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

D1.2 Study Evaluation Report 1

Deliverable Editor:	**Dr. David Tarrant (Open Data Institute)**
Other contributors:	**Mandy Costello (Open Data Institute)**
	Simon Bullmore (Open Data Institute)
	Steffica Warwick (Open Data Institute)
	Aba-Sah Dadzie (Open University)
	Jean-Louis Lievin (ideXlab)
Deliverable Reviewers:	**Chris Phethean (University of Southampton)**
	Alex Mikroyannidis (Open University)
Deliverable due date:	**31/07/2015**
Submission date:	**31/07/2015**
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	24/06/2015	David Tarrant	Added executive summary
0.2	02/07/2015	Mandy Costello	Added background, targets, data collection points
0.3	03/07/2015	Simon Bullmore and Steffica Warwick	Added qualitative results section
0.4	04/07/2015	Mandy Costello	Added expert identification section
0.5	07/07/2015	David Tarrant	Added online survey, automated data collection and dashboard section
0.6	08/07/2015	Mandy Costello	Edited document
0.7	09/07/2015	David Tarrant	Added new exec summary and conclusion
0.8	22/07/2015	David Tarrant	Added in review corrections
0.9	23/07/2015	Mandy Costello	Reviewed added comments
0.9a	24/07/2015	David Tarrant	Added sections from market analysis
0.9b	30/07/2015	David Tarrant	Added Figure 14 along with explanation Changed figures to be more B&W friendly
1.0	31/07/2015	Aneta Tumilowicz	Final QA

Table of Contents

List of Tables

List of Figures

1. Executive Summary

"Surviving in the new data driven economy requires a new set of skills; that of a data scientist. Managers must identify and prioritise knowledge in order to secure new data talent, and train the existing workforce. The problem is the lack of evidence about which subset of data science skills are most in demand and what specific training is required."

This document outlines the initial findings and recommendations from the first phase of the demand analysis survey as outlined previously in D1.1 (Study design document). The demand analysis survey focuses on both qualitative and quantitative data collection in order to ascertain the current gaps in data science training and the opportunities to convene the activity within Europe as proposed initially by the EDSA project.

Initial results from the demand analysis survey, collected over the course of only 2 months, already give clear indication of some of the training needs common across Europe. There are also emerging differences between views of data scientists and managers of data scientists about the best ways to deliver training, the former preferring eLearning with the latter expressing preferences for face-to-face training and webinars. Early automated collection of jobs data related to data science also demonstrates the complexity of analysing this field across Europe, for example in the ways that different European languages describe the various specialist data science skills. This is emphasised by the dominances of established areas like statistics compared to emerging skills like infographics.

The following data has so far been collected as part of the demand analysis survey:

- 11 qualitative interviews from 4 countries.
- 13 responses to online survey from 7 countries.
- Over 40,000 data points from LinkedIn jobs across Europe.

Based on our early findings from face-to-face interviews, the project is taking the right approach to meet the development needs of data scientists. There are areas that need to be explored further in our demand analysis, in particular how to address an emerging need for data scientists to have a range of business skills, specifically the ability to effectively communicate with businesses.

At this stage, our interviews provide further evidence of a lack of suitable training options to meet demand, and a lack of clarity about the best way to develop skills. The result is that organisations are tending to hire, rather than develop skills, which is seen as an expensive but essential investment strategy in a domain that offers them competitive advantage.

Improving the choice of training and guiding learners to the best approaches are core to this project, and could provide companies with the confidence to make investments in developing rather than acquiring people. To make this happen, the interviews we have conducted provide us with some initial areas to focus on, and ways we can improve the approach to get more insight.

Responses to the online survey suggest that there is still a great need for skills expansion in all areas of data science. Additionally, there is a demand for a mixed-mode approach to training delivery consisting of face-to-face as well as online courses and webinars. At this stage, the language training is provided in does not seem as important to respondents, with only one response stating that is was essential to have a non-English translation.

The online survey has also generated some interesting links between existing skills and perceived importance of other skills. From the responses gathered, it would seem that the most essential skills for a data scientist to have are more of the same skills that people in the organisation already have. This could also be seen as a risk, with organisations becoming too focused on a specific set of skills, and not expanding overall data science capability.

Automated collection of job adverts from LinkedIn across Europe shows that a consistent sets of skills that can be attributed to data science roles has yet to emerge. There are a number of established skills areas where demand is high across Europe, including business intelligence, data engineering and cloud computing. Other areas, such as data visualisation, interaction and artificial intelligence are emerging trends with steady growth, but not yet to the high levels of other areas. In the current datasets we have identified a large “skew” towards the UK, which suggests a terminology problem when talking about skills, which needs to be investigated further. Using additional datasets and exploring refinement in the next stages of the project will enable further analysis and an accurate depiction of Europe overall.

Finally, the demand analysis explores the need for a European Data Science Academy to emerge from this project in order to convene a community and provide opportunities for training the next generation of data scientists. We cannot yet define what this academy should be and how it fits with the existing network of communities and course providers.

What is already clear is the need for EDSA to convene and curate content and materials, and to guide data scientists and managers to successful strategies for growing data science capacity for themselves and for their organisations.

2. Background

2.1 Summary of demand analysis

The objective of the demand analysis is to evaluate:

- The current level of data science skills across industries and countries in Europe.
- The provision of data science training currently available.
- The rationale for adopting data science training within an organisation.
- The key data science skills required by organisations.
- The factors that affect the adoption of data science training within an organisation.
- The demand for a European wide data science initiative

The results and findings of the demand analysis will inform the project of the latest view of data science skills and capability in Europe throughout the lifecycle of the project. The study will also make recommendations to ensure impactful delivery of the project activities by informing:

- The curriculum and courses that will be developed as part of WP2.
- Successful approaches for delivery of training as part of WP3.
- Opportunities for engagement with the sector as part of WP4.
- The formation of a European data science initiative as part of WP5.

Furthermore, a demand analysis dashboard for Europe has been produced to visualise the results of the study. The data will be presented on the dashboard in an engaging way that encourages exploration of the results, and contribution of further data. This will also help to shape revisions of the curriculum further into the project, so that the curricula adapts to the latest demands. The curricula in D2.1 (Data science curricula 1) are the first versions of this based on this initial demand analysis and market analysis conducted by the University of Southampton on the data science education market in Europe. This initial research has also provided insight for the demand analysis, for example on key data science search terms. The curricula in D2.1 will be adapted based on further insights from the demand analysis and subsequently, the dashboard.

2.2 Summary of methodology

Below, Figure 1 recaps the chosen methodology for the demand analysis as outlined in D1.1. Both primary and secondary data has been, and will continue to be collected throughout the project to ensure a thorough view of the data science skills and capacity landscape in Europe is captured and explored, through in depth qualitative responses, comparable quantitative data, and trends from automated data collection.

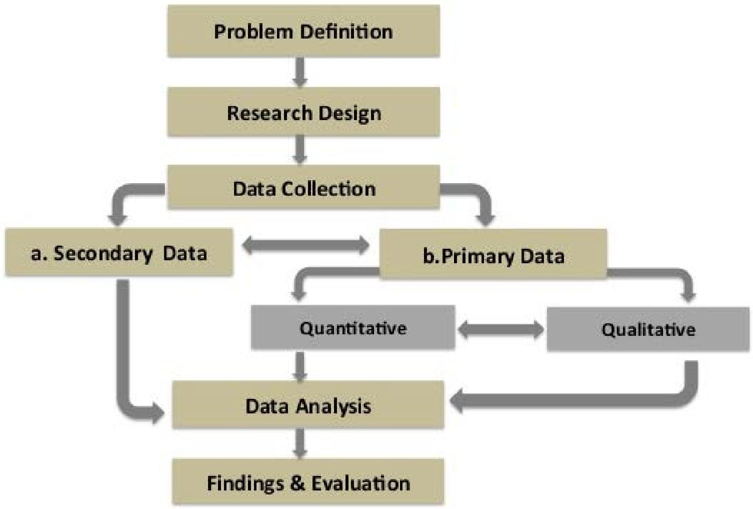


Figure 1: Overview of demand analysis methodology

2.3 Summary of data collection

Several techniques have been used to collect primary and secondary data:

Primary data has been collected from: one-to-one interviews, responses to the online survey and a focus group.

Secondary data has been collected from web services, including LinkedIn to enable analyses of trends across Europe.

While the methodology for the study will remain consistent to ensure comparability of results, it is anticipated the approach to data collection will evolve. As new areas are explored as part of the project's ongoing engagement in the sector the study will develop to include new sources of data and insights. This will be particularly relevant to the secondary data collection, where integration and exploration of additional or existing data sets to the study will present opportunities for further analysis, complementing the primary data collection.

Each data collection point is explored further in this document, with details of the approach taken, progress up to M6 and areas for further exploration for the remainder of the study. Later in the summary, we explore the initial key findings from the data collection so far.

3. Targets and key performance indicators

The demand analysis study targets are split into the two main deliverables for the work package, at M6 and M18. Due to the longevity of the study, the first phase of the project, up to M6, was designed to allow evaluation and testing of the approach and materials that were produced. Following the review of the evaluation phase at M6, the approach will be adjusted as required, and amendments made to any materials to ensure the successful continuation of the study, before a wider European launch.

In this section, we will first review these initial targets. We will then outline new targets and key performance indicators for the next phase of the study.

3.1 Initial study targets – M6

Progress targets and key performance indicators were outlined in D1.1, to enable progress of the study to be monitored.

M6 progress targets have all been achieved. These include:

- Testing of the one-to-one interview design.
- Completion of a number of one-to-one interviews and transcriptions.
- Development and testing of the online survey.
- Collection of online survey responses.
- Integration of the online survey with the EDSA dashboard.
- Completion of one focus group.
- Deployment of automated data collection platform for trend tracking.
- Deployment of data science expert identification platform.
- Analysis of initial data.
- Launch of the dashboard on the EDSA website.

M6 key performance indicators outlined the targets for data collection. Table 1 below, outlines the target set in D1.1, and the figures obtained to date:

Table 1: Key performance indicators – M6

KPI	Target (M6)	Actual figures (M6)
Size of network (qualitative analysis)	3 per partner country	11 interviews (UK 5, Germany 2, Sweden 2, Netherlands 1, France 0, Slovenia 0, non-partner 1) 13 online survey responses (UK 1, Germany 7, Sweden 1, Netherlands 1, France 1, Slovenia 1, non-partner 1)
Number of focus groups	1	1
Number of sectors	10	8
% of EU business registry sectors	20%	38%
Importance of sectors /	20%	Information and communication

Importance of sectors covered measured by market size and growth forecast		– 1,192,677 – 5% Professional, scientific and technical services – 1,160,350 – 4.8% Administration and support services activities – 790,000 – 3.3% Education – N/A Public admin and defence – N/A Human health and social work activities - N/A Transportation and storage – 1,250,000 – 5.2% **18.3% total** (based on market size statistics in non-financial economy) Figures in EUR millions (Source Eurostat[1]) Financial and insurance activities = N/A
Number of EU states	Partner countries	Estonia, France, Germany, Italy Netherlands, Sweden, Slovenia, UK.
% split of Corporate / SMEs	No Target	N/A
% split of managers / Data scientists	No Target	N/A

3.2 Next study targets – M18

Following the initial evaluation phase to M6, a wider European launch of the study will follow.

M18 progress targets are:

- Evaluation of interview design.
- Evaluation of online survey design.
- Promotion of the EDSA dashboard and online survey.
- Further development of the EDSA dashboard with the project partners.
- Development of further visualisations for the dashboard.

[1] Eurostat - http://ec.europa.eu/eurostat/statistics-explained/index.php/Business_economy_-_structural_profile#Further_Eurostat_information

- Integration of additional secondary datasets.
- Further one-to-one interview responses, to match M18 target.
- Collection of online survey responses, to match M18 target.
- Further focus groups, to match M18 target.
- Data analysis of all collected data.
- Final version of the EDSA Dashboard.
- Executive summary, results and recommendations.

Table 2, below, outlines the key performance indicators for M18:

Table 2: Key performance indicators – M18

KPI	**Figures (M6)**	**Target (M18)**
Size of network (qualitative analysis)	26	168 – 6 sectors per member state
Number of focus groups	1	4
Number of sectors	8	17
% of EU business registry sectors	38%	80%
Importance of sectors	18.3%	80%
Number of EU states	8	All
% split of Corporate / SMEs	No Target	60%/40%
% split of managers / Data scientists	No Target	60%/40%

3.3 Reach of demand analysis

In order to create impactful production and delivery of the project's activities, we will strive to build a full picture of the data science landscape in Europe, by collecting data from a wide range of EU business sectors, across all EU member states.

In order to achieve this, the project partners will work together to engage with data science communities across Europe in a number of ways:

- Recommendations from our existing partner network.
- Recommendations from key stakeholders.
- Recommendations from previous participants of the interview and online survey.
- Recommendations from the Advisory Board.
- Promotion of the EDSA Dashboard and online survey.
- Engagement in community events, conferences and meet-ups.
- Use of the ideXlab expert identification platform, outlined in section 3.4.

Further information on these engagement plans can be found in the projects dissemination plans and reports developed as part of WP4, and outlined in D4.3 (Real-world and online community engagement plan).

Figure 2 shows our current country reach, and Figure 3 outlines the engagement planned through deploying these reach methods over the next 3 months.

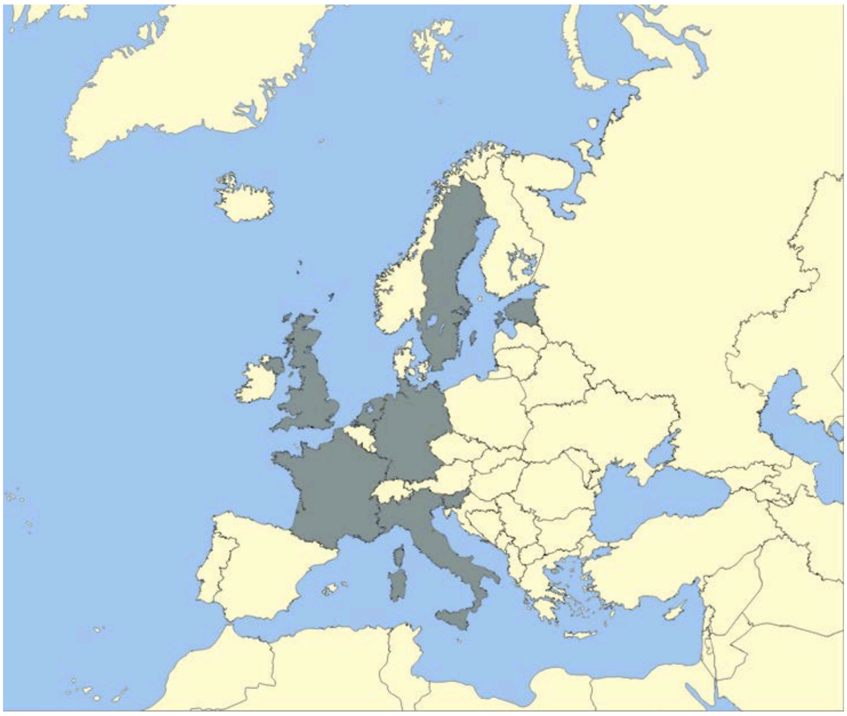


Figure 2: Current country reach

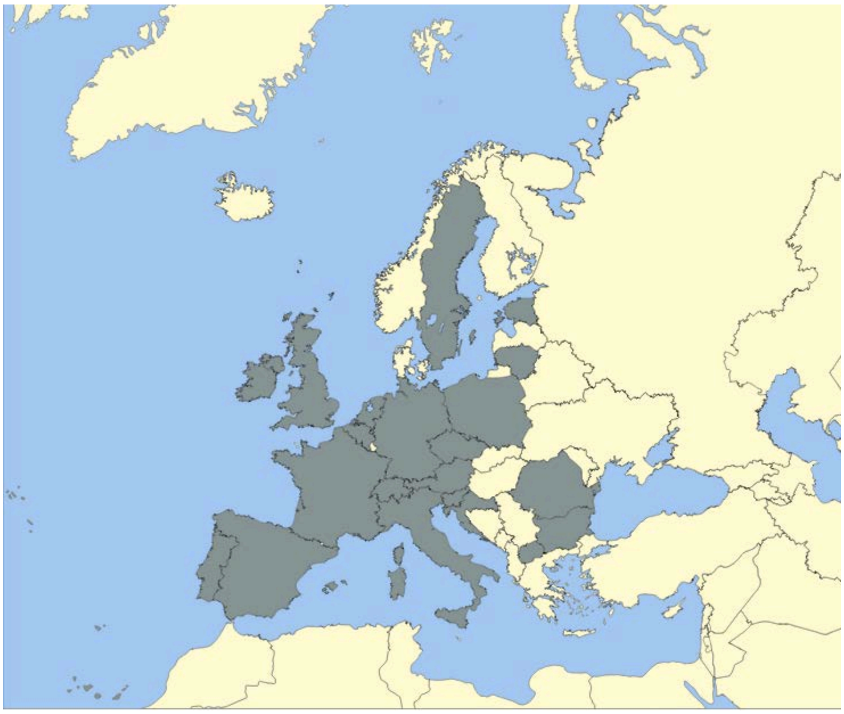


Figure 3: Planned country reach

Maps courtesy of Madman2001/Wikimedia Commons and Phil Archer.

3.4 Automated expert identification

To increase the reach of the demand analysis, a specialist tool will be used, supported by ideXlab, to identify key stakeholders. This will complement the real world and online engagement approach we will take, as outlined in section 3.3.

3.4.1 Methodology

For the purpose of the demand analysis, we define a data scientist as an individual who combines more than one of the key skills areas outlined in the study. The initial phase of the demand analysis has eight of these key areas that have been derived from Drew Conway's Data Science Venn diagram[2], one of the most widely accepted definitions of data science. More on this methodology can be found in Section 2.2 of D1.1 (Study design document)[3].

- Math and statistics
- Machine learning
- Domain expertise
- Data skills
- Advanced computing
- Visualisation
- Scientific method
- Open culture

Each of these areas can be sub-divided into specific skills, with note that these areas can also be a skill, for example machine learning.

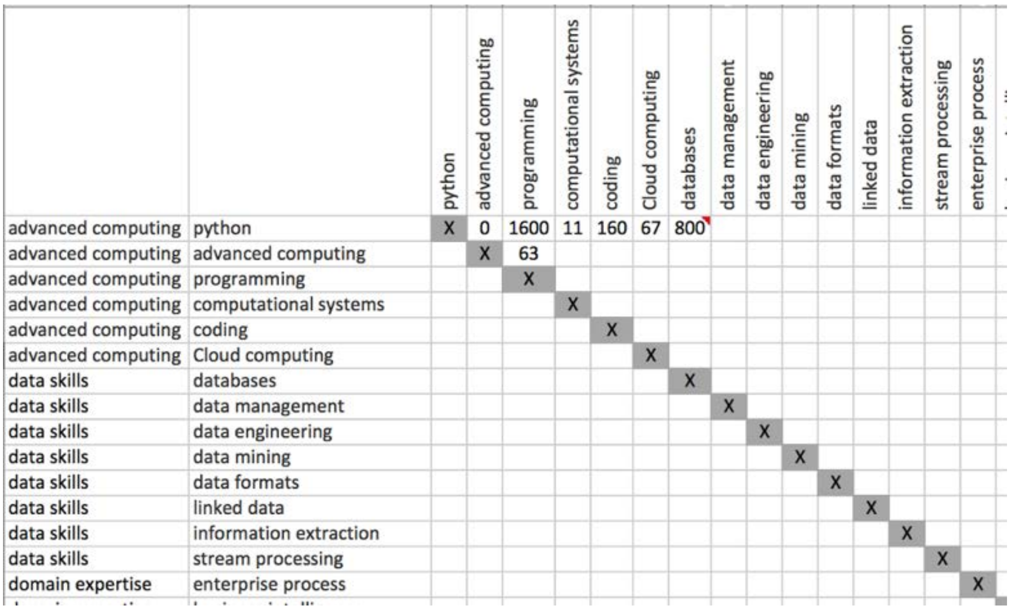
The ideXlab platform will be used to search for individuals with a combination of these skills. The more an individual has skills in different areas, the more likely they will define themselves as a data scientist.

Once identified, these individuals will be contacted, and asked to participate in the study or other aspects of the project. Furthermore, the 'additional skills' data can be analysed to assess if our definition of a data scientist should be adjusted.

[2] Data Science Venn diagram (Drew Conway) http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[3] Tarrant et al, *European Data Science Academy – D1.1 Study Design Document*, 2014

3.4.2 Sample data



		python	advanced computing	programming	computational systems	coding	Cloud computing	databases	data management	data engineering	data mining	data formats	linked data	information extraction	stream processing	enterprise process
advanced computing	python	X	0	1600	11	160	67	800								
advanced computing	advanced computing		X	63												
advanced computing	programming			X												
advanced computing	computational systems				X											
advanced computing	coding					X										
advanced computing	Cloud computing						X									
data skills	databases							X								
data skills	data management								X							
data skills	data engineering									X						
data skills	data mining										X					
data skills	data formats											X				
data skills	linked data												X			
data skills	information extraction													X		
data skills	stream processing														X	
domain expertise	enterprise process															X

Figure 4: Sample data for ideXlab platform

Each query is an AND combination between two skills, for example python AND "cloud computing". When a skill contains more than one word, quotation marks are used.

The result of a query provides:

- An estimate of the total number of experts worldwide that have published scientific papers containing the combination of keywords in their titles, abstracts and keyword sections.
- The name of the expert.
- Suggested keywords, derived from the analysis of the publication abstract texts; these keywords could lead to the discovery of new emerging skills to be added to the key areas of data science defined for the study. These skills could also inform additional areas of focus for curriculum development.

A CSV file can then be extracted, and enriched with further data separately to include the location and sector of each expert.

3.4.3 Next steps

A specific application will be developed to run on top of the ideXlab platform, with the methodology outlined above. The ideXlab platform contains data harvested from publications that can be searched in order to find experts in key areas. Users can perform queries to generate lists of experts in each area that can be exported as CSV. In order to find data scientists, the platform will need to be extended to allow the intersection of skills as opposed to the single term search. The application will have the following functionality:

- Automated generation of queries following Figure 4
- For each query, automated generation of all experts to an extractable list

Further investigation will be conducted before full implementation. The key questions to be addressed before full implementation will be:

- Extraction of expert additional skills data.
- Identification of ranking algorithm of experts.
- Scope and scale of new skills analysis.
- Automatic collection of publications leading to identification of additional skills for the expert.

4. Data collection

4.1 One-to-one interviews

One-to-one interviews provide an opportunity to explore topic areas, and collect in-depth responses about data science within a sector and country. Furthermore, engagement with key stakeholders on a one-to-one basis allows for further exploration of the opportunities for EDSA activities to make impact.

4.1.1 Implementation

The interview questions were developed as part of the study design. Following this, a semi-structured script was created to guide the interview (an example of the script can be seen in Appendix A). This semi-structured script is designed to assist the interviewer to obtain explicit permission to record and use the resultant material from the interviewee, before leading into a more free form interview conversation led by key questions and areas for potential exploration. The ODI and other consortium partners carried out the initial interviews, and evaluated the effectiveness of the questions. This initial phase was essential to evaluate the interview process and ensure that the interview questions were clear to the interviewee, and the nature and intention of the question was clearly communicated.

To ensure continuity to the interviews, partners were provided with 'interview guidelines', (see Appendix B), to enable them to conduct interviews in comparable conditions. The guidelines were created to equip partners with background information on the demand analysis, provide sampling criteria to select potential interviewees, provide questions and prompts for the interview and guidance on conducting the interview. Furthermore, a document was created to share with individuals who were invited to an interview to ensure consistent communication with all participants (see Appendix C).

4.1.2 Analysis methodology

At this stage of the study we adopted a thematic analysis methodology of the transcribed data. Summaries for each transcribed interview were produced in order to draw out high-level themes. Respondent answers were also aligned to the study questions. This process was undertaken manually. However, as the volume of interviews develops we will adopt methods more suited to larger qualitative datasets, including the use of software tools. This may include Thomson Reuters Open Calais[4]. Open Calais processes unstructured text to locate entities, topics, events, relations and social tags, and where possible, link these to other linked open data on the web.

4.1.3 Next steps

The ODI and partners will evaluate the interview structure and questions before further interviews are conducted. We will maintain the questions in order to analyse results with the early collected data, however additional questions or question prompts may be added where commonalities in discusses have been found in a particular area to explore trends further. We will now strive for further country and sector coverage, utilising our existing and growing engagement in the area as outlined in section 3.3. We will also contact those potential participants identified in appendix P1 and P2 of D1.1

[4] Thomson Reuters – Open Calais - http://new.opencalais.com/

4.2 Online survey

The online survey has been developed and is accessible through the EDSA dashboard, through a 'click to contribute' button, as well as a stand-alone survey for use outside of the dashboard. The survey is designed to capture qualitative and quantitative data on the key questions from the interviews, to allow analysis and comparison of larger amounts of data. The online survey can be completed anonymously. However users are encouraged to further participate in the study, and are asked to give contact details, if they wish, to allow EDSA to follow up on other research elements or project activity.

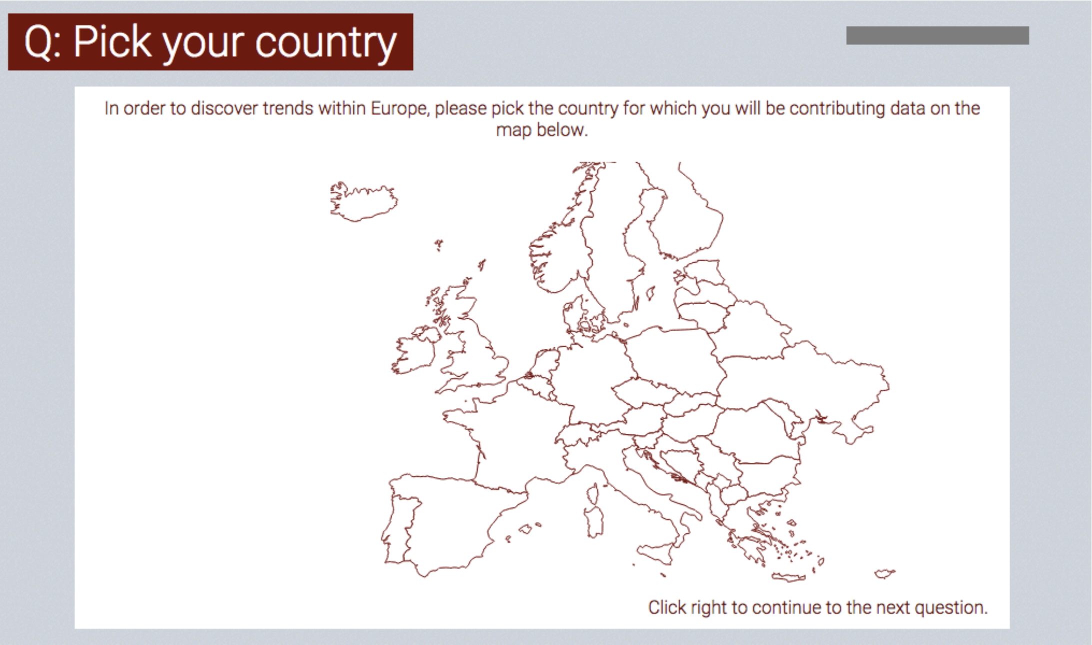
4.2.1 Implementation

The online survey has been designed to be short and engaging and has been developed to allow users to customise the survey as well as to give responses to set questions. Depending on a user's answers, the survey changes to allow the user to contribute further quantitative data. To allow this to happen the survey has been custom built using standard HTML5 to be both responsive to different devices and user input. Survey responses are collected in the JSON format before being processed and anonymised in order to output a set of CSV files for analysis.

4.2.2 Survey design

The survey has been designed to be short and engaging, with interactive elements to encourage the user further, moving away from traditional online surveys designs. Figures Figure 5,Figure 6 and Figure 7 below show some of the interactive elements of the online survey, including a clickable map, drag and drop and sliding scale function.

The survey is designed to showcase what a data science project should be about, and uses the latest web technologies to provide intuitive ways to quickly answer questions and save the user time. For example, rather than asking a user to select their country on a drop-down list, users click on the map of their country.



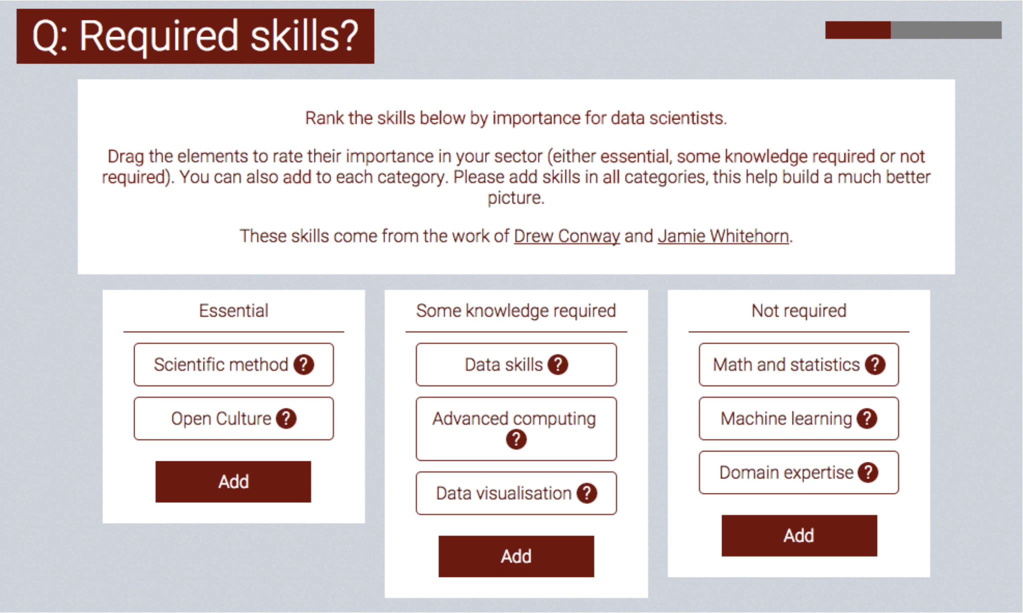
Q: Pick your country

In order to discover trends within Europe, please pick the country for which you will be contributing data on the map below.

Click right to continue to the next question.

Figure 5: Online survey 'pick your country' screen

The ‘Required skills’ question asks users to rate the importance of several keys areas in a data science role - from ‘essential’ to ‘not required’. If a user is unclear on any area, there is help available via the question mark icons. In order to rate the importance of each skill a user simply drags the skills between each box. This approach is much more visual than a rating list. Users can also add further skills to any box.



Q: Required skills?

Rank the skills below by importance for data scientists.

Drag the elements to rate their importance in your sector (either essential, some knowledge required or not required). You can also add to each category. Please add skills in all categories, this help build a much better picture.

These skills come from the work of Drew Conway and Jamie Whitehorn.

Essential
- Scientific method ?
- Open Culture ?

Some knowledge required
- Data skills ?
- Advanced computing ?
- Data visualisation ?

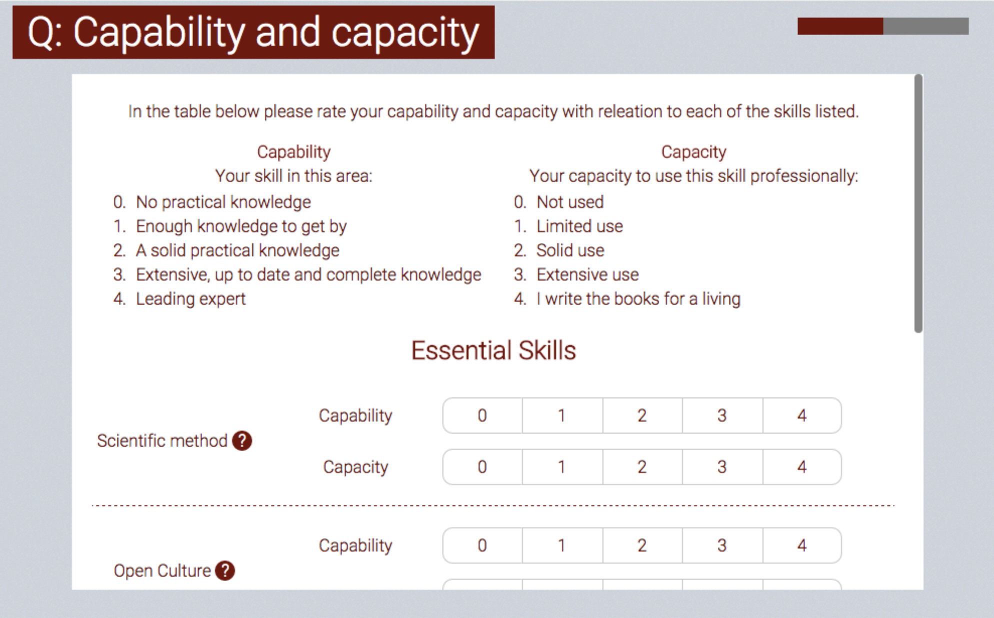
Not required
- Math and statistics ?
- Machine learning ?
- Domain expertise ?

Figure 6: Online survey ‘drag and drop’ function screen

Once a user has answered the required skills questions they are asked to rate the capability and capacity in their area to carry out these skills. As a reminder:

- Capability: Having the required skills
- Capacity: The ability to apply the skills in a professional environment

This question is answered using a scale from 0-4, as seen in Figure 7 below:



Q: Capability and capacity

In the table below please rate your capability and capacity with releation to each of the skills listed.

Capability
Your skill in this area:
0. No practical knowledge
1. Enough knowledge to get by
2. A solid practical knowledge
3. Extensive, up to date and complete knowledge
4. Leading expert

Capacity
Your capacity to use this skill professionally:
0. Not used
1. Limited use
2. Solid use
3. Extensive use
4. I write the books for a living

Essential Skills

Scientific method ?	Capability	0	1	2	3	4
	Capacity	0	1	2	3	4
Open Culture ?	Capability	0	1	2	3	4

Figure 7: Online survey ‘sliding scale’ function screen

4.2.3 Analysis methodology

Data collected from the online survey will be anonymised and published as open data in CSV format to allow anyone to analyse the data. The data will be presented via the dashboard in the form of a cross filter. This cross filter will allow users to quickly see an overall view of the data or be able to zoom in by sector, country, role or even by skill. Due to the amount of data required for an effective cross filter, this will be developed fully by M18. Initial analysis presented in section 5 of this report looks more broadly at answers to each question without considering the country or sector differences that affect the answers.

4.2.4 Next steps

The online survey is now available through the EDSA dashboard, which will be launched at M6. There are several ways we will encourage data contribution. Firstly, the dashboard and online survey will be promoted through EDSA media channels. Secondly, the dashboard will be promoted within each partner's existing network, the EDSA network and the Advisory Board. Thirdly, as the survey is easily accessible and responsive in design to multiple devices, partners representing EDSA will be able to use the survey to collect responses on an ad-hoc basis, at events outlined in D4.3. (Real-world and online community engagement plan) including: conferences, meet-ups and data science themed events. Furthermore, individual partner events and meet-ups, regularly hosted by Persontyle and the ODI will provide an opportunity to share the online survey. Following the gathering of more data, further analysis and development of the interactive presentation of the data will be priorities.

4.3 Automated data collection

To assist in establishing the demand and capacity for data science across Europe a number of automated harvesting services are being used throughout the project. Additionally, a number of datasets from third parties who have already carried out research in similar areas are being considered for inclusion. This section outlines the implementation and analysis methodology for the project's own data harvester, which uses key terms aligned to the face-to-face and online surveys to look at wider demand for jobs in key areas of data science. The results of the initial analysis can be found in Section 5.3.

4.3.1 Implementation

To establish the demand for jobs in data science across Europe, a harvester was developed to search LinkedIn for jobs in each country across Europe. To align with both the face-to-face and online survey, the same eight key areas form the basis for the searches:

- Machine learning
- Advanced computing
- Data skills
- Domain expertise
- Open culture
- Math and statistics
- Data visualisation
- Scientific method

Additionally, each of these terms is embellished with other sub-terms for the category. In total, there are 46 terms searched for. In order to establish demand across Europe, all 46 terms have been automatically translated into 31 languages, with a data harvest per country, in the dominant language of that county. One harvest is performed per day, meaning each day 2162 datasets points are collected.

Each dataset corresponds to a specific term, and includes:

- Number of jobs advertised containing that term.
- Geo-location of the jobs within the country.
- Top employers advertising for those jobs within the country.

All data is stored in CSV files in the dashboard Github repository[5] and updated daily. All files are made available under an open license for anyone to access, use and share.

The harvester is implemented in PHP and is also available in the dashboard Github repository along with all the reference files containing the terms and translations.

4.3.2 Analysis Methodology

Data collected from the harvester will be analysed by term category, time of advertisement, location, and if possible, sector. Due to the difficulties in deducing sector from employers, it is anticipated that this part of the analysis will be more challenging. Initially, analysis will focus on the temporal element in the data, treating the spatial attribute (geo-location and, by derivation, language) as an additional lens (filter) through which it is possible to examine the non-spatial data (skillsets). The aim is two-fold: to identify effective methods for first revealing temporal patterns in the data, and then relating these back to the evolution of data science skills and demand. The second goal is, to clarify our target end user characteristics and identify key stories that can be told from the data. We expect, as a result of this process to identify additional data requirements, structure and content for completing these tasks.

4.3.3 Next steps

Following the initial analysis, there are a number of next steps related to data collection and the analysis methodology. An initial challenge with the early analysis was that the data collected was temporal in nature and other data not detailed enough for in-depth analysis of trends. Collecting more datasets may help to address this. However, having additional datasets that don't correspond to each other will equally make analysis difficult. Therefore, the next stages are three fold:

1. Collect more datasets and evidence from others who have performed similar studies and use these to inform (2).
2. Refine the collection of data to fill in key gaps identified through initial analysis.
3. Expand analysis to reflect new capabilities granted by enhanced data collection.

4.4 EDSA Dashboard

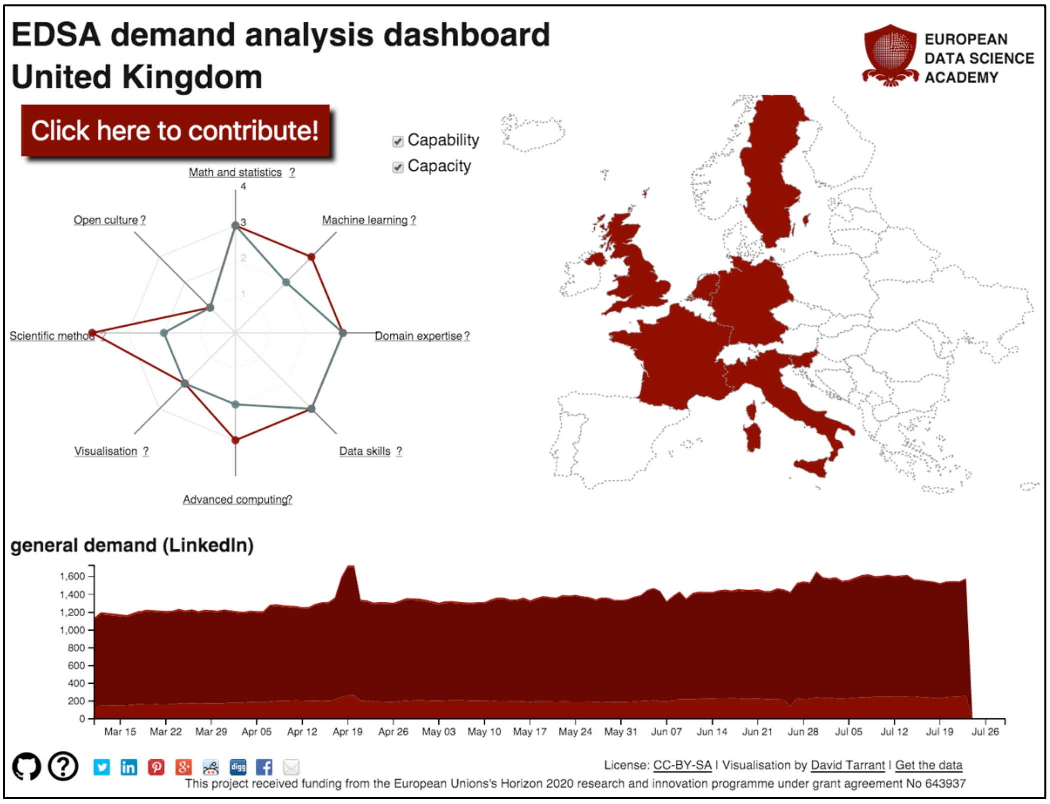
The EDSA dashboard is designed to give users an engaging view of results from the demand analysis. The initial version of the dashboard has been designed to allow users to select countries in order to view data on skills capacity and demand from job advertisements corresponding to the different areas of data science. It was proposed that the dashboard should consist of four main aspects:

[5] https://github.com/davetaz/EDSA/tree/gh-pages/data/harvester

1. Skills analysis radar diagram of the eight main skills areas.
2. A trend tracker displaying the automatically collected data.
3. A topic map of key points from qualitative interviews.
4. A link to training courses available in each area (supplied as part of WP2).

Further development of the dashboard's functions and visualisations are planned following M6.

4.4.1 Implementation



EDSA demand analysis dashboard
United Kingdom

Click here to contribute!

License: CC-BY-SA | Visualisation by David Tarrant | Get the data

This project received funding from the European Unions's Horizon 2020 research and innovation programme under grant agreement No 643937

Figure 8: The EDSA dashboard (M6)

Figure 8 shows the EDSA dashboard at M6 of the project. Here, the UK has been selected on the map and the capability/capacity diagram for the country combining focus group and survey data is shown. Clicking on the titles of the axis of the radar diagram brings up the trend graph for jobs in that area. The current dashboard is designed to provide the highest-level overview of collected data and provide users the opportunity to contribute their own data, through a link to the online survey.

The dashboard is entirely implemented in HTML5 and d3.js is used for all visualisations. The code is open source and available via the Github link. All data is open data and available via the "Get the data" link on the dashboard. Finally, the dashboard is also hosted for free in Github, meaning that it is sustainable beyond the life of the project.

4.4.2 Next steps

Following the launch of the dashboard at M6, the next steps will be to:

- Develop the dashboard's functions and visualisations.
- Update the dashboard to reflect results of all online and face-to-face survey results.
- Collect more data.
- Enhance the analysis of the data.
- Update the dashboard to include more in depth browsing of collected data by sector.

- Enhance the dashboard to allow comparison between countries and sectors rather than a single selection of data.

In order to achieve this, all project partners will be encouraged to make use of open source libraries so results can be directly integrated where possible.

5. Results

In this section we outline the results of the initial data collection, including the results of the one-to-one interviews, online survey and automated data collection.

5.1 One-to-one interview findings

Firstly, we explore the findings from the initial one-to-one interviews. A summary of the results for each question is given below. An overall summary can be found in section 5.1.1. Findings from the final question of the interview on the demand for a European wide data science initiative can be found in section 6.

Question 1 - What is the impact of data science on organisations?

Demand for data science is growing and is being taken seriously by all organisations surveyed. Organisations know they now have access to more data, and better technologies to work with data. Additionally there is also a clear understanding of the value of data science at all levels within the organisation. Challenges exist finding the right combination of skills in enough people.

"Some of our partner companies that came to us doing nothing are now investing significant sums into data science because they've realised the value of it."

The lack of skills in organisations has opened up a market for freelancers, consultancies and training organisations to which managers are turning to meet needs that cannot be met internally. This is especially true for smaller organisations. Larger organisations have been able to meet demand by putting together teams, paying for learning and by subcontracting work.

Question 2 - Is a new set of data science skills required?

Although Data Science is a relatively recently recognised profession, with the exception of machine learning, the spectrum of data science skills has existed in many organisations for many years, most frequently across different people in different teams. The difference now is that organisations want to find the full spectrum of skills in one place - their data science people and teams.

"*I don't think the skill-sets themselves are actually new, I think it's just how we expect now the same person to have all the skills, whereas in the past we may have different people having different bits of these skills.*"

Furthermore, as data science is now closely connected to business needs, data scientists require improved business skills to help them work with managers. Business skills enable them to better understand what managers need and better communicate results.

Question 3 - What approaches are being taken to expand data science capacity in organisations?

Organisations are not doing enough to train up their own staff, despite a clear demand for data science skills.

Beyond academic qualifications, relatively few respondents have undertaken further formal training provided by their organisations. Hands-on approaches and self-guided training, for example MOOCs and evening classes, are the most frequently chosen options. However participants note that this is not a regular practice as there is a lack of knowledge, at both an individual and manager level, about where to find training relevant to an individual's skills gap and level of experience.

Respondents pointed to a variety of training options they had undertaken including paid for classroom and online courses, free online courses, on the job mentoring, MOOCs, books, community events and university courses. With a variety of options available, and little clarity about which options offer the best impact, managers and individuals alike are confused about what to choose.

Organisations are therefore tending to meet skills needs through hiring rather than development as they feel this gives them more control over the acquisition of skills that are key to competitive advantage. Put simply, as data science becomes a tool to beat the competition businesses are reluctant to place their trust in what they perceive to be unproven or unclear training approaches.

Question 4 - On a scale of 1-5, where 5 is excellent, how do respondents rate their strengths in the following areas of data science?

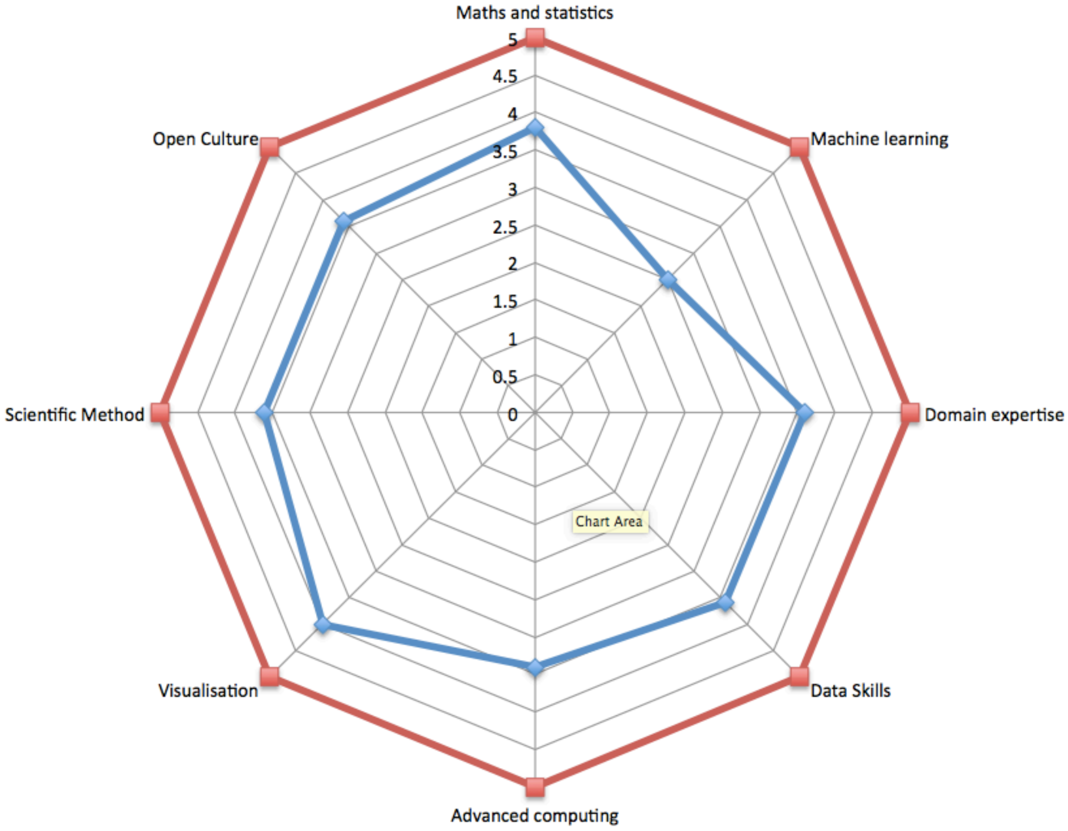


Figure 9: Results of question 4 of face-to-face interviews

Figure 9 provides an overview of how the combined responses of participants map to the 8 key areas. At this early stage, machine learning appears to be a weakness, however all areas require a focus, and further data is required in order to produce actionable evidence.

Question 5 - What are the essential skills for individual data scientists?

A new finding from the interview data is the need for data scientists to have good business skills in order to be able to understand business needs and to be able to effectively communicate insights. This is different from domain expertise in that business skills are the skills required to make domain expertise effective.

The demand analysis will therefore look in more detail at whether these skills are standard business skills that can be met through existing training channels, or whether the business skills data scientist require are unique in any way. Finally, we will explore if EDSA should seek to develop new courses and content to meet business skills needs.

Question 6 – How easy is it to find appropriate data science training?

We found that there appears to be a correlation between the difficulties of finding training with the difficulty of finding skilled people. This early finding needs to be explored further.

The best way to develop skills is by getting hands-on experience, as data science must be learnt in a commercial as well as an academic environment. That said, respondents who had taken data science related degrees pointed to the value of their academic learning.

The on-going development challenge is finding the right type of training for individuals, because training needs are unique to each individual. This is particularly challenging in data science, a new discipline without established competency standards and curricula.

“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...there is no 'one-fits-all' training course.”

Question 7 – How difficult is it to find skilled people in each of the 8 skills areas?

There are challenges in finding skilled people in each domain. One of the key problems is finding people with domain expertise, which is typically acquired through experience. A number of respondents suggested that the problem with domain expertise is that it needs to be combined with business skills such as communication, as discussed above.

However because respondents at this stage were mostly data scientists, without recruitment responsibilities this question needs to be explored further with managers in the next stage of the demand analysis.

Question 8 – Are there any sector specific challenges?

“*Data scientists at Facebook are computing scientists, who have a statistics background. A data scientist at a consumer goods firm probably wouldn’t develop a platform because they’re not working in real-time recommendation engines but they’re going to have a really, really heavy statistics background”*

Early responses suggest there may be differences in the data science skill domains that are most important across different industries - for example, the retail industry may emphasise deeper mathematics skills.

However, at this stage no specific sector differences in skills emphases have arisen and no specific challenges for sectors. This may be because many of the key skills are the same regardless of sector.

"I think a lot of things that we do, for example, basic understanding of statistics, hacking skills, those things are agnostic to the industry that you're in."

However we also recognise that this may be due to the way this question has been put to respondents and we will therefore be separating the question into two parts "What specific skills are important in your sector" and "What challenges do data scientists who work in your sector face?".

Question 9 - What are the most important factors in successful training for organisations?

The most important factor for successful training is hands-on experience. Training must therefore compliment and build on practical experience.

Live business training programmes were identified by some participants as an efficient way to learn in a realistic environment, where complex subtleties become more apparent and the larger scope and repercussions for errors speeds up the learning process.

Workshops are also noted as useful settings for collaboration and hands-on practice. However, participants noted that successful training should not be too time exhaustive. MOOCs were suggested as helpful methods for students to learn in their own time and find modules specific to their own skills gap.

Accreditation would be useful as a recognised stamp of quality, especially in a business environment.

5.1.1 Summary and recommendations

Based on our early findings, the project is taking the right approach to meet the development needs of data scientists.

There are areas that need to be explored further in the demand analysis, in particular how to address an emerging need for data scientists to have a range of business skills, specifically the ability to effectively communicate with businesses.

At this stage, our interviews provide further evidence of a lack of suitable training options to meet demand, and a lack of clarity about the best way to develop skills. The result is that organisations are tending to hire rather than develop, which is seen as an expensive but essential investment strategy in a domain that offers them competitive advantage.

Improving the choice of training and guiding learners to the best approaches are core to this project, and could provide companies with the confidence to make investments in developing rather than acquiring people. To make this happen, the interviews we have conducted provide us with some initial areas to focus on, and ways we can improve the approach to get more insightful answers.

5.2 Online survey findings

In this section we explore the findings from the initial online survey responses. An overall summary can be found in section 5.2.2. The online survey was developed following the design of the demand analysis study approach in D1.1. This initial phase, up to M6, enabled the testing and the development of the survey between the project partners, before a wider European launch. This section looks at the results from the initial 13 responses to the survey, for each question.

Required skills

Users are asked to rank the importance of skills in a data scientist's skillset. Figure 10 below, shows the distribution of answers for the eight key areas, introduced in D1.1.

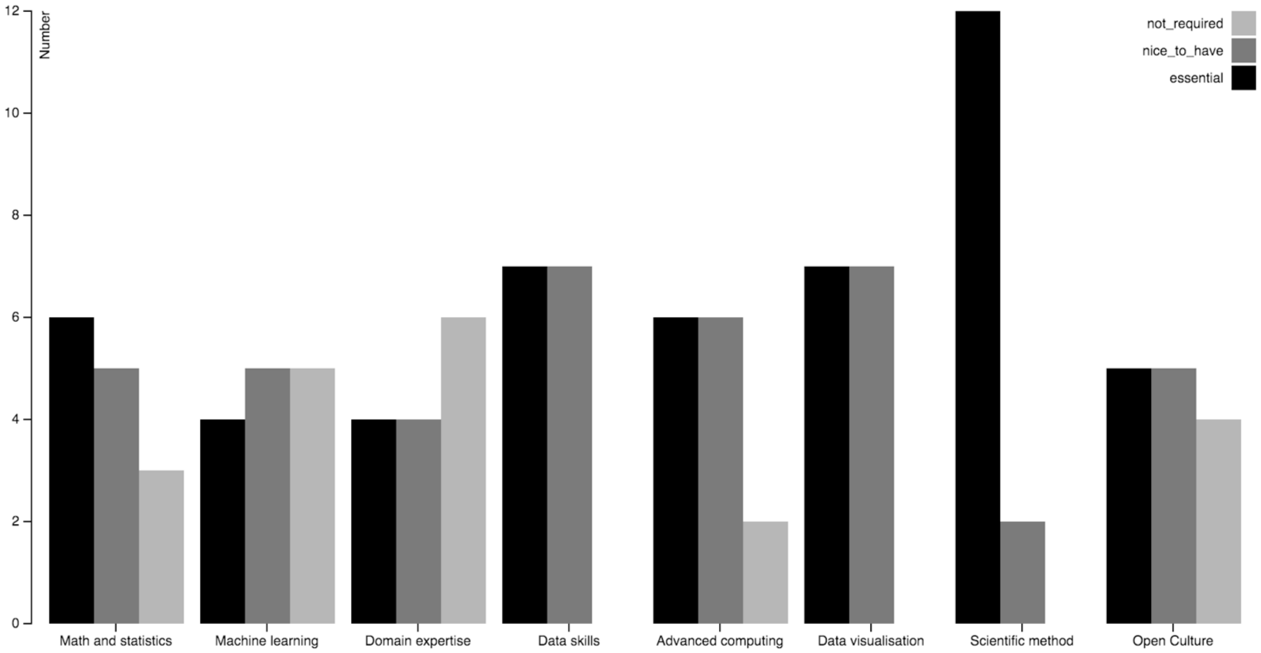


Figure 10: Essential skills for a data scientist

Figure 10 shows one clear trend - expertise in scientific methodology is considered an essential skill for a data scientist. As some responses from this initial small sample were received from partner institutions with an emphasis on academia, perhaps this could be anticipated.

Other important areas are 'data skills' and 'data visualisation', to which no one selected 'not required'. Interestingly, 'math and statistics' received several "not required" responses, which contradicts the traditional strong link this area has shared with data science in other research[6]. The same is also seen in the results for 'domain expertise', which could suggest that data scientists are seen as 'contractors' who can be assigned to any company or problem without the need to fully understand the domain.

As this early stage, and with a limited number of contributors, it is not expected that these findings are reflective of all sectors and countries.

[6] Timeline of data science – http://davetaz.github.io/EDSA/timeline.html

In addition to the key areas in the survey, contributors suggested a number of other areas. These are outlined in below.

Table 3: Additional skills for data scientists

Essential	Nice to have	Not required
Team work	Telling stories with data	Cobol programming
		Waterfall project planning

Capability and capacity in skills

In addition to rating the importance of skills, participants are asked to rate their capability and capacity to use this skill effectively. The radar chart in Figure 11 shows the combined average result from the survey results collected.

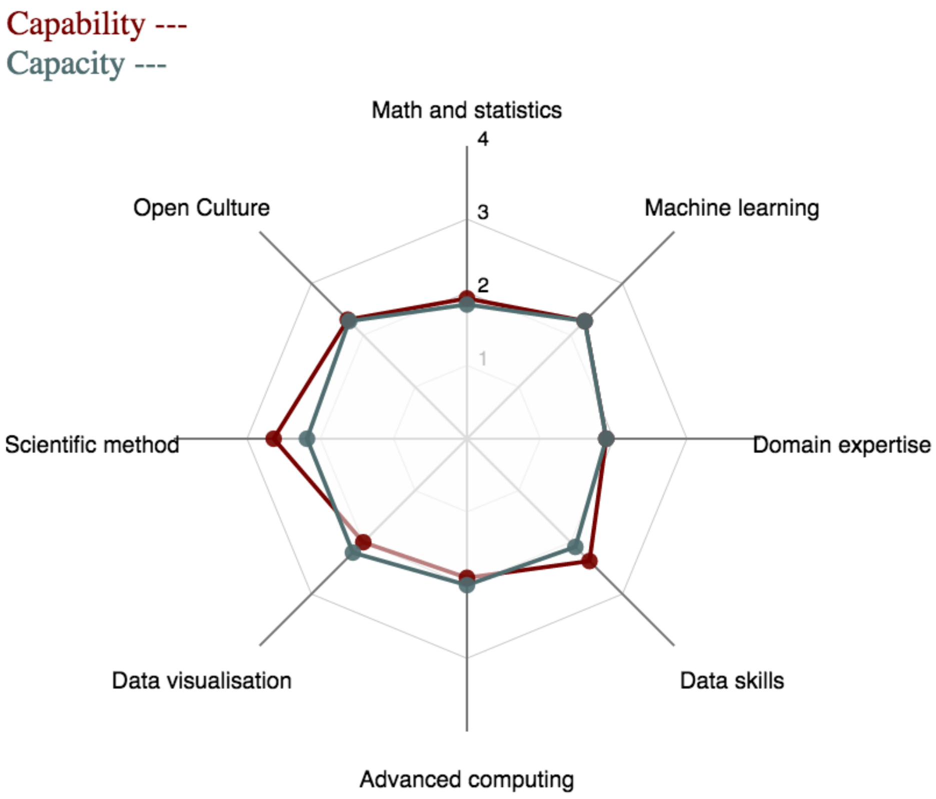


Figure 11: Capability and capacity from online survey

Figure 11 shows a clear correlation between the skills rated as essential in the previous question, scientific method, data skills and data visualisation, and the capability and capacity in these areas. Although marginal at this stage, these results are higher. This suggests that the essential skills are those that the organisation or sector currently has greatest knowledge of. This suggests the acceptance and needs of other skills is not as clear when knowledge of these skills and their benefits is not directly known. Further collection and analysis of results is needed to see if this is a trend across Europe.

Training delivery requirements

One of the final questions of the survey asks users to rate the important characteristics of successful training and delivery. From the initial responses, 'face-to-face' training is seen as the most essential approach, with webinars and eLearning closely behind. Interestingly, 'translation from English' received mainly "nice to have" responses, perhaps due to those responding working in a country familiar with English as a dominant language. At this early stage, the result already demonstrates a clear alignment with the intention of the EDSA project to focus on face-to-face training, eLearning and webinars.

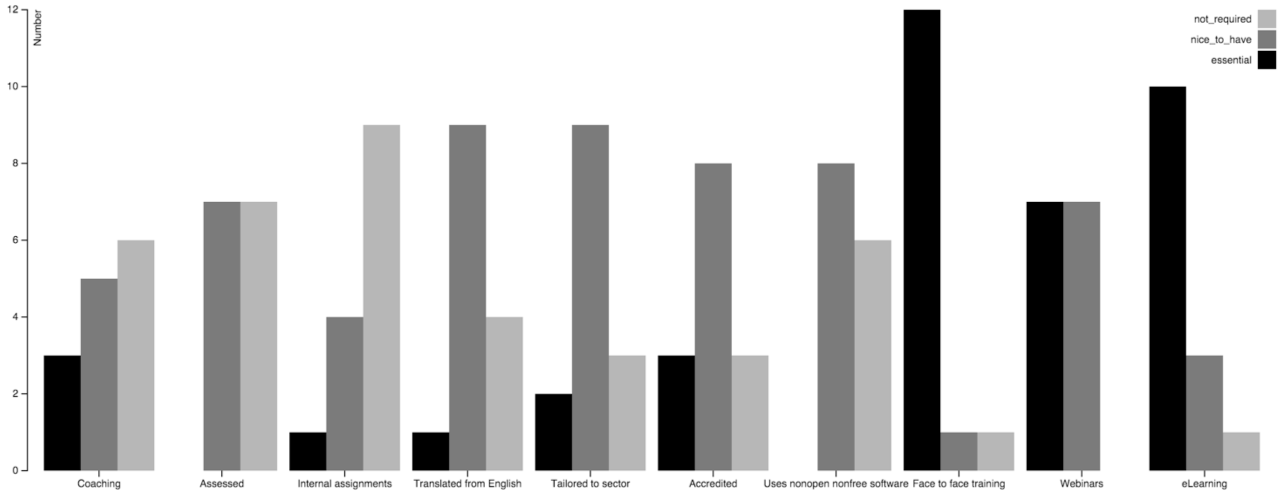


Figure 12: Training requirements for data science training

5.2.1 Summary and recommendations

The online survey has already introduced some interesting findings. Perhaps most intriguing, is the link between demand and existing knowledge. It would seem that the most essential skills for a data scientist to have are more of the same skills that people in the organisation already have. This can be found by comparing the answers from the "Required skills" question in Figure 10 and the results of the "Capability/Capacity" question in Figure 11.

Figure 11 also shows early suggestions that there is still a great need for skills expansion in all areas of data science. Combined with the results in Figure 12, there is clearly opportunity to deliver face-to-face as well as online courses and webinars. At this stage, the question of language does not seem as important to respondents. However, this is an area that will need to be explored further following a wider launch of the survey.

The survey will be launched to a wider audience after M6. Due to the custom nature of the survey, it was essential to first test its functions with a limited number of respondents in order to discover any issues with implementation. The issues are listed in the project dashboard Github repository [7]

It is intended that the survey be rolled out in stages to communities of interest as well as targeted at specific countries and sectors to allow us to gather as many results as possible.

[7] EDSA Github - https://github.com/davetaz/EDSA/issues?q=is%3Aissue+is%3Aclosed

5.3 Automated data collection findings

This section outlines the results of initial automated data collection from LinkedIn. There are a number of key findings:

- There is a steadily growing trend for job adverts in data science across all skill areas.
- Traditional disciplines and job roles, such as maths, statistics, business intelligence and databases, still dominate the majority of adverts.
- Some emerging disciplines are seeing a great number of adverts, including big data, tableau, and interaction. These adverts emphasise the growing importance of visualisation and data skills.
- The UK dominates all searches and must be removed in order to reveal other trends.
- Automated translation of key terms can be unreliable and should be explored further.
- Searching in the dominant language of the country does not always reveal all the jobs as many are advertised in English.

In addition to these findings a more in depth analysis has been carried out by Aba-Sah Dadzie from the Open University, who was able to take the project's open data and perform her own in depth analysis. The results of this have been kindly shared with the project and are outlined in section 5.3.1.

5.3.1 Automated data collection analysis

Initial analysis was focused on the temporal element in the data, looking at the trends over time and the dominance of English "gb" in the field.

Figure 13 shows the trend in three timeline plots for the period '11 Mar 2015' to '09 Jun 2015', focusing on demand for the top level skill "data visualisation", for 'gb' countries. Trends are similar across all sub skills for all but D3.js, which records no counts. There is a small peak early in the plot, and a sharp rise from 17th Apr to the 20th, peaking on the 19th. Beyond this there is a gradual rise for the rest of the period, with a few small dips.

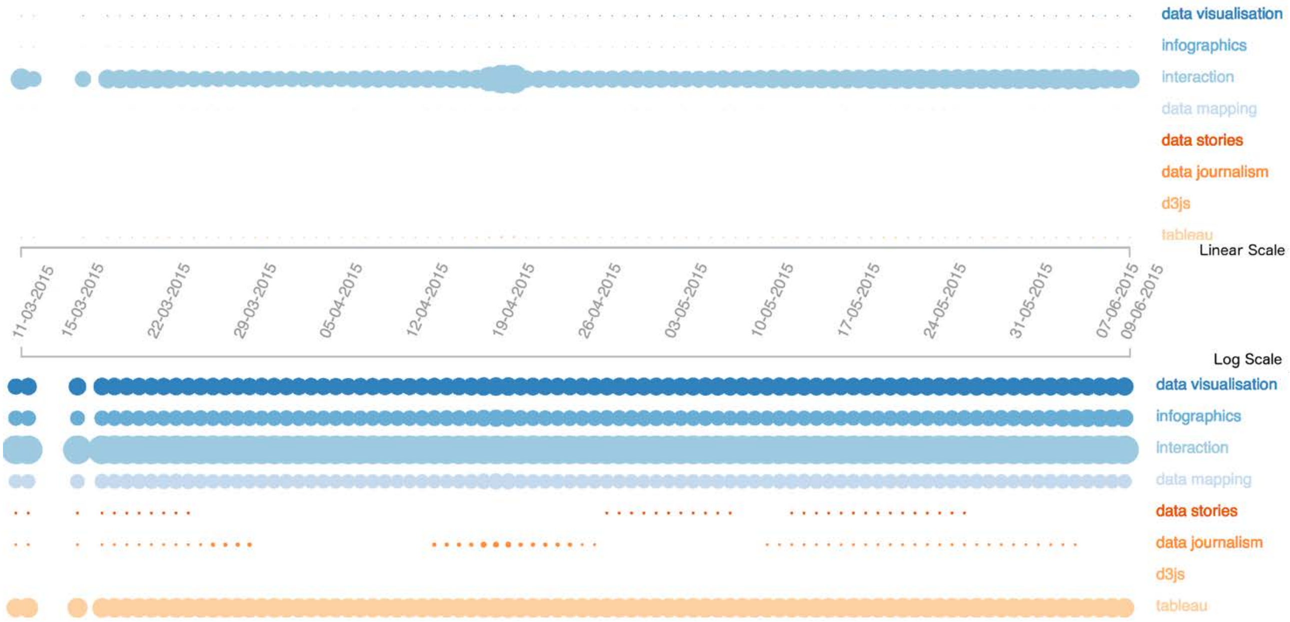
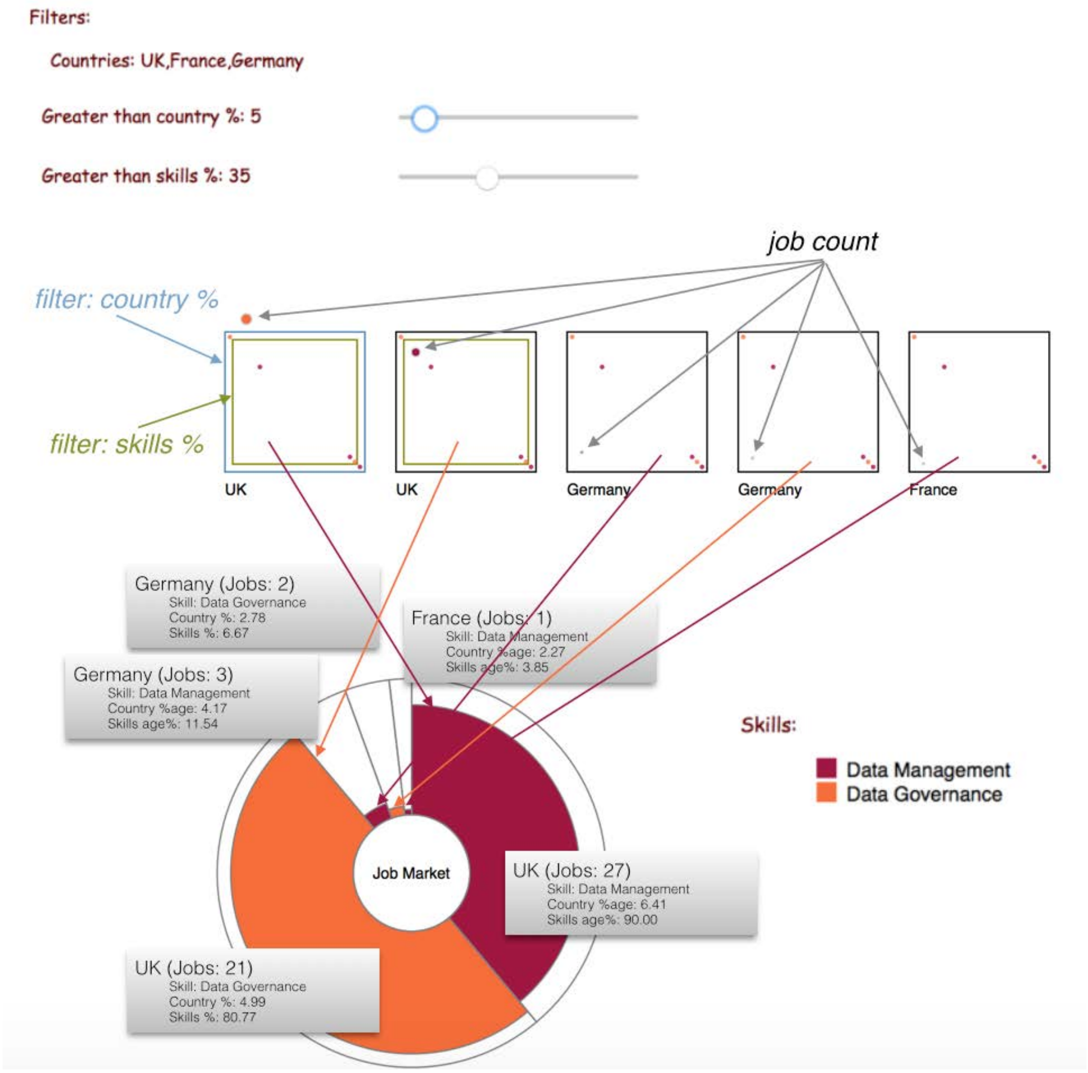


Figure 13: Daily demand for 'gb' jobs, linear and log scale

One skill ‘interaction’ records counts up to twenty times greater than all others, suppressing, as a result, trends in the others. Using a log scale (seen at the bottom plot of Figure 13) suppresses this dominance showing similar trends in other sub skills. By stacking the journal plot with the linear scale on top of the normalised plot we obtain two gains. We are able to examine relative trends for each skill, still within the context of overall demand patterns, but with little increase in cognitive load.

We looked next at a data snapshot, taken in April 2014, aggregated for five countries, excluding the UK, on eleven key topics in data science. While the picture of demand will have changed in 2015, the variables to be examined remain the same. We use small multiples to examine multiple attributes simultaneously, with the overall result shown in Figure 15.

Using small multiples allows the investigation of trends in the data and removal of dominant statistics. Figure 14 demonstrates how small multiples can be used to filter a big dataset to identify particular aspects and trends in that data.



Filters:

Countries: UK,France,Germany

Greater than country %: 5

Greater than skills %: 35

Figure 14: Small multiples example

At the top of Figure 14 are the filters, here set to only include three countries (UK, France and Germany). The other two filters are adjustable and change the way the data is displayed. The first filter (Greater than country), can be adjusted to show if a particular demand for a skill in a country is greater than a certain percentage of the overall demand. If this criteria matches then an outer blue box is drawn around the country and skill shown below. In Figure 14 it can be seen that more that 5% of the overall demand within the country for data science jobs mention the key word “Data Management”. The second filter will draw an inner box on the diagram if the demand exceeds a certain percentage of the demand across the whole of Europe. In Figure 14 we can see that over 35% of the demand in Europe for both “Data Management” and “Data Governance” is in the UK.

There are two sets of dots inside the box plots shown in Figure 14. Those on the diagonal, from top left to bottom right, represent the overall demand from jobs in that country for the selected skills. The second dot, which is singular, on the left (marked job count) shows the job count for the particular skill being represented in that plot.

Figure 15 removes the UK and shows the demand with the filters set at ‘>30%’ within each country and ‘>25%’ for Europe. Figure 15 stacks, from left to right and top to bottom, mini plots showing, in descending order, raw demand for each skill for each country, followed by skill and country percentage. Dynamic sliders are used to compare skill distribution within a country (see blue outer border) and each skill as a percentage of the total demand for all countries (in the dataset, including the UK – olive inner border).

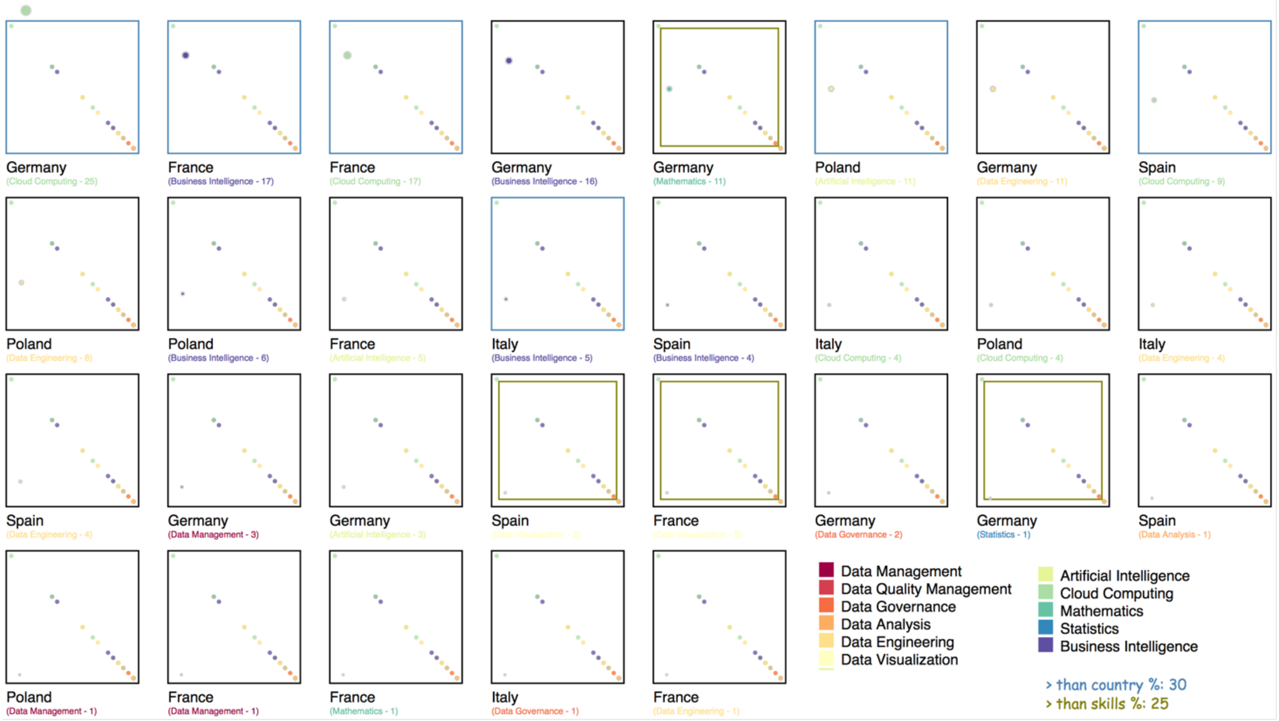


Figure 15: Small multiples to investigate trends across three attributes – skill count and percentage per country, and skill distribution (%) by country. Data for the UK is removed.

Blue borders are found predominantly at the top, showing top-heavy demand for selected skills, such as Cloud Computing in Germany. This is mirrored in the colour-coded line plot for the full dataset overlaid on each mini plot, which shows a long tail with very small counts per skill and country. Olive borders are more randomly distributed, for example, 50% of the demand for Statisticians is in Germany, while the other 50% is in the UK.

It should be noted that while the methodology is correct, this analysis should not be considered significant until further data is collected.

Figure 16 uses a space-filling technique, aster plots and a variant of a nested pie chart, to examine further skill distribution within each country, including the UK, using again the small multiples technique.

Figure 16 shows, on the top left, an overview of total demand for each country, then distribution by skill. Area maps are used to count for each slice. For the first aster, height maps to number of skills per country, up to 11, and for all others, skill percentage, over all countries. While the UK dominates all others, the individual country plots reveal a degree of similarity in distribution of skills. Business intelligence, data engineering and cloud computing are in demand across all, followed by artificial intelligence, which is highest in Poland. One skill, data quality management, records one count in only one country, the UK – so slim that only by thickening the borders of each inner slice, to provide an additional visual cue, is it recognised. Here, we see the power in multiple perspectives – this is highlighted in the matrix plot (see Figure 15) for the full dataset with the skills slider set to the maximum (100%).

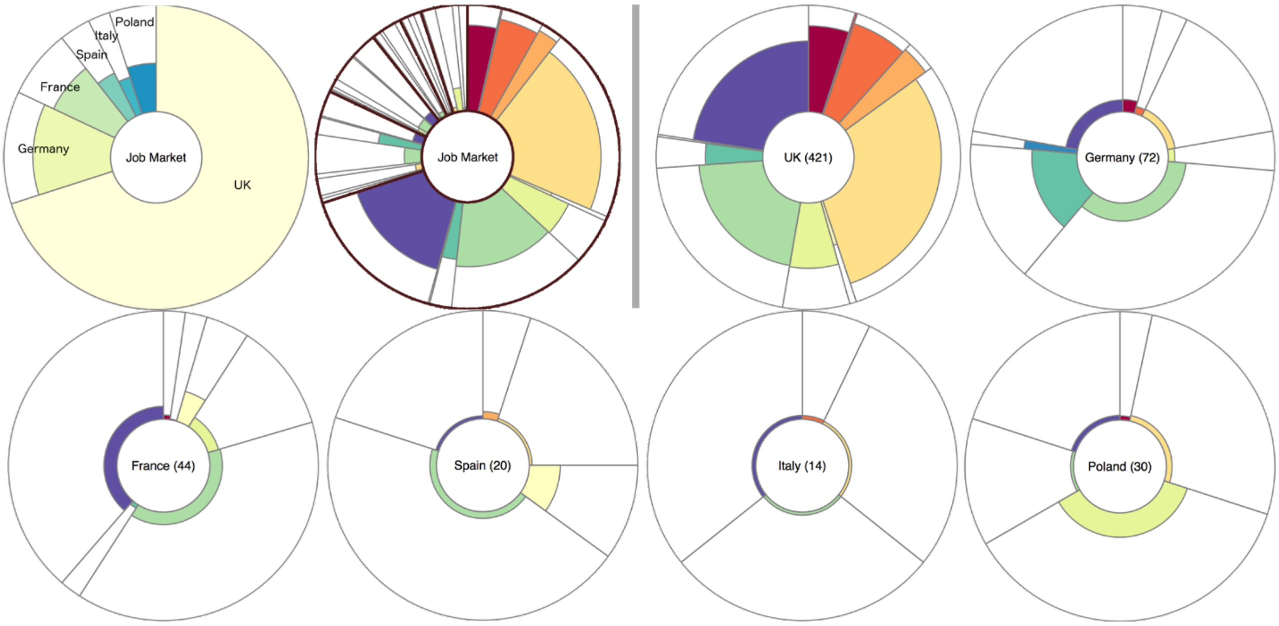


Figure 16: Aster plots for job demand for top 6 countries

Although the UK dominates much of the data, removing it or reducing its visibility shows clear patterns across parts of Europe. It should be noted that the data does appear skewed. The skew may be due in part to differences in terminology usage and interpretation across regions or the translator used. Furthermore, the data collected to this point comes from a single web source.

5.3.2 Summary and recommendations

Data science and its related skillsets and profiles are still emerging. The automatically collected datasets reinforce this finding. There are clear established areas where demand is high across Europe, in areas including business intelligence, data engineering and cloud computing. Showing a clear need for data skills with contextual knowledge. Other areas, such as data visualisation, interaction and artificial intelligence show emerging trends with steady growth but not to the high levels of other areas. There is also a large "skew" towards the UK, which shows that perhaps there is a terminology problem when talking about skills that needs to be investigated further. Using additional datasets and exploring refinement of the existing dataset will enable further analysis and an accurate depiction of Europe overall.

6. The demand for a European Data Science Academy

The final question of the one-to-one interviews aims to access the demand for a European-wide data science initiative - what sort of initiatives would respondents like to see emerge to help fill the skills gap? Responses to this question will be used to inform the project's work in WP5 - the creation of the European Data Science Academy.

The question enables us to gather information about existing initiatives on a national and international level, to assess the position a European initiative could take. Respondents are also asked what that initiative should offer to address the data science skills gap in Europe.

6.1 Summary and recommendations

"There's a lot of community meet-ups in the sort of hubs of this, so in London and other parts of Europe there's a lot of hubs you can go to regular weekends and things, but where we are we're a bit out of the way of all that."

Broadly, respondents were in favour of the implementation of a European Data Science Academy in that it should provide a broader choice of trusted courses and content, delivered in a variety of formats. However, respondents were also wary of the number of communities and related communities who are also doing similar activities and stressed the importance of not replicating or attempting to supersede existing valuable communities. The main difference between some existing communities and EDSA is that many communities focus on a specific skill and not on the ability to grow knowledge at the same time. The interviews revealed a demand for a community where people can learn across sectors, domains and expertise, share experiences and boost research possibilities.

To further address the demand for data scientists in Europe, the training must also look to facilitate a culture shift or attitude change, to make people and businesses in Europe more data-driven in decision-making processes.

It is still early to clearly define what the EDSA should be, and how it fits with other communities. What is clear is the need for EDSA to bridge all the various skills without specialism and convene key materials that guides both data scientists and managers to successful strategies for growing data science capacity in their role.

7. Evaluation and conclusion

The demand analysis is at an early stage, however initial results suggest there are gaps in data science training that specifically address more than one discipline or skill simultaneously. The initial results also suggest that any future provision of training through the EDSA structure will also need to be multidisciplinary. It is clear that a number of communities and sources of training already exist covering specific skills in significant depth. The key for the data science community, and EDSA is to distil the curricula and provide clear guidance to learners about the best ways to access courses and develop skills, whatever their level, and whatever their preferred way to learn.

In addition to this study, the EDSA Market Analysis conducted as part of WP2 provides similar insight from curricular available in the UK. This analysis results in three conclusions:

1. The majority of data science courses focus on specific skills with only a small number of courses available that give an introduction to data science. There are no shorter courses that offer more than an introduction to data science.

2. Programming languages, tools and technologies are wide spread and varied in use. Frequently used tools include: Python, Hadoop, R, Perl and Java.

3. These courses offer places and collaborations with companies in specific technological sectors. Examples include: Google UK, BBC, IBM, Amazon UK, KPMG Digital, Unilever, Starcom and Thomson Reuters.

Foundation courses in data science and big data will provide overviews of the topics and contain clear pointers to additional materials in each area. Additionally the “Finding stories in data” course focusses on the ability to communicate effectively, an area of weakness for existing data scientists according to the demand analysis study. Expanding from these foundations, the curricular, outlined fully in WP2, will address key gaps in training in the areas including data visualisation, computational thinking and stream processing. In all cases, interactive online resources, with rich teaching notes will enable the diverse implementation and delivery of the curricular across Europe online and through face-to-face training.

There is still much to do - more data is needed from countries and sectors across Europe to meet the project’s targets. Not gathering enough data is the biggest risk to the project, however it may not affect the conclusions from the demand analysis. This is something that will be evaluated throughout the next phase to ensure an accurate representation of views is presented, even if not backed up with every required interview. One key area of focus will be the sector coverage, particularly into sectors that are not primarily technological. This effort has already started with surveys completed with representatives from transport and agriculture sectors, and interviews in new sectors, including finance, already planned during the next phase. Expanding the reach of the online survey and inclusion of datasets from other projects will also provide more information to derive findings, ensuring an accurate and wide reaching result.

Data Scientists and Data Managers

The McKinsey report on “big data”[8] made clear that the lack of trained data scientists and analysts is only part of the problem. A shortage of 1.5m managers willing to employ data scientists is also an obstacle to industry getting value from data. EDSA should therefore address this challenge. This is reflected in the early results collected from the demand analysis. Job advertisements are still dominated by established and trusted skills, and recruiters seem unwilling to take risks on hiring people with newer skills. This is further evidenced with the analysis of online survey responses which shows that people assess the skills they already have as being as more important than the skills they may need, but are less familiar with.

To further address the demand for data scientists in Europe whilst addressing the challenges of the role becoming more widely valued, the project must therefore also consider how to support a culture shift and attitude change, to make people and businesses in Europe more data-driven in decision-making processes.

Business skills and different ways of learning

The project is taking the right approach to meet the development needs of data scientists. However, there are areas that need to be explored further in our demand analysis. In particular, how to address an emerging need for data scientists to have a range of business skills, specifically the ability to effectively communicate with businesses. This will help further with the instillation of culture change throughout a business to increase the number of managers willing to take risks.

Likewise, there is an emerging need for a complete coverage, multi-mode approach for training, beginning in higher education and continuing through professional and on-the-job development.

Valuable and positive feedback on EDSA and the demand analysis approach

So far, the demand analysis has seen very positive responses from those involved regarding the methodology and data collection techniques. The online survey has been well received and demonstrates the need to customise an experience in order to engage a contributor – a data-science approach. Responses from interviews have also been very positive and thought provoking, both for the project as well as the participants which has led to the establishment of a highly respected industry advisory board for the project.

Overall there has been positive support for a European Data Science Academy, with the caveat that the project may not be the first in this area, given its growing popularity and we should be careful to join and convene rather than override and control.

[8] Big data: The next frontier for competition - http://www.mckinsey.com/features/big_data

8. Appendices

Appendix A.- *Interview script*

Firstly a huge thank-you for agreeing to take part in this survey.

This call will be split into four parts:

1. First will we provide some background on the project
2. We will then ask you for some background information.
3. We will then ask you a series of questions that should lead to a free conversation, we ask you to be as honest as you can.
4. We will then wrap up and outline the next steps.

Background

This survey forms part of the demand analysis research for the European Data Science Academy project, funded by the European Commission.

The purpose of the project is to establish the areas of demand for data science training in Europe, which will inform the construction of new circular and training. The project will also look at the opportunities for a European initiative to spearhead data science training through Europe in order to fulfil the perceived demand on data scientist roles.

Your experiences are vital to help us establish the current view of data science and where these gaps in capability are.

Are there any immediate questions you would like answered?

Introduction

This survey forms part of the demand analysis section of the project. The purpose of the demand analysis is to evaluate the level of data science skills across Europe. The study will:

- Identify gaps in current training
- Validate the demand of skills training
- Analyse what specific training is needed across sectors and countries
- Discover the reasons for adopting data science training in organisations and the factors that affect successful development of skills
- Provide recommendations for curricula and course development

Any questions about the survey?

At this point it is important to note that we will be recording responses for transcription and later analysis. You will also have the option to request a copy of this transcription and recording, although we cannot permit changes. You can also opt to be contacted about the results so you can see what happened after the Analysis. At no point will you be identifiable to anyone outside of the project or it's funders with whom we are mandated to provide copies as evidence of our research. All analysis will result in anonymous, aggregated data that cannot be used to identify any individual. The aim of the project is to identify and reveal trends only. All anonymised derived data will be made available under an open license for others to use and benefit from.

Do you have any immediate questions?

Before we begin recording, do you still agree to take part?

Record.

Thank-you

This is _____ from the ODI and I'm joined today by _____ from _____ in _____.
Before we begin the survey can I ask you to confirm your name and state that they have agreed to take part in this survey?

Can I also ask you to confirm that you have agreed for the interview to be recorded and for the transcription to be used for analysis within the bounds of the project only?

Thank-you.

Firstly, can I ask you to describe your organisation briefly, the countries and sectors within which it operates?

Can I ask you to describe your role within the company?

Thank-you. This will help us ensure that we are collecting data from a wide variety of countries, sectors and roles.

Now let's move onto the survey.

This survey looks to establish the demand for data science training. Analysing gaps and opportunities. Can I ask you to describe the **impact of data science in your organisation? Q1**

(Question 1): What is the impact of data science on your organisation?

<u>Question prompts:</u>

- Is it changing the roles that people fulfil?
- How is the demand and challenge being addressed in your organisation, and has this been successful?

(Question 2): Are a new set of skills required?

<u>Question prompts:</u>

- What are these skills?
- Why is this skill so important to you?
- Do you see this skill as being fulfilled be a new role or expansion of existing knowledge?

(Question 3): What approaches have you taken to expand data science capacity in your organisation?

<u>Question prompts:</u>

- Have you or your staff attended any data science (or data) courses?
- Do any courses or providers stand out (and why)?
- Would it be useful to have more providers and courses and what would make these stand out?
- What other approaches do you take to developing skills? i.e. coaching, internal assignments

Part 2

The second part of the interview focuses on the definition and skills of a data scientist

Let's take a look at Drew Conway's Venn diagram you should have been sent ahead of the interview. For the study, we have used this Venn diagram, initial interviews and focus groups to establish initial key areas of data science - a total of eight areas, to obtain in depth skills information for analysis. You can find these eight key areas on the second page of the document.

Next we will look at these eight key skills and validate their importance.

(Question 4): On a scale of 1-5 (where 5 is excellent) how would you rate your strengths in the following areas of data science:

- **Math and statistical knowledge**
- **Machine learning**
- **Domain expertise**
- **Data skills**
- **Advanced computing**
- **Data visualisation**
- **Scientific method**
- **Open culture**
- **Other (if interviewee answers have identified additional 'core' skills)**

Question prompts:

- Are there any other key areas of data science for you?
- Do you have problems at scale?
- Are you more comfortable with certain types of data?
- Are there new skills you would like to acquire?

(Question 5): Considering the role of a data scientist as a single individual, how essential would you rate each of these skills to have for this person (Essential / Some knowledge required / not required)

- **Math and statistical knowledge**
- **Machine learning**
- **Domain expertise**
- **Data skills**
- **Advanced computing**
- **Data visualisation**
- **Scientific method**
- **Open culture**
- **Other (if interviewee answers have identified additional 'core' skills)**

(Question 6): Using the same eight categories from (4) rate each in terms of difficulty when finding appropriate training or skilled people.

- **Math and statistical knowledge**
- **Machine learning**
- **Domain expertise**

- **Data skills**
- **Advanced computing**
- **Data visualisation**
- **Scientific method**
- **Open culture**
- **Other (if interviewee answers have identified additional 'core' skills)**

Question prompts:

Can you expand on your key areas and why these have been more challenging?

(Question 7): Do you have any sector specific challenges to add to either list?

Question prompts:

Can you expand on these to offer details of the types of training you would like?

(Question 8): What are the most important factors in successful training for your organisation?

- **Face-to-face, webinars, eLearning, hybrid**
- **Duration**
- **Language**
- **Relevance to sector**
- **Accreditation**
- **Technologies used in course**
- **Technology used for delivery**
- **Internal assignments**
- **Coaching**
- **Others....**

(Question 9): What sort of initiatives would you like to see (if any) emerge to help fill the skills gap?

Question prompts:

- What do you feel the key roles of this institute should be?
- How do you feel a European Data Science Institute would help?

Wrap up - Finally, thank-you for taking part in the survey.

Would you like to be contacted further about the projects progress via the EDSA mailing list?

Would you like a copy of the recording or the transcription via email?

Would you like to be contacted about the results of this demand analysis survey?

We will now end the recording. Thank-you

Appendix B.- *Interview guidelines*

Summary of the demand analysis

The following section gives an overview of the demand analysis, which can be used as background information for interviewers.

Purpose of the demand analysis

The purpose of the demand analysis is to evaluate the level of data science skills across Europe. The study will:

- Identify gaps in current training
- Validate the demand of skills training
- Analyse what specific training is needed across sectors and countries
- Discover the reasons for adopting data science training in organisations and the factors that affect successful development of skills
- Provide recommendations for curricula and course development

What data are we collecting?

Primary data will be sourced directly from study participants via:

- One-to-one interviews
- Online survey
- Focus groups

This will result in the collection of both qualitative and quantitative data that can be analysed in combination with secondary data collected automatically from web services. This will consist of data from job sites and expert networks where trends in the evolution of skills and their proliferation in sectors and community networks can be analysed automatically

Where will the data be collected?

The results of the primary and secondary data collection will be analysed in order to produce an interactive dashboard view of Europe. This dashboard will allow users to filter the data collected by two key factors: Country or region and sector. Users will then be able to obtain a quantified specific skills gap on a topic level (e.g. statistics, machine learning) as well as links to courses offering training in these skills.

What will we do with the data?

While the quantitative analysis and automated studies are designed to provide a wide collection of results, the qualitative interviews play a key role in exploring topics in-depth. This depth will be essential to contextualise other results and identify key opportunities to develop training tailored for a particular audience and their learning requirements.

Interviews and focus groups:

- Initially, data will be analysed, anonymised and published via the EDSA Dashboard.
- Later, other in depth analysis included thematic analysis of the key discussion points will be conducted, anonymised and used to advise the findings and recommendations in the final study evaluation report, and the EDSA dashboard.

Online survey:

- Data will be analysed, anonymised and published via the EDSA dashboard.

Setting up an interview

The below section outlines all considerations for providing contacts for, or conducting interviews. Please advise the ODI team about any interviews that you intend to conduct. This will ensure that we do not send out multiple invitations.

Selection criteria

Please consider the following criteria when selecting individuals to invite to interview:

Criteria 1 - Role

- Practicing data scientist **or**
- Senior Manager/Leader (focused on skills development)

Criteria 2 - Country

- From an EU member state

Criteria 3 - Industry/sector

- A range of sectors and industries will provide us with valuable insight.

Conducting the interview

Providing the ODI with contacts for interviews

If you are unable to conduct the interview, please pass on the contact details of the interviewee, and make any introductions to the ODI team. We will be working with an experienced subcontractor who can carry out the interview, record and transcribe the content.

Partner conducting an interview

If you would like to conduct the interview, you must record the interview (audio only). This is a requirement of the project. Once you have recorded the interview, please send the audio file to the ODI team. We will work with our subcontractor to transcribe the interview for analysis.

Translations

Our subcontractor is able to conduct the interviews in multiple languages and translate recordings that are made - please advise the ODI team before the interview so that we can ensure that translation can be arranged by our subcontractor.

Recording the interview

Recordings can be captured in any digital format.

Various plugins for Skype are available. Details (depending on your requirements) can be found on the official Skype website.

https://support.skype.com/en/faq/FA12395/how-can-i-record-my-skype-calls?

Interview questions and topic guide

There are nine interview questions. These are outlined in the below table. We have also provided 'question prompts' these are included to encourage further in-depth discussion as appropriate:

- Part 1: The impact of data science and existing capacity and training.
- Part 2: The definition and skills of data science.

The 'key skills' interviewees are asked to consider in part 2 have been defined as part of the study design. Before part 2 of the interview, please introduce how these have been obtained. Detailed information of this process has been included in Appendix 1.

Part 1

(Question 1): What is the impact of data science on your organisation?

Question prompts:

- Is it changing the roles that people fulfil?
- How is the demand and challenge being addressed in your organisation, and has this been successful?

(Question 2): Are a new set of skills required?

Question prompts:

- What are these skills?
- Why is this skill so important to you?
- Do you see this skill as being fulfilled be a new role or expansion of existing knowledge?

(Question 3): What approaches have you taken to expand data science capacity in your organisation?

Question prompts:

- Have you or your staff attended any data science (or data) courses?
- Do any courses or providers stand out (and why)?
- Would it be useful to have more providers and courses and what would make these stand out?
- What other approaches do you take to developing skills? i.e. coaching, internal assignments

Part 2 - Interview information:

The following part of the interview focuses on the definition and skills of a data scientist as set out in the model for the demand analysis (A1)

The interviewee should be introduced to the Drew Conway's data science Venn diagram and each of its areas as well as the projects additional areas of interest (A1)

Following this introduction, the interviewer should recap the questions and answers so far and surmise how their answer fits or expands upon these core skills. The interviewee may also wish to add more to their answers at this point.

Any additional skills that do not fit the projects outlined model should be added as answers to the following questions.

(Question 4): On a scale of 1-5 (where 5 is excellent) how would you rate your strengths in the following areas of data science:

- **Math and statistical knowledge**
- **Machine learning**
- **Domain expertise**
- **Data skills**
- **Advanced computing**
- **Data visualisation**
- **Scientific method**
- **Open culture**
- **Other (if interviewee answers have identified additional 'core' skills)**

Question prompts:

- Are there any other key areas of data science for you?
- Do you have problems at scale?
- Are you more comfortable with certain types of data?
- Are there new skills you would like to acquire?

(Question 5): Considering the role of a data scientist as a single individual, how essential would you rate each of these skills (from Q4) to have for this person.

Please group the skills into the following three categories:

- **Essential**
- **Some knowledge required**
- **Not required**

(Question 6): Using the same categories from (4) rate each in terms of difficulty when finding appropriate training or skilled people.

Question prompts:

Can you expand on your key areas and why these have been more challenging?

(Question 7): Do you have any sector specific challenges to add to either list?

Question prompts:

Can you expand on these to offer details of the types of training you would like?

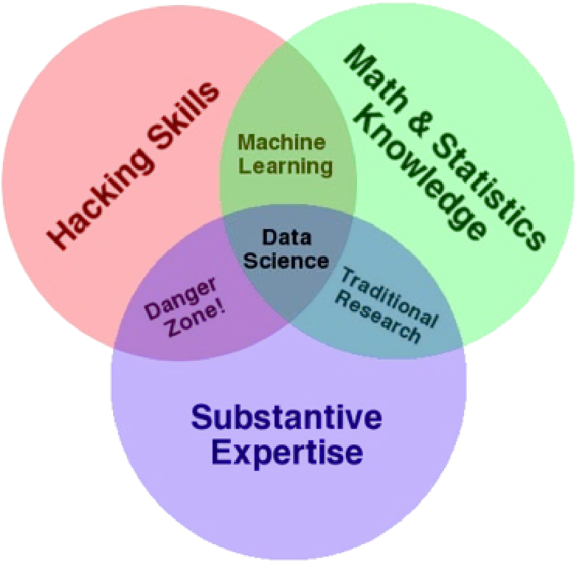
(Question 8): What are the most important factors in successful training for your organisation?
- **Face-to-face, webinars, eLearning, hybrid** - **Duration** - **Language** - **Relevance to sector** - **Accreditation** - **Technologies used in course** - **Technology used for delivery** - **Internal assignments** - **Coaching** - **Others….**
(Question 9): What sort of initiatives would you like to see (if any) emerge to help fill the skills gap? Question prompts: - What do you feel the key roles of this institute should be? - How do you feel a European Data Science Institute would help?

A1 Defining key data science areas

Before conducting part 2 of the interview, please introduce Drew Conway's Venn diagram, and explain for the study, we have used this Venn diagram, initial interviews and focus groups to establish the key areas of data science in the interview - a total of eight areas, to obtain in depth skills information for analysis.

Defining data science - how we have defined the key areas of data science for the purposes of the demand analysis

Historically there has not been one canonical definition for the term "data science". Today the most widely adopted definition comes from Drew Conway who presents a Venn diagram of the data scientist skills.



Having carried out a number of initial interviews as well as a focus group with the UK Government Data Science group, it was established that these key areas should be increased to eight. Separating these areas makes it easier to establish the specific capability rather than assuming individuals are able to combine the skills.

The eight key data science areas outlined in the interview - definitions and applications

Key area of data science	Definition	Data science application
Math and statistical knowledge	The theory and methods used in collecting, analysing and interpreting data to generate reliable robust conclusions.	Important to establish if collected data is reliable and establish how it can be analysed. Knowledge of distributions, averages and z-scores are key skills required.
Machine learning	The construction and study of algorithms that enable computer systems to learn from data.	Ability to train a computer to find trends in data, e.g. flooding risks.
Domain expertise	Having authoritative knowledge of a specific area or topic.	Essential in order to know the true meaning of data and impact of potential application and risks involved.
Data skills	The ability to collect, store, manage, process and clean data in a variety of types and formats.	In order to map a dataset will require at least two or three sources of data in different formats. The ability to clean, transform and combine data is essential.
Advanced computing	Selecting and using the right tools, techniques and algorithms to work with and analyse data. Includes programming and managing computer systems such as cloud and big data systems.	In order to remove the boundaries set by applications such as excel, it is necessary to have a knowledge of how to build and apply your own solutions.
Data visualisation	The ability to present data in an appropriate visual format that helps people understand its significance.	The ability to create powerful, customised infographics to tell powerful stories, such as the impact of an earthquake.
Scientific method	Rigorous methods of research in which problems are identified, hypothesis formulated and empirically tested and results openly published in a reproducible format.	Important to ensure that the approach is reliable and that an application that addresses flood risks can be taken seriously.
Open culture	A culture or way of working that promotes the spread of knowledge by allowing anyone, at an early stage, to access, use, adapt and share data, information and knowledge, without restriction.	Working openly allows community contribution, open communications about impact and use of open and online tools that can rapidly speed up projects.

Email template

Dear XXX,

Can you give 30 minutes to help the next generation of data scientists in Europe?

As a practicing data expert, helping us understand your experience will enable us to improve data science training across the EU. During the interview, we will ask about your experience, how you develop your skills and how you think things can be improved.

Your insights will help others develop their data science skills, and help the EU stay competitive in this crucial domain. Please reply to this email and let me know if you can help.

The interviews are being conducted in support of the European Data Science Academy project (EDSA). Over the next three years EDSA will:

- Analyse what sector specific skillsets are required for data scientists across Europe
- Develop modular and adaptable data science curricula to meet these needs
- Deliver training supported by multiplatform and multilingual learning content
- Evaluate how effective the new training and content is.

The EDSA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643937.

You can find out more about EDSA at http://edsa-project.eu/.

Interviews will be recorded and transcribed to enable the EDSA research team to generate insights about the demand for data science skills and to help develop new learning programmes to meet this demand.

- The recording and transcription of your contribution will not be published. It will be securely stored and only used for analysis purposes, and provided to the European Commission as evidence of the study.
- Results derived from your contribution will be completely anonymised before being used in any project outputs. They will not be attributed to you, or your organisation.
- You can request a copy of the transcription and recording.
- You may at any time request your contribution to be deleted

If you have any questions please reply to this email.

Kind regards,

Appendix C.- *Interviewee document*

EUROPEAN
— DATA SCIENCE —
ACADEMY

European Data Science Academy (EDSA) - Interview

As a practising data expert, helping us understand your experience will enable us to improve data science training across the EU. Your insights will help others develop their data science skills, and help the EU stay competitive in this crucial domain.

What is the European Data Science Academy?
The interviews are being conducted in support of the European Data Science Academy project (EDSA). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643937. You can find out more about EDSA at http://edsa-project.eu/.

Over the next three years EDSA will:

- Analyse what sector specific skillsets are required for data scientists across Europe
- Develop modular and adaptable data science curricula to meet these needs
- Deliver training supported by multiplatform and multilingual learning content
- Evaluate how effective the new training and content is.

Why are we conducting interviews?
Interviews will enable the EDSA research team to generate insights about the demand for data science skills and to help develop new learning programmes to meet this demand.

During the interviews we aim to:

- Identify gaps in current training.
- Validate the demand of skills training.
- Analyse what specific training is needed across sectors and countries.
- Discover the reasons for adopting data science training in organisations.

What will EDSA do with the recording of your interview?

- The recording and transcription of your contribution will not be published. It will be securely stored and only used for analysis purposes, and provided to the European Commission as evidence of the study.
- Results derived from your contribution will be completely anonymised before being used in project outputs. These outputs will include an interactive dashboard view of Europe available via the EDSA website. The dashboard will allow users to filter the view by country or region and sector, to obtain a view of a skills gap on a topic level as well as links to courses offering training in these skills.
- Results will not be attributed to you, or your organisation.
- You can request a copy of the transcription and recording.
- You may at any time request your contribution to be deleted