as an essential skill, another 38 percent think it is a desirable one. With only three percent not viewing data interpretation and visualisation skills as required, this seems to reflect on a widespread expectation for data scientists to not just analyse data or research hidden patterns in organisational operations (e.g.

[54] http://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html; http://www.ibmbigdatahub.com/blog/do-data-scientists-need-data-management; http://www.computing.co.uk/ctg/news/2456542/ubm-data-scientist-a-background-in-data-analytics-not-data-management-is-key-to-being-a-data-scientist

from sales and marketing data). Beyond this, data scientists should also have an ability to interpret and visualise data in concise and meaningful ways. From a functional perspective, such skills relating to "storytelling" reflect on many companies' demand for data scientists to communicate relevant insights to organisational decision-makers to influence and guide their decision making. From a job profile perspective, this is turning data scientists into agents to progress data-driven business from within the organisation.

On the list of essential skills, expertise in Maths and Statistics (54 percent), Big Data (47 percent), as well as Machine Learning and Prediction (46 percent) further complete the list. All three domains receive very high ratings of around 90 percent when adding the "desirable" votes. Nonetheless, it appears remarkable that skills in machine learning and prediction are not more widely regarded as essential, given their prominence in the discussions about the benefits of data science. This might have two interrelated reasons. Firstly, it can be seen as a reflection of the sample composition which included a variety of professionals working in the data science domain. In practice, many of the respondents might perform roles more closely related to data management or engineering, thus rendering advanced predictive techniques less required from their point of view. Additionally, it might also be a reflection of European organisations' perceived reality in implementing data science. Rather than fully using advanced predictive analytics and machine learning, many organisations are still in the process of embedding data science practices into their daily operations. Practically, they would thus rely on slightly more heuristic abilities, e.g. in maths and statistics or data collection and analysis, to understand and expand their data base.

The remaining three skills domains ("Business Intelligence and Domain Expertise"; "Advanced Programming and Computing"; "Open Source Tools and Concepts") were rated as essential by no more than one third of the respondents; nevertheless, all three domains received high "desirable" rating between 53 and 60 percent. Generally, for all three domains, lower rating might be somewhat unsurprising since they represent expertise areas that are often already covered by relatively more classical roles, particularly business intelligence analysts as well as data engineers and architects. Drawing from the results of the interviews, there may be some additional reasons for the lower approval ratings in the three domains. For example, business domain knowledge is frequently seen as a complementary skill-set which is often acquired on the job. While data scientists, therefore, do not need in-depth business expertise initially, they should still have an affinity for business as well as a strong curiosity to comprehend and analyse business related problems.

Perhaps the most striking result of this question is the relatively low rating for both advanced computing and open source skills, given how intensely modern data science seems to rely on advanced, open-source based computing. However, data scientists often simply need solid computing and programming capabilities to build upon, that are good enough to solve practical problems when they appear. With regard to advanced computing and programming skills, this makes those skills non-essential, but widely desirable, as reflected by 60 percent of respondents.

The almost even share of respondents seeing expertise in open source tools as either essential or not required might reflect on a wider divide between data scientists. Users working with proprietary systems might perceive such knowledge as redundant, while users of open source tools could perceive this knowledge as either essential or desirable, as long as it is good enough to understand and apply fundamental principles on a day to day basis.

Responses to the question: How would you rate the following skills for a data scientist?
(N = 651)

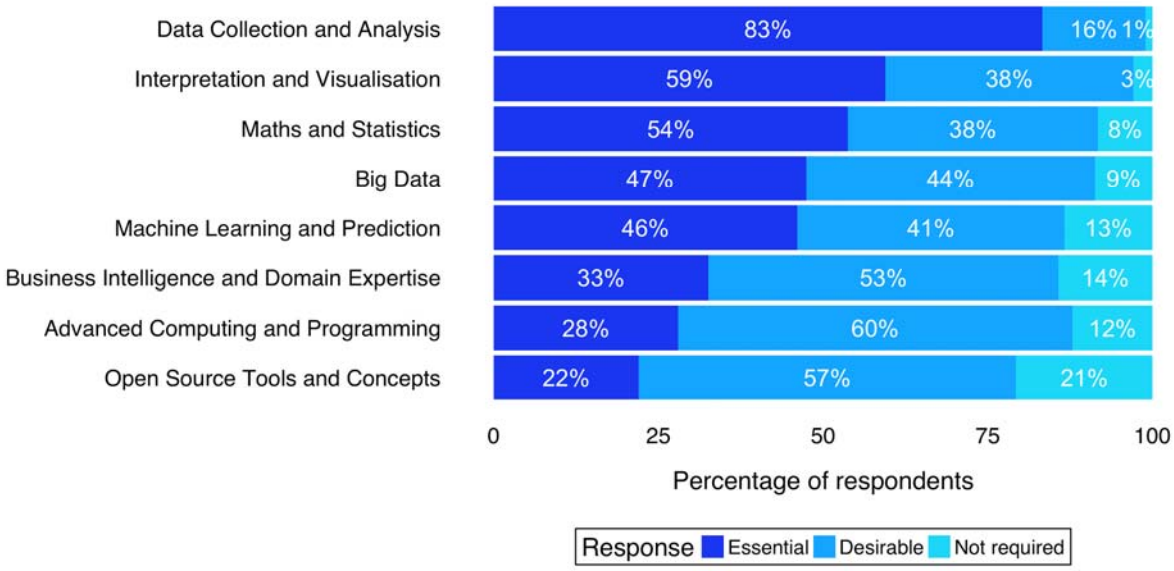


Figure 5: *Skills that a data scientist should have*

Assessment of existing skills

We asked respondents in data science roles to assess their own skills on a scale from 1 (very poor) to 5 (very good) (see figure 6). To receive a supplementary perspective, we also asked managers to rate their team's skills (see figure 7).

Beginning with the self-assessment of data science practitioners, we can generally observe a relatively strong rating for all skills areas. This group is most confident about their interpretation and visualisation skills. More than three quarters of the 355 respondents for this question said that they rate their capacities in interpretation and visualisation of data as either very good (score: 5) or good (score: 4). Only 8 percent rate their own skills as poor or very poor.

A similar pattern exists for skills in machine learning and prediction: Here, roughly two thirds of respondents rate their own abilities as very good or good. Only 9 percent say they are very poor or poor. However, for this skillset, more than a quarter of surveyed data scientists ranked their expertise as intermediate. Furthermore, almost half the respondents ranked their business intelligence and domain expertise as either good or very good. A fifth of data scientists nevertheless also thought they were poor or very poor; again about a third of respondents said their skills were "ok".

For the remaining skills in "Big Data", "Open Source Tools and Concepts", "Data Collection and Analysis", "Maths and Statistics" and "Advanced Computing and Programming" it is noteworthy that their distribution is relatively similar. For each of them, 35 to 39 percent of respondents ranked their abilities as good or very good, while poor or very poor self-assessments range from 29 to 33 percent. Differences between intermediate ratings are slightly higher, floating between 27 and 37 percent for skills in "Data Collection and Analysis" and "Open Source Tools and Concepts".

While survey participants were least self-assured in the latter domain, data collection and analysis skills seem to be most controversial among survey participants. With only 27 percent of data scientists ranking their skills as intermediate, 39 percent thought of them as either good or very good. At the same time, a third of respondents also said their skills were poor or very poor. Hence, while being the most needed and, arguably, genuine capacity domain for data scientists, it appears that two camps exist. One which claims to be specifically competent and another one which is rather reluctant about their own expertise.

Comparing the results further with the previous questions, it also appears that skills in interpretation and visualization are not only highly needed for data scientists. Data scientists also seem to be largely confident about these capacities. Interestingly, however, respondents of this question also seemed to rate their skills higher in domains such as advanced computing, machine learning, and business intelligence. The same skills had just been ranked as less needed in the previous questions.

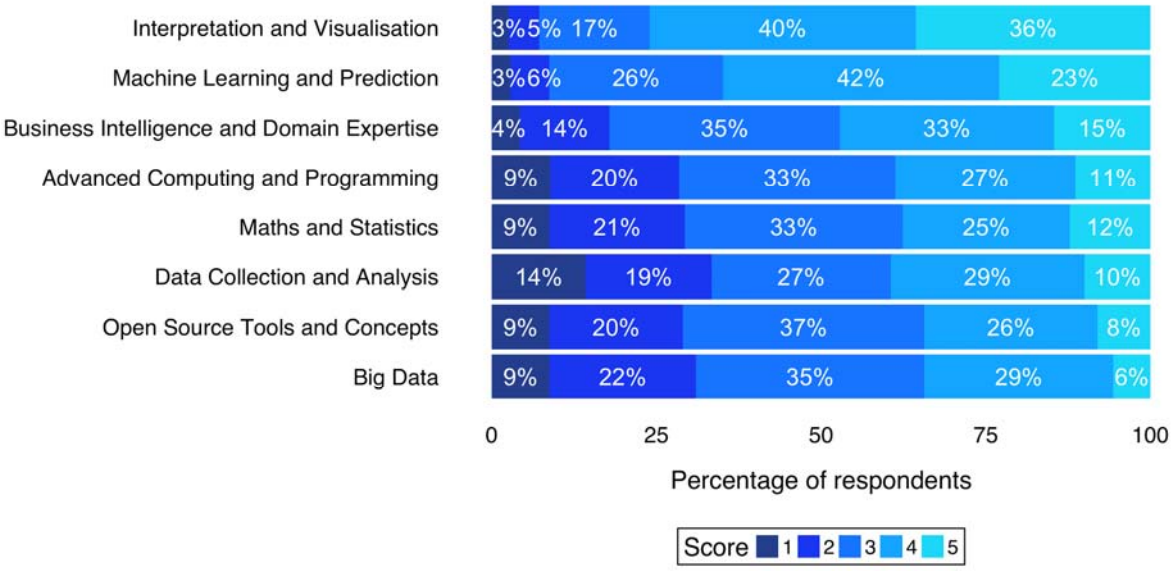


Figure 6: *Self-assessment of own skills by data scientists (N = 355)*

Turning to the manager's' assessment, the first noteworthy result is that they seem to be slightly more optimistic about their teams' skills than their subordinates themselves (see figure 7). While the results for very good and good skills are roughly the same, the lowest positive skill rating is 43 percent, given for skills in "Open Source Tools and Concepts", compared to 38 percent in the data scientists' self-assessment. Additionally, it also appears that, to some extent, managers rate their teams' strengths differently from how data scientists' rate their own strengths.

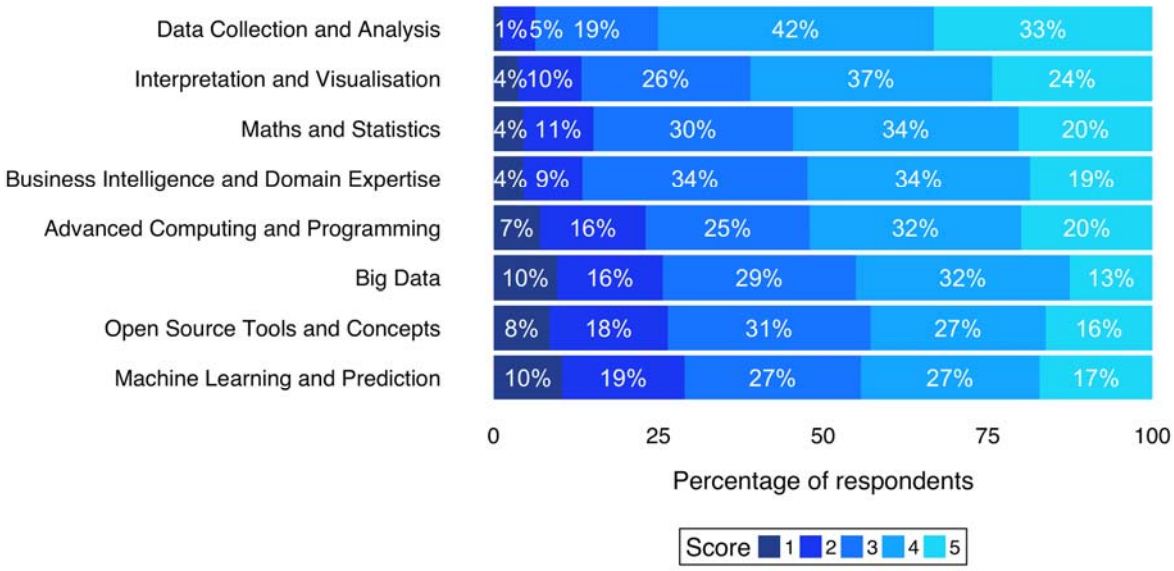


Figure 7: *Assessment of team's skills by manager (N = 278)*

Firstly, managers seem to be most confident about their team's ability in "Data Collection and Analysis". Three quarters of managerial respondents ranked these skills as good or very good among their teams. Only six percent said they were poor or very poor; 19 percent ranked them as intermediate. Interestingly, this is a very different result from the previously evaluated self-assessment of data scientists on this skillset. On a precautionary note, it needs to be highlighted that the two sub-samples mostly did not come from the same organisations. This means that there are only a few cases surveyed in which both data scientists and managers come from one organisation. Hence, these numbers cannot be understood as evidence for varying perceptions of skills within one organisation. Instead, they could be seen as indicators for differing perceptions of existing skills and current development needs. According to this interpretation, managers seem to think their teams are largely well-equipped for the data collection and analysis tasks they face.

More in line with the self-assessment of data scientists is the managers' high ranking of data interpretation and visualisation skills. Placed second in both ranks, 61 percent of managers rated these skills as good or very good among their teams; an additional quarter thought their teams were least familiar and competent users. Nevertheless, 14 percent of managerial participants also thought their teams were poor or even very poor in interpreting or visualising data. This share is thus almost twice as high as from data scientists' own perception. In sum, this might reflect high managerial expectations on data scientists being progressive intra-organisational communicators for data-driven insights, a demand, which might not be fully met by the current skills of data science professionals.

52 to 54 percent of managers had good or very good perceptions about their team's skills in "Maths and Statistics", "Business Intelligence and Domain Expertise", as well as "Advanced Computing and Programming". Strikingly, these three domains also occupied the same ranks in the self-assessment. Even though presented in a different order, this seems to indicate an overlapping perception between

managers and data scientists. Looking on the more sceptical ratings, 14 and 15 percent of managers ranked their team's expertise in "Business Intelligence and Domain Expertise" as well as "Maths and Statistics" as poor or very poor. While this completes the impression of rather confident managers in these two expertise areas, almost a quarter of managers also said their team's "Advanced Computing and Programming" skills were poor or very poor. These skills had been ranked as less needed in the previous section, this perception however also seems to overlap with the self-assessment of data scientists.

Lastly, managers also were most critical about their team's skills in "Big Data", "Open Source Tools and Concepts" as well as "Machine Learning and Prediction". A relatively low share between 43 to 45 percent ranked these as good or very good and 27 to 31 percent gave intermediate ratings. However, in each case, more than a quarter of managers also said these expertise areas were poorly met by their teams. This might indicate general development needs for teams to apply big data technologies and make proficient use of open source technologies, including the production of customised solutions for different organisations.

Most remarkable is the low evaluation of machine learning and prediction skills. While data scientists had been rather confident about their own skills, ranking them second among all domains, managers appear to be less confident, ranking this as the least developed expertise area. Certainly, the largest share of managers still have positive perceptions of these skills among their teams, but 29 percent think that skills are poor or very poor. Comparing percentages with those of data scientists, the managers' share of negative assessments is nevertheless 20 percent higher, while the share of positive evaluations falls 21 percent behind. "Machine Learning and Prediction" skills were seen as moderately needed, however, the large differences between data scientists' and managers' perceptions might still suggest different expectations between the two groups. Particularly managers might expect their teams to perform stronger in this domain in order to tackle more advanced, predictive data analytics tasks.

Technologies, tools, and languages for training

Data science is deeply dependent to the emerging provision of digital technologies, tools and languages, understanding how they can be used is therefore vital for data scientists. Hence, we asked survey participants which technologies, tools and languages should be included in training for data scientists. In addition to 16 predefined answers, which were selected based on potential curriculum contents[55], users could also enter additional options in a free text field.

The very high number of 180 unique answers resulting from this question underlines the impression that data science is a fragmented profession which relies on the provision of a variety of different tools and technologies. We grouped together these additional technologies which had been mentioned at least twice by users; Appendix III.b details these in full. Figure 8 presents instead only those categories which had been mentioned by at least 15 of the 575 respondents for this question.

[55] Predefined answer categories were AWS, Spark, Hadoop / MapReduce, MongoDB, Open Refine, QMiner, Apache Flink, Apache Storm, ProM or Disco, NoSQL / Cassandra, R, Python, Javascript / JQuery, D3 / nvD3, Java, and z-scores.

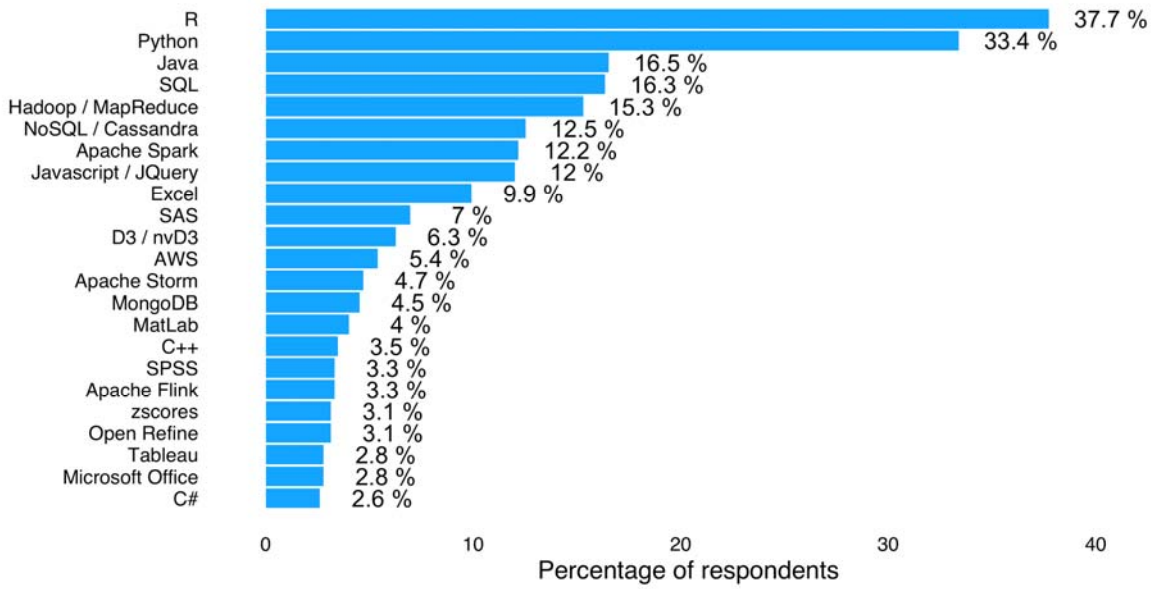


Figure 8: *Technologies, tools and languages to be included in data science training*

While knowledge in advanced computing and programming, as well as open source tools and concepts, was relatively low ranked among respondents, more than a third want to see general purpose programming languages such as R and Python included in data science training. Java should be covered according to an additional 16.5 percent of respondents. Almost the same share of participants would also like to see SQL included. The inclusion of this special-purpose programming language for data held in relational databases seems to reflect on the general importance of relational database management for the work of data scientists. Other programming languages are less in demand. 12 percent of respondents thought that Javascript and its respective jQuery library were important enough to be included in data science training. C++ and C# were only requested by 3.5 and 2.6 percent respectively. MatLab is the only proprietary programming language in this list, mentioned by 4 percent of users.

On a relative scale, numerous respondents also valued open-source frameworks for the distributed storage and distributed processing of very large data sets on computer clusters as very important. Specifically, Hadoop and MapReduce stand out in this context with 15 percent of respondents stating that these tools should be included in data science training. 12 percent also thought that Apache Spark should be covered by such training. Less than 5 percent of respondents also thought that Apache Storm and Apache Flink need to be taught as part of a good training course. Into this picture also fit the 5 percent of participants who want to see the distributed, cloud-based computing services of Amazon Web Services covered by data science training.

Beyond this, non-relational, distributed NoSQL-database management systems such as Cassandra or MongoDB are seen as relatively important. However, while Cassandra was requested by more than 12 percent of respondents, only about 4 percent required the MongoDB. While data scientists should therefore have a chance to be trained on data management solutions, around 3 percent of respondents additionally think that the data cleaning tool OpenRefine should supplement good training.

Compared to the perceived importance of data interpretation and visualisation skills for data scientists, visualisation technologies are relatively low in demand for training. Only about 6 percent require the Javascript-based data visualisation libraries D3 and nvD3 to be taught as part of training. Around 3 percent of users also consider Tableau's visualisation solutions worth including.

Some users also want to see general database and statistics software systems such as Excel and SPSS covered by data science training. In the case of Excel, a tenth of respondents required this as part of data science training; only 3 percent required SPSS. 7 percent of users also wanted training with the advanced data analytics software of SAS. Furthermore, an additional roughly 3 percent required instruction on Microsoft Office.

In summary, this snapshot of the demand for different data science training components seems to reflect the great importance which data science professionals currently put on open source, highly flexible and customisable analytics solutions. The required contents range from general programming languages, to database management systems, distributed computing frameworks, and visualisation solutions. While proprietary, often more traditional solutions such as Excel, SPSS and Tableau are important to some users, their relevance ranks substantially below the previously mentioned contents.

In our interview conversations, several data scientists and managers confirmed that open-source tools are crucial for their work. Their flexibility and openness allows users to exchange and adapt components, building analytics systems that are truly customised to users' needs. From a training point of view, three phone survey respondents also explained that courses offered by a generalist academy such as the EDSA should not focus primarily on specific tools. Rather, giving course participants an overview and enabling them to conduct product comparisons, would create added value. While tool specific training can easily be acquired from technology providers, the EDSA should center on building users' expertise on how to use different tools along the data value chain. This knowledge is conceptual and thus very useful as a guiding framework for users to more easily keep up to speed with new tools and technologies.

Training methods

Particularly relevant for the EDSA's curriculum development are the respondents' preferred training methods. Online course providers such as Coursera[56] or Udacity[57] have attracted much attention over the last few years, however our results suggest that other training modes deserve more consideration. A generally striking finding is that, when combining "essential" and "desirable" votes, almost all training modes are in demand by at least 80 percent of respondents.

Two aspects deviate from this pattern. Least requested is training which is translated from English into other languages.[58] 54 percent of all participants consider training which has been translated from English as not required, only 18 percent think it is essential. After removing responses from participants

[56] https://www.coursera.org/
[57] https://www.udacity.com/
[58] We also picked up a similar pattern in our interviews, discussed in section 4.1.3. Furthermore, we discuss the potentials for future work related to this in section 6.2.3.

in the UK and the Republic of Ireland, the shares barely change with 56 percent saying that non-English training is not required and only 17 percent stating it is essential.

A likely reflection of the popularity of open source technologies in the previous section is the finding that training for non-open, non-free software is the area with the second lowest demand. 42 percent of respondents think it is irrelevant. At the same time however, 13 percent think that training with proprietary tools is essential.

As can be seen from figure 9 below, online-based, asynchronous training methods such as eLearning and Webinars received the highest positive ratings of around 90 percent when counting "essential" and "desirable" votes together. This sum however, is inflated by almost two thirds of "desirable" votes, indicating that other methods might be more effective for skills progression.

In our supplementary conversations with data-science team managers and learning professionals, we found that online learning methods are indeed popular because they can be consumed on demand in asynchronous, flexible ways. Yet, in-person training methods were considered more effective when it comes to progressing skills. Therefore, it seems less surprising that both eLearning and Webinars are considered as essential by only 35 and 27 percent of respondents respectively.

Instead, face-to-face training and coaching were rated as essential by 47 and 42 percent; an additional 42 and 47 percent think they are desirable. While such training requires more coordination and time investment, our qualitative interviews confirmed that these in-person courses were considered to be particularly effective when they are combined with real-world assignments and tasks based on company or sectoral data. Accordingly, internal assignments also rank high among participants. 45 percent think such assignments are essential, another 43 percent think they are desirable.

High ratings for sector specific content and assessed trainings complement this picture. 38 percent of respondents think these are essential. Sector specific trainings are seen as desirable by 48 percent of respondents. Additionally, every second participant views assessed trainings as desirable. In summary, participants seem to prefer scheduled, in-person and hands-on training, which integrates industry- or company-specific data and includes assessments to compare trainees' achievements. To complement these characteristics, an accreditation of courses is desired by 57 percent, with almost a quarter of participants believing it is essential. A fifth of respondents however said that accreditation is unnecessary.

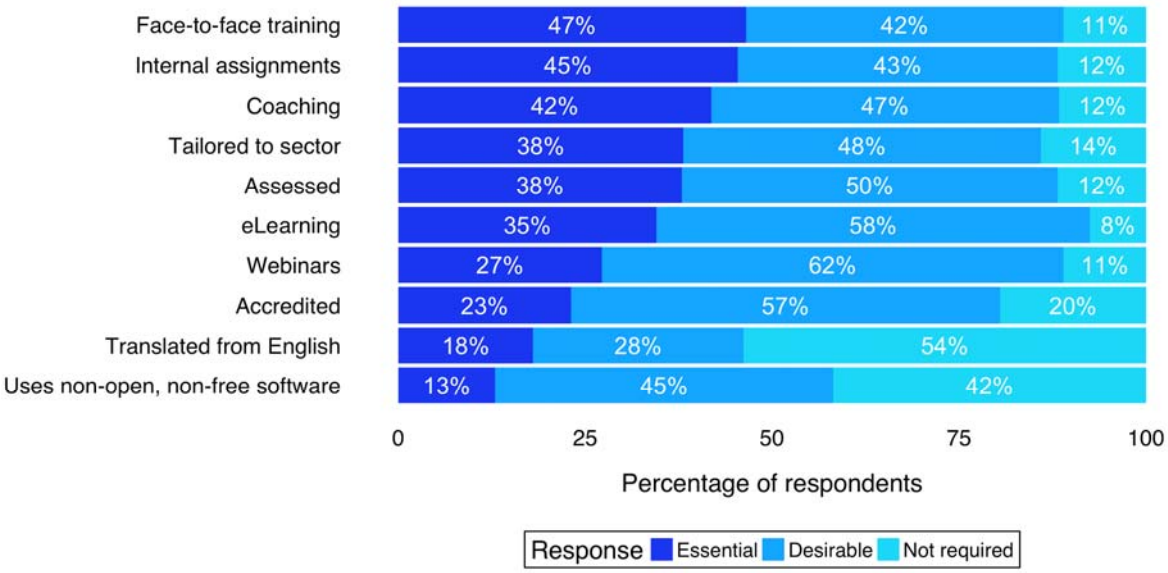


Figure 9: *Preferred training methods of data science professionals*

Additional skills

Finally, we asked survey respondents which additional skills data scientists should have. In total, 190 gave free-text answers. Their responses were categorised and aggregated based on a grounded theory approach into 38 different categories. The most frequently mentioned categories were:

- **Communication and presentation skills**. These were mentioned in 38 responses (20% of all answers). Answers typically referred to a need for data scientists who are able to present and communicate their analyses and findings in an easily accessible manner to a variety of different audiences. In particular, data scientists should be able to provide concise results to business decision makers and non-data-science teams (e.g. sales and marketing teams). This again seems to reflect on the expectation of data scientists not just serving analytical functions, but being catalysts of data-driven business strategies and operations.
- **Industry and business domain knowledge**. This expertise area was mentioned in 27 responses (14% of all responses). Answers typically referred to data scientists needing a solid understanding of their business and wider industry environment (e.g. in health care, bio and life sciences, manufacturing).
- **Teamwork**. This characteristic was mentioned in 19 responses (10% of all responses). Answers relating to teamwork usually mentioned a need for data scientists who are easily able to collaborate with other team members, but also colleagues from other departments. Data science tasks are often complex and are completed by multiple individuals, thus necessitating social and collaborative skills. Particularly the cross-organisational service function, which data-science teams serve in many organisations, require an ability to adapt to different workflows and -cultures.
- **Data management**. Such skills were mentioned in 17 responses (9% of all responses). Answers typically referred to increased or improved general data management skills being in demand. Data warehousing, (meta-)data architecture and data cleaning are required from data science professionals.

- **Social skills**: Closely relating to teamwork, these skills were mentioned in 15 responses (8% of all responses). Answers regularly referred to data scientists needing good interpersonal skills, not just to work with colleagues and other teams, but also to create and drive desired impact across the organisation.

Other frequently mentioned skills related to knowledge in database management (mentioned in 13 answers) and sound analytical capacities (mentioned in 12 answers). All other categories were mentioned less than 10 times.[59]

In total, this indicates that soft skills are needed in addition to technical and analytical skills, which dominate most current trainings. Instead, we argue that soft skills deserve more coverage in data-science trainings. As mentioned previously, data scientists are often hired with high expectations regarding their abilities to transform business practices. This implies a (perhaps limited) leadership function which cannot be fulfilled without particular social and interpersonal skills. As a result, these skills, in combination with effective communication and presentation skills are strongly in demand. As we will later explore in our interview findings, these skills separate good analysts from the transformative professionals which many businesses seek at the moment.

4.1.3 Interview results and analysis

To provide the main narrative for this demand analysis report, we have interviewed 108 data-science professionals on their ideas and approach to data-science training. In the following section, we analyse the results from our discussions with these industry-leading experts from across Europe.

Impact of data science

Data science has a profound impact on many organisations and their environment, at least this is the underlying assumption of the public discourse on this emerging space. But even where it exists, this impact might be far from uniform. We therefore first asked participants what impact data science has had on their organisations until now.

From 108 interviewees answering this question, 71 stated that data science already had a largely positive impact on their organisations. In particular, data science had enabled not simply a better analysis of business operations and production processes, but their automation. Increased productivity, driven by lower production costs and times are one impact, innovation is another one. A number of respondents said that, through deeper data analytics, their organisations were able to understand problems in business operations better, enabling them to solve problems in more structured and effective ways. In some cases, this has led to improved internal procedures, for example, when hiring new staff.

More importantly, data-science techniques have also facilitated new customer facing products; 11 study participants mentioned that data science had a direct impact on customer facing products, while 30 said that data science had been used to provide analytics of existing products and their future demand.

[59] For a full listing of all mentions, see Appendix 5

In line with these responses, the majority of respondents took a primary interest in the "data" aspects of data science. Only 13 respondents, half of which were from the science and education sectors across Europe, focused on how data science can progress education or research.

Additionally, data science also appears to remain linked to a sandboxed group of tasks in organisations. Only 17 respondents acknowledged a wider organisational change as part of the impact of data science. This group highlighted how data science had quickly expanded and spread across their organisations, leading to a transformation of traditional roles and responsibilities and necessitating improved data capabilities and data literacy across their organisations.

While this indicates a relatively deep organisational impact of data science, 35 respondents also voiced neutral opinions. However, the reasons for this are very different. Seven out of this group said that data science is their organisation's core business, thus data science has had a foundational, rather than transformative impact on their businesses. Nine respondents from across Europe and different sectors instead said that they currently don't see an impact, but expect such in the future, without specifying what this impact could be. Four other respondents claimed that their organisations had just discovered data science, with more people being hired and more teams working in the space.

New skills in demand

Data scientists are often proclaimed as a new breed of professionals, but as we have seen from the evaluation of our quantitative data in section 4.1.2, many of the skills associated with data scientists do not necessarily appear new when considered individually. We therefore asked participants whether data science really requires professionals with new skills, and if so, what kind of skills.

Remarkably, almost all respondents agreed that data science professionals do not need new skills. As an exception, only "machine learning" was recognised as a newly emerging skill by 9 respondents. One respondent also added that, strictly speaking, machine learning is not a new skill, rather it is one that has not been widely recognised as such until now.

Generally, it appears that respondents put an emphasis on data scientists as strong technical and analytical additions to the workforce. 67 interviewees said that technical skills are important; 42 respondents also require strong statistical and analytical skills. Six other respondents mentioned the importance of a variety of different scientific skills in both researching and collecting data. Three of these respondents declared that data scientists need skills which can be best developed during self-led PhD studies. As one respondent expressed it, data science tasks pose very similar challenges. Data scientists "are expected to create new insights and science", not just to replicate the techniques they learned through formalised study programmes.

The ability to blend different techniques and to apply them to practical problems is thus a primary feature of data science professionals. Accordingly, 14 participants underscored the importance of a new mindset coming with data scientists. While the different skills and expertise areas exist already, data science brings them together in an interdisciplinary, unconventional fashion. Data scientists can therefore be described as composite professionals, who need to be willing to learn, and who must have an open mindset to solve problems in new, flexible ways. As two of our respondents formulated it, a "data scientist must be open to asking experts in other areas" and "be willing to adapt quickly in order to keep up".

Arguably the most significant aspect of this mindset is the ability to solve business problems by re-employing and combining existing technologies. Because companies' data is increasingly "bigger and it requires a completely different way of interacting with it", data scientists are often asked to build new systems, algorithms and front-end services to enable added-value analyses. In most cases, these new technologies build on already existing developments, then integrate them with other tools to maximise impact.

If rolled out consistently, the resulting impacts deeply affect not just daily business operations, but also business strategies. Hence, in line with the results of our survey, 18 respondents mentioned that leading data scientists will also need a range of soft skills going beyond their technical and analytical capacities. In particular, they need to be business savvy and have strong communication and presentation skills to secure the support from senior management. To create impact across an organisation, it will not suffice for data scientists to remain analysts, rather they need to bridge the gap between technologists and decision makers within a company. A data scientist from a UK startup required an even more active role for data scientists, stating that they must have the "soft skills required of interacting with businesses and guiding the people responsible for making decisions".

In the process of making organisations more data driven, sector-specific (business) knowledge is even seen as more important than technical skills in some sectors. Respondents from the health, energy and finance sectors in particular saw specific experience as highly important. Specifically, in these sectors, experience is valuable to understand how to best clean and manage data, but more importantly to understand and ensure the reliability of data. As one respondent from the energy sector said, "[technology] helps control and check data...experience prevents problems".

Course attendance by data scientists and their teams

The next part of the interview focused on how individual data scientists and data-science teams acquire their skills. We asked data science professionals and managers of data science teams whether they have attended any courses to develop their skills.

Of the 46 data science professionals we spoke to, only seven had not attended any formal course. An overwhelming majority of 39 respondents mentioned they had attended formal training. As part of professional development, online courses seem to play an important role: 15 respondents revealed that they had attended online courses, benefiting from the flexibility which these courses offer in terms attendance and skills entry level. Coursera in this context seemed to be most popular provider with mentions raised by 9 participants. Other course providers mentioned were Udacity, Datacamp, edX, Iversity, S2DS and Udemy.

The team managers we spoke to gave similar responses. Out of 47 managerial participants, only 6 revealed that their teams had not completed any data science training within their current role. Partly this was because they were not able to find the right kind of courses or because their organisations hire externally to fill skills gaps. The teams of the remaining 41 managers had all taken part in formal training.

The focus of these courses varied. In general, organisations we spoke to are seeking to advance their knowledge in data and analytics or specific areas linked to data science, for example, in statistics. However, the courses which team members attend rarely list "data science" in their title as they tend to focus on specific domains.

Notably, 16 respondents indicated that they had touched upon all essential data science skills already during their undergraduate, graduate and PhD studies. This also reflects our conversations with managers and learning professionals. In many cases, data scientists are hired as experts with an existing skillset. This skillset may be refined and enriched with business skills for example, but businesses seem to rarely invest in developing their staff's data science skills from an early stage. One respondent highlighted that traditional programmes, such as computer engineering, mostly cover some relevant aspects of data science; however, since they are normally not designed with a focus on the latter, they also often miss some crucial perspectives, e.g. on the exploitation of large and big data.

In addition to technical and statistical training, an emerging training domain for teams lies in communication and business skills. Our interviews with managers and learning professionals revealed that soft skill training in these domains is increasingly seen as an essential adjunct to ensure that data science teams cannot only deliver evidence and insights, but progress business operations and strategies. To achieve this, teams need to be trained in the "power of persuasion" as one respondent formulated it. Accordingly, training to develop communication and presentation skills as well as business domain knowledge were frequently mentioned, specifically by managers and learning professional. These are necessary for data science professionals to be able to effectively transfer their insights to other teams (e.g. sales and marketing) and senior management staff. Crucially, data science teams normally depend on the latter to adapt business operations and strategies, thus setting their organisations on course to harness the potentials of data driven businesses.

Until now, data science managers have relied on a mix of different supply modes. 17 managers disclosed that they focused on internal training, delivered through their own organisations. 15 have also trained their teams through online courses, making also use of MOOCs. Coursera was also popular among managers, mentioned eight times. 14 managers furthermore stated that their organisations had contracted external trainers from Apache, Cloudera, ExperTeach, Integra or IBM to deliver training to their teams. Of these, six had brought trainers on-site, five sent their teams off-site and three used both approaches.

Apart from mere course labels, managers highlighted the difficulties in identifying high quality training. Knowing which courses are taught by true experts who can add value to a team's skills is very difficult. This is even true in cases where the requirements for training have been clearly defined. A trusted, neutral platform which provides such information, ideally in a standardised or at least comparable format, is lacking at the moment.

New approaches to training

As previously seen, participants indicated that data science professionals need to be continuously learning and adapting. This might be one underlying reason why innovative learning formats have pioneered the delivery of formal and informal training, especially in the data science domain. MOOCs and numerous community-led forums are arguably the most important examples for this trend. We therefore also asked data science professionals and their managers whether they had taken new training approaches to expand skills. It is worth highlighting that in our conversations, data scientists, their managers and learning professionals all highlighted the great importance of informal, self-guided learning. Some even said that the majority of skills development in data science is achieved through these means.

Against this general background it is hardly surprisingly that the majority of individual data scientists we spoke to are using alternative learning approaches to expand their skills base. Only nine declared that they have not until now, one of them saying that this was because he was unable to find anything useful "unless you want an IT degree".

Out of the remaining 38 data scientists who use alternative learning approaches, 18 used self-learning, eight were involved in peer-learning and 9 used both these techniques. Among the self-guided learners, 16 expand their skills mainly through reading, particularly data science blogs and open access journals and papers. At least one respondent however noted that while this approach may help to maintain or develop knowledge, it is less effective in acquiring skills. For example, one respondent explained that statistics is not well covered in Bulgarian education, hence she took some time to compensate for this by reading on statistics and machine learning. However, maintaining and deepening this knowledge is difficult while working in a business where the immediate application of these skills is not required. Related to this, one manager also mentioned that it was sometimes hard to keep team members on track with project work while they are reading to expand their skills. Making sure that self-trained skills are acquired in ways that are beneficial for work related purposes is a complex challenge and one of the pitfalls of self-guided learning.

Four other respondents combined reading with hands-on self-training, which mostly meant the continuous testing of new data science tools to remain up to date with current developments. However, for some practitioners, this still does not provide a strong enough practical experience. Three data scientists followed a more applied learning approach to expand their hands-on experience wherever possible. This might also involve searching for application cases outside of work, e.g. by volunteering for charities.

Managers made similar claims on the training approaches of their teams. While 32 teams had tried new training approaches, only five had not done so. Among the approaches taken, informal training dominated again. 23 managers said their teams had worked out internal approaches to train themselves, not making use of external support through coaches or trainers. In this group, daily on-the-job training proved to be most important, mentioned by eight managers. While this was mostly not undermined through a specific team structure, one team had implemented a daily rotation to ensure that all team members acquire and maintain broad skills which are executed across the team. In another case, the company was pushing for a cultural shift to make data sharing across the organisation easier. In the long term it was hoped that this will encourage extensive use of data across the organisation, eventually enriching the data literacy of employees in all departments. One large company had implemented a global forum to facilitate the dissemination of skills across geographically dispersed teams.

13 managers said that their teams had to some extent structured their informal skills sharing. In seven cases, this had been done through internal peer-to-peer learning workshops and presentations, primarily to share skills across teams. Six other teams had implemented mentoring and coaching models. While these proved to be helpful in upskilling new and junior staff, it also required commitment from senior staff to ensure that sessions followed a clear purpose. Interestingly, very few teams had implemented more formal learning approaches. Three managers said their teams relied on internal assignments and only one team used internal webinars.

Among the nine teams that used external sources to develop skills, conferences were the most important approach to progressing data science skills. Six managers said they are regularly sending their team

members to conferences to help them to remain up-to-date with current industry developments, share skills, present their work, and network with peers.

As mentioned in the previous section, some organisations have reverted to hiring as their approach to skills acquisition. Six managers noted that profiled hiring was an essential guarantee to ensure the right skills level. One manager of a small business pointed out that his company's approach was to be ruthless in recruiting and retaining only those people with the data skills required. In his previous role he was involved in internal training, but sees detrimental risks with this approach. Most importantly, he felt companies should not train people in areas such as statistics if they do not have a background in this area already. Nevertheless, other more applied skills such as the use of specific tools, can be delivered through internal training.

Finding training

The size of the opportunity has led to a real mushrooming of data science training courses over recent years. As a result, the market for data-science training is crowded and sometimes confusing. Accordingly, we asked data scientists whether they had any difficulty in finding training.

Out of 42 respondents, 26 declared that they had indeed faced problems. 19 gave a wide range of reasons for this, which can be structured roughly into three groups. A first set of problems relates to the discovery of courses. Six respondents said that the sheer proliferation of courses turned filtering relevant ones into a very valuable asset. Generally, more than enough information is available online and can be searched with popular search engines. However, the whole process would be much easier if information was both more structured and made accessible through one access point.

A second, related challenge is to find specific course information. Often, data scientists and their teams have specific requirements which are, among others, defined by the desired course content, duration and location. 10 respondents mentioned that filtering and comparing this more detailed information was a subsequent challenge after finding courses in the first place.

A third, logical problem domain is to compare the quality of relevant training. Five respondents mentioned a lack of information that helps users to understand which courses fulfill individual and objective quality standards. As quality varies, potential trainees may want to know who teaches a course, whether instructors are acknowledged experts in their field, if training is hands-on and if a course makes effective use of the newest technologies. The practical problem is that many courses use very similar descriptions and "data science buzz words", whilst providing limited information on what participants can really expect. Objectively, three respondents also mentioned that there is a lack of comparison standards and, thus, transparent quality assessments. Particularly, respondents from private sector backgrounds said that it is easier to justify the costs of a course if its quality can be judged and is high according to a transparent measure.

Interestingly, eight of the ten respondents who indicated that they had no problem in finding training alluded to a related set of problems. More difficult than finding training is understanding what training is required for a data scientist. While there are a lot of resources to expand individual data science skills, "the difficulty is finding or understanding where the individual is in terms of his own skill-set and where he needs to develop, and then finding the resources to plug into those gaps." Generally, there is a lot of information available, but there is no one-fits-all training course. Hence, understanding individual

training demands is essential, and being able to match this with the right kind of training is what helps individuals to truly progress.

Finding skilled data science workers

For managers, identifying skilled data scientists is an essential challenge. We therefore asked managers about their view on the key problems in finding skilled data scientists. From a demand side perspective, this question allowed us to explore what characteristics managers like to find in good data scientists and how they evaluate them. This knowledge is useful for our curriculum development as it helps us to understand which are the quick wins that training should exploit.

34 out of 46 managerial participants said that there are substantial challenges in finding skilled data science workers. From a labour market perspective, seven interviewees said that the most pressing issue is simply the high demand for data scientists which is currently not met by supply. As an immediate result of this asymmetry, competition for skilled professionals is fierce, and the resulting high costs for hiring data scientists might inhibit some employers from expanding their teams. 12 participants also had the impression that the intensity of the problem depends on which country or sector employers are trying to hire within. In our discussions, managers said that finding skilled data scientists was more difficult in Eastern European countries and for sectors which do traditionally not have strong relations with the ICT sector. The latter might also apply to businesses in waste management, construction industries and tourism, all of which returned lower response yields in this study.

More generally, some managers raised awareness of the fact that data science is a new domain, which naturally limits the supply of skilled workers. However, four respondents also added that, while demand is rapidly expanding, the traditional education system is not failing to provide both the quantity and the quality of data scientists needed. Accordingly, a large group of managers mentioned that data scientists often lack one or multiple skills. Finding potential hires with a well-balanced skill set which includes, for example, solid expertise in data visualisation, statistics or database administration is difficult.

Against this background, organisational skills development is still mostly led by external hiring. Ten managers said their organisations are currently focusing on hiring to upskill their organisation's data science abilities. Candidates' learning abilities, but also cultural and team fit are essential aspects in this practice from managers' points of view. Only two respondents revealed that their teams used examinations to investigate skills of potential team members.

Eight teams instead put their focus on training staff. Three managers of this group even revealed that they had given up on searching for external applicants in order to focus on the upskilling of current team members. Five managers said that due to rising demand for data science related tasks in their companies, they have to expand their teams. Nevertheless, their focus lies on "hiring quick learners who can be trained on the job". In sum, this seems to show that the current surge in the demand for data scientists puts managers and organisations under pressure. To find well rounded data scientists who can not only perform technical and analytical tasks, but also excel at leading their organisation's data driven transformation is a unicorn hunt.

Whether or not data scientists need specific domain expertise at the point of hiring seems to be disputed among managers. While four managers said that they had problems finding data workers with good business domain expertise, others explicitly refused this idea. "The role of the data scientist is not to actually have any authoritative knowledge on any specific topic, but to be able to understand and benefit

from the authoritative knowledge of any other person." The discussions about which specific skills data scientists need in a role might also be further complicated by a lack of understanding about the domain on behalf of senior managers, as was mentioned by three managers. The key challenge is for companies to understand what they need in a data scientist, in order to find the right person. Whether this is a domain expert or a generalist is obviously an important part of this consideration.

A highly important factor in this discussion is also the lack of soft and managerial skills which many managers assign to data scientists. As stressed by 10 managers and learning experts, data-driven management has often proven difficult to implement with traditional senior management. Hence, impactful data scientists need to have distinctive social and influencing abilities which do not just allow them to "crunch numbers" but "build alliances with and influence those people who make strategy or operational decisions." According to participants, these aspects are however missing from almost all data science curriculums.

Training delivery

How the critical data science skills can be acquired best is an important question. Effective training modes lead to better skilled workers that are fit to tackle the real-life challenges of organisations across sectors. Judged by its dominance in the public discourse, online training for data scientists seems to be exceptionally popular. Conversely, the evaluation of surveys has already shown a preference for face-to-face training. This suggests that while online training is usually more accessible and can be integrated more easily with daily work responsibilities, in-person training modes might be more effective. To assess this discussion further, we asked data science professionals on their ideas and preferences for effective data science training.

Out of 100 relevant responses on the topic, 43 mentioned face-to-face training in our conversations. The remaining 57 participants referred to online courses, particularly e-learning (mentioned 16 times), webinars (10 mentions), webinars and/or e-learning (6 mentions) as well as MOOCs (3 mentions). At first sight, this seems to document a greater awareness for online training modes – nevertheless, from a qualitative perspective, this does not imply that respondents consider these as more effective.

Instead, of the 43 respondents who talked exclusively about face-to-face training, 30 expressed strong preferences for face-to-face training. As the main reason, they cited that this training mode is "definitely best for communication and business understanding". Five respondents also explicitly mentioned the potential of coaching. Learning from peers with advanced knowledge in specific data science domains can be a very effective mode to deepen existing skills and blend them with relevant business domain context. Fitting well with the later aspect, three respondents also highlighted on-the-job training as an effective mode to acquire skills.

Weighing different options of face-to-face and online training, 15 respondents who referred mainly to face-to-face training agreed that in-person training is the more effective training mode, but could be supplemented by online training to add more flexibility. One dimension of this is the use of online training for basic training, but face-to-face training for more advanced lessons. As one respondent said, "online delivery is useful in the early stages to present the problem and data, but face-to-face sessions are important to help people understand the data and move forward." Also, "online tutorials and resources can be useful but face-to-face is most effective, especially for younger students with less self-discipline." Additionally, online supplements can serve as "refreshers" with self-guided practice tasks. While these answers show a strong preference for face-to-face training, participants also noted some

foundational problems with this training mode. Among them are less flexibility (compared to online training), higher time commitments for participants, as well as location restrictions for off-site training. One respondent claimed in this context that "while webinars are most useful because they are easiest to access and gain management approval for, face-to-face training is most effective. But their location presents a barrier to attending training courses." Flexibility therefore, is an important consideration, particularly for individuals and teams that acquire skills through part-time courses accompanying regular work days. Online courses and webinars give people the freedom to adapt them to their own learning needs. Concise courses that consist of stand-alone elements can be particularly useful for teams who need to work towards tight deadlines and thus might need to interrupt learning sessions frequently.

From our conversations with participants who referred mostly to online training it is notable that 26 respondents (i.e. almost half of them) seemed to actually prefer blended learning. While arguing that online training is very important, these participants also noted that interaction with peers and trainers is a crucial element in improving the learning effect. Generally, some respondents noted that this interaction could be mediated online, e.g. by combining webinar sessions with one-to-one instructions and interactive feedback sessions in virtual classrooms. Others however preferred supplementary face-to-face sessions again, e.g. by adding problem-based learning sessions in teams to MOOC courses.

From our conversations, it appears that group tasks could also offer a solution to tailor training more to specific sectors. Similar to the previously cited disagreement among professionals on whether data scientists need to have specific domain knowledge, there is also some division on whether training should be tailored to specific sectors. Even respondents who first claimed that they cannot see added value in training tailored to their sector, went on to declare that sector specific knowledge can best be acquired from working with the people in an organisation. Others added that workplace integration of training is key. According to our interviewees, this could be achieved by following three principles. First, training on statistical concepts and theories as well as technical tools should follow a general framing; second, hands-on tasks and examples should be tailored to sectoral audiences; third, where possible, group assignments should be completed together by functional teams. Particularly this last principle seems to allude to the results of our survey and previous interview questions, where respondents indicated that teamwork and collaboration are important social skills that are mostly not covered by current data science training.

The goal of data science training should be to prepare for real-life tasks. As one participant expressed it, "practical training is key, real-world experience is what makes a data scientist effective from day one". Hence 15 respondents raised assignments as an important component for data science training. In any case, these should be strongly oriented towards a practical application, best enhanced by an applied teaching style and some labs work. Similarly, where assessments form part of training, respondents proposed practical formats such as project style assessments. On a broader reflection, assessments, e.g. implemented as exam-style tests and take home reports, were a topic in eight of our conversations, seven of them discussed assessments as an important part of training. Most importantly, participants said that assessments help to track and evaluate individual's progress regarding course content. Additionally, some respondents saw them as a means to keep motivation and attention high.

After course completion, participants have an interest in displaying their abilities in some form. Certificates are usually an important instrument to do this and thus were a topic in five conversations, however not all respondents were convinced of this approach. Some respondents said that certificates

can indeed be a valuable way to display certain skills, although others thought that a proven commercial mindset is much more important. Rather than saying "I have done a programme" respondents should show how they have worked on real-world projects. In this context, one participant also proposed a shift from certificates to badges to show evidence of honed and practiced skills rather than intellectual knowledge acquired through formal learning. On a related note, five managers and learning professionals also discussed accreditation. In line with our quantitative results, they saw accreditation of curricula as an interesting supplement, but usually not a priority. A core problem here is that the standards for accreditation are often not clear to industry practitioners, which limits the added value of accreditation as a sign for assessed course quality. This could however change if transparent standards were established that are acknowledged by both learning and industry experts.

A similarly divided picture also exists with regards to the discussion on English versus native language training. To increase the accessibility of data science training across Europe, the EDSA set off to offer more training in non-English languages. However, from our conversations with data science professionals, managers, and learning professionals, we learned that this is only considered useful in some specific cases. Five participants who favoured translations into native languages, argued that this this would make training more effective, engaging and accessible. One learning professional speculated that native languages would be more effective in non-technical training, e.g. on soft skills and teamwork. Particularly in those domains, professionals rely on a fine tuned understanding of language to facilitate interpersonal communications. Non-native speakers might however simply lack this fine grained knowledge. With regards to technical training, full translations could also be useful. For example, in addition to English webinars, the availability of scripts in users' languages would constitute a valuable addition.

29 other participants considered the discussion as rather unimportant and backed English as a general course language. The main rationale for this was that the wider data science community tends to operate in English anyway. Not only is most data science literature and conferences in English, but sometimes entire teams are adopting it as their internal business language. Therefore, any course not in English would risk being disconnected from this ecosystem. Given these considerations, respondents from the private sector also frequently added that English was simply the more cost effective course language.

4.2 Desk research on data science courses

To supplement our primary data research for the demand analysis, we also conducted an extensive survey of the current data science training supply across Europe. In particular, we focused on the provision of courses from higher education institutions and professional training suppliers.

4.2.1 Supply of training across Europe

Across 23 EU member states, we were able to identify 456 courses, roughly evenly split with about 48 percent offered by academic institutions (i.e. 221 courses) and 52 percent provided by professional

training suppliers (i.e. 235 courses). While the number of courses on offer in different countries varies substantially, this also implies that most EU member states have some training offers in the domain (see table 14). The United Kingdom (139), Germany (46), France (22), Ireland (16), and Spain (16) deliver the highest number of courses. The discrepancies in total numbers between the UK and these other high scoring countries is remarkable. UK suppliers alone offer almost three times as many courses as German ones. This suggest that the access to data training is very unequal across Europe. Public and private sector institutions in some countries, such as the UK, Germany and France, have made investments into data science training, leading to a higher availability of training.

Table 14: *Course provision of country*

Country	**Academic providers**	**Professional providers**	**Total**	**% of total**
United Kingdom	80	59	139	30.48
Multiple, Europe & Worldwide	0	82	82	17.98
Germany	13	33	46	10.09
Multiple, Europe	7	39	46	10.09
France	20	2	22	4.82
Ireland	16	0	16	3.51
Spain	15	1	16	3.51
Denmark	4	9	13	2.85
Italy	11	2	13	2.85
The Netherlands	11	1	12	2.63
Austria	7	1	8	1.75
Belgium	3	5	8	1.75
Portugal	4	1	5	1.10
Sweden	5	0	5	1.10
Greece	4	0	4	0.86
Romania	4	0	4	0.88
Czech Republic	3	0	3	0.66
Finland	3	0	3	0.66
Cyprus	2	0	2	0.44

Hungary	2	0	2	0.44
Lithuania	2	0	2	0.44
Slovakia	2	0	2	0.44
Croatia	1	0	1	0.22
Luxembourg	1	0	1	0.22
Poland	1	0	1	0.22

At the same time, five countries are missing from this list and seem to not offer data science training as specified by our search criteria. These include Malta, Latvia, Slovenia, Bulgaria, and Estonia. Particularly for the latter three, this comes as a surprise. Estonia and Bulgaria have recently gained international praise amongst others as leading countries in digital government or as startup hubs[60]. In addition, Slovenia supplied a high number of responses for our survey. This shows that industry practitioners are working in the country, nevertheless our research also suggests that at the same time education might be lagging behind. A general caveat to these results is that our research is not representative and only provides a selective snapshot based on the selection criteria, search terms and search methods we applied. Hence, while we did not identify them through our research, it is still possible that data science trainings exist in those countries. Even if this was the case, however, the fact that we were not able to discover those offers, raises questions on their discoverability and accessibility.

A further interesting finding is that the second highest number of courses comes from international providers which make their services available across the world. These account for 82 globally available professional training courses. An additional 39 professional development courses are available in multiple European countries. This suggest that professional training providers are particularly active in offering multi-country courses. Their relevance in the professional training market is also underpinned by the fact that more than half of all professional training courses are offered in at least two countries. To realise the business opportunities in data science training, professional trainers appear to rely on an international outlook.

Alternatively, most courses which are offered in only one country come from an academic background. This is, for example, the case in the UK, which currently lists 80 academic courses and 59 professional ones. Instead, Germany, Denmark and Belgium are the only countries where more professional than academic courses are available. Generally, we found that the establishment of academic courses also seems to overlap with the founding of data science institutes, particularly in the UK, but also in other European countries. The underlying motivation for this might be the exploitation of data science as a fundraising opportunity, which private sector donors are currently especially willing to invest in. Notably, some universities also list private sector companies such as SAS or IBM as funding partners. The fact that many courses, particularly those offered by universities, involve industry placements and

[60] http://www.forbes.com/sites/federicoguerrini/2016/04/14/is-sofia-bulgaria-the-real-digital-capital-of-the-new-markets/#367537597bba

internships seems to generally support the impression that industry proximity and application are increasingly important elements in academic data-science training.

4.2.2 Classification of training courses

Looking at the refined categorisation of courses (table 15), we find that Masters courses are most in supply from academic suppliers. About 46 percent of all identified courses are part of Master degree programmes. It is noteworthy that Masters courses were typically aimed at professionals with a few years of experience as well as recent graduates. While the former seems to be particularly true for France, UK courses were typically aimed at graduates. Generally, most Masters courses require high levels of technical competence to enrol. Additionally, few programmes offered foundational or pre-Masters study programmes to help candidates from more diverse (i.e. non-technical) backgrounds to fulfil these criteria. According to our assessment, none of the Masters courses are targeted at beginners.

Table 15: *Count of classified data science training offers*

	Degree classification	**Total**	**% of total**
Academic course providers	Masters	194	42.54
	Bachelors	14	3.07
	Certificate	4	0.88
	Diploma	4	0.88
	Bachelors, Masters	2	0.44
	Doctorate	2	0.44
	Executive Course	2	0.44
	Academic Expert	1	0.22
Professional training providers	Short Course	218	47.81
	Modular Course	10	2.19
	Workshop	4	0.88
	Traineeship	1	0.22

Reflecting on the interdisciplinary nature of data science, our survey features a few courses at the intersection between data science and another discipline. Apart from courses relating to different business domains, there are many courses available in geoinformatics. Additionally, some entries mark courses at the intersection of health and data science. Equally, from the opposite point of view, data science courses often offer modules on bioinformatics, computational biology or geoinformatics. Hence, the academic disciplines that seem to engage most with data science are health and life sciences, geography and business.

For undergraduates, there are very few opportunities to enroll in degree courses. Only 3 percent of the courses we identified led to bachelor degrees. While we identified and analysed only 14 degrees, we nevertheless discovered some patterns regarding their structure. Most notably, the first year of these

courses typically provides a basic foundation in mathematics and computer science, specific data science modules are only taught after this.

For both graduate and undergraduate programmes, the modules that feature within most data science courses tend to be core mathematics, core programming, data mining and machine learning. Particularly in comparison to professional development training, higher education courses are more likely to provide courses under broad topical headings (e.g. Distributed Computing, Data Mining, Computer Vision) than professional training courses, which focus more on particular platforms, technologies or programming languages. In addition, knowledge discovery and knowledge management also appear as relatively common modules.

Within training courses for professionals, 228 short and modular courses make up half of our total courses sample. As our interviews confirmed, short courses and modular training offer a much higher degree of flexibility, allowing professionals to dig deep into specific topics. Accordingly, most professional course providers offer very focused courses that go into depth about how particular tools, such as Hadoop, R, Python or SPSS, can be used by businesses. This is different from holistic higher education courses which tend to focus more on underlying theories and overviews across different toolsets. Additionally, course providers typically encourage businesses to bring their own case studies and/or data.

Another factor which makes professional training courses more flexible are their broadly varying teaching and interaction modes. Classroom-based (Instructor led, public), video live stream, online-self paced, or in-house corporate training (instructor led, private) appear only as the standard forms of delivery with an abundance of mixed-mode training being available as well. Additionally, many US-based providers offer the option to join livestreams of classroom-based courses, making them accessible to Europeans as well. Most course providers furthermore organise ad hoc training and/or customise them.

Among professional training providers, the number of different courses on offer vary widely as well. For example, the most active providers have more than 20, mostly platform-specific courses on offer. These cover for example Apache Pig, Apache Mahout; it is also worth mentioning that the courses offered seem to reflect demand by industry and therefore are not as fixed in terms of schedules as Higher Education courses. Moreover, the provision of these courses is typically quite dynamic, with new ones appearing as dictated by industry demand.

Unsurprisingly, different levels of courses are offered, ranging from beginners to advanced users and from hands-on to theoretical content. For example, there are many courses on programming and using big data/data science tools but also some programmes for managers and leaders in business to help them understand what data science can do for their organisations. These latter courses typically do not require programming experience.

As an example for training delivery, Cloudera is a recognised certification provider with standardised course content with decentralised delivery through a global network of partners. Microsoft, SAS, MapR and Hadoop HortonWorks also use this model of delivery partners to train users on their platforms. These platforms often offer a standardised training course with an industry-recognised certification track.

4.2.3 Training languages

Course delivery and the efficiency of communicating contents is crucially dependent on language choices. Additionally, the availability of non-English training is also a matter of accessibility for parts of the population which do not speak English as a native language. In our research, we however found that more than 70 percent of courses are taught in English; an additional 14 percent are taught in both English and another language (see table 16). With only about 15 percent of courses exclusively taught in languages other than English, there is a great dominance of courses taught in English. On the one hand, this distribution to some extent results simply from the large group of international professional training providers. On the other hand, however, it also reflects on industry wide preferences which, according to our interview results, do not favour the delivery of native-language training, except in some very specific domains. However, while some of the experts we spoke to suggested teaching soft and communication skills for data scientists in their native languages, this does not seem to be the current approach of academic and professional training providers. Rather, native languages courses are split across countries and frequently cover technical and statistical aspects.

Table 16: *Course languages*

Language of courses	**Total**	**% of total**
English	321	70.39
non-English	67	14.69
Both English and other	64	14.04
Data Not available	4	0.88

In summary, our survey and analysis has uncovered a rapidly emerging, but also very crowded training marketplace. In this market, a strong supply of technical training seems to be ensured. However, soft skill and communications trainings are missing. Additionally, assessments of the contents and quality of data science courses on offer deserve more attention in order to help students and organisations identify the training they need.

4.3 Job posting analysis

Our analysis of job postings was guided by a set of questions to construct an accurate picture of the demand that leads towards a context-sensitive definition for what skills a data scientist requires. We therefore asked:

- Which data science skills are demanded most?
- When and where do relevant job postings come from?
- Which skills are mentioned most often in various countries and sectors?
- Given a core skill which other skills are often required in addition?

The insights gathered through these questions feed into the design of individual courses and programme curricula for data science. They can also help defining new job roles that target individuals with the capabilities to meet those jobs' requirements.

Skills networks

We first used a visual exploration approach to obtain an overview of the data distribution. This involves looking at the trends that result from data peaks, troughs, dense regions, as well as outliers and islands of data, to identify both anomalies and regions of interest (ROIs) to study in further detail. We then used detailed quantitative analysis, particularly both visual and formal SPARQL querying, to answer more specific questions and reveal further detail.

In figures 10 and 11, we illustrate an analysis guided by the knowledge framework which has been implemented for the job postings. More specifically, figure 10 displays an analysis of a region of interest in an interactive network diagram. The diagram visualises the similarities in job postings in the United Kingdom, Germany and France, based on commonly listed skills. Particularly “python”, “data mining” and “data science” stand out as frequently mentioned skills

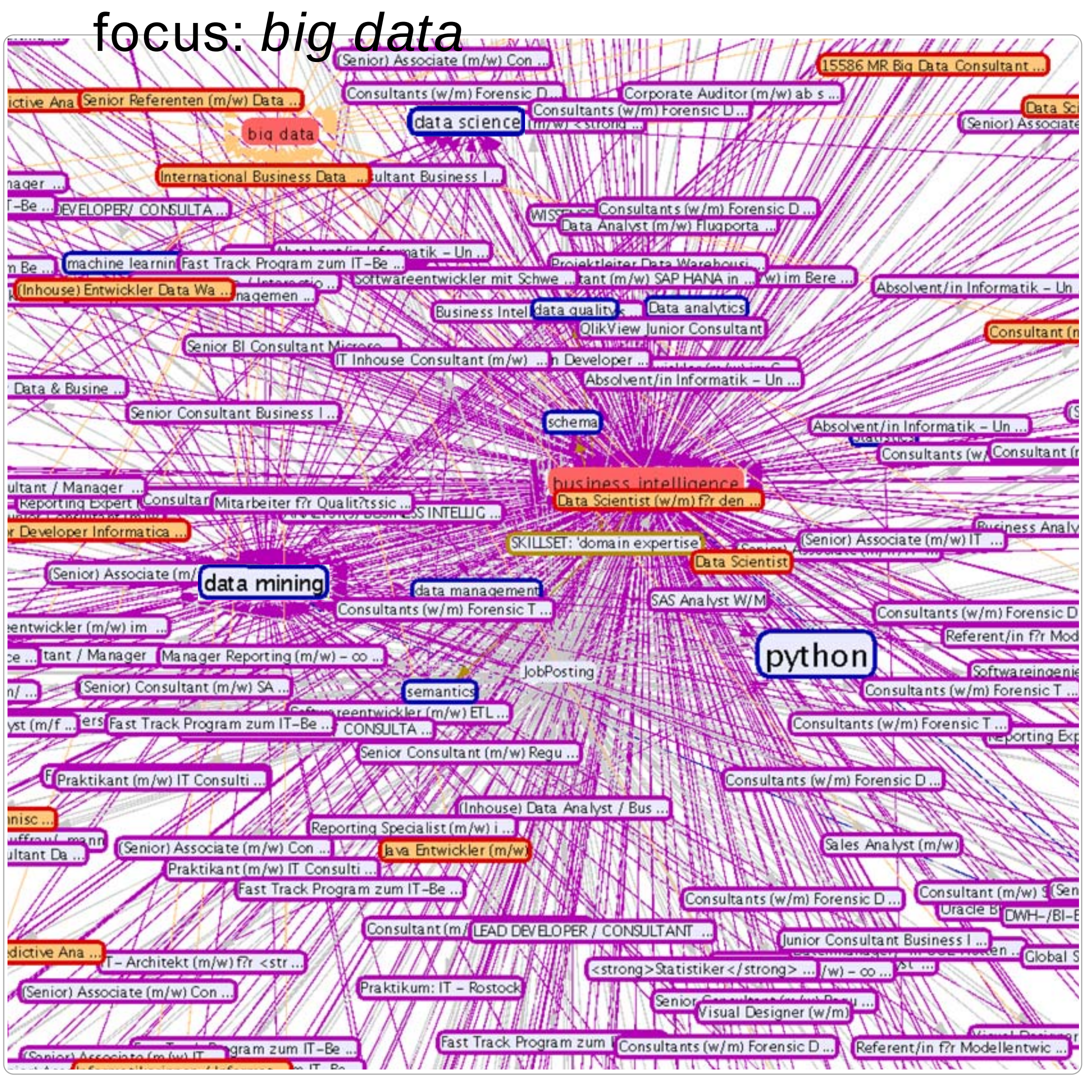


Figure 10: *Job posting network graph*

Zooming into an ROI in a dense network containing ~140,000 job postings (magenta border) in the UK (~90K), Germany (~21K) and France (~35K). Similarity between postings is based on skills (blue border) mentioned in each. Top, left, "big data" has the mouse focus, highlighting (in

amber) all postings that require this skill. Skill node size maps to total frequency of mention; 'python', 'data mining' and to a smaller extent, 'data science' stand out in this view.

As the next step, figure 11 filters out the dense layer of job posting nodes (in magenta) to reveal seven skills clusters (encircled in magenta) corresponding to the skill sets that were derived from Drew Conway's Venn Diagram. In this image, the node size is matched with the frequency of mentions across the entire dataset. Snapshot (a) applies a date filter, highlighting with pink borders those skills mentioned in October 2015. In (b) and (c) *statistics* and *data mining* respectively are given the focus (highlighted in red). This centres the layout on each, hiding all other nodes that do not co-occur with them (marked here with three dots). *Python* stands out as the single most frequently and persistently occurring skill, followed by *business intelligence* and *data mining*. Importantly, as will be illustrated further, this layout highlights both co-occurrence and exclusively occurring skills. These two factors are important to determine core skills that are important across a variety of domains as well as skills that are related to only one or a few aspects of data science. From the curriculum's development perspective, these are important design factors

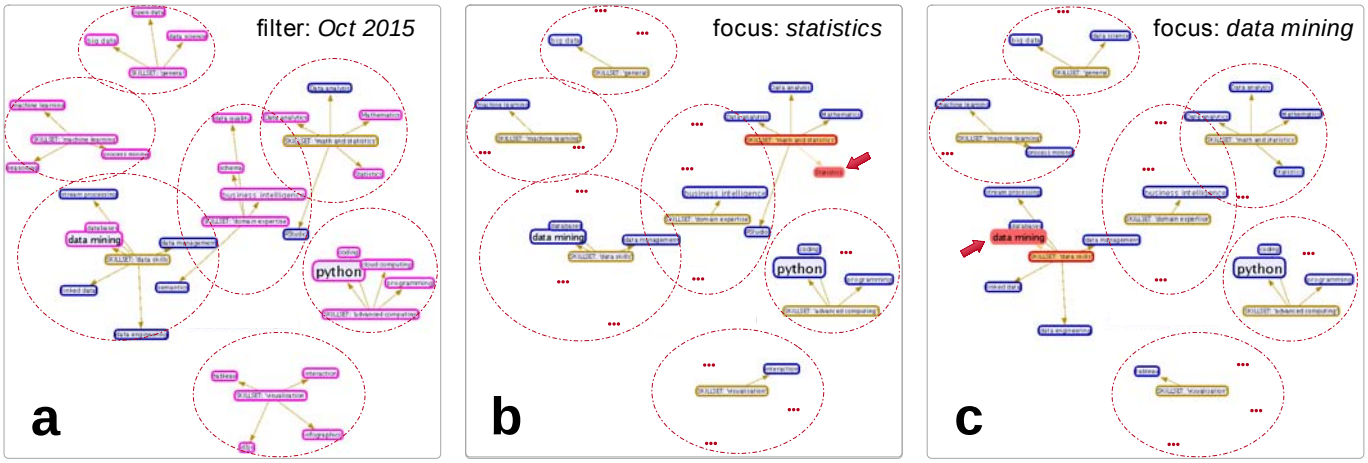


Figure 11: *Spring based layout of skill co-occurrence*

Spring-based layout used to visualise frequency of mention and co-occurrence of skills in new job postings. To aid analysis the large number of postings (See figure 10) from which the skill sets are extracted are hidden from the view

Frequency of skills

Looking at the total frequency counts, table 17 displays the top 20 skills for the entire dataset. Skills that fall into the initial direct or semantic filter set are highlighted. When considering all data, python is only ranked 11th and thus occurs far less frequently than figure 10 initially suggested. The top two skills, database and statistics also saw high co-occurrence with other skills across the subsets represented in figures 10 and 11. Through alternative visualisations we found that this spike, which is double the size of the next peak, results from high frequencies of these skills in job postings from the UK. Poland and Germany also see two and one spike respectively for Python, both over two non-overlapping time periods.

Considering the full dataset, the top three peaks for the smaller skill set are *data mining* (12th in Table 17), *big data* and *programming*, respectively. Merging *programming* with the smaller peak seen for *coding* would raise the former over *data mining*; both terms fall under *software development*, which falls 4th overall. *Business intelligence* and *machine learning*, two additional peaks due to the UK skew, while still seeing high mention, drop to 4th and 8th, respectively. Two additional spikes in the visualisation skill set - *interaction* and *tableau*, seen also in Figure 3, were discounted as further analysis showed these to be due to language use in French not consistent with their definition within this scope.

Table 17: *Top 20 skills by frequency of mention for data set containing ca 316K job postings across Europe*

	Skill	Count
1	*database*	37,995
2	*statistics*	35,189
3	project-management	34,272
4	*software-development*	34,255
5	design	25,958
6	*data-analysis*	23,465
7	sql	22,705
8	leadership	20,749
9	computer-science	19,495
10	*analysis*	18,079
11	*python*	17,433
12	*data-mining*	16,882
13	java	15,817
14	artificial-intelligence	15,663
15	finance	15,552
16	android	14,335
17	sdlc	11,733
18	*analytics*	11,303
19	sales	10,871
20	javascript	10,375

Table 18 shows the frequency of mentions for the top six countries for these skills, comparing them with other frequently mentioned skills in the field. Note that the counts here are postings listing the skill, not total frequency of mention. The UK and France top all lists, the two largest data subsets by location; we therefore also report this as a percentage of the total number of postings per country.

With these findings in mind, we however need to note two key limitations. Both of them contribute to significant skew in the data, with approximately 28% of the postings coming only from the UK. English terms are often used as-is, even in non-English job postings; however also because the filter terms do not always translate literally in other languages relatively fewer counts are recorded from non-English job postings. A second limitation results from the coverage of target portals, whose operation is restricted to a sub-set of EU countries.

As we noted in section 3.3.5 the terms and conditions and terms of service of content providers' APIs also changed over time, limiting our data collection and reuse, as well as truncating content. The second most detailed sub-set, France, with 11% of the total, was collected along with Germany before a change in the LinkedIn API access limited detail in data collection.

Data Science		
United Kingdom	22,008	25%
France	17,332	50%
Ireland	4,600	24%
Germany	3,854	18%
Sweden	3,842	28%
Poland	3457	19%

Data Mining		
United Kingdom	22,827	26%
France	17,427	50%
Switzerland	6,358	43%
Germany	5,107	24%
Ireland	4,671	25%
Sweden	3947	28%

Machine Learning		
United Kingdom	21,090	24%
France	17,517	50%
Ireland	4,320	23%
Germany	3,976	19%
Sweden	3,720	27%
Poland	3416	19%

Analytics		
United Kingdom	21,740	24%
France	17,349	50%
Ireland	5,049	27%
Poland	4,228	23%
Sweden	3,982	29%
Germany	3923	18%

Data Visualisation		
United Kingdom	18,309	21%
France	17,326	50%
Ireland	4,137	22%
Germany	3,773	18%
Sweden	3,464	25%
Poland	3273	18%

Statistics		
United Kingdom	27,259	31%
France	22,886	66%
Germany	9,110	43%
Ireland	5,863	31%
Romania	4,804	26%
Poland	4747	26%

Business Intelligence		
United Kingdom	19,999	23%
France	17,333	50%
Ireland	4,504	24%
Germany	4,282	20%
Poland	3963	22%
Sweden	3849	28%

Cloud		
France	17,336	50%
United Kingdom	17,208	19%
Germany	4,155	20%
Ireland	4,019	21%
Sweden	3,423	25%
Poland	3226	18%

Hadoop		
United Kingdom	19,834	22%
France	17,508	50%
Ireland	4,017	21%
Germany	3,907	18%
Sweden	3423	25%
Poland	3214	18%

Python		
United Kingdom	26,346	30%
France	17,570	50%
Germany	5,728	27%
Poland	4,780	26%
Ireland	4,024	21%
Romania	4007	22%

Java		
United Kingdom	23,503	26%
France	17,923	51%
Poland	5,218	29%
Germany	4,964	23%
Ireland	4,591	24%
Sweden	3748	27%

Scripting Language		
United Kingdom	18,315	21%
France	17,355	50%
Germany	4,436	21%
Ireland	4,309	23%
Sweden	3604	26%
Poland	3544	19%

Table 18: *Comparison of selected skills across locations by frequency: Comparison of selected skills across location, showing frequency of mention in postings for the top 6 countries, and the percentage this is of the total number of postings for each country*

While we expect differences in trends across time and location we must consider these limitations and the resultant skew when drawing conclusions over the entire dataset or even a single location.

Data at source is often "dirty", compounding differences in granularity and accuracy. As part of our analysis task we "clean" the data, merge entries where necessary and discarding those that fail to meet further essential processing requirements. There is therefore some discrepancy between the original data counts and those reported on querying the knowledge store. The more detailed sub-set for the Netherlands, for instance, has largely failed the automated upload so that the current count is 525, out of the 9,571 postings obtained. Work is on-going to identify the source of the errors and feed this into refining the data acquisition process.

Other work in progress to improve data acquisition includes the construction of a dictionary of terms, as part of the ontology population process, first to consolidate synonyms in the English term list, such as highlighted for *software engineering*, *programming* and *coding*. This contributes also to categorisation and ranking of the complete filter set along two dimensions: skill *sets* and *types*, where the latter distinguishes between skills as *capability* (e.g., machine learning), *product* (e.g., Hadoop) and/or *tools* (e.g., Python). We will feed this into on-going work on building corresponding multi-lingual dictionaries, including both formal and colloquial terminology, to increase precision and recall during data acquisition across non-English speaking areas of the EU.

5. Discussion and recommendations

Having explored and analysed the results of our studies in the previous sections, we now want to discuss our essential findings with a functional focus. Against the background of the previous section, we have therefore developed seven recommendations to inform and progress EDSA's curriculum and training offer. We generally propose that the consortium should explore options to implement these and, where possible, agree on concrete steps to realise them as part of EDSA's work. However, to what extent this can be done is subject to further discussion and agreement by the consortium partners.

EDSA's current focus is on producing high-quality, multilingual, multimodal training materials to cover the key curriculum topics. The multi-modal training offer includes classroom based courses, ebooks, MOOCs and online videos. We argue that EDSA's curriculum and training approach is generally in line with industry demands across Europe. Therefore, the curriculum's current focus on comprehensive technical and analytical training should be maintained. But to enhance EDSA's innovation potential, we suggest integrating technical and analytical training into a more holistic approach to skills development.

This proposition is based on a core result of our studies: Our analysis on the current course supply as well as interviews does not suggest a large training supply gap in technical or analytical data science training. While technical and analytical data science training is still in strong demand with various untapped improvement potentials, general market supply in those domains is strong.

Therefore, when it comes to further developing EDSA's technical and analytical training, the project can offer a more targeted and innovative service. More targeted by aligning training needs to industry demand. More innovative by providing comprehensive, blended training on open source tools.

The data suggests that increasing the impact of data science will also come from solving other problems. In particular, interviews and surveys showed that soft skills and training navigation are turning into challenges as important as technical and analytical challenges. EDSA is perfectly positioned to address these. Given EDSA's function as a research and innovation project, we suggest focusing on product innovation that addresses these training needs which are currently unaddressed by other suppliers.

Our concept builds partly on the expansion of EDSA's current curriculum in order to enhance supplementary skills that can catalyse the impact of data science in organisations. Additionally, we suggest that the EDSA consortium takes measures which can make easier both the discovery and identification of needed skills and suitable training. Our seven recommendations are briefly outlined in table 19 below.

Table 19: *Recommendations for EDSA curriculum development*

Title	Intervention level	Summary description
1. Holistic training approach	General training approach	Refine EDSA's training approach and curriculum cycle to strengthen data science skills for data science teams and data literacy across various units of each organisation.
2. Open source based training	Existing curriculum design	Continue current technical and analytical training based on open source technologies; apply cross-tool focus to deliver overarching training.
3. Soft skills training	Expansion of curriculum	Integrate soft skill training to increase performance and organisational impact of data scientists / data science teams.
4. Basic data literacy training	Expansion of curriculum	Develop basic data literacy training for non-data scientists to improve basic skills across organisations and facilitate uptake of data-driven decision making and operations.
5. Blended training	Course delivery	Develop blended training approaches including sector-specific exercises and examples to increase effectiveness of training delivery.
6. Data science skills framework	Training approach and delivery	Implement a data science skills framework to structure skills requirements, assess skills of data scientists, and identify individual skills needs.
7. Navigation and guidance	Training market	Develop quality assessment of third party courses; provide navigation support to identify relevant training from EDSA and third parties.

5.1 Holistic data science training

From the analysis of the quantitative and qualitative data collected through the course of our studies, we found that a strong supply of technical and statistical training exists. Additionally, our survey, interviews and review of current job postings confirmed that the demand for technical skills continues to be high. Having approached the study with a primary focus on these skills, we thus found our initial assumptions verified.

However, beyond this, results from open ended questions in our survey and interview conversations introduced another, less recognised dimension to the picture: In current industry discussions it is often assumed that data scientists can transform organisations through the production of objective evidence. In this rationalist paradigm, influence emerges from rational arguments which are produced based on scientific standards. From this perspective, even the term "data science" can be seen to imply two layers of objectivity: "Data" as an objective, impartial artefact resulting from human interactions with or transactions in the outside world. In the modern world, these can include tweets, data on online

purchases or one's web browsing history. In addition, "science" carries the notion that objective and impartial analysis is applied to these artefacts. The result of this should be objective, impartial evidence - convincing in itself.

The problem which many interviewees noted in our conversations is that it is much more difficult to make "data science" work in organisational and business practice than this line of thinking suggests. While many businesses and public sector organisations discuss the implications of becoming data-driven organisations, they face a holistic challenge. From our extensive demand analysis, we argue that data science lies at the heart of this challenge - but that it must be understood more broadly to maximise its impacts.

As a first, broad recommendation, we therefore suggest applying a more holistic training approach. According to this, EDSA work should continue to focus on expanding the technical and analytical skills of data scientists. However, the consortium should also build offers to improve the effectiveness of data science teams in organisations as well as to identify effective training more easily. For EDSA's curriculum and its wider related works, this implies three general pillars of development:

1) **EDSA needs to work on extending the skills of data scientists and non-data scientists.** Acquiring and being able to employ hard, sound, technical and statistical skills lies at the heart of data science training. But being effective as a data scientist requires the additional abilities of collaborating within and across teams, convincing through stories and narratives (rather than mere data), and influencing organisational leaders to make the right decisions. On the other side non-data scientists within organisations and businesses are required to have a basic degree of data literacy in order to be able to understand and - more importantly - critically reflect on the findings from data scientists.
2) **EDSA needs to differentiate its course offer to ensure that the most effective training means are used**. A vast amount of training is currently delivered by online providers. However, both surveys and interviews suggest that face-to-face elements can make training more effective. We therefore recommend that EDSA diversifies its curriculum delivery in order to explore innovative training modes. As highlighted by both survey and interview findings, study participants demanded particularly face-to-face trainings, blended learning formats, and sector specific assignments. For a scalable compromise between these different dimensions, we suggest EDSA explores options to diversify its training delivery, e.g. through a blended learning approach.
3) **EDSA should offer training advice.** The most effective data science training will have little impact if users cannot discover it or assess their own training needs falsely (and thus settle for unsuitable trainings). As we found from our desk research, the European market for data science training is strongly evolving and already crowded. As highlighted by team managers and learning professionals in our interviews, effectively training data scientists and building a data-driven organisation builds on two assumptions: first, that both individuals and organisations understand their own training needs; and second, that they make the right decisions about effective training and skills development. EDSA should address these challenges by offering individuals and organisations ways to assess their current skills, needs, and guidance towards appropriate training.

Together, these pillars form the foundations of a holistic approach to data science training. Its core assumptions are that data science does not just depend on technical skills, but making convincing arguments. Additionally, the audiences to whom data scientists speak need to have some basic data

literacy in order to be able to consume and work with the evidence that they are presented. Lastly, a wider challenge lies in helping individuals and organisations steer through the emerging data science training market, making the right decisions to meet their individual needs.

In the following six recommendations, we discuss which individual measures should be taken to implement this holistic training approach.

5.2 Technical and analytical data science training

With regards to EDSA's existing offer, our analysis suggests that the curriculum is largely in line with current high level demand trends. Requests for technical and analytical data science training are strong across Europe. Good data scientists appear as professionals who are strongly equipped with technical, analytical, and business skills. From a technical perspective, our survey and interviews together suggest that wide knowledge on different open source technologies is seen as key to using data science tools. Consequently, teaching wider skills of different open source tools should form an important element of data science training. Analytical training should instead focus on students' abilities to convey relevant findings effectively to different business audiences, especially through visualisations.

Additionally, data science professionals need to possess strong business acumen with an ability to quickly, flexibly and creatively apply their skills to a variety of real-life business situations. Facing this demand across various European industry sectors, EDSA's curriculum appears well positioned to boost Europe's data science skills.

As a general result with regards to EDSA's existing curriculum, we suggest maintaining and further refining the major technical and analytical parts of the curriculum. Training based on open source technologies (e.g. Hadoop, Python and R) should constitute one of the main pillars of EDSA's offer. At the same time, as raised by our interviews, EDSA's training focus should also feature a broader cross-technology and cross-tools content. While in-depth training for specific tools seem to exist in abundance, EDSA should offer courses that display how different tools can be used together, exploiting synergies of different systems. Strong cross-tool and cross-topical training can help users to understand how different data science domains relate to each other and how they can be used in combination.

From a procedural perspective, the ongoing refinements of courses should be guided by the analysis of course demand analytics (e.g. traffic data for individual modules) and learning analytics (e.g. analysis of course completion rates, dropouts etc.). As suggested in the EDSA Charter[61], regular course reviews, guided by EDSA's demand and learning analytics, should provide the basis for a competitive long term offer of EDSA.

5.3 Building soft skills

[61] EDSA Charter (D5.3) is currently under review and unpublished. The document will be accessible through EDSA's website: http://edsa-project.eu/downloads/deliverables/

One of the most remarkable patterns in both interview conversations and surveys is the emphasis that respondents put on softer skills.

Data interpretation and visualisation skills were highly rated skills for data scientists in our survey. As our interviews point to, these skills serve a crucial function in making core findings and evidence accessible to other functional teams and decision makers within organisations and businesses. However, to achieve impact, data scientists need to be able to understand demands and needs across teams - and align both their analysis and communication of results accordingly.

As we found in our interview conversations, managers and data scientists see this as a crucial way for data science to inform data-driven businesses. They therefore demand effective training to tackle this deficit.

Both team managers and data scientists were confident about their own and their team's data interpretation skills. However, at the same time, they noted in interviews that communication skills should be more strongly developed among data scientists. While they possess often excellent programming, numerical and statistical skills, understanding for what other teams need from them is not always optimal. This may lead to presentations being overloaded by technical details, rather than actionable insights. Additionally, from managers' perspectives, a lack of business focus leads to managing data science teams sometimes being like "herding the cats" as described by one manager. Our survey of data science courses across Europe also found a substantial lack of soft skill training, which is at the moment almost completely absent from course offers.

As a result of this finding and to tackle this deficit, we suggest integrating soft skills training into EDSA's curriculum. The focus of this is threefold: first, to develop audience-specific communication skills; second, to increase capacities in storytelling and finding data narratives; and, third, to build capacities to influence strategic management and leadership.

5.4 Providing data literacy training for non-data scientists

In addition to the previously mentioned measures, EDSA should also explore options to develop basic data literacy training for non-data-scientists. Our conversations with managers and learning professionals revealed that data literacy is often lacking across organisations. Even in cases where data science teams communicate evidence compellingly, a lack of understanding by those being communicated with may impede effective decision making based on the evidence provided. Specifically, members of other functional teams, such as sales or marketing teams, or senior management, might not understand the utility and limitations of analyses, leading to falsely informed decisions.

In more extreme cases, organisations might not even be interested in understanding data analytics because the deluge of algorithmically analysed data speaks for itself[62]. In such organisations, data science teams exist merely as units to uncover seemingly sound, actionable statistical evidence for other teams. Assuming that they operate based on "facts", those teams will however normally not question -

[62] http://www.wired.com/2008/06/pb-theory/

or even understand - the details of analyses. From an organisational perspective, such an organisational approach seems convenient because it does not necessitate wider investments into skills development - apart from building a separate data science team. Nevertheless, the lack of basic data literacy and a basic understanding of data science could substantially undermine an organisation's ability to become truly data driven.

A lack of data literacy implies poor understanding for how evidence was gathered - and thus little control on whether this work has been carried out effectively - carrying with it a risk of invalid conclusions. In the best case, strategical and operational decisions made without understanding the context of evidence are just ill-informed (but still effective); in the worst case, they will be ineffective or even harmful - with decision makers not even understanding why this is the case. In both outcomes, there is no reliable return on skills investments as organisations are not becoming data driven.

A lack of broad data literacy can lead to the further siloing of data analytics in one or a few cross-functional teams within an organisation and missed opportunities. In such a setting, other teams might have little to contribute to data analytics, leading to important insights being missed. Even if this does not lead to harmful consequences, it might still mean that the innovation potentials of data analytics will not be exploited as effectively as they could be.

Basic data literacy should thus be spread more broadly across organisations. This is to increase their collaborative, data-driven innovation potentials as well as to distribute control over data analytics, and improve the commissioning of work.

We believe that the most basic aspects of this can be addressed through the creation of a data literacy online module for non-data-scientists. This module could provide an introduction to the basic concepts and technical aspects of data science, furthermore focusing on introductions to statistics, data management and processing, as well as ethics and business strategy creation. To create this module, it should be mostly possible to repackage already existing course contents. An alternative solution would be to curate a directory of suitable third party offers and guide interested professionals towards those.

In general, it should be noted that students would not need to learn in-depth aspects of machine learning or data architectures. Instead, course contents should enable them to build the skills that will allow them to act as a sounding board for the insights gathered by data science teams.

5.5 Exploring options for blended-learning training

Quantitative and qualitative results from the demand analysis suggest that there is a great appetite among both data scientists and their managers for more effective training. While online training, particularly through MOOCs, as well as self-guided ad-hoc learning seem to have established themselves as the most popular training approaches, neither are ideal solutions: Rather, managers are looking for more customised solutions, ideally with some face-to-face elements. Whether these need to be implemented in-person or can be mediated online should be part of an experimental product development process. Additionally, blended learning offers should include sector specific hands-on exercises, both for teams and individuals.

While a relatively large proportion of interviewees preferred face-to-face trainings, focusing exclusively on face-to-face delivery would in our opinion not fit with EDSA's emphasis on scalability. Instead, we therefore recommend further developing the EDSA curriculum, using blended and problem-based learning approaches. As a minimum option, this should integrate problem-based assignments into existing modules. To further refine EDSA's offer, project partners should consider contributing sector-specific practice examples to individual courses. Our research shows that while general lecture and seminar content should remain focused on generalist contents, practical exercises should be tailored to more specific sectors. For example, assignments and practice examples could use sector specific data and task scenarios. As a compromise between generalist training and sectoral domain expertise, this approach would allow the maintenance of EDSA's current theoretical and lecture contents, whilst expanding their sectoral relevance through tailored, practical exercises.

To go beyond this minimal revision, partners could prioritise the integration of blended-learning content into the existing curriculum. Given the strong preferences of study participants for face-to-face learning, EDSA partners should ideally trial the offline delivery of some course contents. In the project's context, it is however problematic that these would incur substantial fees for course participants. To limit these expenses, project partners could also experiment with customised online classes, creating a virtual classroom experience. In this setting, assignments and other class materials could be distributed and collected at certain intervals in order to assess students' performance and give feedback on their progress.

5.6 Establishing a data science skills framework

In our interviews, team managers and learning professionals voiced concerns over the problems of understanding individual and organisational training needs. As has already been expressed at the very beginning of this report, data science is an emerging, heterogeneous profession. When understanding their own existing skills as well as their development targets and needs, both individuals and organisations thus face complex challenges. Unfortunately, these conflict with a simple, linear development paradigm because the precise skills needs of data scientists vary greatly as we found.

In this context, we suggest EDSA partners develop a flexible and multidimensional skills framework for data scientists. This should help them to self-assess their own skills and make decisions about their learning needs. The EU-funded EDISON project[63] is currently conducting similar work to establish a competence framework for data scientists based[64]. Since initial findings suggest that EDISON comes to similar conclusions regarding the skills that data scientists require, we suggest an exchange of ideas and findings in order to align further initiatives where possible.

Generally, we also suggest that the development process could start with a comprehensive review of the high-level technical and non-technical skills areas of data science. In our interviews with managers and learning professionals we found some criticism that current expectations regarding the skills of data

[63] http://edison-project.eu/
[64] http://edison-project.eu/sites/edison-project.eu/files/attached_files/node-29/edison-cf-ds-draft-cc-v06.pdf

scientists are overloaded. A goal of this work could be to identify, if possible, skills clusters that allow a more nuanced view on the different profiles of data scientists in practice.

The individual areas of the skills framework and the resulting profiles need to be tested against data scientist's real skills. For this, the consortium could create an online survey which lets users self-assess their own skills and maps them against the profiles laid out by the skills framework. Where they lack skills, users could also receive recommendations for EDSA or third party training. Other, more sophisticated approaches could involve the creation of semi-automatic tools where users can upload their CVs which would then be text mined and analysed; scores would be assigned to different skills, mapped against profiles and other users' data, with course recommendations to improve skills.

From an organisational perspective, a skills framework could be useful as it would help to characterise existing team skills more clearly. Organisations could thus assess collective and individual skills assets and assess more easily which ones need further development - and who should ideally receive training.

5.7 Providing navigation and guidance

The point of departure for our last recommendation is the feedback from both managers and data scientists on their difficulties in finding appropriate training. As our analysis has shown, there seems to be no lack of technical and analytical data science training. Instead, potential students and organisations are struggling to navigate and filter the sheer abundance of training opportunities. However, finding those that fit one's own needs and also deliver high-quality learning experiences is a crucial task. As an impartial and transparent advisor, EDSA could help European data scientists to find the training they need. Accordingly, we suggest that the consortium explores the options to integrate course navigation and guidance functions into its work.

The core motivation of this is to guide data scientists and organisations to high quality courses offered by EDSA and third parties. This implies two interconnected sets of measures:

1) **First, the consortium should develop and pilot a transparent evaluation framework to assess the quality of online and face-to-face training**. This framework should be created along the lines of the EDSA design and delivery values, creating a uniform quality standard for data science training in Europe. This work could build on and expand the consortium's current work on a course endorsement process.[65] Additionally, in order to match training based on their content, the framework should also define a set of training characteristics which can be used to gather structured information on addressed skills but also on the skills length of courses, costs, delivery formats, etc. To test the practical feasibility, consortium partners should work with an initially limited number of third-party providers to assess course quality and other characteristics. This work could potentially be supported by EDSA's network of advisors and ambassadors; furthermore it will help to build a training delivery with third party content

[65] EDSA has created an application process for course endorsements which is accessible online: https://docs.google.com/a/theodi.org/forms/d/1qpNSzHKBSiT1vUB1evpE7EGDYTv0HGsxIRL2pOn0Flo/viewform?edit_requested=true.

providers, as envisioned by the EDSA Charter[66]. Evaluated training offers that match the EDSA quality requirements should be listed as third-party training resources on EDSA's project website. To raise market awareness for these high quality training, but also EDSA's own brand, project partners should furthermore consider developing "EDSA badges" or training certificates so that users can easily identify compliance with EDSA's quality standards. Successfully assessed course providers would then receive a badge or certificate for display on their website.

2) **Second, the consortium should expand its efforts in guiding users to relevant data science training.** As reported in the technical report attached to this analysis, EDSA's job posting dashboard currently includes a function which guides users to data science training relevant to skills mentioned in specific job postings. This feature could be expanded to allow users to filter training listed on EDSA's website based on various criteria, including skills, costs, delivery formats etc. Consortium partners should also explore options to integrate the skills self-assessments proposed previously and the navigation of relevant training offers. This would allow a more dynamic navigation approach, potentially helping users to identify training aspects which they would not have discovered if self-assessments and training navigations were kept separately

If successful in a pilot, EDSA could continuously expand this offer, by partnering with more third-party suppliers. The consortium should also work to integrate external courses into EDSA's curriculum, ensuring the cohesiveness of EDSA's course offer.

We are convinced that these measures together tackle the current core challenges in Europe's data science training market. If implemented consistently, they will help to not only fill supply gaps, but facilitate informed choices on training investments. Lastly, they also contribute to further expanding EDSA's role as a leading, standard-setting data science authority in Europe.

[66] The EDSA Charter maps the consortiums sustainability plans and commitments to build a distributed European Data Science Academy. The document is currently under review and unpublished, but will later be accessible on EDSA's website: http://edsa-project.eu/downloads/deliverables/

6. Conclusions

6.1 Conclusions

Modern digital economies need workers who can process, analyse, and make sense of exponentially growing piles of data. Accordingly, organisations across the globe from all sectors increasingly rely on a new group of professionals to increase their productivity and deliver better services. Often these experts are referred to as data scientists. The opportunities for employees, employers, and educators are vast - and thus a whole ecosystem has emerged to teach data scientists.

Nevertheless, beyond all the hype and fanfares, it appears that many details of this new professional phenomenon are still not well understood. Many of the existing problems involve seemingly basic challenges, such as defining the profile and skills for data scientists in specific positions or finding appropriate trainings.

Against this background, we took off to explore the current demand for data science skills in Europe's various industries. From this, we also sought to learn what trainings should be offered in order to accommodate this demand - and, subsequently, which options exist for EDSA to develop a sustainable, high-impact offer.

Through our mixed-mode study, connecting a series of qualitative and quantitative data collection modes, we retrieved in-depth insights from more than 690 data science professionals and managers across all EU member states as well as Switzerland, Norway, Iceland and Serbia. To additionally back up our research, we also reached out to high level managers and learning professionals to discuss how their organisations approach data science training. Furthermore, we conducted four focus group workshops to test training demand patterns with practitioners. To triangulate our findings through a more refined contextual view, we gathered rich secondary data. This included the analysis of 456 data science trainings offered by universities and professional training suppliers across Europe as well as 316k jobs postings across Europe.

Among practitioners from a variety of sectors and countries, our study attracted great interest and was met with enthusiasm. However, for some sectors and countries we found it difficult to identify and acquire participants: In particular, we would have liked to conduct more interviews in Slovakia, the Czech Republic, Hungary, Cyprus, and Luxembourg. Similarly, some industry sectors with traditionally less intense ICT usage patterns were difficult to survey and thus returned a lower number of responses. This is particularly the case for the agriculture, mining and quarrying, real estate, as well as water supply and waste management industries. Given the substantial additional efforts that seem to be required to acquire additional study participants from these domains, we suggest to conduct specific studies to better understand the needs of early adopters.

Across Europe, we find that the demand for data scientists is strong. However, as the findings from our job postings data suggest, demand differ across countries with Western European countries leading. Skills in statistics and programming, particularly using Python, are strongly required. Additionally, we found from our surveys and interviews that general data collection and analysis as well as data interpretation and visualisation skills are the most desired capabilities for data science professionals. Other skills, such as in maths and statistics, big data, machine learning and prediction, business intelligence and domain expertise, advanced computing and programming, as well as open source tools

and concepts are seen as either essential or desirable by more than three quarters of all respondents. Interestingly, we also find that the managers of data science teams seem to be more optimistic about their team's skills than data scientists themselves.

Taking a look at the training approaches of respondents, it appears that data scientists often integrate their ongoing learning efforts into their daily work routines. Accordingly, self-driven ad-hoc learning is widespread; in our interview conversations it appeared that these learning modes are particularly useful for improving technical skills on the many, often open source based technologies which data science professionals use on a daily basis. At the same time, there however also appears to be a great interest in more refined and, according to respondents, more effective training approaches. These should involve at least some face-to-face elements and be accompanied by sector specific assignments. Hence, while domain specific knowledge is important to many of our study participants, we suggest to embed sector specific training as a lateral, practical element. Fully sector-tailored training seems to be instead not required from most respondents' perspective.

From our analysis, we can also see that while there appears to be a rich market supply for technical and analytical skills training, the demand for softer skills training is largely unmet. As we have learned from our interviews, data scientists in practice are required to be much more than sophisticated data wranglers. Rather, particularly managers and learning professionals are enviosioning influencers who can drive the data-driven transformation of organisations. Naturally, this requires not just technical skills, but strong skills in teamwork, communication, presentation, and leadership. Beyond this, some managers and data scientists expressed concerns as to whether members of non-data science teams (e.g. from marketing and senior management) are able to fully comprehend new analyses. A general level of data literacy is thus required to ensure that teams across the organisation do not only understand analyses, but can innovate based on them.

Drawing these findings together, it does appear less surprising that not just managers and learning professionals but even data scientists struggle with defining the specific skills and profile for data scientists. As we learned from our conversations with participants across sectors and countries, data scientists are a rather heterogeneous group of workers whose specific skill requirements depend on the organisational setting they are operating in. Hence, while data scientists and team managers appear to be rather confident with regards to their own and their team's skills, both also said it was difficult to make predictions on how they compare with their capacities and which skills they should further develop. From an organisational perspective, the related task of specifying the training needs for teams and individual members is even more complex. In addition to this, finding relevant, high quality training is a difficult - ironically, because of the sheer amount of training on the market.

For EDSA's further development, these findings are interesting as they direct the project consortium's attention towards a more holistic training paradigm. As we outline in our recommendations, a holistic training concept should add soft skills and general data literacy training to the existing technical and analytical training. Additionally, it should help individuals and organisations to understand where their skills need improvement - and where they can find suitable trainings to address these needs.

From this recommendation for a holistic EDSA training approach, we derived six additional recommendations: Given our study results, we think that EDSA should further continue its focus on technical and analytical training based on open source technologies. However, while we find that this training is in line with industry demand, we suggest to also develop and integrate soft skills trainings

for data scientists as well as general data literacy trainings for non-data-scientists. To improve the learning experience and impact, we propose that EDSA should invest in blended learning formats that use additional sector-specific examples and assignments. Finally, as a supplement to EDSA's core training offer, we recommend the development of frameworks to assess the skills of data scientists as well as the contents and quality of training. While both measures can function as stand-alone features, their strongest impact will emerge when deployed together. Data scientists will be able to understand their own skills level, and be guided to the most suitable and effective training on the market.

Together, these recommendations provide directions for EDSA's long-term development beyond the project's end. Based on the insights gathered through our studies, we are convinced that these recommendations will help progress and establish EDSA's competitive offer. Crucially, they will help to address development areas which have been largely missed by current training suppliers.

6.2 Future work

In summary, we find that our methodological approach and its implementation served the explorative purposes of our study well. Given the early stage of the project, we were able to identify the demand for crucial skills demands across Europe in a variety of industry sectors. With these new insights, additional material and methodological questions have arisen as well. Hence, we would like to conclude this analysis with a brief discussion of potentials for future work. As with our recommendations, these could provide a starting point for discussions on further research through the consortium.

6.2.1 Deepen sectoral and country research

As we discussed at the beginning of section 4 and in our report on KPI reach, country and sectoral coverage remains relatively challenging in some cases. While we did not find major differences between the demands for data scientists in different industry sectors, we would still suggest exploring in more detail some of the industry sectors and countries that were only marginally covered by our data.

With regards to data acquisition and sampling, this would likely require substantial additional efforts: As we found, identifying and recruiting study participants in some Eastern European countries as well as more traditional industry sectors was challenging. Speaking to further experts from sectors such as agriculture, water supply and waste management or real estate could provide useful insights into how early adopters from traditional industries are coping with the application of data science - and where they struggle. Uncovering specific evidence on the adoption in these sectors could eventually help to progress data science in sectors with a strong, immediate real life impact.

The same case applies to the weaker sectoral coverage of some Eastern European countries. For a balanced long-term development of Europe's economy, it is important to understand how these member states are faring - and where they might need specific training initiatives. More field research in e.g. Slovakia, the Czech Republic, and Hungary could potentially provide these insights.

6.2.2 Increase total sample size

In order to make more robust claims about the statistical relations between different skills domains and their demand in various industry sectors and countries, we suggest increasing the sample size substantially. As we mentioned previously, with a coverage of 28 EU member states and 19 Eurostat-defined industry sectors, our sample has a very high dimensionality. This implies that subsamples will often be rather small - which impedes the feasibility or robustness of conducting more advanced statistical analyses, such as regression or network analyses, on those data. It would thus be worthwhile to expand the current sample through additional data collection. Methodologically, this can be largely based on the current study design. However, in terms of coverage, additional data collection should focus on the acquisition of more data from countries and sectors that are currently underrepresented.

To facilitate the acquisition of suitable study participants, future projects should also explore in more detail how publicly accessible data from social or professional networks can be used to identify participants. Bearing in mind the potential sampling bias resulting from such approaches, one possible approach would be to scrape Twitter profile data from the Twitter API, filtering for users who either mention data science and related keywords in their biography or tweets.

At this point, we however also need to critically reflect on a learning from our work: It appears that data producers are increasingly protecting, both through legal and technical means, data which was previously accessible. In our case, LinkedIn changed its terms of use and also introduced technical means to impede web scraping mid-way through our data collection. Generally, we find this is a problematic trend which can dramatically reduce the availability of publicly accessible data for research. In particular, we expect that this can negatively impact research into highly dynamic, new economic, social and political phenomena which are usually not well covered through traditional statistics. From the perspective of researchers, the imminent risk of data providers banning access also introduces additional obstacles to conduct studies. On the one hand, requests for access to data are frequently not answered. On the other hand, the volatility regarding data access implicitly also requires researchers to permanently conduct safeguarding measures, such as keeping a record of the terms of use for specific online services.

6.2.3 Explore the benefits of non-English data science training

Interestingly, our survey and interview participants did not express great interest in data science trainings translated into other languages. Our research suggests that non-English training is not seen as a cost effective investment because most tools and technologies used by data scientists are documented in English. Hence, as soon as data scientists need to explore new domains and tools which they have not used before, it is very likely that these are documented mainly in English. In short, it thus appears that the data science ecosystem is mainly built in English. In this context, in-depth technical and analytical training which is not taught in English would be more disconnected from wider industry developments.

In our interviews, a majority of respondents expressed relatively strong preferences for English as a course language. However, this does not answer questions on the niches where native languages might still be useful to convey relevant data science expertise. For example, basic trainings on general data literacy, introductions to data science or soft skills could still be useful if taught in other languages than

English. In particular, they would make basic data science training accessible to people with limited knowledge of English. Additionally, they could also help soft skill students to better understand interpersonal as well as cultural clues and notions. Often, non-native speakers miss these important aspects if instructed in other languages. Hence, there might still be the case for more effective, non-English trainings.

However, to further explore the validity of such arguments, further research is required. Our current data on non-English trainings does not cover all relevant aspects and sufficiently granular data to back up these claims. For future research, we thus propose to explore in more detail where non-English data science trainings can add substantial training value - and where not. To provide rich qualitative insights we would suggest research designs that rely primarily on extensive, semi-structured interviews.

6.2.4 Conduct further analysis based on learning analytics

A further domain of potential new research is to explore the effectiveness of various training formats through learning analytics. As discussed, online learning currently seems to dominate the market even though it is not seen as the most effective method. Therefore, as a compromise between integrating face-to-face and sector specific assignment elements as well as ensuring scalability, we recommend the consortium develops blended learning formats. While our research indicates that managers and data scientists would find these training measures far more effective, we think that closer monitoring is required to make final claims about how good various learning formats are in different settings.

The consortium already explores data gathered from the learning analytics systems currently adopted by the EDSA consortium as part of work package 2 and 3. For online training, this data can be gathered through tracking tools to record useful information, such as how long students engage with specific modules, drop-out rates etc. Where possible, similar data should also be collected for face-to-face training. Comprehensive data analysis of learning analytics data should be able to inform continuous, more refined course and curriculum design. This would not just lead to a more refined understanding of how well specific courses are received, but also which training methods are really more effective under specific circumstances.

7. Appendices

Appendix 1. Achieved coverage of network KPI

Country	Number of interviews conducted	Number of sectors covered
Austria	2	2
Belgium	3	3
Bulgaria	9	5
Croatia	2	2
Czech Republic	2	2
Denmark	2	2
Estonia	3	3
Finland	2	2
France	5	4
Germany	9	6
Greece	3	3
Hungary	2	2
Ireland	3	3
Italy	3	3
Latvia	2	2
Lithuania	2	2
Luxembourg	2	2
Malta	7	5
Netherlands	3	3
Poland	3	3
Portugal	2	2
Republic of Cyprus	2	2
Romania	2	2
Slovakia	2	2
Slovenia	3	2
Spain	9	6
Sweden	5	5
United Kingdom	12	8
Non-EU member states		
Serbia	0	0
Switzerland	1	1

Iceland	1	1
Norway	0	0

Appendix 2. Interview and survey questions - final design

Part 1: Personal details and professional background

Note - could be filled out before interview by interviewer

Q1.1: Country

Question: Which country do you work in? [or Select the country you work in]

Answers:

- List of EU countries:
  - Austria
  - Belgium
  - Bulgaria
  - Croatia
  - Republic of Cyprus
  - Czech Republic
  - Denmark
  - Estonia
  - Finland
  - France
  - Germany
  - Greece
  - Hungary
  - Ireland
  - Italy
  - Latvia
  - Lithuania
  - Luxembourg
  - Malta
  - Netherlands
  - Poland
  - Portugal
  - Romania
  - Slovakia
  - Slovenia
  - Spain
  - Sweden
  - UK

Q1.2: Role

Question: Which of these best describes your primary role:

Answers:

- a data scientist
- someone who manages data scientists

Additional note:

- Question determines which interview pathway is followed

Q1.3: Size of organisation

Question: What type of organisation do you work for? [or Select the size of the organisation you work for]
Answers:

- Individual
- Micro (<10 employees)
- SME (10 to 250 employees)
- Large (250+ employees)

Q1.4: Sector
Question: Which sector do you work in? Please choose the one which best describes your main focus.
Answers: list of sectors.

- Agriculture
- Mining
- Manufacturing
- Energy
- Water and waste management
- Construction
- Transport
- Accommodation and food services
- Media
- Data and information systems
- Finance and insurance services
- Real estate
- Professional services
- Scientific and market research
- Business administration services
- Tourism
- Public administration and defence
- Education
- Human health and social work
- Arts, recreation and entertainment
- Consumer services
- Government and public sector
- Wholesale and retail

Part 2: Qualitative questions
Interview only - qualitative section

Q2.1: Impact of DS
Question: What impact has data science had on your organisation?
Answers: Free
Prompt: Can you tell me how this has impacted your organisation? Changes to roles?

Q2.2: New set of skills for DS
Question: Does data science as a profession require a new set of skills?
Answers: Free

Prompt If yes: What are these skills?
Prompt If no: What existing skills are required?

Q2.3a: Have <u>you</u> attended courses
Question: Have you attended any data or data science courses to expand your own skill set?
Answers: Free
Prompt: Can you identify particular course names/types/providers

Q2.3b: Has <u>your team</u> attended courses
Question: Have members of your team attended any data or data science courses to expand capacity in your organisation?
Answers: Free
Prompt: Can you identify particular course names/types/providers

Q2.4a: Have <u>you</u> taken other approaches
Question: Have you taken any other approaches to develop your skills?
Prompt: for example coaching, assessments etc.

Q2.4b: Has <u>your team</u> taken other approaches
Question: Have you taken any other approaches to develop the skills of your team?
Prompt: for example coaching, internal assessments etc.

Q2.5a: Key challenges <u>finding training</u>
Question: What are the key challenges in finding training in the skills needed to stay ahead in data science?
Prompt: Do you find it hard to locate the right training? Is the training good?

Q2.5b: Key challenges <u>finding skilled workers</u>
Question: What are the key challenges in finding skilled people for your organisation to stay ahead in data science?
Prompt: Do you find it hard to employ the right people? Why? Do you take advantage of the skills you have?

<u>Part 3: Quantitative questions</u>
Both interview and survey
Q3.1: What skills should a DS have
Question: Considering the role of a data scientist as a single individual, how essential would you rate each of following skills for that person:
Skills:

- Big Data
- Machine Learning and Prediction
- Data Collection and Analysis
- Maths and Statistics
- Interpretation and Visualisation
- Advanced Computing and Programming
- Business Intelligence and Domain Expertise

- Open Source Tools and Concepts
- *Any other key or sector-specific skills?*

Categories:

- Essential
- Desirable
- Not required

Additional note: Has qualitative 'other' answer

Q3.2a: Rate <u>your</u> strengths

Question: How would you rate your strengths in each of the areas:

Skills:

- Big Data
- Machine Learning and Prediction
- Data Collection and Analysis
- Maths and Statistics
- Interpretation and Visualisation
- Advanced Computing and Programming
- Business Intelligence and Domain Expertise
- Open Source Tools and Concepts

Scale:

Number	1	2	3	4	5
Level	Very Poor	Poor	Ok	Good	Very Good
Meaning	Little or no knowledge/ expertise	Experimental /vague knowledge	Familiar and competent user	Regular and confident user	Leading expert

Q3.2b: Rate <u>your team's</u> strengths

Question: How would you rate your team's strengths in each of the areas:

Skills:

- Big Data
- Machine Learning and Prediction
- Data Collection and Analysis
- Maths and Statistics
- Interpretation and Visualisation
- Advanced Computing and Programming
- Business Intelligence and Domain Expertise
- Open Source Tools and Concepts

Scale:

Number	1	2	3	4	5
Level	Very Poor	Poor	Ok	Good	Very Good

Meaning	Little or no knowledge/ expertise	Experimental /vague knowledge	Familiar and competent user	Regular and confident user	Leading expert

Q3.3: Technologies, tools and languages

Question: Which data technologies, tools and languages would you like to see included in data science training?

Answers:

- AWS
- Spark
- Hadoop / MapReduce
- MongoDB
- Open Refine
- QMiner
- Apache Flink
- Apache Storm
- ProM or Disco
- NoSQL / Cassandra
- R
- Python
- Javascript / JQuery
- D3 / nvD3
- Java
- z-scores
- *Any other key or sector-specific tools?*

Categories:

- Essential
- Desirable
- Not required

Additional note:

- Has qualitative 'other' answer

Q3.4: Training methods

Question: What training delivery methods do you consider to be important factors in successful training?

Answers:

- Face-to-face training
- Webinars
- eLearning
- Translated from English
- Tailored to sector
- Accredited
- Uses non-open, non-free software

- Coaching
- Assessed
- Internal assignments

Categories:

- Essential
- Desirable
- Not required

Additional note:

- Has qualitative 'other' answer

Part 4: Wrap up

Split by survey/interview

Q4.1 survey: Comments

Question: If you have any other comments relating to data science skills and requirements in your organisation and sector please feel free to include these below.
Answers: Free form

Q4.1 interview: Vision for EDSA

Question: What would your vision for a European Data Science Academy be?
Prompts: Key roles? Mission?

Instructions on interview pathways

Question pathway for each persona - survey and interview.

Survey

9 questions

Practitioner pathway:

Part 1: Q1.1, Q1.2, Q1.3, Q.1.4
Part 2: None
Part 3: Q3.1, Q3.2a, Q3.3, Q3.4
Part 4: Q4.1 survey

Manager pathway

Part 1: Q1.1, Q1.2, Q1.3, Q.1.4
Part 2: None
Part 3: Q3.1, Q3.2b, Q3.3, Q3.4
Part 4: Q4.1 survey

Interview

10 questions (+4 by interviewer)

Practitioner pathway

Part 1: Q1.1, Q1.2, Q1.3, Q.1.4
Part 2: Q2.1, Q2.2, Q2.3a, Q2.4a, Q2.5a
Part 3: Q3.1, Q3.2a, Q3.3, Q3.4

Part 4: Q4.1 interview

Manager pathway

Part 1: Q1.1, Q1.2, Q1.3, Q.1.4
Part 2: Q2.1, Q2.2, Q2.3b, Q2.4b, Q2.5b
Part 3: Q3.1, Q3.2b, Q3.3, Q3.4
Part 4: Q4.1 interview

Appendix 3. Report on research company KPI compliance

Total counts

Variable	Target	Current	Achieved?
Interviews	56	56	Yes
Survey	500	500	Yes

Interviews by countries

Country	Target	Current	Achieved?
Austria	**2**	**2**	**Yes**
Belgium	**2**	**2**	**Yes**
Bulgaria	**2**	**3**	**Yes**
Croatia	**2**	**2**	**Yes**
Republic of Cyprus	**2**	**2**	**Yes**
Czech Republic	**2**	**2**	**Yes**
Denmark	**2**	**2**	**Yes**
Estonia	**2**	**2**	**Yes**
Finland	**2**	**2**	**Yes**
France	**2**	**2**	**Yes**
Germany	**2**	**2**	**Yes**
Greece	**2**	**2**	**Yes**
Hungary	**2**	**2**	**Yes**
Ireland	**2**	**2**	**Yes**
Italy	**2**	**2**	**Yes**
Latvia	**2**	**2**	**Yes**
Lithuania	**2**	**2**	**Yes**
Luxembourg	**2**	**2**	**Yes**
Malta	**2**	**2**	**Yes**
Netherlands	**2**	**2**	**Yes**
Poland	**2**	**2**	**Yes**
Portugal	**2**	**2**	**Yes**
Romania	**2**	**2**	**Yes**
Slovakia	**2**	**2**	**Yes**
Slovenia	**2**	**1**	**No**
Spain	**2**	**2**	**Yes**
Sweden	**2**	**2**	**Yes**

UK	2	2	Yes

Interviews by industry sectors

Eurostat code	Sector	Target	Current	Achieved?
A		2	0	No
B		2	1	No
C		2	5	Yes
D		2	3	Yes
E		2	0	No
F		2	1	No
G		2	3	Yes
H		2	1	No
I		2	3	Yes
J		2	7	Yes
K		2	5	Yes
L		2	1	No
M		2	7	Yes
N		2	2	Yes
O		2	3	Yes
P		2	5	Yes
Q		2	5	Yes
R		2	1	No
S		2	3	Yes

Balance of sectoral coverage

Lower bound	3%
Upper bound	15%

Sector Code	Current value	Current Percentage	Achieved?
A	6	1%	No
B	3	1%	No
C	38	7%	Yes
D	34	6%	Yes
E	8	1%	No
F	13	2%	No

G	**25**	**4%**	**Yes**
H	**19**	**3%**	**Yes**
I	**17**	**3%**	**Yes**
J	**91**	**16%**	**No**
K	**33**	**6%**	**Yes**
L	**9**	**2%**	**No**
M	**78**	**14%**	**Yes**
N	**38**	**7%**	**Yes**
O	**34**	**6%**	**Yes**
P	**37**	**7%**	**Yes**
Q	**36**	**6%**	**Yes**
R	**19**	**3%**	**Yes**
S	**18**	**3%**	**Yes**

Regional coverage

Region	Current value	Current Percentage	Achieved?
Northern Europe	**166**	**30%**	**Yes**
Eastern Europe	**101**	**18%**	**Yes**
Southern Europe	**150**	**27%**	**Yes**
Western Europe	**139**	**25%**	**Yes**

Coverage of roles

Role	Current value	Current Percentage	Lower bound	Upper bound	Achieved?
Data Scientist	308	55%	35%	45%	No
Manager	248	45%	55%	65%	No

Organisational coverage

Size	Current value	Current Percentage	Lower bound	Achieved?
SME (10 to 250 employees)	**194**	**35%**	**30%**	**Yes**
Large (250+ employees)	**335**	**60%**	**40%**	**Yes**

Appendix 4. Country grouping according to UN-defined European regions

Please note that regions were assigned according to the following UN classification: http://unstats.un.org/unsd/methods/m49/m49regin.htm

Country	Region
Austria	Western Europe
Belgium	Western Europe
Bulgaria	Eastern Europe
Croatia	Southern Europe
Republic of Cyprus	Southern Europe
Czech Republic	Eastern Europe
Denmark	Northern Europe
Estonia	Northern Europe
Finland	Northern Europe
France	Western Europe
Germany	Western Europe
Greece	Southern Europe
Hungary	Eastern Europe
Ireland	Northern Europe
Italy	Southern Europe
Latvia	Northern Europe
Lithuania	Northern Europe
Luxembourg	Western Europe
Malta	Southern Europe
Netherlands	Western Europe
Poland	Eastern Europe
Portugal	Southern Europe
Romania	Eastern Europe
Slovakia	Eastern Europe
Slovenia	Southern Europe
Spain	Southern Europe
Sweden	Northern Europe
UK	Northern Europe

Appendix 5 Full ranking list of technologies, tools and languages to be covered by data science training

Note: This list includes a comprehensive listing of all categories (both pre-defined and added via free text form) that have been mentioned at least by two different survey participants.

Tool	Count	Percentage
R	217	37.7
Python	192	33.4
None	125	21.7
Java	95	16.5
SQL	94	16.3
Hadoop / MapReduce	88	15.3
NoSQL / Cassandra	72	12.5
Apache Spark	70	12.2
Javascript / JQuery	69	12
Excel	57	9.9
SAS	40	7
D3 / nvD3	36	6.3
AWS	31	5.4
Apache Storm	27	4.7
MongoDB	26	4.5
MatLab	23	4
C++	20	3.5
Apache Flink	19	3.3
SPSS	19	3.3
Open Refine	18	3.1
zscores	18	3.1
Microsoft Office	16	2.8
Tableau	16	2.8
C#	15	2.6
QMiner	13	2.3
Scala	13	2.3
Oracle tools	11	1.9
ProM or Disco	11	1.9
Visual Basic	11	1.9
Microsoft Access	10	1.7

Microsoft tools	10	1.7
C	9	1.6
Linux	8	1.4
QlikView	8	1.4
HTML	7	1.2
PHP	7	1.2
SAP tools	7	1.2
VBA	7	1.2
.NET	4	0.7
Apache tools	4	0.7
Hive	4	0.7
Microsoft Power BI	4	0.7
Unix	4	0.7
CSS	3	0.5
Fortran	3	0.5
GitHub	3	0.5
Google Analytics	3	0.5
Julia	3	0.5
Lambda	3	0.5
Microsoft Powerpoint	3	0.5
MySQL	3	0.5
Pandas	3	0.5
SAP Business Object	3	0.5
Apache Samoa	2	0.3
HANA	2	0.3
Jupyter	2	0.3
Kafka	2	0.3
Mathematica	2	0.3
Microsoft VisualStudio	2	0.3
MicroStrategy	2	0.3
Minitab	2	0.3
OpenStack	2	0.3
Orange Data Mining	2	0.3
Pablo	2	0.3
Plotly	2	0.3

PostGIS	2	0.3
Weka	2	0.3
XML	2	0.3

Technical Report

Project acronym:	**EDSA**
Project full name:	**European Data Science Academy**
Grant agreement no:	**643937**

Technical Report

This document is part of a research project funded
by the Horizon 2020 Framework Programme of the European Union

Table of Contents

Table of Contents....108
List of Tables....109
List of Figures....109
1. The Demand Analysis Dashboard....111
2. Dashboard Design....112
2.1 User Specification....112
2.2 The Knowledge Framework....113
2.2.1 The Skills and Recruitment Ontology....114
2.2.2 Third-Party Resources reused in SARO....118
2.2.3 Example of Use- Skill Correlation....119
3. User and Task Perspectives....121
3.1 Map Views....123
3.2 Timeline Views....123
3.3 Statistical Analysis Views....123
3.4 Intermediary Component - Switching between Modules....123
3.5 User Perspectives Design Sketches....125
4. Demand Data....129
4.1 Data Acquisition Pipeline....130
4.2 RDF Data Store....133
5. Ontology-Guided Visual Exploration, Analysis and Knowledge Acquisition in the Dashboard....137
5.1 Visual Analysis Tools and APIs....137
5.2 Usability Evaluation - Beta Version June 2016....137
5.3 Analysing Data Science Skills Demand....138
6 Discussion....146
6.1 Data Acquisition....146
6.2 Hosting the EDSA Dashboard as a Live, Interactive, Online Tool....146
6.3 Development Resource....147
Appendix A: Usability Evaluation Tasks & Questionnaire....148

List of Tables

Table 1: *Total number of job postings by country across all extraction all sources and extraction methods used between Nov 2015 and Jun 2016 - including historical data from 2013.* ---------- 132

List of Figures

Figure 1: *initial, high-level knowledge structure and framework used to capture user requirements. to guide data collection for and design of the EDSA demand analysis dashboard.* ---------- 113

Figure 2: *Structure of SARO - the Skills and Recruitment Ontology.* ---------- 117

Figure 3: *Concept specification for SARO - see Figure 2.* ---------- 117

Figure 4: *Legend for ontology in Figure 2 and Figure 3.* ---------- 118

Figure 5: *Populated sample based on the SARO ontology.* ---------- 120

Figure 6: *Design sketch for the Demand Dashboard Intro View.* ---------- 121

Figure 7: *User & Task Models.* ---------- 122

Figure 8: *Methodology followed to explore the design space for the EDSA demand dashboard.* ------ 122

Figure 9: *Design sketch for the Policy-Maker Perspective.* ---------- 125

Figure 10: *Design sketch for the Job Seeker/Trainee Perspective.* ---------- 126

Figure 11: *Design sketch for the Skills Profile Form.* ---------- 127

Figure 12: *Design sketch for the Expert/Practitioner Perspective.* ---------- 128

Figure 13: *Requirements for Data Capture & Analysis.* ---------- 129

Figure 14: *Data acquisition and enrichment pipeline* ---------- 130

Figure 15: *Wikification for a job posting* ---------- 131

Figure 16: *Output from the JSI Wikifier* ---------- 131

Figure 17: *Output for the posting shown in Figure 16 encoded as RDF/XML and enriched with geolocation information* ---------- 133

Figure 18: *Sample SPARQL query. Note this has been formatted for readability - the SPARQL endpoint requires URL encoding to be posted successfully over HTTP.* ---------- 135

Figure 19: *Result of query in Figure 18. Note this is edited for readability to show values only, without dataTypes* ---------- 136

Figure 20: *Job demand overview for the policy-maker, showing the results for a search for the skill 'statistics'* ---------- 139

Figure 21: *The policy-maker view in the EDSA dashboard* ---------- 140

Figure 22: *In descending order, frequency of mention of skills that co-occur with those in the filter in Figure 21.* ---------- 141

Figure 23: *Further detail for the only job posting found in Portugal for the query in Figure 21- see also Figure 23.* ---------- 142

Figure 24: *Top job matches and learning resources (including conferences and other academic events) matching the skills selected in the job-seeker perspective, based on the search filters in Figure 21.* -- 142

Figure 25: *A posting of interest selected from the results of a filter overlaid on the overview for the SkillSet viewer.* ---------- 144

Figure 26: *Overview showing data aggregated by location and time for the 40 skills* ---------- 145

Figure 27: *Filter as for Figure 21 to retain only those aggregates with skills that co-occur with those of interest* ---------- 145

1. The Demand Analysis Dashboard

The second version of the demand analysis dashboard was released in month 18 of the project. This version incorporates feedback from the consortium obtained during demonstrations of the dashboard at conferences and meetings with the scientific community and policy-makers, and heuristic evaluation with experts in Data and Computer Science external to the EDSA consortium. The following subsections of the technical report, outline the dashboard design and summarise the outcomes of the initial evaluation.

The live version of the dashboard is hosted by the Knowledge Media Institute (KMI) at the Open University on the EDSA project's wesbite[67]. Additionally, we also host a beta version for update and interim testing[68]. The latter is pushed to the live link when key additions or upgrades are completed. This update procedure will persist to the end of the project.

[67] http://edsa-project.eu/resources/dashboard
[68] http://dashboard.edsa-project.eu/beta

2. Dashboard Design

2.1 User Specification

Information on user requirements were derived based on the interviews carried out with data science practitioners during the initial stages of the demand analysis reported in D1.2[69], and the outcomes of analysis with the initial data collected. This initial specification is further described in (Dadzie & Domingue, 2015)[70]. Figure 1 shows this captured as a knowledge framework and, to support reuse and extension, through conversion to a top-level and more detailed sub-ontologies (see also Figure 2 and 3).

Key to the redesign of the user interface (UI) is a move from data centric visual representations to user and task centric views. This is reflective of feedback collected from outside the EDSA consortium. Users expressed difficulty recognising the functionality available for exploration and analysis. Such users, unlike consortium members, were not provided with specific context knowledge to make optimal use of the original layouts that focused on providing a picture of demand from the data perspective of job postings. An additional feature in the revised design is a clearer image of the contribution of different components to the overall demand, including the perspective of practitioners in the field, which only partially overlaps the demand data provided by job postings.

Section A in Figure 1, shows a breakdown of the target user types. The current design replaces the presentation based on data and tools with a set of frames containing three perspectives focused on tasks typical of each of the three main targets within this scope:

1. The Policy-Maker (figure 9)
2. The Job Seeker/Trainee Data Scientist (figure 10)
3. The Expert / Practitioner Data Scientist (figure 12)

More details on these three user types and the tasks they would typically be expected to carry out using the dashboard can be found later in the report.

[69] http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-P-D12-FINAL.pdf
[70] Dadzie, A. and Domingue, J. (2015) Visual Exploration of Formal Requirements for Data Science Demand Analysis, Workshop: Visualizations and User Interfaces for Ontologies and Linked Data (VOILA 2015) at ISWC 2015, Bethlehem, Pennsylvania, USA.
Location of the publication- http://oro.open.ac.uk/44263/

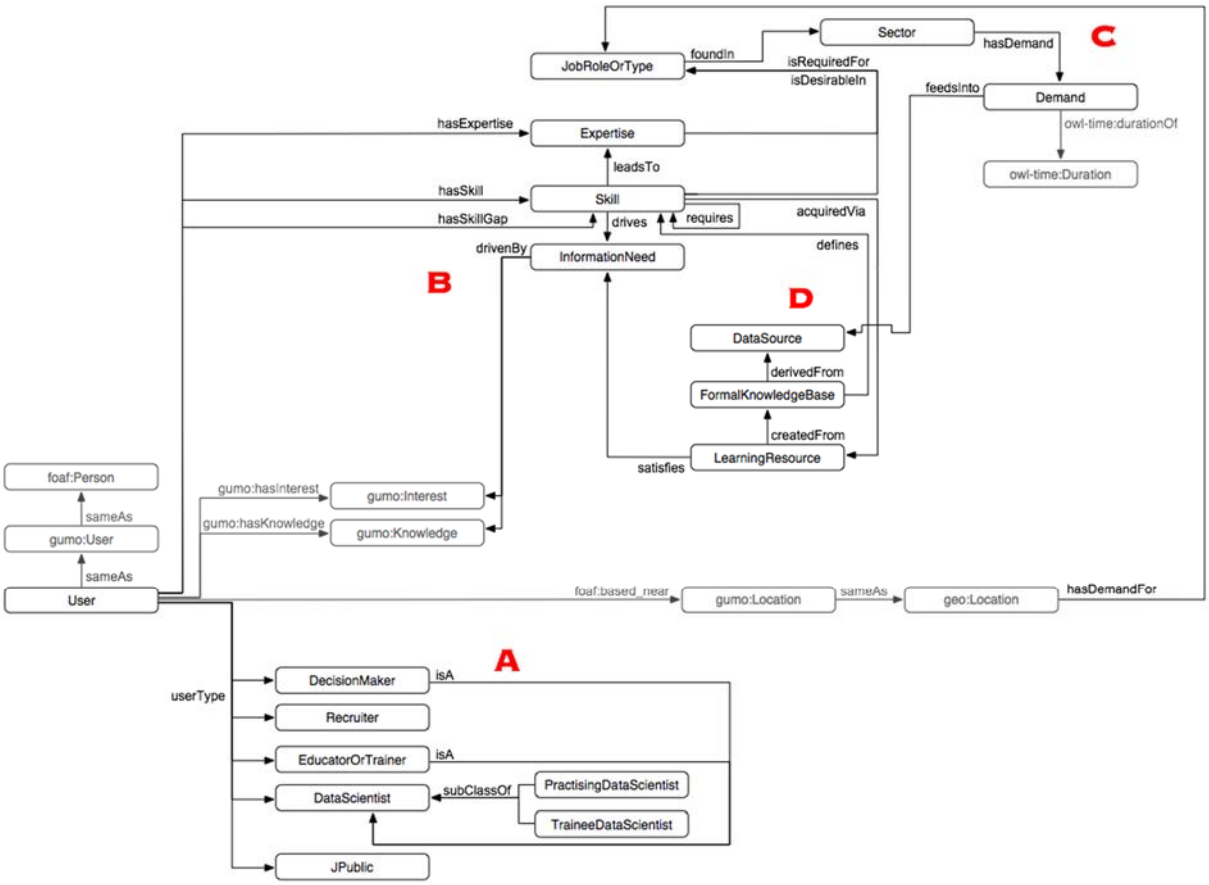


Figure 1: *initial, high-level knowledge structure and framework used to capture user requirements. to guide data collection for and design of the EDSA demand analysis dashboard.*

The initial data collection and analysis exercise highlighted additional requirements for data collection. Regions B in Figure 1 above, distinguishes requirements for mapping existing skills in the workforce from those for defining skill requirements for a specific role in a new job posting. Section 5 of this report discusses in more detail the requirements identified for data acquisition, both for reuse within the dashboard (region D in Figure 1) and for reuse and sharing beyond this use case.

2.2 The Knowledge Framework

Exploratory and guided detailed analysis of data science demand within the scope of the project dashboard follows an ontology-driven approach. This serves a number of purposes, including:

1. to provide a structured framework for capturing the design space (see section 2.1)
2. to guide data collection, linking to third party resources and reuse of both data and the results of analysis (see section 4)
3. to guide the discovery of related information hidden within the demand data and aid navigation through the data as size and complexity increase (see section 4)
4. to support translation of the input data into information and, subsequently, enriched, contextual knowledge, and therefore support effective answering of end users' questions (see section 4.1 and 5)

The initial ontology, based on the user and data specifications, used the base Unique Resource Identifier (URI) http://www.edsa-project.eu/edsa# aligning with the project title. This has since been extended

as SARO, the *Skills and Recruitment Ontology*, hosted at: http://eis.iai.uni-bonn.de/vocab/saro/index.html, with base URI http://vocab.cs.uni-bonn.de/saro# and prefix saro. Below we outline key concepts and relations in the current version - at April 2016. Note that the same relationships will be used to map SARO to previously extracted data, to follow best practice for data and ontology reuse and to support project goals to release openly EDSA data and results as linked data.

Branding the ontology as SARO increases reusability beyond the project scope, to include the more general skill and occupation classifications across industry and in domains other than data science, to support and facilitate the analysis of job function. The objective is to allow those high-level policy- and decision-makers in government and private enterprises
who advertise new job roles to reliably assess the value and relevance of specific skills to a domain or industry sector. This in turn supports the job seeker or practitioner in mapping a job role and description to their skill set, and where necessary identify the need to acquire new or upgrade existing skills and capabilities.

2.2.1 The Skills and Recruitment Ontology

SARO is being developed for practical use within the project dashboard, as outlined above, but additionally also for reuse:

1. to track demand for skills within a domain and/or industry sector and capability to meet this demand
2. as a support tool for policy and decision makers in recruitment and in-house training and development
3. for practitioners in a domain, trainees and the academic community and industry based training/service providers. Additionally, to support recognition of the need to retrain in existing and/or acquire new skills in order to remain relevant and competitive within industry.

At the top level (see Figure 2 and 3), SARO is centred around seven core concepts:

1. saro:User
2. saro:JobPosting
3. saro:Skill
4. saro:ProficiencyLevel
5. saro:Qualification
6. saro:Curriculum
7. saro:AwardingBody

saro:User

Of the original five user types (see Figure 1) we focus on three for a more generalised specialism/domain:

1. **saro:Practitioner and saro:Trainee** – as domain specialists, the saro:Practitioner is able to assess competencies with respect to a skill set within their industry sector, domain or job role. They are therefore well positioned to identify skill gaps and the direct impact of such gaps. This user type may also seek learning resources to help them develop or update their skills, either to increase effectiveness within their current role or to enable a transition to another role or domain.

2. **saro:EducatorOrTrainer** – develops learning resources for one or more related skills and competencies.
3. **saro:DecisionMaker** – within an enterprise, the decision-maker is responsible for the definition of new roles and corresponding *essential* and *desirable* skills. They will also influence training of new and existing employees. At a higher level the saro:DecisionMaker also refers to the policy maker who influences policy with respect to development of key industrial sectors, and therefore, allocation of resources for the training of key workers as part of a long-term strategy or to meet a specific, immediate need.

saro:JobPosting

This refers to a job advert listed by a specified so:hiringOrganization. While we restrict our analysis for practical implementation reasons to online postings, the definition covers any listing that includes the key concepts defined within a saro:JobPosting.

It extends the JobPosting concept in schema.org (prefix 'so'), and defines essential attributes including:

- The job role (jobRoleOrType)
- Job description (via so:description)
- Date posted (via so:datePosted)

saro:jobLocation is linked to the posting via the relationships: saro:hasDemandFor / saro:hasCapacity / saro:hasCapability and so:jobLocation - describing an address or local location description. To allow merging of the same location referred to by alternate names each posting also refers to a saro:geoLocationUri which points to a geoNames ID.

The saro:JobPosting also defines other concepts and relationships to allow the collection of other useful metadata including industry sector, salary and working hours.

saro:Skill

A saro:JobPosting links to a set of inferred and explicitly specified saro:Skills, using the relation saro:listsSkill or saro:requiresSkill. A specific saro:Skill may also link to another saro:Skill using saro:coOccursWith to define co-occurrence in a saro:JobPosting.

Another key relation is saro:frequencyOfMention, used to specify the number of mentions (occurrence) of a saro:Skill in a saro:JobPosting.

saro:Skill extends the *European Skills, Competences, Qualifications and Occupations* (ESCO) Ontology[71]; esco:Concept, which categorises skills or competencies as job-specific or transversal (cross-sector). SARO further extends these as:

[71] Smedt, Vrang, & Papantoniou, 2015 - ESCO Ontology: http://data.europa.eu/esco/model

- saro:JobSpecificSkill: representing technical or domain-specific skills related to a particular sector (we provide examples in ICT), further subclassed into:
    - saro:Product: competence using a particular product, e.g., *Hadoop*.
    - saro:Topic: capability in a domain and/or role-specific topic required to achieve observable result, e.g., *Data Analytics*.
    - saro:Tool: competence in the use of a tool specifically for carrying out technical tasks, for example, a specific programming language (e.g. *C*, *Java*) or database type (e.g.,*MySQL* - saro:Product of type *SQL*, *MongoDB* - saro:Product of type *NoSQL)*, or at a higher level of abstraction relational vs non-relational databases.
    - saro:TransversalSkill: sector and occupation-independent skills foundational to personal development, often referred to as "soft" skills, such as *team-work* and *communication*.

saro:ProficiencyLevel

The proficiency level for a saro:Skill. We consider the required proficiency level indicated in job postings a core concept as this is instrumental in identifying skills core to a sector or domain, and therefore, skill gaps and job market needs.

saro:Qualification

A saro:Practitioner or saro:Trainee may progress towards achieving a qualification in the form of formal certification awarded by an authoritative awarding body. Skills acquired on the job may also result in in-house, less formal, qualifications. SARO through ESCO also builds on the European Qualification Framework (EQF)[72], allowing reuse of EQF standards. Further, ESCO ensures traceability between qualification, awarding body and related occupations and skills/competencies.

saro:Curriculum

Formed based on the set of skills a learning institution aims to develop in its students.

saro:AwardingBody

An official or otherwise recognised institution certified to provide proof of the acquired skills and competencies in relation to a given standard following formal assessment.

[72] https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97

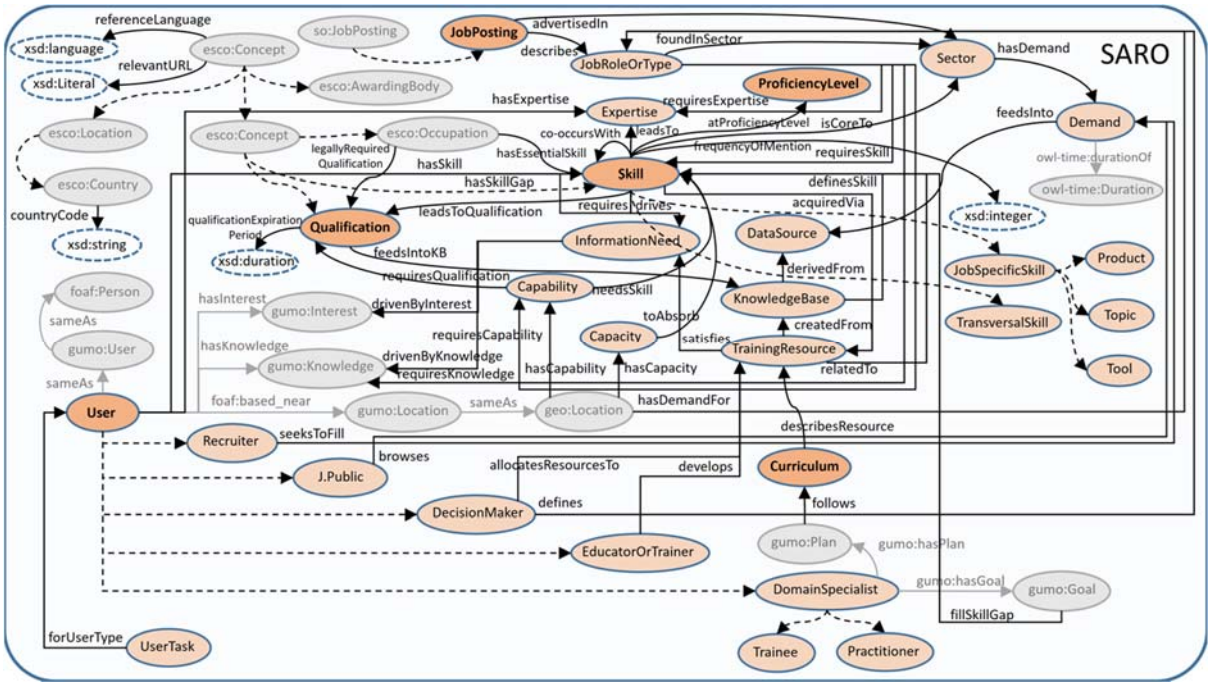


Figure 2: *Structure of SARO - the Skills and Recruitment Ontology.*

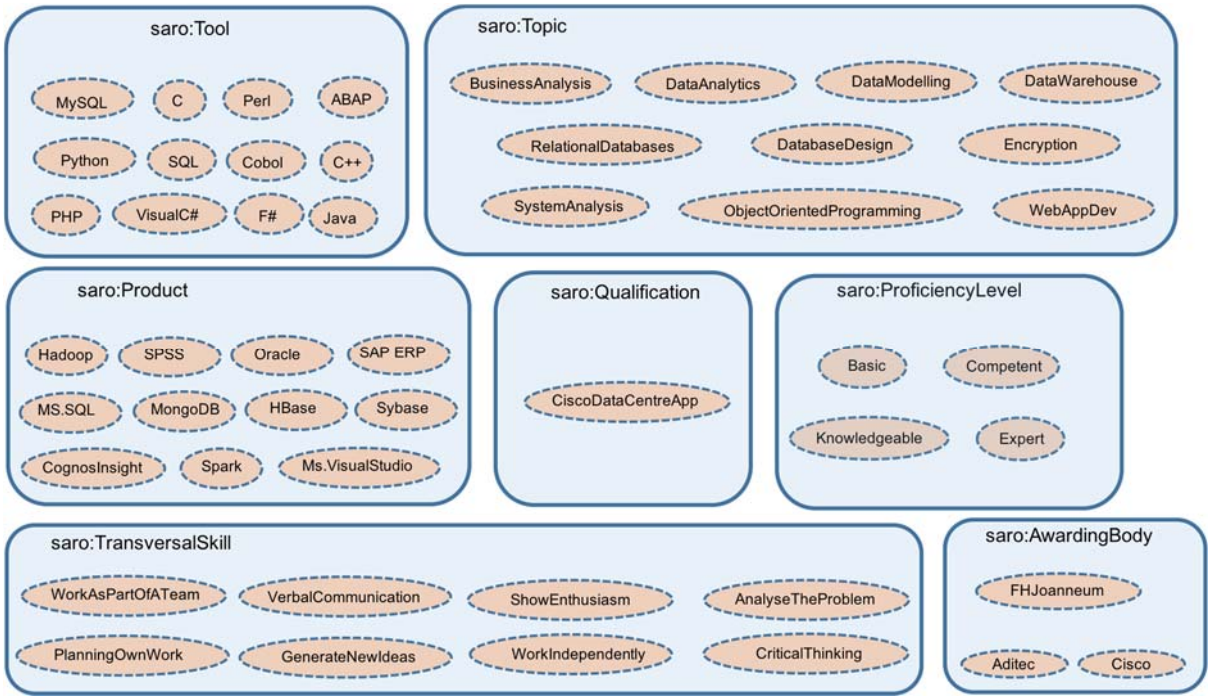


Figure 3: *Concept specification for SARO - see Figure 2.*

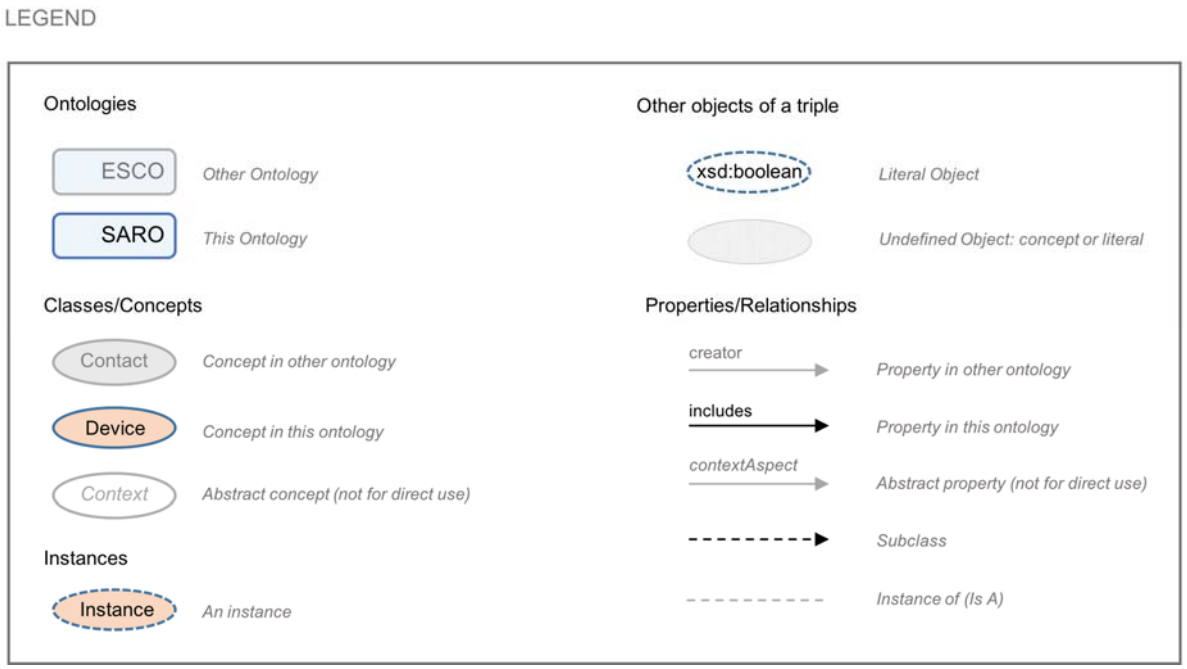


Figure 4: *Legend for ontology in Figure 2 and Figure 3.*

2.2.2 Third-Party Resources reused in SARO

The project aims to follow best practices in ontology design and data acquisition and reuse, as well as release as close to the gold and green open-access models as possible[73]. Further, to enable release of data and results as linked data, the ontology design process and modelling strategy required reuse of relevant, existing models, standard vocabularies and ontologies, as well as other resources.

In the past two decades several attempts have been made to design knowledge models representing information. The ESCO ontology focuses on the EU labour market, describing skills and qualifications specific to the region. ESCO covers a large subset of the project's requirements, bar the representation of real job postings. The latter are however defined by *Schema.org*, which also defines the skill, occupation and qualification concepts. We therefore reuse directly or extend where appropriate concepts and relations in ESCO and Schema.org, to model more fully the domain knowledge.

Additional models that fed into SARO are the *European e-Competence Framework* (e-CF)[74], the UK-based *Labor Market Information for All* (LMI4All) database[75], and extending beyond the EU, the US Department of Labor's *Occupational Information Network* (O*NET) project[76]. SARO was designed to provide a comprehensive representation of the knowledge required to define and interpret job postings in the context of skills, competencies and qualifications needed to fulfil a role.

[73] http://www.ncl.ac.uk/openaccess/green-gold
[74] https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97
[75] LMI4All: http://www.lmiforall.org.uk
[76] O*NET OnLine: https://www.onetonline.org

2.2.3 Example of Use- Skill Correlation

SARO serves also to guide use of the resources it describes; a key element of the skills analysis feeding into the construction of SARO is to determine skill correlation and ranking within skill sets overall, and in a named domain or industry sector. Therefore, as a first step we use skill (term) frequency and co-occurrence within each posting and the complete dataset to weight skills. Figure 5 illustrates the extraction of metadata for a job posting, annotated to highlight skill frequency.

<http://www.edsa-project.eu/jobposting/**JobPosting_ID_1**>

a
<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#JobPosting> ;

<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#describes>
<http://www.edsa-project.eu/jobposting/JobPosting_ID_1/Database%20and%20ETL%20Consultant%20-%20BODS,%20SQL%20Server> ;

<http://schema.org/**jobLocation**> "Manchester, UK" ;

<http://schema.org/**hiringOrganization**> "Venturi Limited" ;

<http://schema.org/**datePosted**> "2014-07-08 02:49:39" .

<http://www.edsa-project.eu/jobposting/JobPosting_ID_1/Database%20and%20ETL%20Consultant%20-%20BODS,%20SQL%20Server>

a
<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#JobRole> ;

<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#**requiresSkill**>
< **http://www.edsa-project.eu/skill**/JobPosting_ID_1/SQL>, < http://www.edsa-project.eu/skill/JobPosting_ID_1/MongoDB> .

< **http://www.edsa-project.eu/skill**/JobPosting_ID_1/SQL>

a <http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#Tool> ;

<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#**frequencyOfMention**> "**3**" .

< **http://www.edsa-project.eu/skill**/JobPosting_ID_1/Hadoop>

a
<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#Product> ;

<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#**frequencyOfMention**> "**2**" .

< **http://www.edsa-project.eu/skill**/JobPosting_ID_1/MongoDB>

a
<http://www.semanticweb.org/elisasibarani/ontologies/2016/0/saro_ontology_populated_instances#Product> ;

Figure 5: *Populated sample based on the SARO ontology.*

3. User and Task Perspectives

Figure 6 below, shows the design sketch of the introduction page for the second version of the dashboard. This provides a minimalist overview of the demand landscape using a map view coupled with a timeline. Colour and size are used to encode density and type, based on skill of demand per location. This is accompanied by a brief text description of the dashboard - data and functionality, and links to the three user-centred perspectives and a fourth pane with links to key modules and a data browser.

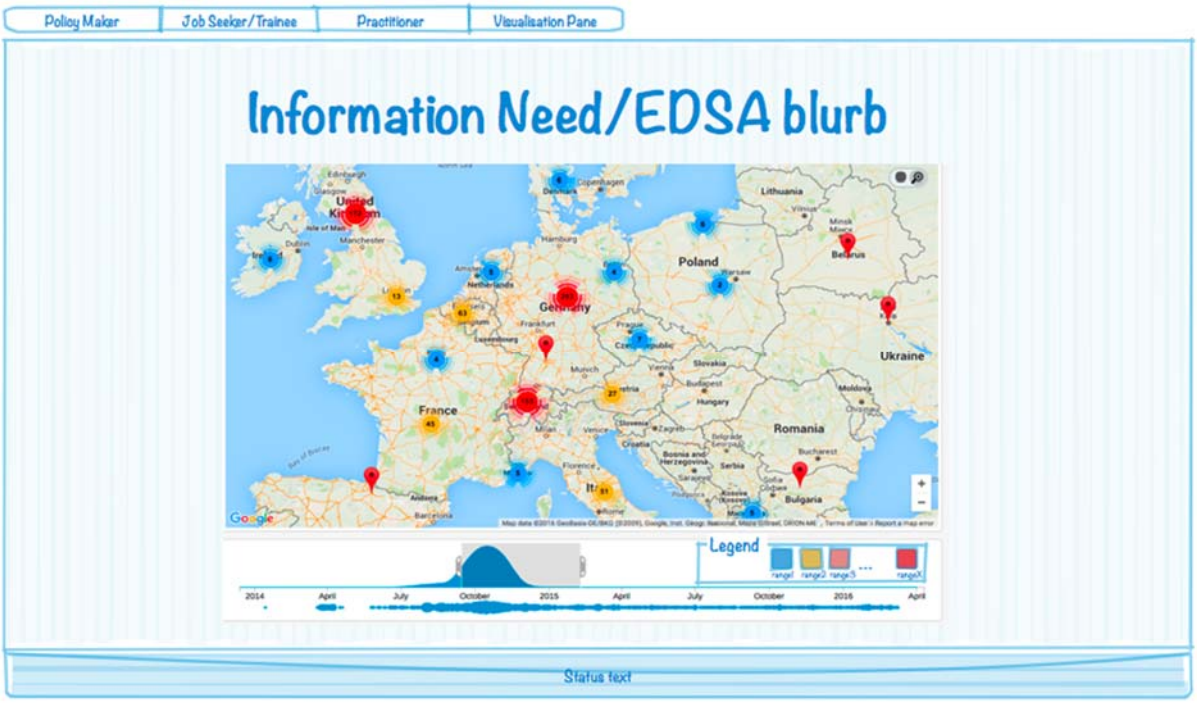


Figure 6: *Design sketch for the Demand Dashboard Intro View.*

Figure 7 provides more detail for the user design first seen in region A, Figure 1, with the information seeking tasks the dashboard supports for the key target user types. The EDSA knowledge framework focuses on two key concepts - JobPosting and Skill. The dashboard focuses on the requirements of three of the five target UserTypes defined; each of Figure 9, Figure 10 and Figure 12 sketch the perspective provided for the policy/decision-maker, the Job seeker/trainee and the Data Scientist, illustrating options for exploring demand trends using closely coupled modules. In each view, the central area at point of loading contains the module that addresses the key user requirement identified during ongoing research, and links to additional modules for detailed analysis of regions of interest and selected data.

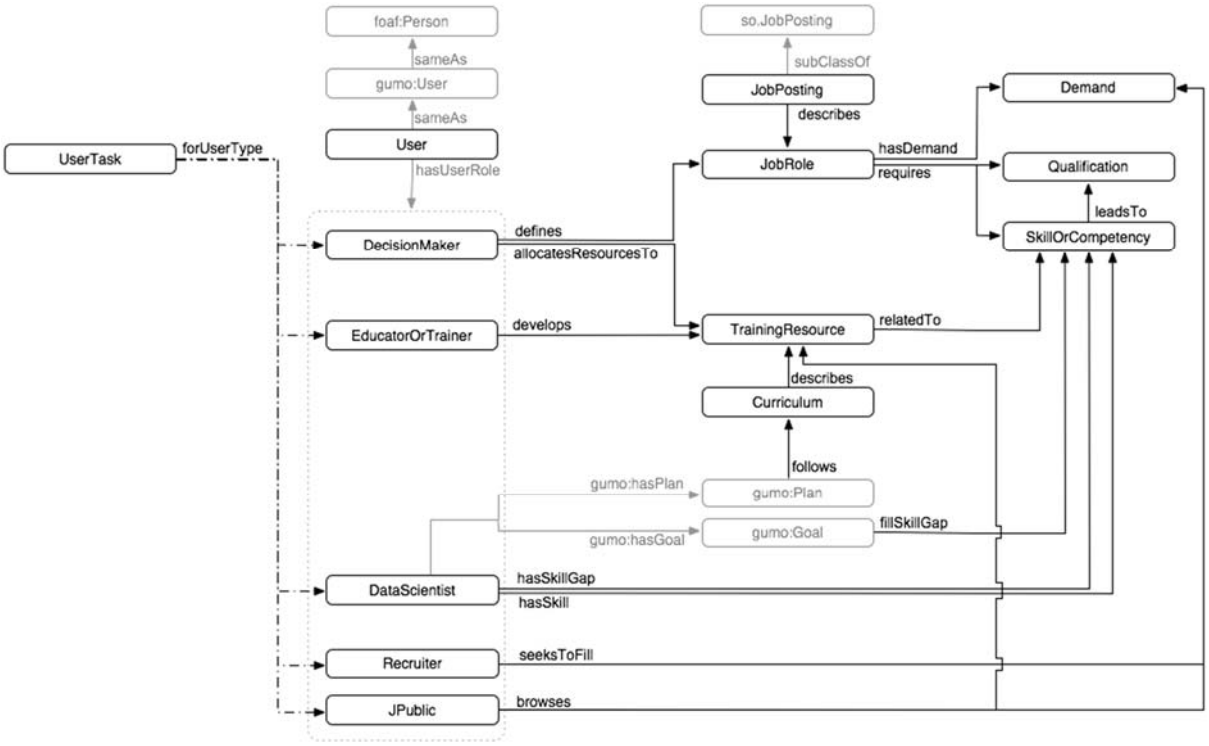


Figure 7: *User & Task Models.*

All visual analysis modules contributed to the dashboard must expose features for exploring the demand data along at least one of the following indicators (facets):

1. time
2. location

Other indicators of interest (see also Figure 13) are:

1. skill
2. domain / industry sector
3. (working) language; this typically maps to geographical location

A number of visualisation options have been, and continue to be explored for use in all perspectives, following the methodology in Figure 8.

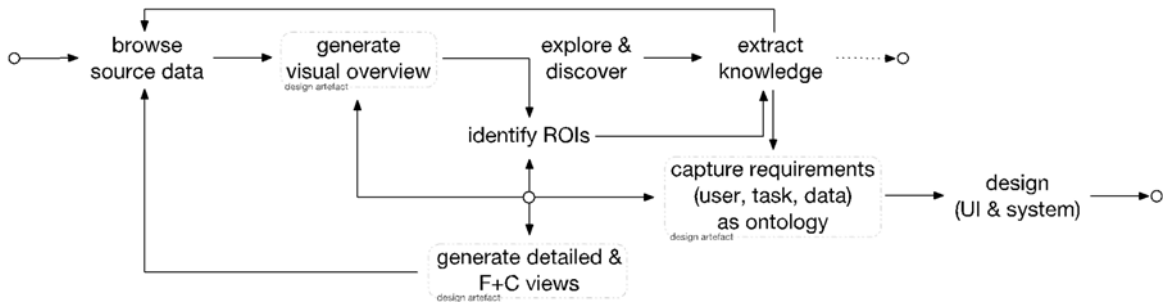


Figure 8: *Methodology followed to explore the design space for the EDSA demand dashboard.*

3.1 Map Views

Available in all views but at different levels of detail, the map layouts provide overviews which display locations of new jobs posted. Beyond a size threshold, data is aggregated by time or other user-selected options such as skill or skillset, with the level of detail for location mapped to semantic zoom.

The user may focus on the details for a region of interest on the map, or by switching to another visualisation type using selected indicators, with additional functionality to discover more detailed analysis dependent on indicator and visualisation type.

3.2 Timeline Views

A basic timeline slider filter may be built into or coupled with any module through the intermediary component, providing interactive exploration of trends for an indicator of interest over time.

Additional standalone timeline modules explored included the use of:

- Line graphs or dot plots exploring a small number of data attributes (dimensions)
- Spatio-temporal views using ThemeRivers [77][78][79]
- High-dimensional analysis using parallel coordinates; to allow an infinite number of dimensions to be visualised simultaneously, restricted only by screen real estate and processing power[80][81][82][83]

3.3 Statistical Analysis Views

In each user view, support for simple statistical analysis is provided using basic charts and pre-specified calculators, for example, job posting counts per location or skill type.

3.4 Intermediary Component - Switching between Modules

[77] G. Sun, Y. Wu, S. Liu, T. Q. Peng, J. J. H. Zhu and R. Liang, "EvoRiver: Visual Analysis of Topic Coopetition on Social Media," in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1753-1762, Dec. 31 2014. doi: 10.1109/TVCG.2014.2346919

[78] L. Byron and M. Wattenberg, "Stacked Graphs – Geometry & Aesthetics," in IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pp. 1245-1252, Nov.-Dec. 2008. doi: 10.1109/TVCG.2008.166

[79] Susan Havre, Elizabeth Hetzler, Paul Whitney, Lucy Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 9-20, January-March, 2002 doi: 10.1109/2945.981848

[80] Inselberg, Alfred, The plane with parallel coordinates. The Visual Computer 1(2): 69-91 (1985) doi: 10.1007/BF01898350

[81] A. Inselberg, "Visualization and knowledge discovery for high dimensional data," User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on, Zurich, 2001, pp. 5-24. doi: 10.1109/UIDIS.2001.929921

[82] Inselberg, Alfred, Parallel Coordinates: Visual Multidimensional Geometry and Its Applications, Springer-Verlag New York, Inc. 2009. book - no doi

[83] C. K. Hung and A. Inselberg, "Visualizing Multidimensional Relations with Parallel Coordinates," Information Technology: Research and Education, 2006. ITRE '06. International Conference on, Tel-Aviv, 2006, pp. 261-265. doi: 10.1109/ITRE.2006.381579

While working in a selected pane or perspective, a different module may be brought to the fore either by explicit selection or as a result of user action. For example, selecting to view details for a specific region of interest may redraw the map using a different level of detail or switch to a statistical chart. Alternatively, choosing to carry out analysis of a skill set based on the results of a search in the policy maker perspective may switch to the practitioner view, as the latter hosts the skill centric analysis as the main module. The job seeker/trainee perspective on the other hand supports the construction of a skills profile (see also Figure 11) with the aid of a map showing distribution of job and skill demand. This module may be accessed from the policy maker and practitioner views after narrowing down to a selection of job postings and choosing the option to carry out a skill match.

To support transition between modules and views, the independently built modules are being coupled, to allow exchange of state and information, source data and the results of analysis. This is achieved in practice by requiring each module to communicate with an intermediary component, transparent to the end user.

The intermediary component exposes methods for passing source data and result sets between modules, and for retrieving additional data using object IDs mapped to data in the RDF store. Template SPARQL queries reduce latency in retrieving information from the store and also ensure that objects created in the independent modules are based on the concepts as defined in the underlying ontology; first to provide a simple method for validating data entries, and - second - to support the user in maintaining a consistent mental model as they explore the data and increase their understanding of the bigger picture and detail in regions of interest.

3.5 User Perspectives Design Sketches

Policy/Decision maker

The policy/decision-maker view is centred on the geographical landscape of demand, coupled with a time slider, with options for filtering by skill and location. The map is coupled with statistical charts plotting top skills, locations and the time range for the results. Additional widgets include a word cloud showing skill frequency of mention. Figure 9 displays the design sketch for the policy-maker perspective.

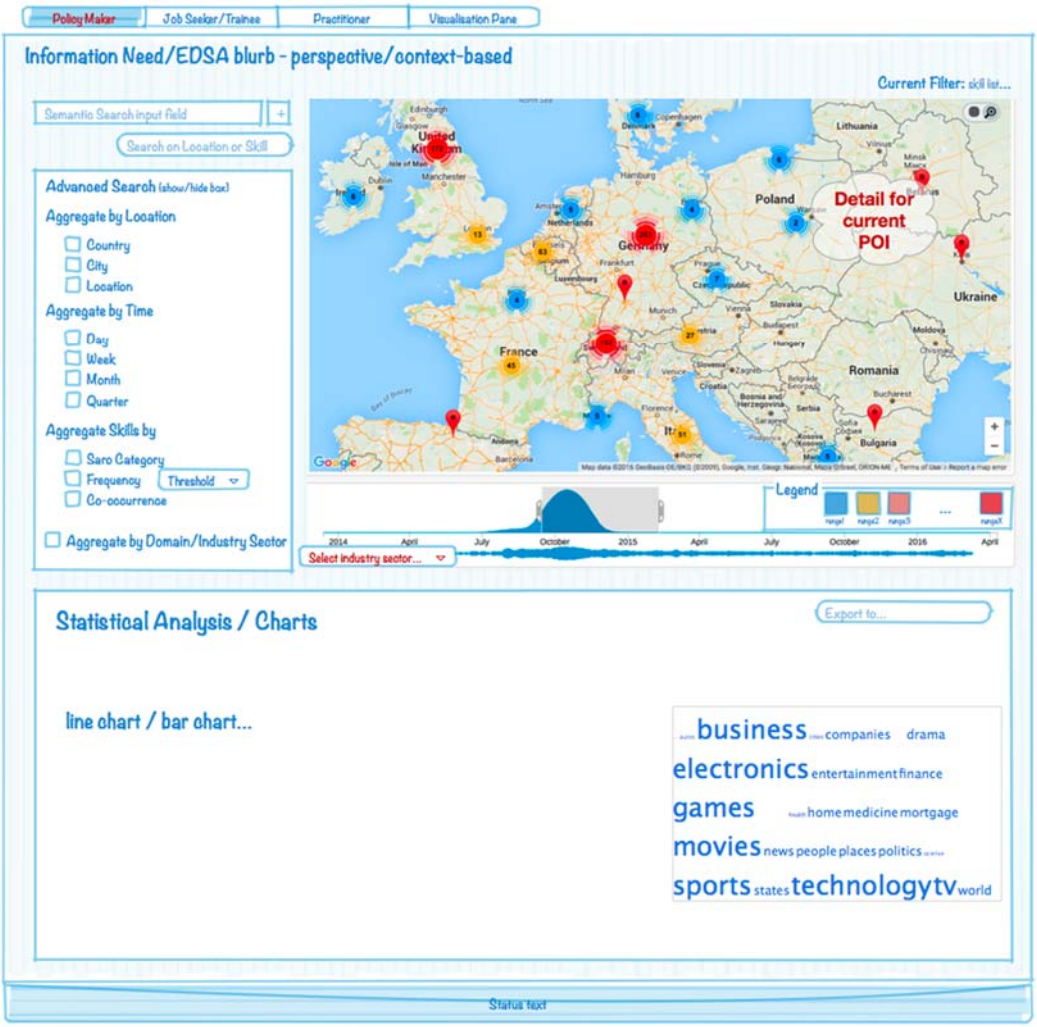


Figure 9: *Design sketch for the Policy-Maker Perspective.*

Trainee/Job Seeker

Similarly, the trainee/job-seeker view provides as a map of job demand as default, with functionality for filtering or drawing an overlay based on the skills profile generated from user input (see Figure 11). This in turn may be used to trigger the retrieval of detail for matching jobs and training courses to help the job seeker meet more closely the requirements for job roles of interest. Figure 10 displays the design sketch for the policy-maker perspective.

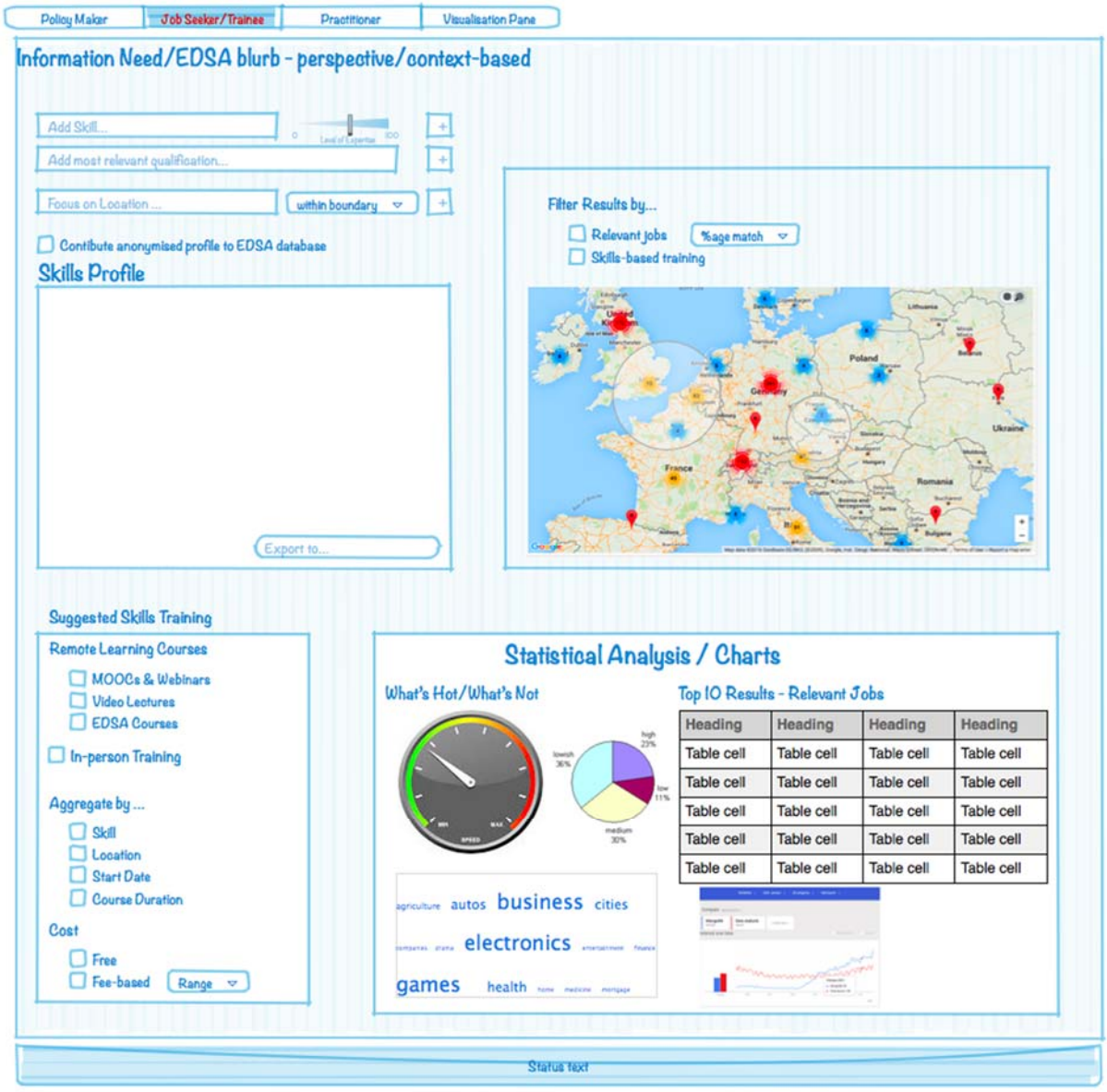


Figure 10: *Design sketch for the Job Seeker/Trainee Perspective.*

While the skills profile generator is aimed at the trainee/job seeker, we envisage use also by the practitioner seeking to review their skill base, or as an aid for decision makers at management level planning in-house training, expansion of, or creation of new roles and teams. This module development is currently in progress. The results of the evaluation with domain experts will contribute to determining which factors should feed into the generation of skills profiles. These factors are still expected to evolve with the picture of demand and changes in technology.

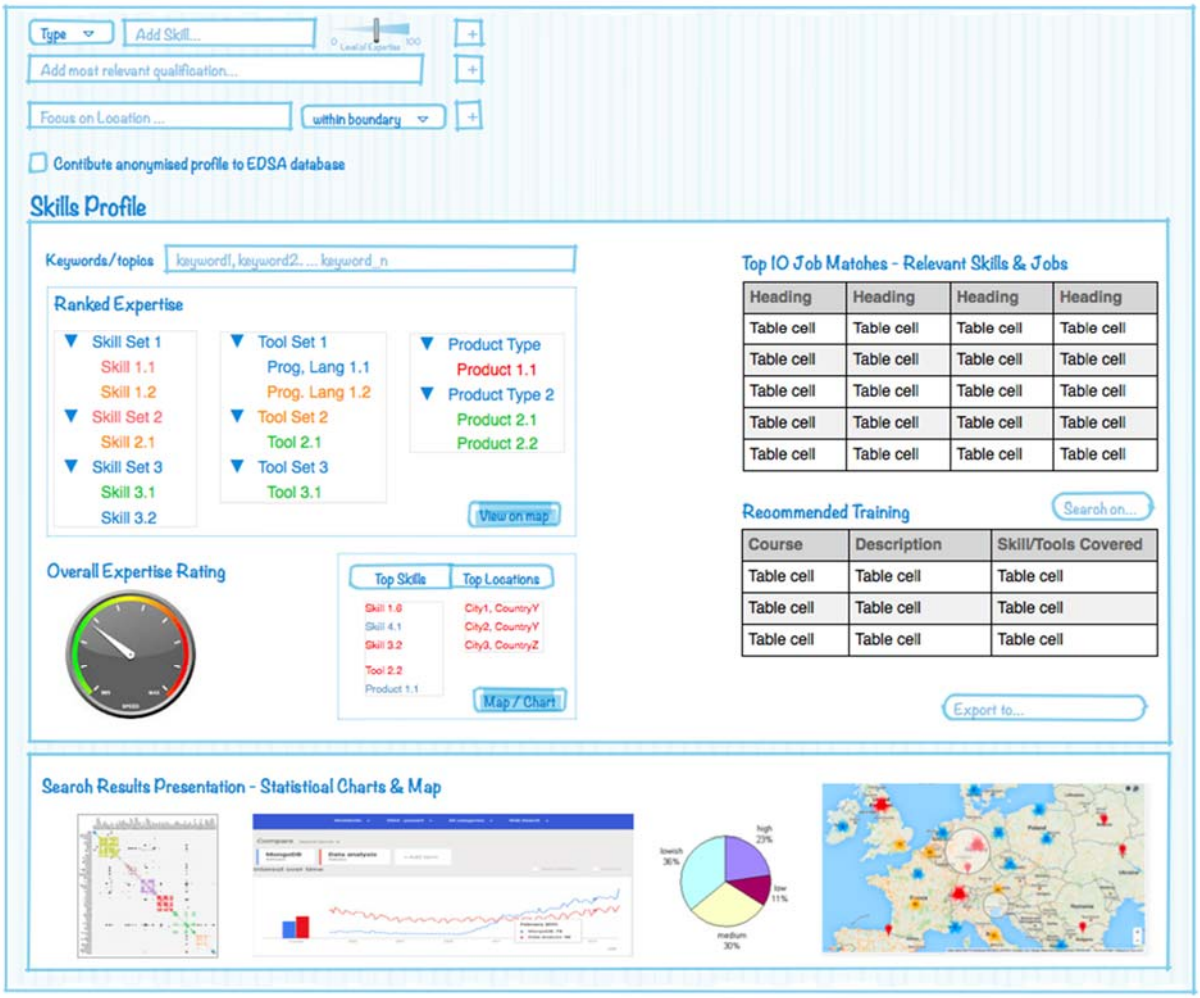


Figure 11: *Design sketch for the Skills Profile Form.*

Practitioner

The practitioner's view allows for collection of information to provide an alternative perspective on skill demand, to that obtained by only analysing job postings, in order to move closer to the ground truth in determining the picture of demand. This perspective therefore also provides functionality specifically for browsing and analysing skill occurrence and correlation within the demand data. Figure 12 displays the design sketch for the policy-maker perspective.

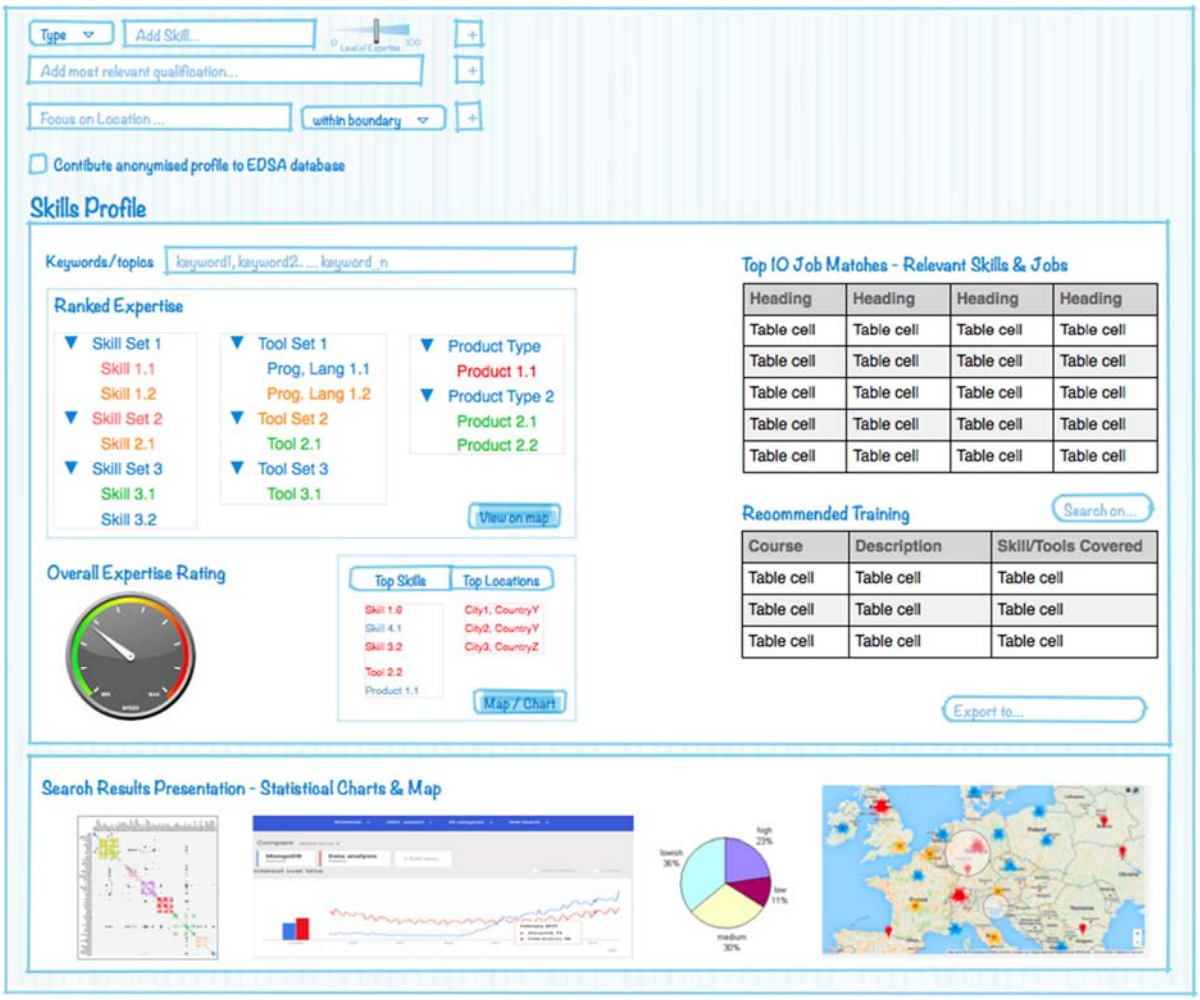


Figure 12: *Design sketch for the Expert/Practitioner Perspective.*

4. Demand Data

Figure 13 provides a top level view of the data requirements for the skill and job demand analysis, highlighting key information that must be collected with the job postings, including:

- Location of the job, to identify patterns due to geographical location and, correspondingly, working language/s.
- Duration of the post, to track temporal trends and windows in the data.
- Industry sector, and therefore potential demand impacts due to specific domain requirements, policy and organisational culture.

Additionally, the posting itself must contain sufficient detail to infer capability and capacity. Capability in this sense refers to the workforce available to fill these posts; capacity refers to the ability of the market to absorb a workforce in the named domain - in this case data science, as defined in the SARO ontology.

Based on the outcomes of the analysis, measures of the gap between capacity and capability of the job market are to be provided, with supporting evidence, addressing some key overall aims of the project, including:

- Providing guidance for policy makers in resource allocation to fill the skill gap, based on the overall demand landscape and the requirements of different, key industry sectors and geographical locations.
- Providing evidence to guide the design of new or adaptation of existing courses to fill the gap recognised.

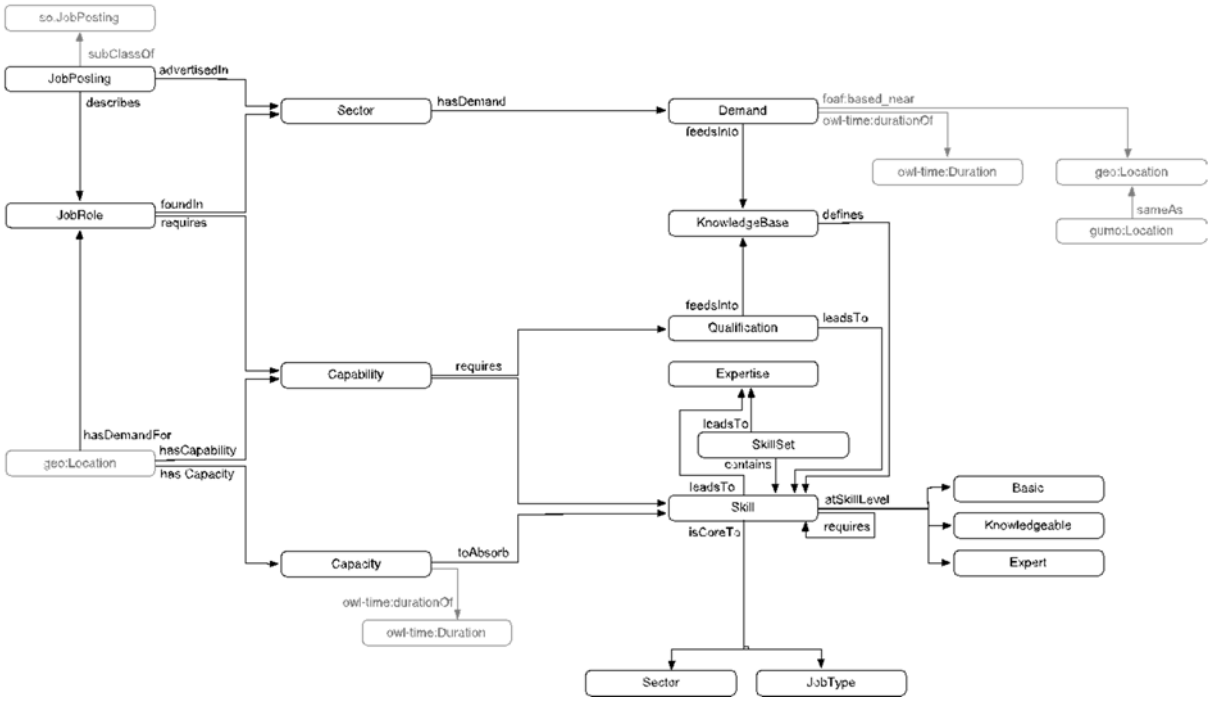


Figure 13: *Requirements for Data Capture & Analysis.*

4.1 Data Acquisition Pipeline

A number of job portals exist that aggregate job postings by location, sector or domain, job type, applicant qualifications and skill set or type. This includes domain specific portals and those targeting specific job types, such as jobs.ac.uk[84]. This portal advertises postings in academia and industrial research. More vertical portals cover a wide range of classifieds, such as Trovit[85], a leading search engine for classified ads in Europe and Latin America. Available in 13 languages, it provides a search engine for real estate, cars and other products as well as jobs.

We used those portals and specific employer sites to facilitate collecting the 'ground truth' data around the demand for data scientists, and also to evaluate the support we provide for analysis and management of big data.

Following the data acquisition and enrichment pipeline in Figure 14 we first mine data either using dedicated APIs such as the Adzuna API[86] or custom web crawlers. This is formatted as json, to aid further processing and enrichment.

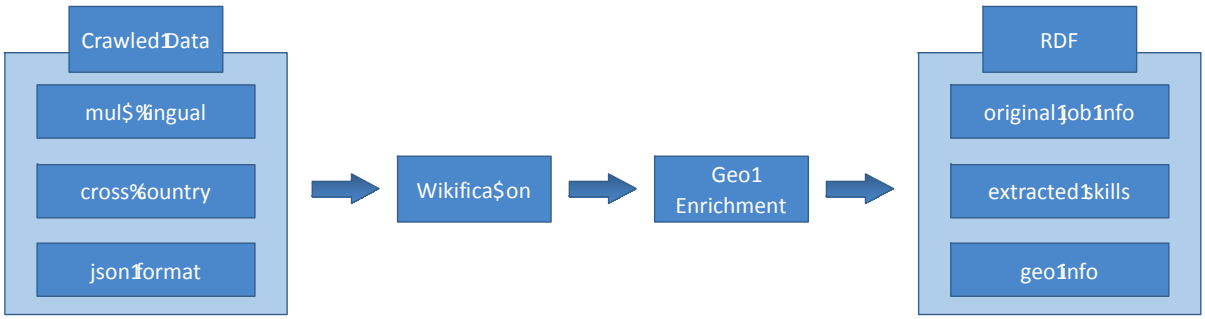


Figure 14: *Data acquisition and enrichment pipeline*

The next step in the pipeline, *wikification* - identifying and linking textual components to corresponding, disambiguated Wikipedia pages (Ratinov, Roth, Downey, & Anderson, 2011)[87], is carried out using the JSI Wikifier[88], developed at the Artificial Intelligence Lab of the Jožef Stefan Institute, University of Ljubljana, Slovenia[89]. Wikification of the data mined enabled large-scale semi-supervised text annotation. Further, the JSI Wikifier supports cross-linguality and multi-linguality for extracting and

[84] jobs.ac.uk
[85] http://www.trovit.com/
[86] https://developer.adzuna.com/overview
[87] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 1375-1384.
No doi - ACM url: http://dl.acm.org/citation.cfm?id=2002472.2002642
[88] http://wikifier.ijs.si
[89] http://ailab.ijs.si/

annotating relevant information from the postings in different languages across the EU. The results are aligned with corresponding concepts defined in SARO using name matching (compare Figure 16 with Figure 17).

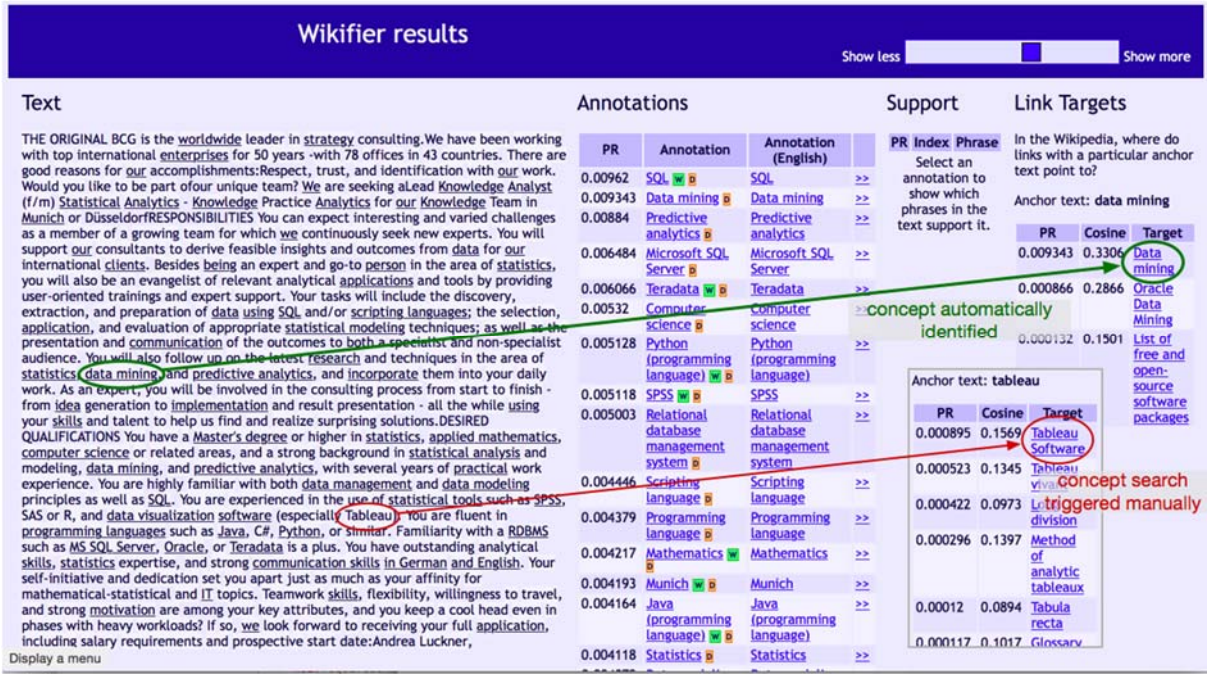


Figure 15: *Wikification for a job posting*

All concepts automatically identified are underlined in the text on the left, with values for PageRank (PR).The anchor text, "Tableau" (encircled in red), which was not automatically annotated, is used to trigger a concept search, which returns "Tableau software" as the top hit.

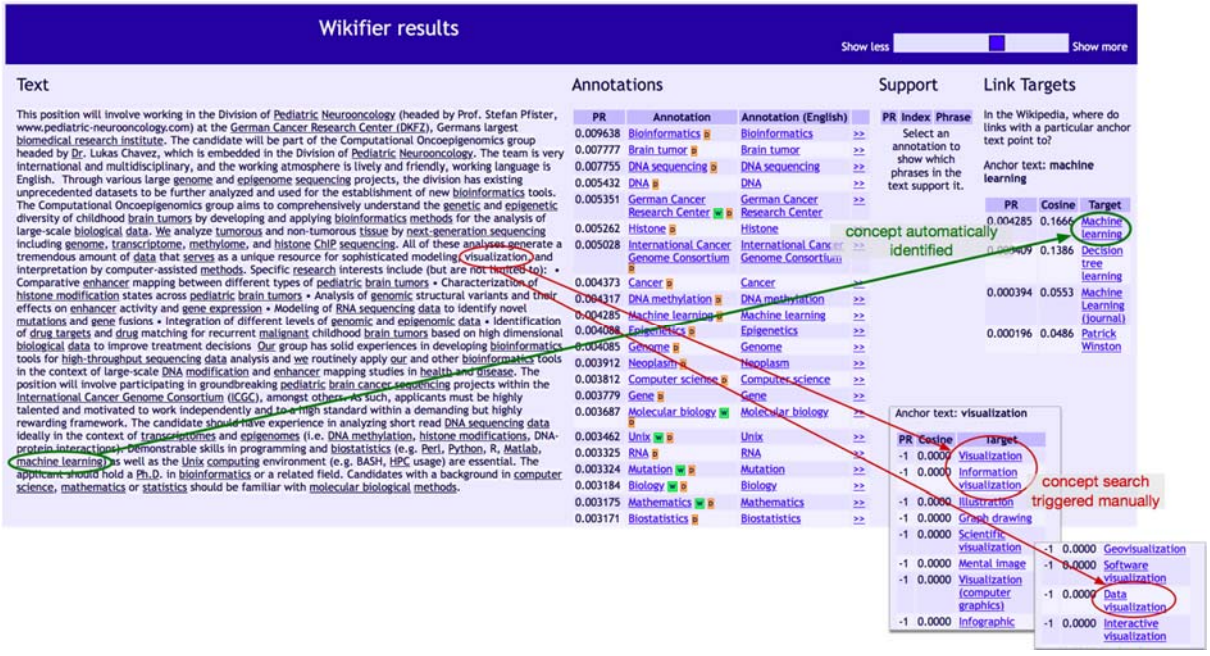


Figure 16: *Output from the JSI Wikifier*

The image shows an automatic identification of concepts for a job posting (e.g. encircled in green - "machine learning"); on the lower, right-hand corner the overlay shows additional concept recognition triggered by a manual search from "visualization" (encircled in red).

Finally, to support the identification of location sensitive trends, the postings are enriched using the GeoNames ontology[90], to include latitude and longitude and the unique corresponding GeoNames ID and location name - the output is encoded as RDF/XML.

While search portals tend to be restricted to operation in specified countries, coverage often extends beyond these, allowing the capture of posts advertised for countries other than those formally specified. Table 1 shows the total number of postings extracted for 22 out of the 28 countries in the EU. Of these countries, two have totals under five, and a further three under 50. As noted throughout the project, the UK dominates, at almost 90,000, more than 2.5 times greater than the next highest, France, at around 35,000.

Table 1: *Total number of job postings by country across all extraction all sources and extraction methods used between Nov 2015 and Jun 2016 - including historical data from 2013.*

Country	**Number of Job Postings Extracted**
Austria	4,694
Belgium	10,764
Bulgaria	44
Czech Republic	6,232
Denmark	7,026
Estonia	2
France	34,881
Germany	20,938
Hungary	7,225
Ireland	18,919
Italy	11,543
Malta	38
The Netherlands	9,571
Norway	33
Poland	18,159
Portugal	21,784
Romania	17,946
Slovakia	3
Spain	9,210
Sweden	13,870
Switzerland	14,616
United Kingdom	89,343
Total	**316,841**

[90] http://www.geonames.org/ontology/documentation.html

4.2 RDF Data Store

An RDF store using Jena Fuseki 1.1[91] is hosted at the Knowledge Media Institute (KMI) at the Open University, UK, with read only, open access . All data used in the demand and skills analysis for the dashboard will be stored in the single repository, to allow a single point from which to access data for analysis within the project and as made available from the dashboard. The server will be supported by the Open University beyond the project, with opportunity for continued support from the European Data Science Academy as part of work package 5.

Figure 17 shows the output for the posting being wikified in Figure 16, after enrichment and annotation[92].

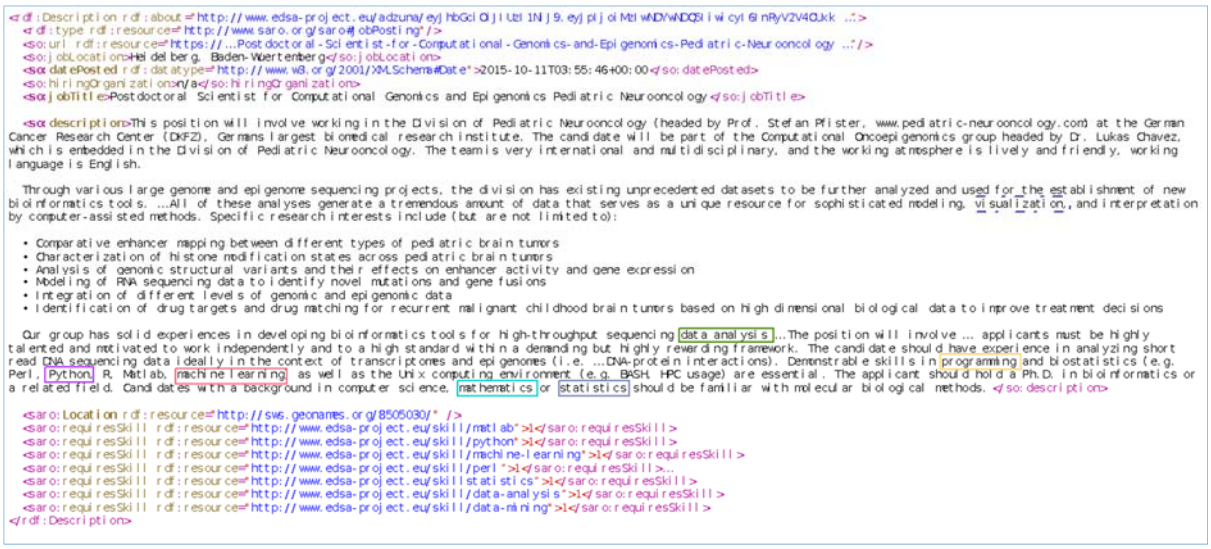


Figure 17: *Output for the posting shown in Figure 16 encoded as RDF/XML and enriched with geolocation information*

Latitude and longitude are not shown here - these, along with more geolocation detail may be retrieved on demand using the geonames ID. Recognised skills are extracted and annotated based on the ontology, specifying also frequency of mention. Other metadata extracted from the posting is similarly annotated. The colour-coded overlay highlights the skills annotated at the bottom of the extract. Note that the description text as shown is truncated.

The SPARQL endpoint is available at http://dashboard.edsa-project.eu:3030/data/databases/rdfstore/edsa/query.
The output may be rendered with a user-specified stylesheet and/or exported as text , json (default), xml or csv/tsv

[91] https://jena.apache.org/documentation/fuseki2/
[92] The detail for the full posting may be retrieved from the RDF store using the below complete ID, by querying the public SPARQL endpoint for the data store: http://www.edsa-project.eu/adzuna/eyJhbGciOiJIUzI1NiJ9.eyJpIjoiMzIwNDYwNDQ5IiwicyI6InRyV2V4OUxkUjR5Um1fMEpPemJHSUEifQ.TucWkRyC74dXpyaIjhBqCb24fwXNzw3og12RFZ0G0kI

A query "queryText", to be formatted as "outputFormat" using style sheet "stylesheetUri" would be posted as:

http://dashboard.edsa-project.eu:3030/data/databases/rdfstore/edsa/query?query=queryText&output=outputFormat&stylesheet=stylesheetUri

Figure 18 shows a query to retrieve detail for the job posting with URI:
http://www.edsa-project.eu/adzuna/eyJhbGciOiJIUzI1NiJ9.eyJpIjoiMzcwNjUzMjM5IiwicyI6IkNDMHRtTjVoU21TM3VkeWZnN04zcFEifQ.eNEsY3Wx7qJtDZW7VW_0fwVXVR8krUUV8E__NRAyi5I.

The results are shown in Figure 19 formatted as json, the default formatting returned if no value is specified for the output query parameter.

```
PREFIX  schema: <http://schema.org/>
PREFIX  geo:  <http://www.geonames.org/ontology#>
PREFIX  edsa: <http://www.edsa-project.eu/edsa#>
PREFIX  wgs:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT  ?jobPostingUri ?url ?datePosted ?skillUri ?jobTitle ?hiringOrganization ?jobLocation
?location ?description ?geoLocationUri ?alternateNameEn ?parentCountryUri ?countryCode ?countryName
?latitude ?longitude

WHERE {
?jobPostingUri rdf:type edsa:JobPosting .
?jobPostingUri schema:datePosted ?datePosted .
?jobPostingUri schema:jobTitle ?jobTitle
OPTIONAL { ?jobPostingUri schema:url ?url}
OPTIONAL { ?jobPostingUri schema:hiringOrganization ?hiringOrganization}
OPTIONAL { ?jobPostingUri edsa:requiresSkill ?skillUri}
?jobPostingUri schema:description ?description .
?jobPostingUri schema:jobLocation ?jobLocation .
?jobPostingUri edsa:Location ?geoLocationUri
OPTIONAL { ?geoLocationUri geo:name ?location}
OPTIONAL { ?geoLocationUri geo:parentCountry ?parentCountryUri}
OPTIONAL { ?parentCountryUri geo:name ?countryName}
OPTIONAL { ?geoLocationUri geo:countryCode ?countryCode}
OPTIONAL { ?geoLocationUri geo:alternateName ?alternateNameEn
FILTER langMatches(lang(?alternateNameEn), "en")
}
```

Figure 18: *Sample SPARQL query. Note this has been formatted for readability - the SPARQL endpoint requires URL encoding to be posted successfully over HTTP.*

```
{
  "head": {
    "vars": [ "jobPostingUri" , "url" , "datePosted" , "skillUri" , "jobTitle" , "hiringOrganization" , "jobLocation" , "location" , "description" , "geoLocationUri" , "alternateNameEn" , "parentCountryUri" , "countryCode" , "countryName" , "latitude" , "longitude" ]
  },
  "results": {
    "bindings": [
      {
        "jobPostingUri": { "http://www.edsa-project.eu/adzuna/eyJhbGciOiJIUzI1NiJ9.eyJpIjoiMzcwNjUzMjM5IiwicyI6IkNDMHRtTjVoU21TM3VkeWZnN04zcFEifQ.eNEsY3Wx7qJtDZW7VW_0fwVXVR8krUUV8E__NRAyi5I" } ,
        "datePosted": { "2016-03-24T19:52:18+00:00" } ,
        "skillUri": { "http://www.edsa-project.eu/skill/statistics" } ,
        "description": { "... - création et montage des dossiers commerciaux - analyse et réponse aux dossiers d'appels d'offres - suivi de l'activité commerciale ( tableaux de bords, reporting, )Poste à mi-temps ..." } ,
        "jobTitle": { "ASSISTANT ADMINISTRATIF ET COMMERCIAL (H/F)" } ,
        "hiringOrganization": { "Manpower" } ,
        "jobLocation": { "Aix-les-Bains, Chambéry" } ,
        "location": { "Aix-les-Bains" } ,
        "geoLocationUri": { "http://sws.geonames.org/3038350/" } ,
        "alternateNameEn": { "Aix-les-Bains" } ,
        "parentCountryUri": { "http://sws.geonames.org/3017382/" } ,
        "countryCode": { "FR" } ,
        "countryName": { "France" } ,
```


Figure : Result of query in REF _Ref452991701 \h Figure 18. Note this is edited for readability to show values only, without dataTypes

Figure 19: *Result of query in Figure 18. Note this is edited for readability to show values only, without dataTypes*

5. Ontology-Guided Visual Exploration, Analysis and Knowledge Acquisition in the Dashboard

5.1 Visual Analysis Tools and APIs

The web-based dashboard contains modules built based mainly using the JavaScript library D3.js for visual data manipulation. Sub-modules including some of the statistical charts were built using the Highcharts JavaScript library. Additional functionality makes use of, among others, jQuery and TopoJSON, the extension of GeoJSON. PHP is also used for server-side data parsing and processing.

In addition to parsing and processing, online custom Java software is used to pre-process the input data, among others, to aggregate the data to support interactive visualisation online and the application of filters and layers based on a number of pre-selected criteria, including skill, time and location.

5.2 Usability Evaluation - Beta Version June 2016

In this section we employ tasks 2 and 3 from the set up for the formal usability evaluation[93] to illustrate how the dashboard may be used to obtain a picture of demand. Demand both within the EDSA consortium, as analysis is carried out with this data to answer questions pertinent to the project, and by our target end users. We employ the results obtained in the pilot evaluation with two target end users self identified as working in data and computer science, in France and Spain, to test the functionality available in the working prototype.

Task 2 – Job Role Creation Activity
Task Description: You are relocating to another branch of your company and have been tasked with specifying the job role for your replacement:

- How would you phrase the job title?
- What skills would you expect in successful candidates?
- Where, within the EU, would you place your job advert in order to maximise your chances of finding the right candidate(s)?
- Write a tweet to advertise this position

Completing this task requires the user to browse the demand data starting from the policy maker perspective, switching between modules and views as appropriate.

Participants in the evaluation were required to:

1. List essential and desired skills
2. Identify suitable locations for placing the job advert
3. Provide a role description
4. Write a tweet advertising the new post

[93] Full questionnaire available https://www.dropbox.com/s/ei4fqkoda3vu42a/questionnaire_formal_usability_evaluation_complete.pdf?dl=0

The aim of this task is to identify where and how the dashboard may be used to support:

- The identification of skills seen as core to a job type or role
- How capability varies by location, and potentially language
- Identifying gaps in demand within the job market, typically seen by policy-makers, and practitioners

Task 3 – Job Seeking Activity

Task Description: Your current employment contract is coming to an end and you are looking for new opportunities. Bearing in mind your ability to relocate, where are you most likely to find your ideal job?

- Is this impacted by restrictions in your ability to relocate?
- Would support for retraining change your options?

Completing this task requires the user to browse the demand data starting from the job seeker/trainee perspective, or potentially from the practitioner perspective for end users with significant experience in the field. As per Task 2, end users would be expected to switch between modules and perspectives to carry out whatever exploration or more detailed analysis is necessary to complete the task. To complete this task, participants were required to:

- Identify their top skills and top job matches based on these and other criteria set, including location and when the search would have been carried out
- Identify any training required for updating existing or acquiring new skills listed as essential or desired for their top job matches

The aim of this task is to identify where and how the dashboard may be used to obtain, as per the previous task, the picture of demand across the region. Further, this task examined also tool specific functionality including:

- The setup of filters, usefulness and usability of results
- Usefulness of different search types - one or all of keyword search, skill (as specified terms), posting location and date posted
- Usefulness of data summaries and ability to extract detail of individual postings
- Ability to compare posting content based on skills requirements and demand across time and location

5.3 Analysing Data Science Skills Demand

Starting with Task 2, the Job Role Creation Activity, Figure 20 shows the results of a search for the second of three skills - *statistics* - listed as essential for a high level research job in data science. To replicate the view in figure 20 with updated dashboard data, users should search for the skill "statistics" through the dashboard's policy-maker view[94]. *Statistics* is, incidentally, the second most frequently mentioned skill across the complete dataset.

[94] http://jobs.videolectures.net/policymakers

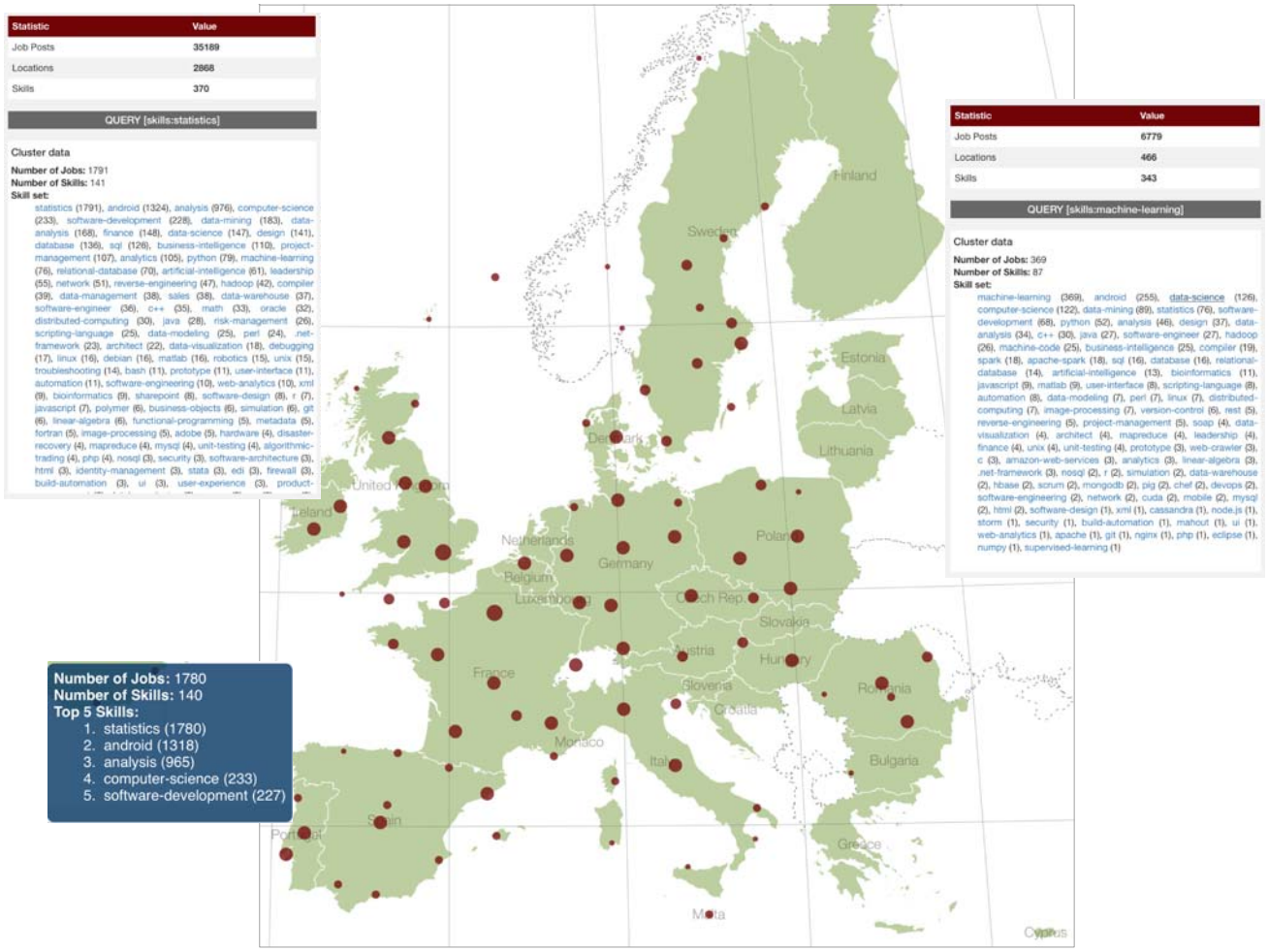


Figure 20: *Job demand overview for the policy-maker, showing the results for a search for the skill 'statistics' The overlay at the bottom shows a summary for a cluster of interest, and that at the top, left, shows further detail for all other co-located job postings along with statistics for skill mention. The overlay top, right, shows detail for a search for 'machine learning', with a much smaller result set.*

Each point on the map represents an aggregate sized based on job count for the city location. The layout in Figure 20 shows relatively broad mention of *statistics* across the EU. Hovering over each cluster shows job count, total number of skills mentioned in the postings and the top five skills required for jobs in the aggregate.

Overlaid also (top right) on the map is the summary for this cluster by searching for the first skill listed as essential - *machine learning* only; while distribution is similar for *statistics* this represents a much smaller result set - almost 7000 posts as opposed to over 35,000 respectively. The number of co-occurring skills for each is however similar, 343 to 370 respectively, out of a total of 537 skills recognised in the complete dataset.

The main search in Figure 20 is further narrowed down in Figure 21, by adding the skills *machine learning* (essential) and *data visualization* (desired) to the previous search. The resulting data matches with 18 locations across Ireland, Italy, Portugal and the UK. Overlaid on the map are the popups for three clusters, including that with the focus in Figure 20. The job match count in the focus cluster (in central Portugal) has gone down from 1784 to one, which requires, in addition to those in the query, five additional skills.

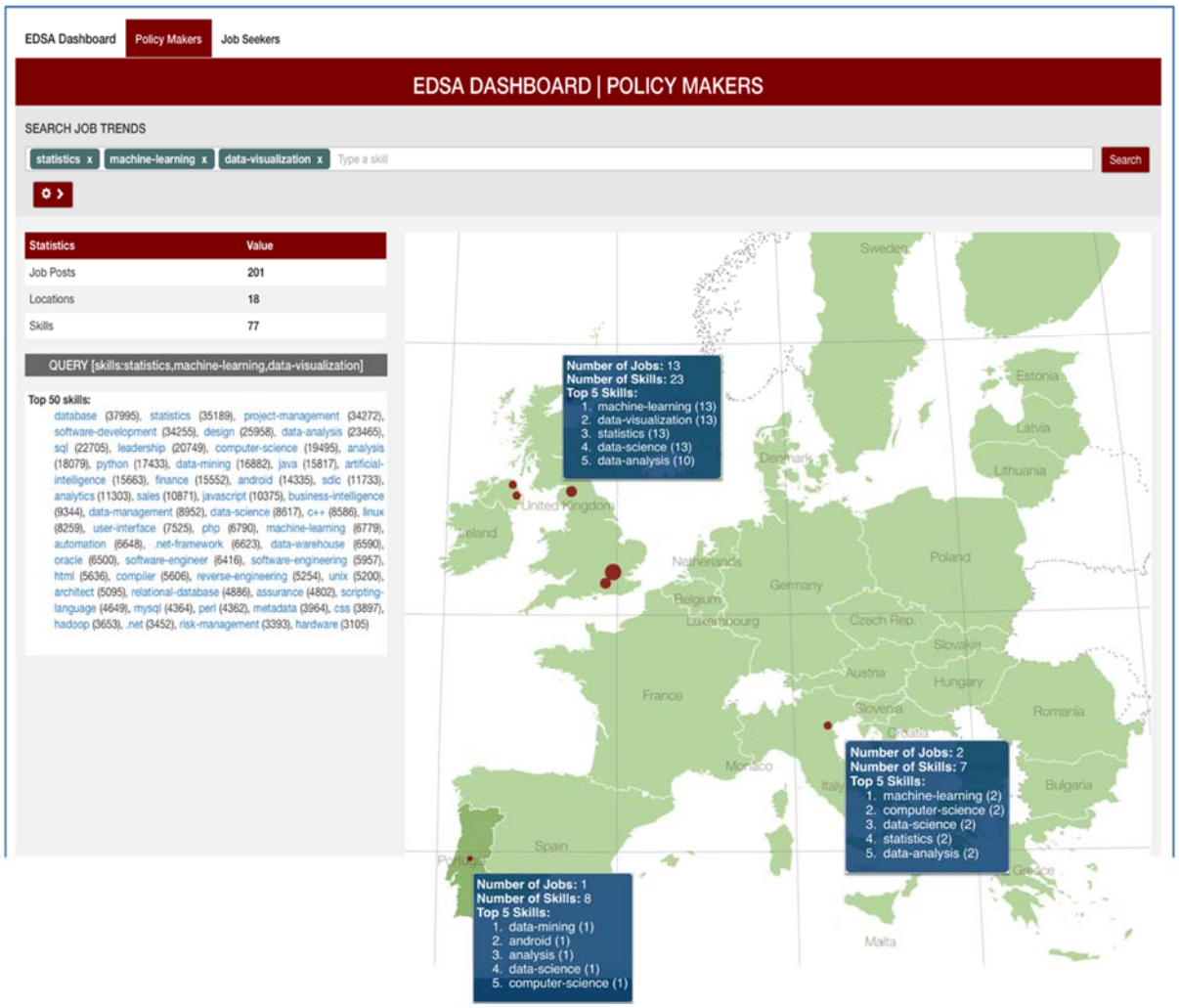


Figure 21: *The policy-maker view in the EDSA dashboard*
This view shows summary information for three matches for the query for the co-occurring skills: 'statistics', 'machine learning' and 'data visualization' - a total of 201 posts in 18 locations across four countries.

Of the 77 co-occurring skills for this query, the top *n* are shown in a coupled histogram (see Figure 22). Relatively high co-occurrence is seen for the three skills of interest, falling gradually for the next six. The following ten skill records mention just under half that seen for the top three, after which skill mention falls away in a long tail. Cross-referencing co-occurring skills at the top end of the scale with others listed as essential or desirable for the new role, in this case *big data* technologies and *mathematical modelling*, or variants thereof, provides some indication of the probability of successfully filling this position and also in which regions advertising the post would be likely to attract the strongest candidates.

An alternative tool for carrying out more detailed skills analysis is the SkillSet viewer described in Figure 25.

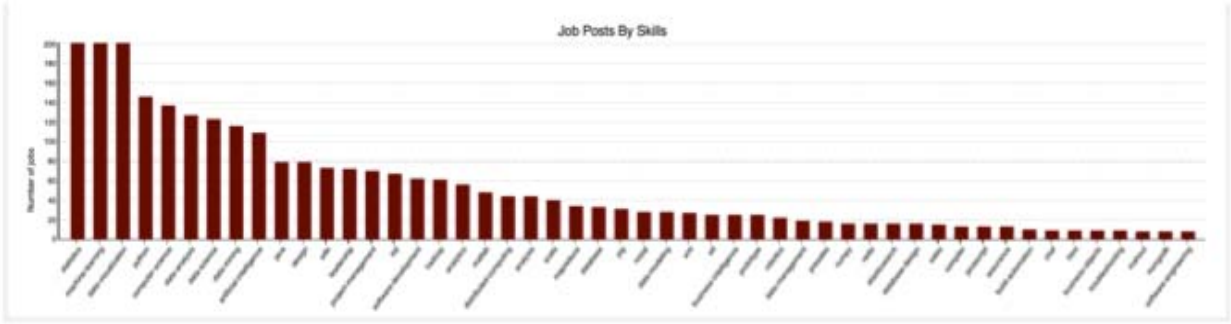


Figure 22: *In descending order, frequency of mention of skills that co-occur with those in the filter in Figure 21.*

Summarising job descriptions in a tweet allowed us to examine what participants saw as key information, providing more insight into skills seen as core to a role. Examining the tweet written to advertise this post we see first soft, then technical skills listed as required for the position - the latter summarised as *data science*; a more specific skill list is provided in the participant feedback, the top three for which direct matches were found in the current data set are used to formulate the search in Figure 21.

> *"[CompanyNameMasked] is looking for a VP of Research. Excellent leadership and communication skills expected, along with important research experience and wide technical versatility in data science.*"

The second tweet, for a different position, reads:

> *"We are hiring! Ontologist in London. More information http://……*

Location information in the first may be derived from the company name. The second explicitly mentions *London*, to reflect the preponderance of results in the UK and London specifically. However, both participants, working outside the UK, noted that the job distribution did not match their knowledge of the field, each expecting to see proportionally larger numbers of postings in their "home" countries. It should be noted that the current June 2016 dataset covers significantly more of the EU than at the time of the pilot, so that a much smaller skew is seen in these snapshots.

Switching to the job-seeker perspective to address the second task (Task3– Job Seeking Activity) we obtain, along with the results on the map, details for matching job posts and links to learning resources matching top skills and other related information sources such as conferences (see Figure 23 and Figure 24).

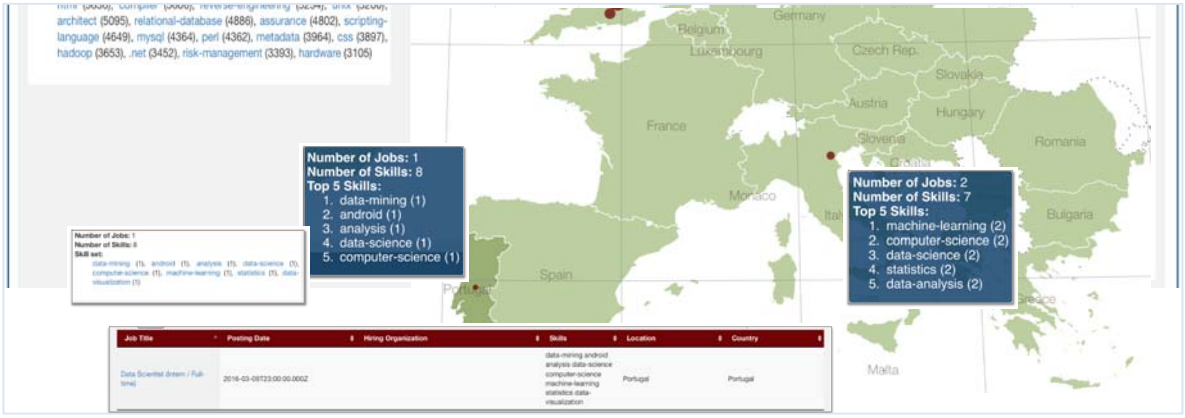


Figure 23: *Further detail for the only job posting found in Portugal for the query in Figure 21- see also Figure 23.*

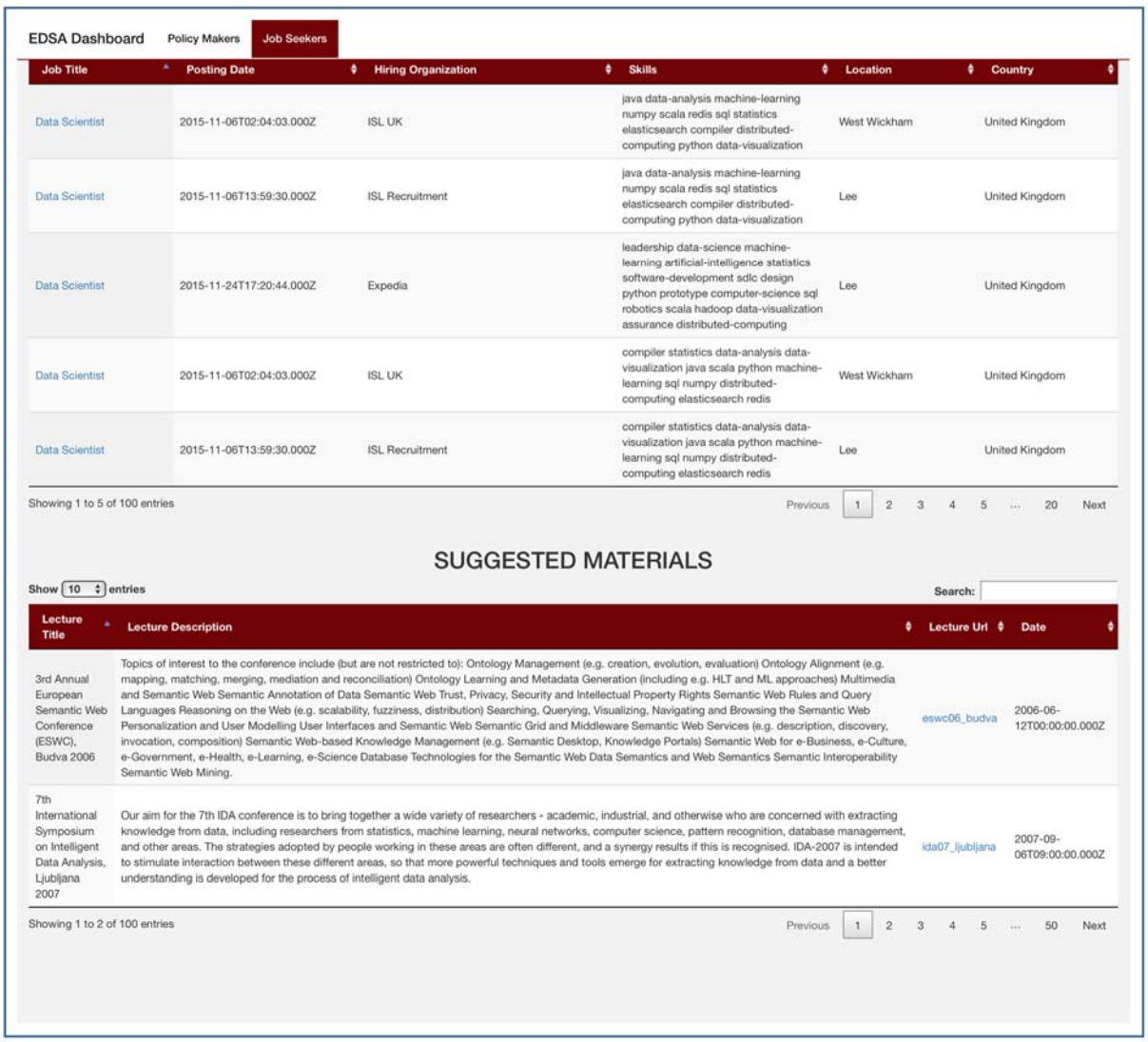


Figure 24: *Top job matches and learning resources (including conferences and other academic events) matching the skills selected in the job-seeker perspective, based on the search filters in Figure 21.*

An alternative approach to exploring skill demand and capability is to use the SkillSet viewer - shown in Figure 25 for a plot of over 12,500 individual job postings in Data Science across Europe, from 13th July to 10th November 2015. The viewer, accessed from the practitioner perspective, uses parallel coordinates to enable interactive comparison and querying of high-dimensional data[95] [96] [97] [98].
Considering each skill as a dimension, frequency of mention is plotted along each vertical axis corresponding to a skill. A randomly coloured polyline is drawn through all axes at the value for the multi-dimensional, multi-attribute data point.

On the far left additional variables of interest are also plotted on additional axes, allowing skill mention to be filtered along one or more of these. Axes may be rearranged and/or hidden to allow focus on a subset of interest and to allow closer comparison between any two variables.
Figure 25 illustrates how the target user may carry out a more detailed comparison of skills from the overview. The overlay hides all skill axes outside the three skill sets of interest: 'general', 'maths & statistics' and 'visualisation'. The data is then filtered by selecting the last three weeks in November 2015 (dragging to draw the red selection area - axis far left). An AND query is formulated by selecting postings with at least one mention of 'd3js'. This filters out all postings outside these filters - *c.f.* the dense plot in the overview for the same region with the much sparse overlay. Of the result set 'big data' shows highest mention, followed by 'interaction', with a much smaller rise for 'd3js'. Other skills co-occurring with the focus fall outside these three skill sets, listed in the popup showing more detail for the posting selected (thicker line in red).

[95] Inselberg, Alfred, The plane with parallel coordinates. The Visual Computer 1(2): 69-91 (1985) doi: 10.1007/BF01898350
[96] A. Inselberg, "Visualization and knowledge discovery for high dimensional data," User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on, Zurich, 2001, pp. 5-24. doi: 10.1109/UIDIS.2001.929921
[97] Inselberg, Alfred, Parallel Coordinates: Visual Multidimensional Geometry and Its Applications, Springer-Verlag New York, Inc. 2009. book - no doi
[98] C. K. Hung and A. Inselberg, "Visualizing Multidimensional Relations with Parallel Coordinates," Information Technology: Research and Education, 2006. ITRE '06. International Conference on, Tel-Aviv, 2006, pp. 261-265. doi: 10.1109/ITRE.2006.381579

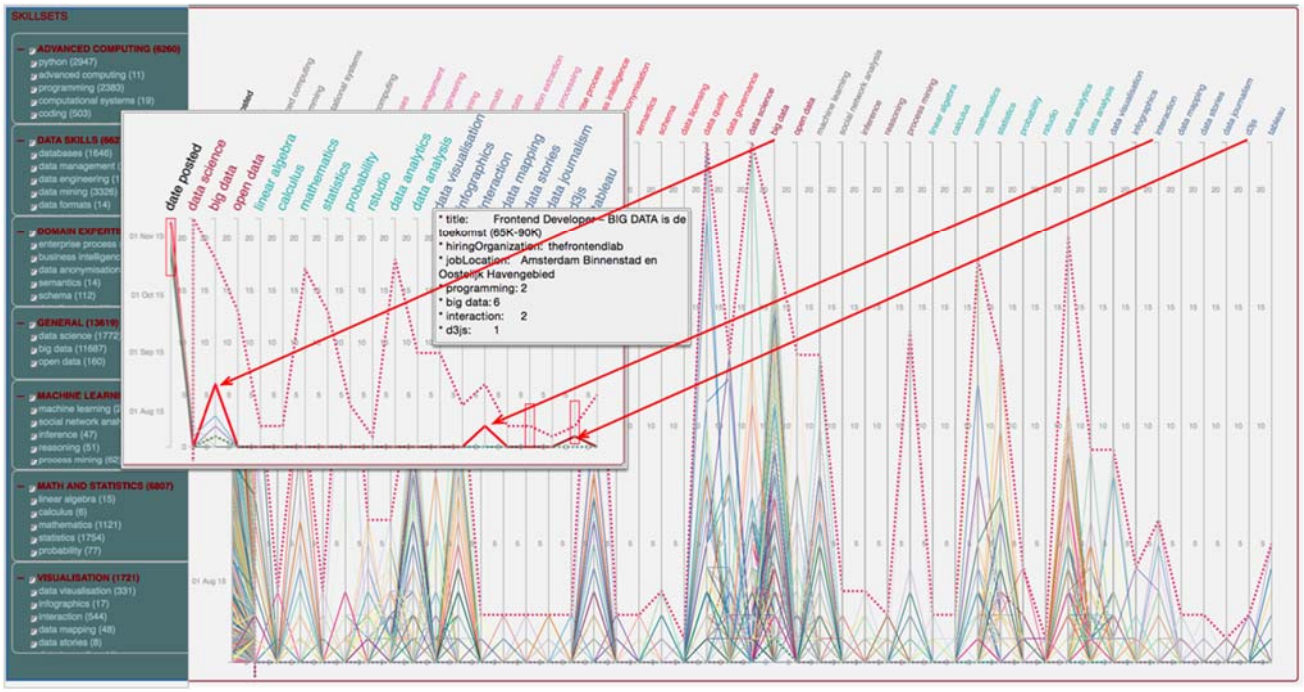


Figure 25: *A posting of interest selected from the results of a filter overlaid on the overview for the SkillSet viewer.*

Beyond a threshold of a few thousand posts across up to about 50 axes, interactivity is compromised. We therefore aggregate the data as previously discussed, on a set of dimensions to provide overviews that highlight attributes of interest. Figure 26 and Figure 27 show the data aggregated by location (country) and time (week starting from the earliest posting date) for the complete dataset containing 308,511 validated postings. The skill axes heights are also lowered to half the value of the highest peak - 'python' at almost 1654 mentions in a single aggregate, compared to the next two tallest peaks, 'programming' and 'big data' at 951 and 942 counts for the aggregate - to reveal smaller peaks hidden by this spike. Detail for a selected aggregate is provided in coupled tables - fetched from the RDF store on demand.

Figure 27 shows the results for the filter in Figure 21. The clearest difference in the pictures seen in Figure 26 and Figure 27, in addition to the small subset of countries with matching posts, is the change in peaks for skills. 'Interaction' and 'tableau', for instance, are required far less frequently with these three skills.

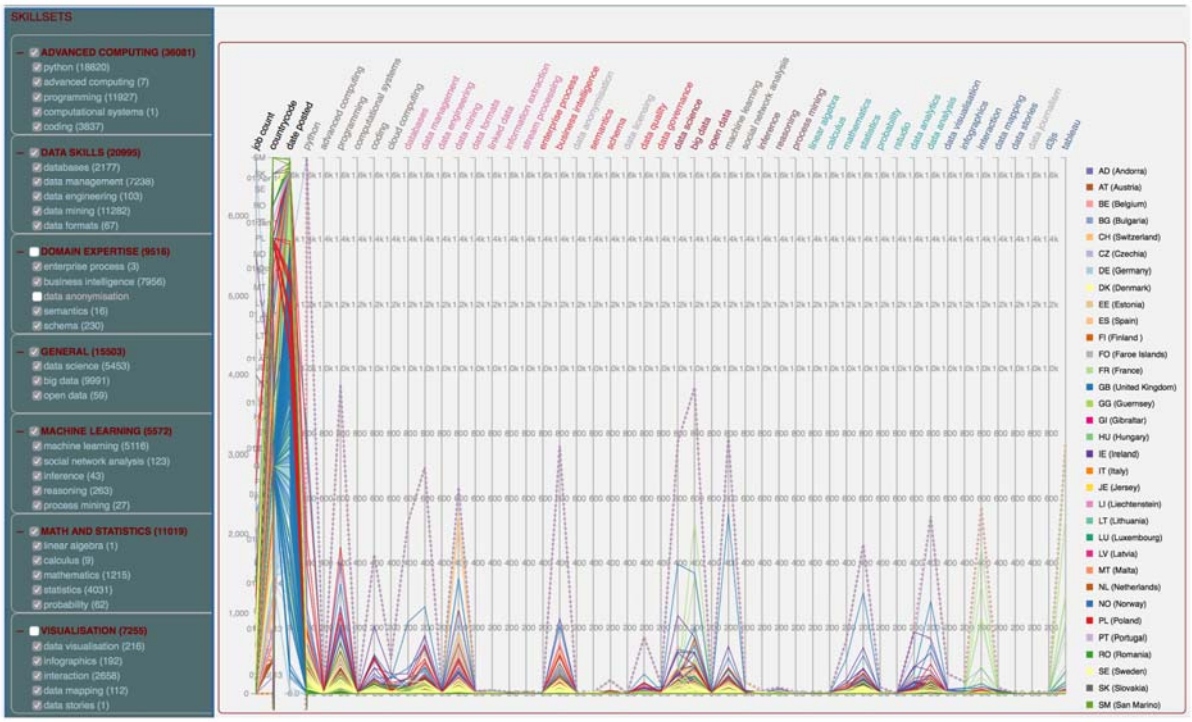


Figure 26: *Overview showing data aggregated by location and time for the 40 skills Skills have been grouped into seven skillsets analysed in EDSA deliverable D1.1 skillset viewer. The broken line in red shows the maximum skill mention across all aggregates.*

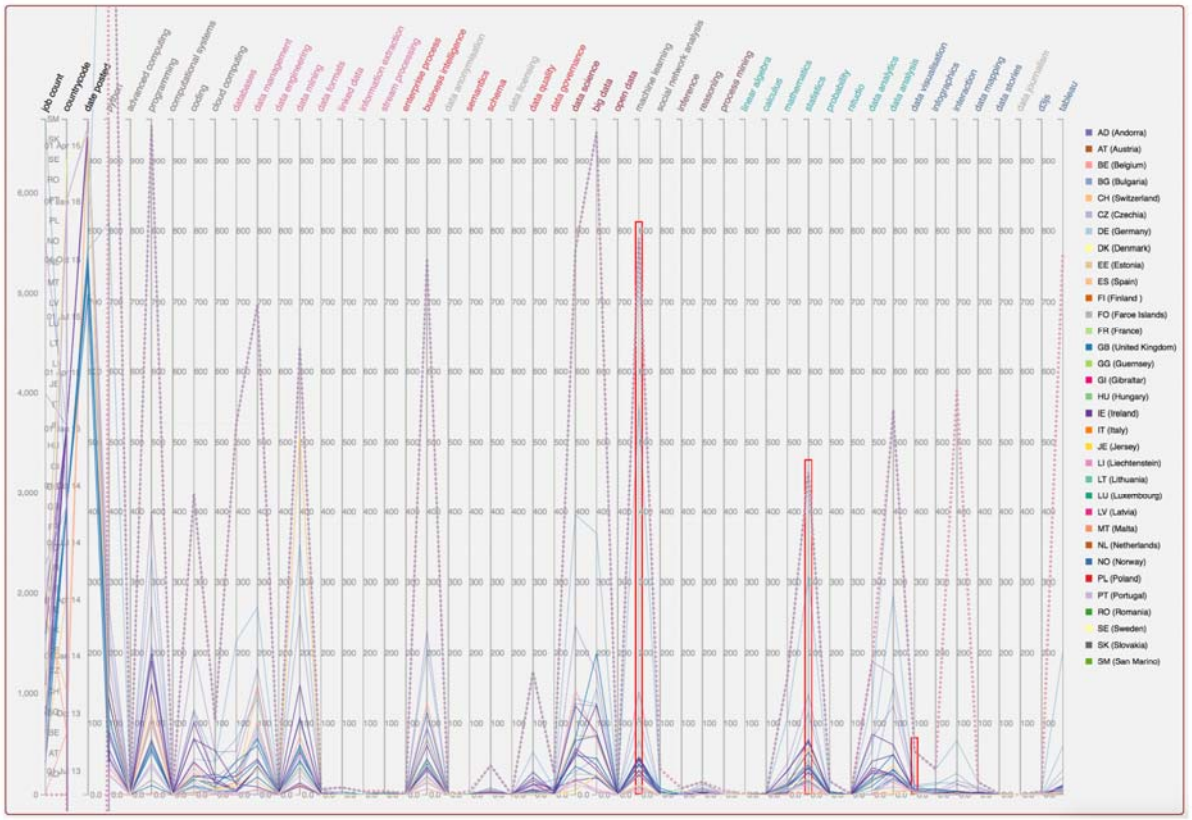


Figure 27: *Filter as for Figure 21 to retain only those aggregates with skills that co-occur with those of interest*

6 Discussion

Construction of the EDSA dashboard continues to follow the user-centred design and development cycle described in Figure 8, envisaged to continue beyond M18 of the project. The design methodology promotes independent tool development, both to make optimal use of distributed resources and to increase usability of the modules built. An on-going task is coupling the different modules to allow data and state to be exchanged more fluidly as users browse the dashboard and its underlying data.
Three key challenges in development have been

1. data acquisition - volume and coverage
2. web-based hosting
3. resource for the development activity

6.1 Data Acquisition

A number of job and other employment resources are available online. Along with developer tools and APIs, including LinkedIn, they form a key source of data. However, changes in the Terms of Service for use of the LinkedIn API and availability of other similar sources have resulted in restricted access to job postings on the scale envisaged, and providing representative, if not exhaustive coverage of the EU. This limitation impacts the breadth of analysis required as part of the design activity, to identify optimal functionality for supporting target end users. Further, the picture of demand obtained is only as reliable as the input data; this was reflected in feedback during heuristic evaluation with end users - while the potential of the tools was clear the skew in the picture of demand saw them reluctant to rely on this as a valid source of data for answering their questions about job and skill demand.

Work is ongoing to identify new sources of data, with permission to release at least the results of our analysis as open, linked data.

6.2 Hosting the EDSA Dashboard as a Live, Interactive, Online Tool

Hosting the dashboard online allows open access, a key requirement both for continuing to collect feedback from target end users and to meet project requirements. However, web-based hosting presents its own challenges, as network latency increases response time especially as the dataset size grows. Further, interactive response is best achieved using client side JavaScript, to allow wider access to open-source visualisation and data processing libraries and support on a wide range of modern web browsers. More computing intensive data (pre-)processing however relies on server-side tools and computing power in order to obtain an optimal balance between useful functionality and user interface response.

While a large number of libraries and APIs are now available for web development functionality available and especially heavy duty data processing support is limited compared to standalone tools using object oriented languages such as Java or C++. As previously discussed, we obtain a balance between wider availability and utility by carrying out data pre-processing offline and generating overviews starting from aggregated data, with options to drill down to more detail as the user narrows to selected regions of interest.

6.3 Development Resource

The challenges in data acquisition are further compounded by limited resource within the consortium, which further restricts the coverage and depth of the analysis being performed, both for job demand from the perspective of the employer advertising new positions and corresponding skills analysis that also involves examining the perspective of the practitioner mapping out their view of capability.

Analysis and development is therefore expected to continue beyond the end of the activity as detailed in the project Description of Work.

Appendix A: Usability Evaluation Tasks & Questionnaire

A documentation of the usability evaluation tasks and questionnaire can be downloaded from the following link:
https://www.dropbox.com/s/ei4fqkoda3vu42a/questionnaire_formal_usability_evaluation_complete.pdf?dl=0