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

D1.1 Study Design Document

Deliverable Editor:	**Dr David Tarrant (Open Data Institute)**
Other contributors:	**Simon Bullmore (Open Data Institute)**
Deliverable Reviewers:	**Mandy Costello (Open Data Institute)** **Ali Syed (Persontyle) / Jean-Louis Lievin (ideXlab)**
Deliverable due date:	**31/03/2015**
Submission date:	**31/03/2015**
Distribution level:	**Public**
Version:	**Final**

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	3rd March 2015	David Tarrant	Initial version
0.2	10th March 2015	David Tarrant	Restructure following D1.1 feedback session
0.3	17th March 2015	David Tarrant	Added dashboard design. Review copy
0.4	25th March 2015	David Tarrant	Incorporated review feedback, expanded background and methodology
0.5	26th March 2015	Mandy Costello, Simon Bullmore	Final review copy
0.6	30th March 2015	Mandy Costello	Final document
1.0	31st March 2015	John Domingue Aneta Tumilowicz	Final QA
1.1	23rd April 2015	David Tarrant, Mandy Costello	Addressed comments from PO, Final document

Table of contents

List of tables

List of figures

1. Demand analysis research methodology

1.1 Executive summary

This document outlines the research methodology for the demand analysis study of the European Data Science Academy (EDSA) project. The purpose of the study is to evaluate the level of data science skills across Europe and identify gaps in training now, and in the future. This study will validate and ascertain demand so that the current strengths and weaknesses and gaps in the provision of data science training can be identified. The study will also make recommendations for curricula and courses that need to be developed to address gaps. Furthermore, the study is intended to discover the reasons for adopting data science training in large and small organisations and other factors that affect successful development of data science skills and capability.

Our approach is grounded in the following description of the challenge for European data science:

"Surviving in the new data driven economy requires a new set of skills; that of a data scientist. Managers must identify and prioritize 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."

Figure 1 below provides a high-level description of the overall methodology for the demand analysis, which will consist of the collection, and analysis of both primary and secondary data. .

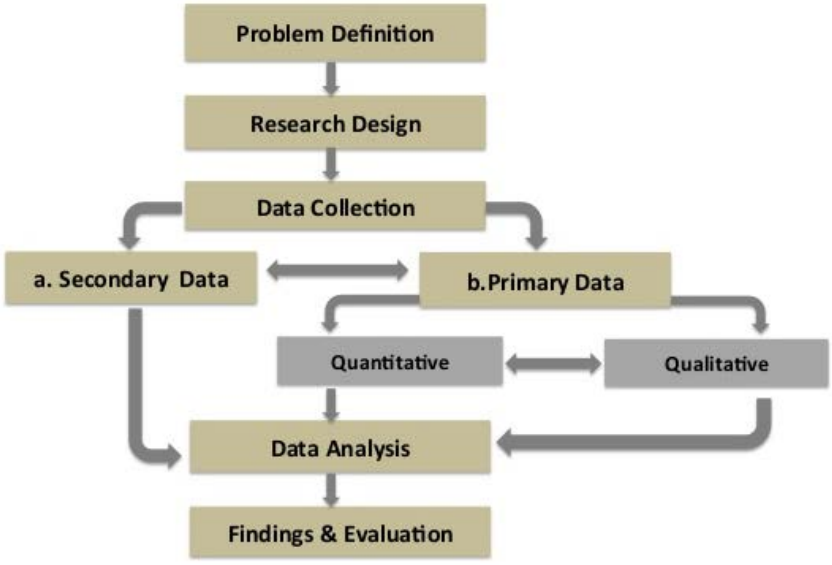


Figure 1 - Overview of methodology for demand analysis

Primary data will be sourced directly from study participants via one-to-one interviews and through online surveys. This will result in the collection of both qualitative and quantitative data that can be analysed in combination with secondary data.

Secondary data will be sourced and 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.

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.

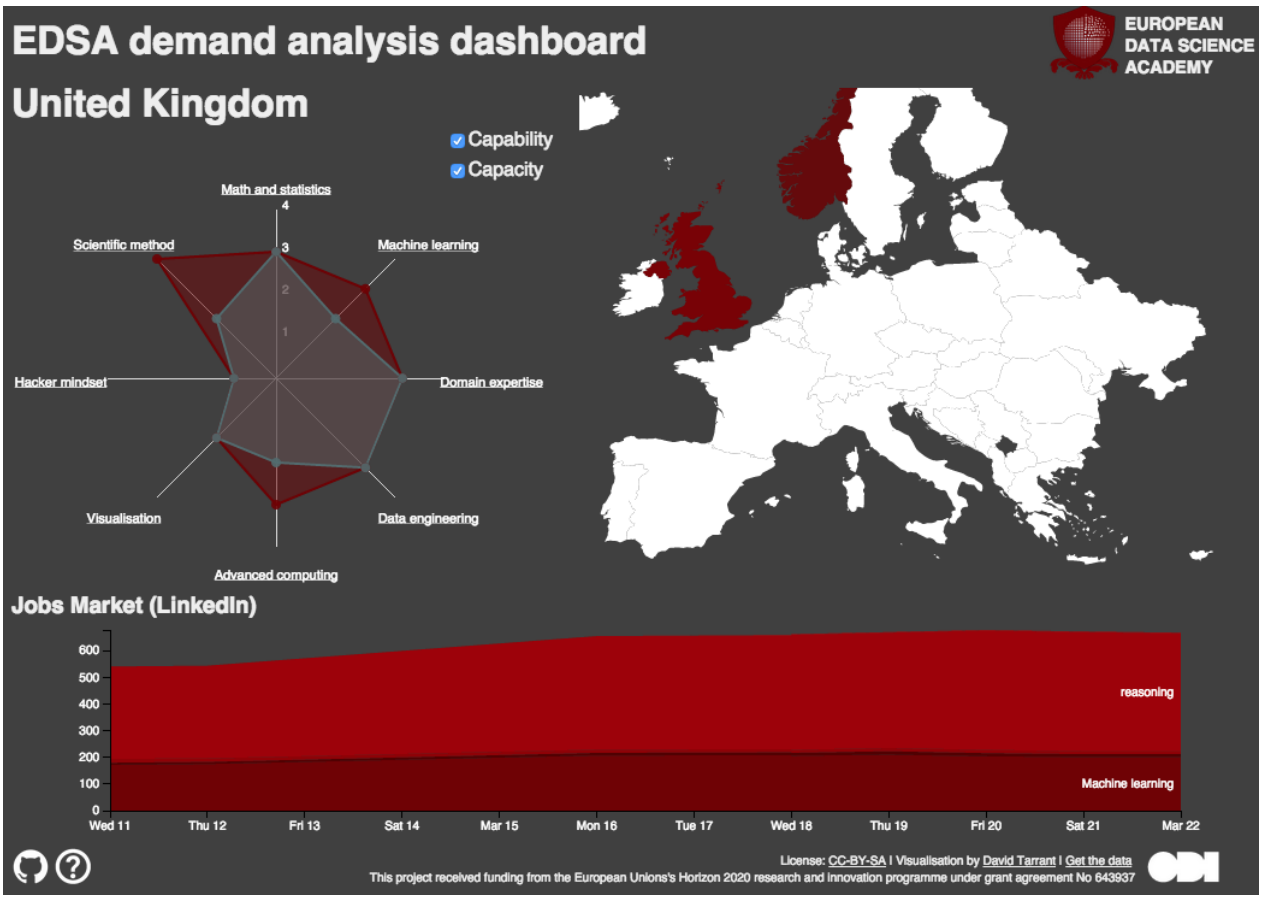


Figure 2 - Proposed EDSA demand analysis dashboard

2. Background

2.1 A brief history of data science

In 2011, McKinsey published a benchmark report[1] warning that the size of the data skills gap will be 160,000 by 2018 globally. While this report brought the issue onto a more global stage, the importance of data science is something with a history dating back to 1977, when The International Association for Statistical Computing (IASC) was established. The aim of the IASC was to "*link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.*"[2]

In 2012, Harvard Business Review termed data science "The sexiest job of the 21st century"[3] driven forward by increased access to technology, education and open data. This increased accessed has quickly opened up the opportunities for everyone to do "data science".

2.2 Defining data science

Historically there has not been one canonical definition for the term "data science". The above-mentioned definitions agree that data science is the extraction of knowledge from data and employs techniques and theories drawn from many disciplines. However, the combination and importance of different disciplines has varied over the years. Today the most widely adopted definition comes from Drew Conway who presents a Venn diagram of the data scientist skills (Figure 3).

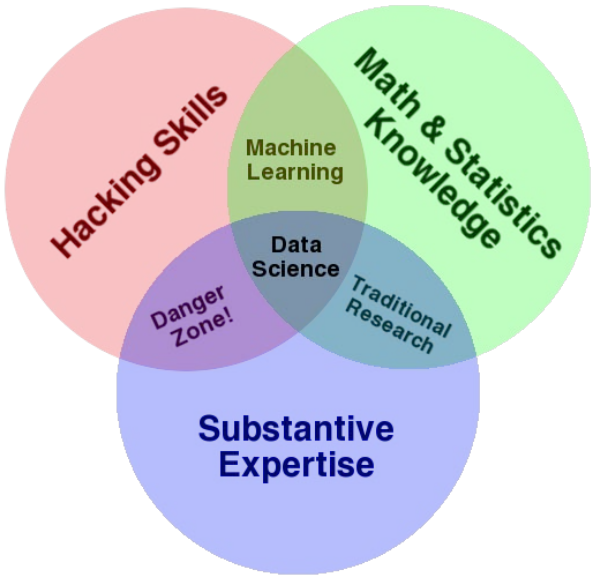


Figure 3 - Drew Conway's data science Venn diagram

[1] Big data: The next frontier for innovation, competition - http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[2] A very short history of data science - http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

[3] Data Scientist: The sexiest job of the 21st century https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Further studies have added a fourth cultural dimension, based on observations that modern data scientists do not fit traditional career pathways[4]. This is also related to the movement towards openness led by the open source and open data activities that now play a key influencing role in data science[5]. Such studies have observed that data scientists emerge from environments where there is no restriction on forming collaborations and combining skillsets in order to solve a real world problem. Additionally, a data scientist may prefer to work in a more agile way, focussing on exploring the unknown as opposed to making and executing a plan. These two aspects are often seen as risks to organisations that are protecting their Intellectual Property and carefully planning projects. With the global uptake in Open Data, the closed culture this produces is beginning to shift towards a more open way of working and some companies even see the closed culture as being higher risk than open[6]. We are still in the early stages of this cultural shift, evidenced by McKinsey stating that there is a lack of managers willing to employ data scientists, perhaps through fear of risk. Thus within the demand analysis it will be important to establish exactly where this problem exists, by country and sector. In order to achieve this, we have expanded Conway's Venn diagram to include a fourth "open culture" aspect to the attributes that make a data scientist:

- Maths and statistical knowledge
- Hacking skills (coding and data management)
- Subject matter expertise and presentation skills
- Open culture

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 four areas should be expanded to eight:

- **Math and statistical knowledge**

The theory and methods used in collecting, analysing and interpreting data to generate reliable robust conclusions.

Data science application: 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.

Data science application: 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.

Data science application: Essential in order to know the true meaning of data and impact of potential application and risks involved.

---

[4] Data Science, Moore's Law, and Moneyball - http://www.harlan.harris.name/2011/09/data-science-moores-law-and-moneyball/

[5] Data Science: What's in a name? - http://blog.revolutionanalytics.com/2011/05/data-science-whats-in-a-name.html

[6] AstraZeneca moves to Cambridge to improve collaboration with leading UK Scientists – http://www.bbc.co.uk/news/uk-england-21833207

- **Data skills**

The ability to collect, store, manage, process and clean data in a variety of types and formats.

Data science application: 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.

Data science application: In order to remove the boundaries set by applications such as excel, it is necessary to have 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.

Data science application: 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.

Data science application: 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.

Data science application: Working openly allows community contribution, open communications about impact and use of open and online tools that can rapidly speed up projects.

Here we have added “Data skills”, “Machine learning” and “Scientific method”, and replaced “Hacking skills” with “Advanced computing”.

The addition of “Machine learning” and “Scientific method” comes from the Venn diagram where there is a combination of the other skills. Separating these areas makes it easier to establish the specific capability rather than assume that people are able to combine the skills. “Data skills” has been added in order to specifically separate this from “Advanced computing” (Hacking skills in the Venn diagram) and reflect that throughout Europe there may be individuals who collect, store, manage, transform and analyse data as part of their daily role and that set of skills are subtly different from those who have computing skills. Finally “Data visualisation” has been added to reflect the importance of being able to present data in a compelling way. During early stage interviews and the initial focus group, there was consensus that this area was of critical importance, significant enough to be added to our study areas.

These eight areas will form the basis of the demand analysis for collection of both primary and secondary data. As the study evolves, it may be necessary to add or remove areas from this set. This will help establish a European view on what should be included in the definition of data science.

These eight areas, with the exception of Open culture, can be clearly subdivided into the many topics proposed for training development as part of the EDSA curricular development. A mapping of these can be found in Appendix 1.

2.3 A growing domain with a growing skills gap

Despite the lack of an agreed definition, the importance of data scientists has risen through a number of high-profile examples of data science at work - powerful demonstrators that process large amounts of data in real-time and present it back through engaging stories. Examples include the history of aviation[7], analysis of new financial sectors[8] and the changing nature of education[9]. These examples combine data from many sources; both open and closed, in order to present a compelling story to the reader. All use complex algorithms and statistics in order to reduce data down to that needed to effectively tell the story. Additionally all of them require substantive domain expertise to interpret and present the data.

Data scientists are however still a rare breed. Beyond the occasional data centric start-up and the data analytics department of large corporations, the skills scarcity is already becoming a threat for many European companies and public sector organisations as they struggle to seize Big Data opportunities in a globalised world. McKinsey estimate[10] that by 2018 the United States will require 60 percent more graduates able to handle large amounts of data as part of their daily jobs. With an economy of comparable size (by GDP) and growth prospects, Europe will be confronted with a similar talent shortage of hundreds of thousands of qualified data scientists, and an even greater need of executives and support staff who will need basic data literacy. The number of job descriptions and an increasing demand in higher education programs and professional training confirm this trend[11], for data science positions in less than a decade, with some EU countries forecasting an increase of almost 100 percent in the demand[12].

---

[7] 100 years of aviation - http://www.theguardian.com/world/ng-interactive/2014/aviation-100-years

[8] Show me the money - http://smtm.labs.theodi.org/

[9] The Singapore Education Story - http://educity.sg/

[10] Big data: The next frontier for innovation, competition - http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[11] Government calls for more data scientists in the UK - http://www.computerweekly.com/news/2240208220/Government-calls-for-more-data-scientists-in-the-UK

[12] Demand for big data IT workers to double by 2017. http://www.computerweekly.com/news/2240174273/Demand-for-big-data-IT-workers-to-double-by-2017-says-eSkills

3. Study methodology

3.1 Study objectives

The EDSA project has been formed in order to develop data science capability throughout Europe. As data science is a broad discipline, it is first necessary to carefully analyse the market, establish demand and see where the gaps are in skills training.

By combining primary and secondary data with expert views we anticipate that the demand analysis will affirm current areas of interest where training is already available. We will evaluate the various training offerings, using this data to help build the European Data Science Institute (EDSI), to lead data science training throughout Europe. Secondly, it is expected that there will be a number of gaps in demand for key areas of data science due to the evolution of the discipline and the stages of this evolution that are affecting the community, such as Visualisation Tools.

This document outlines the methodology for the demand analysis phase of the project. This phase has several objectives:

Objective 1*:* To identify the key areas of data science where training is required, in key sectors across Europe.

Objective 2*:* To connect to industry-leading players for recommendation to join the industry advisory board.

Objective 3*:* To ascertain the level of interest in a European Data Science Institute and factors that would make this successful.

Objective 4*:* To evaluate levels of demand for current data science training available from EDSI project members and other organisations, and gather feedback on industry views on this curriculum.

3.2 Study output

The goal of the study is to enable the production of a detailed demand analysis dashboard for Europe (shown in Figure 2). This interactive dashboard is designed to be both a demonstrator for data science and a way to present users with a compelling way to view and navigate through a large amount of data very quickly.

Initially, 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)
- Apply detailed facets to examine data, and show the strengths and weaknesses in capability, on a country-by-country band sector basis, as well as at European level.
- Understand the availability of related training from EDSI partners.
- Find links to courses offering training based on their needs.
- Contribute data where none has yet been collected.

The dashboard will provide a pan European view, clicking on a country or sector that lacks data will prompt users to engage in the survey or send a recommendation email to a colleague or friend and ask them to contribute data.

3.3 Overview of methodology

A number of different techniques will be used to gather the high quality and relevant data. Primary data will be collected from key stakeholders via interviews, focus groups and online surveys. Secondary data will be collected automatically from online sources and will be used to track market trends. Data will be collected using a combination of qualitative and quantitative techniques and later analysed and added to the dashboard. Adopting a hybrid approach will give maximum flexibility to the analysis.

In summary, three main techniques will be used for gathering data:

- Detailed **qualitative** interviews and focus group studies
- Online **qualitative** and **quantitative** questionnaires
- Automated **studies** of emerging trends

Each aspect is designed to reflect the human sociological model where facts, figures and trends play a key role in influencing opinion. Thus, by deduction, one should be a good reflection of the other. 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.

The automated collection of data from services such as LinkedIn and Monster will play a key role in providing evidence to complement the findings of the data collected via surveys. It will also allow the tracking of key markets over the course of the project and allow a temporal evaluation of the changing nature of survey results. This will allow the study to react to changes in emerging trends including the introduction of any new disciplines to the area of data science in a way a static survey with data collected at a particular point in time will not allow.

4. Qualitative methodology

4.1 Overview of the approach

For the purposes of the qualitative demand analysis we will focus on three main core methods for gathering results – one-to-one in-depth interviews, focus groups and online surveys. Figure 4 (from Power perceptions [13]) outlines effective techniques for collective qualitative data from communities where results are hidden from view. A combination of techniques can ensure the best spread of results and increase coverage. Due to the complexity of supporting online forums and communities, we are proposing to not use these techniques. Each of the other techniques provides good opportunity for overlap. For example, a person identified via a web survey may later wish to partake in a focus group or interview.

Figure 4 - Qualitative data collection techniques

In all cases, qualitative data gathering will involve industry leaders and practicing data scientists. In the original proposal it was stated that only industry leaders would be interviewed. However, only interviewing this group would lead to a one-sided view of data science training and leave out the valuable insights of the data scientists themselves. Learning from the practicing community will be equally essential to establish key success factors and opportunities to build a community of peer-learners who can help bridge the skills gap, and so individuals in this group will also be include in the data gathering.

[13] Qualitative data collection tools - http://www.powerdecisions.com/qualitative-market-research-experts.cfm#.VRpl6WTF9OE

4.2 One-to-one interviews

One-to-one interviews will be critical to get an in-depth view of the effect of data science in each country and sector. The objective of the interviews is to provide real insight into the demand, and the needs of each community.

Members of the advisory board will come from this community of engaged users. Further members of the community will be added later at the point of inception of the European Data Science Institute.

Interviews will be carried out both face-to-face and via telephone by skilled research interviewers. The interviewers will use a **topic guide** (see 4.2.1 below) to ensure that all objectives are expanded upon. Using a topic guide allows an interviewer to both address each objective while allowing space to further explore the responses obtained from each individual.

Interviews will be conducted and transcribed for further analysis for the Study Evaluation Report delivery (D1.3)

Analysis techniques will include thematic analysis of the key discussion points, which will be publishable via the demand analysis dashboard. Other in-depth analysis will be included in the Study Evaluation Report (D1.3)

In order to reduce crossover, some of the key quantitative questions have been combined with the topic guide to ensure usability on the EDSA dashboard.

4.2.1 Interview questions and topic guide

(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 by 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

Additional interview information:

The following part of the interview focuses on our definition and skills of a data scientist as set out in section 2. The interviewee will be introduced to the data science Venn diagram and each of its areas as well as the projects additional areas of interest.

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

A Full list of areas (and breakdown) is available in Appendix 1

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?

4.3 Online survey

Alongside the one-to-one interviews and focus groups, there will also be a short and engaging online survey. This survey will be accessible via the dashboard and enable collection of additional qualitative and quantitative data. We will also send out a number of targeted emails in order to help prompt contribution of data to the platform. Identification of candidates will be carried out through the ideXlab platform, by searching for people who are identified as a data scientist or related to the area (see Section 5.2 for details).

The survey will be designed to present many of the same questions as the interviews, however the responses will be on a numerical scale as opposed to text based. The focus will be to gather data for the study, and therefore the dashboard, while aiding the collection of data in countries where it is not possible to obtain qualitative responses.

4.3.1 Design

Part 1 - User background

The first part of the survey will ask the user their **country, sector, role** and **size of company**. These answers will be compulsory. At this point in the survey, users will be given the option to provide an email address or contact telephone number to enable a follow-up or possible one-to-one interview. The structure of the survey will vary dependent on if the user identifies themselves as a practicing data scientist, industry leader or manager.

Part 2a – Personal skills (for practicing data scientists)

Users who identify as a practicing data scientist will be asked to rate their capability and capacity in a number of key data science skills. This data will be used to ascertain the current capability of

practising data scientists throughout Europe and enable an accurate picture to be created of both capability and capacity to use these skills.

Part 2b - Skills and training analysis (for industry leaders and managers)

Users who identify as an industry leader or manager will be presented with an online version of Q3 and Q4 from the qualitative survey. Additionally users will be asked how many training courses they know exist in the area, what percentage of these are relevant to them and how many they have used.

A free form text field will allow users to enter comments about each area for qualitative analysis.

Part 3 - Priority identification Skills

The final part of the survey will ask users to prioritise the areas where they would like to see more training. The final screen will ask users to prioritise other factors such as method of training, language, accreditation and technologies.

4.3.2 The online survey and dashboard

The online survey will be available directly from the online dashboard. It will be promoted heavily in countries and sector where there is a lack of data, e.g. for particular countries. This is particularly useful as we have the potential to auto-fill some of the questions if people have come from one of these entry points.

After the surveys have been submitted the response will be reviewed before being added to the public and project datasets for inclusion on the dashboard.

4.4 Focus groups

Focus groups will be a key part of the strategy to discover the significant differences within each sector or community in which the focus group is involved.

Each focus group will consist of both managers and data scientists. Participants will be divided into groups, each with a moderator. The combination of large and small groups allows both the exploration of larger subjects in order to form consensus, as well as opportunities to discover the more personal challenges of individuals within groups.

Focus groups will be deliberately run for specific sectors or organisations in order to generate value for both the group as well as the project. Additionally, this will ensure that the group can be open about their experiences. Focus groups are run as half-day training workshops.

Each half-day workshop will have the following basic structure:

Focus group session 1: The beginnings of data science

The opening session allows time for introductions and establishes a common understanding of data science. The session is immediately interactive as participants are asked a number of questions including:

- Why does data science appeal to you?
- Defining data science “What is it to you?”
- How is it different to what you do now?
- What new set of skills are required?

It is of critical importance in this session to set the scene and focus attention on data science and not other more generic skills that are required. Building to the skills question ensures that the most relevant answers are likely to be offered. The Open Data Institute has previously run a series of sessions where the skills question was asked first. It was found that people were unclear what the context of the question was and thus answers had to be discarded. Following this session is a short presentation of the data science Venn diagram from Drew Conway (shown in Figure 3). The group's definition is then evaluated against this representation and each expanded where necessary.

Focus group session 2: Capability and capacity

The second session of the workshop focuses on the current capability of those working with data. In addition to focussing on the required skills, this exercise looks at the cultural capability required in the data science community using the eight areas derived from Drew Conway's Venn diagram (outlined in Section 2.2).

An exercise (see Appendix 2) has been developed to examine each participant's capability in these areas. Participants consider their capability from the levels suggested and produce their own radar diagram. The capability diagram shown in Figure 5 comes from a focus group with the UK Government Data Science group.

During this session participants are also asked to rate the requirement for each skill on a simple three-point scale: essential, some knowledge required and not required. Combining this with the capability and capacity figures will give a clear picture of priorities both on a country and sector basis.

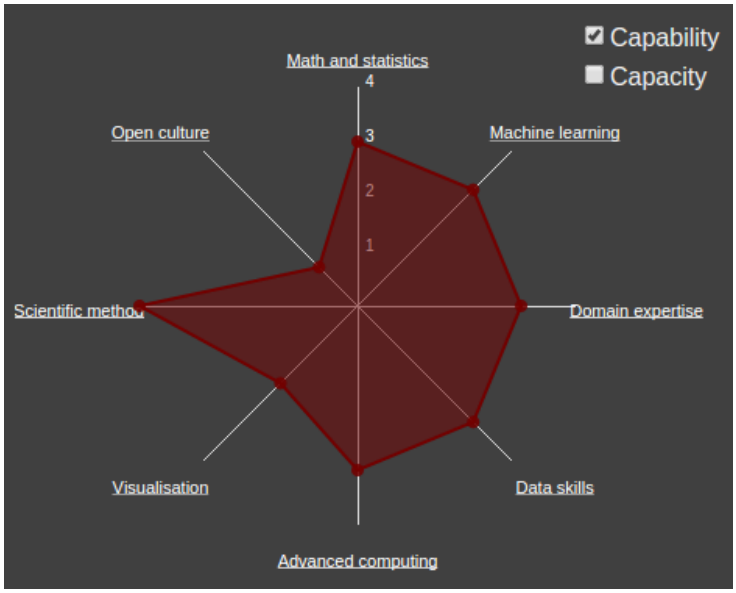


Figure 5 - Capability of eight core skills in UK government data science group

The second stage of the exercise looks at the capacity to use this capability to its maximum potential (Figure 6). This helps identify where people feel they are not supported within their organisation. This is essential to help managers provision better environments in which to practice good data science.

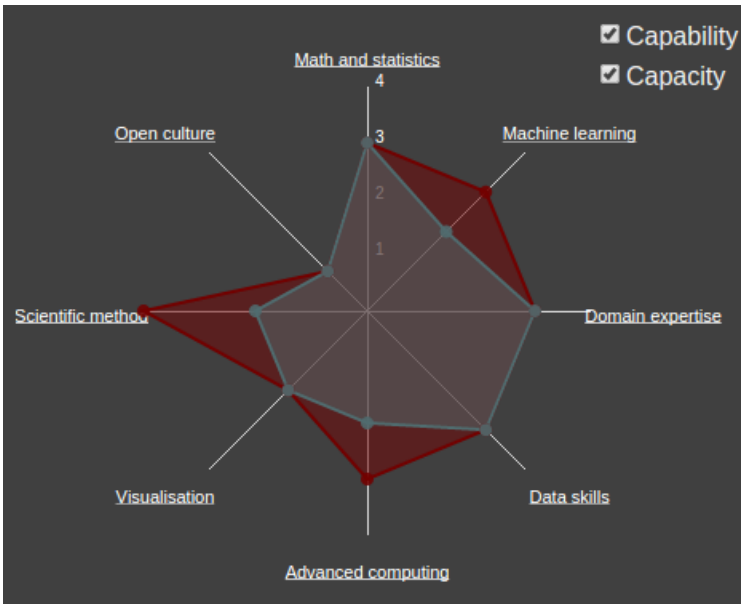


Figure 6 - Capability vs capacity in data science within UK government data science group

It is envisioned that the result of this exercise will be different for each country and sector depending on the core aim. The UK government data science group contains many professional statisticians whose focus it is to use data to inform policy. Careful and reliable modelling is of critical importance when making such high profile decisions.

Focus group session 3: Data science SWOT analysis

The final part of the workshop focuses on a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis of the organisation. Participants are asked to identify specific needs and detailed plans of action to address these needs. Based on previous focus groups work and preparatory interviews conducted these needs are likely to include high-level data science training, specific skills training as well as capacity analysis.

Questions in this session will include:

- What does the SWOT analysis look like?
- Where are the individual weaknesses?
- Where are the organisational weaknesses?
- Top two points in each SWOT category
- What further training would help and who should that training be focused at?

5. Sampling approach

5.1 Sampling across qualitative and quantitative analysis

As a qualitative survey, it is understood that more responses will not necessarily add new data. For this reason we have decided to use a *maximum variation sample* technique to help identify participants. To help with the select of participants, a number of key criteria will be considered before selecting participants.

Criteria 1: Role

For the purposes of the survey we are looking for two key roles to interview:

- Industry managers responsible for developing teams of people on projects. This may include senior or team managers who make decisions on training and capacity building. A list of candidates in these roles can be found in Appendix P1.
- Data scientists both in industry, connected via contracts and working outside industry.

Criteria 2: Sector

In order to study the similarities and differences between sectors, we will aim to select participants from across many commercial and non-commercial sectors.

Criteria 3: EU states represented

From Appendix P1, there is a weakness in representations from Eastern Europe. We anticipate fewer responses from these countries. This can also be demonstrated with the quantitative analysis of the job listings in these countries. On-going development of the demand analysis will seek to redress this imbalance by using the online survey and proactively targeting interviewees.

Criteria 4: Size of organisation

In addition to investigating the differences between countries and sectors, it is anticipated that there are likely to be significant differences between organisations of different sizes.

5.2 Evidence based sampling

The study will use a specialist tool to complement the skills analysis and provide additional contacts at various stages of the project. By connecting to various databases (scientific publication, patents, companies) and applying a set of dedicated algorithms, the tool identifies key experts, or companies related to a set of keywords. For example, a search for some of the skills listed in Appendix 1 such as “machine learning” or “business intelligence” would return a ranked list of experts working in these fields worldwide. In a second step, the tool automatically gathers contact details and enables contact through an automated workflow. We can decide whether the contacted experts should be in the EU or worldwide, depending on the application.

By combining several skills (e.g. “machine learning” AND “business intelligence”), we obtain within seconds a list of experts who are combining both skills (at least in the work they published).

We will therefore use this technique to refine and complement our demand analysis. A first application of this approach will be to refine the demand and expectations of the European Data Science Institute. This will allow us to precisely target those people who may become stakeholders, and understand their motivations or inhibitors.

A second application of the tool will be to question some of these experts to understand the state of their interaction with industries (and as they are scattered around the world this may help us identify qualitative differences between the EU, the US and Asia.)

A third application, later in the project, will be to invite some of these experts to contribute to course material for the project. As some of the course material will be based on fairly rare or recently emerging areas, this is an important asset to guarantee a timely delivery within the context of the project.

5.3 Potential study participants

Potential participants are listed in **Appendix P1**. Note that this appendix is not included in the public version of this document.

5.4 Key performance indicators

Table 1 shows the key targets for the demand analysis, split by the two major deliverables for the work package. At month 6, we will deliver an initial set of results and launch the EDSA dashboard. The dashboard will contain data from the Consortium partner countries, as a minimum and give a demonstration of the outputs of the complete demand analysis. The dashboard will also become a primary mechanism for collecting data via a survey link. This will allow the collection of data from across Europe and many different sectors.

The targets for month 6 are designed to allow our partners to evaluate the effectiveness of the survey design before a wider European launch. By the end of month six, we envisage a set of results gathered from partner countries. While it is expected that a good number of sectors, roles and different types of business will be surveyed, the targets are moderate, so that the focus during this initial stage is to allow evaluation of the methodology before expanding the targets prior to the final report.

Following the initial 6 month deliverable, the methodology will be reviewed and targets updated to reflect potential. The targets for month 18 here reflect the challenges in obtaining responses from further EU member states and relevant sectors. It is anticipated that many more responses will be received however these may not cover all the required sectors in every country, hence why at this stage in the study design, six sectors, per EU member state, has been selected.

Table 1 - Key performance indicators

KPI	Target	As of March 2015
Size of network (qualitative analysis)	24	19
Number of focus groups	4	1
Number of sectors	10	14
% of EU business registry sectors	80%	Unknown
Importance of sectors (%)	80%	Unknown
Number of EU states	12	6
% split of Corporate/SMEs	60%/40%	Unknown
% split managers / Data scientists	60%/40%	66%/33%

6. Quantitative analysis

The automated quantitative analysis will analyse trends across Europe to identify demand for certain skills in industry based upon job postings and skills listing on services such as LinkedIn, Science Jobs, and Monster. Figure 7 below, shows the result of an initial analysis of a number of countries across Europe and already demonstrates significant differences in demand. This visualisation shows data collected from LinkedIn jobs and demand for jobs per country and per area. This snapshot, taken in April 2014, shows high demand for jobs in Cloud computing (pink), business intelligence (brown) and data engineering (green). This demonstrations the demand for skilled people in these areas and the potential for an individual data scientists who hold a combination of all these skills.

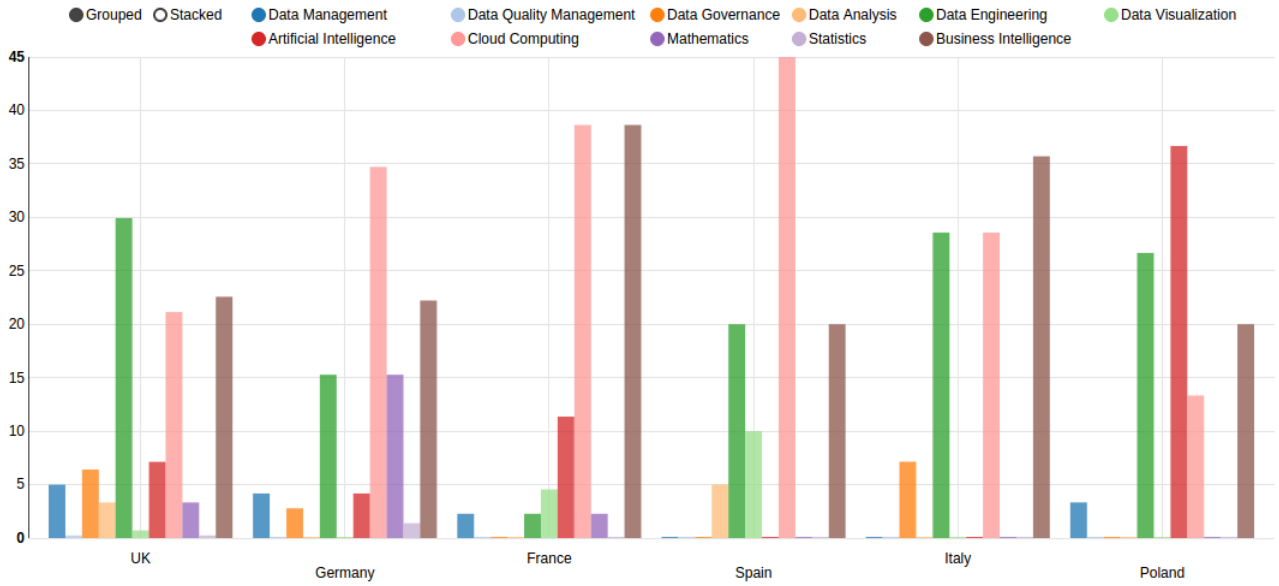


Figure 7 - Early analysis of demand for data science jobs in Europe

This analysis will expand on the initial research done at the beginning of the project and provide a live view of the changing demand over time. Tracking this in near real time will also allow for the early identification of trends and analysis of their causes.

The automated analysis will regularly collect data from each country on jobs and roles related to the six of the eight key areas outlined earlier in the document. This grouping will further allow us to compare the perceived demand to the listing of jobs and role profiles of professionals on services such as LinkedIn.

The secondary aim of the automated collection is to provide a service of immediate value in order to further encourage participation in interviews and focus groups.

Where possible, data collected will be made available as Open Data for others to download and analyse, further demonstrating good practice in data science.

6.1 Initial dashboard design

An interactive dashboard will combine many aspects of the demand analysis together and provide an engaging way to view data for each country in Europe. A number of aspects will be combined in order to allow users to select generic or specific sets of data.

6.2 Interactive map

An interactive map of Europe will allow users to immediately select a single country or region in order to view individual and aggregated data respectively.

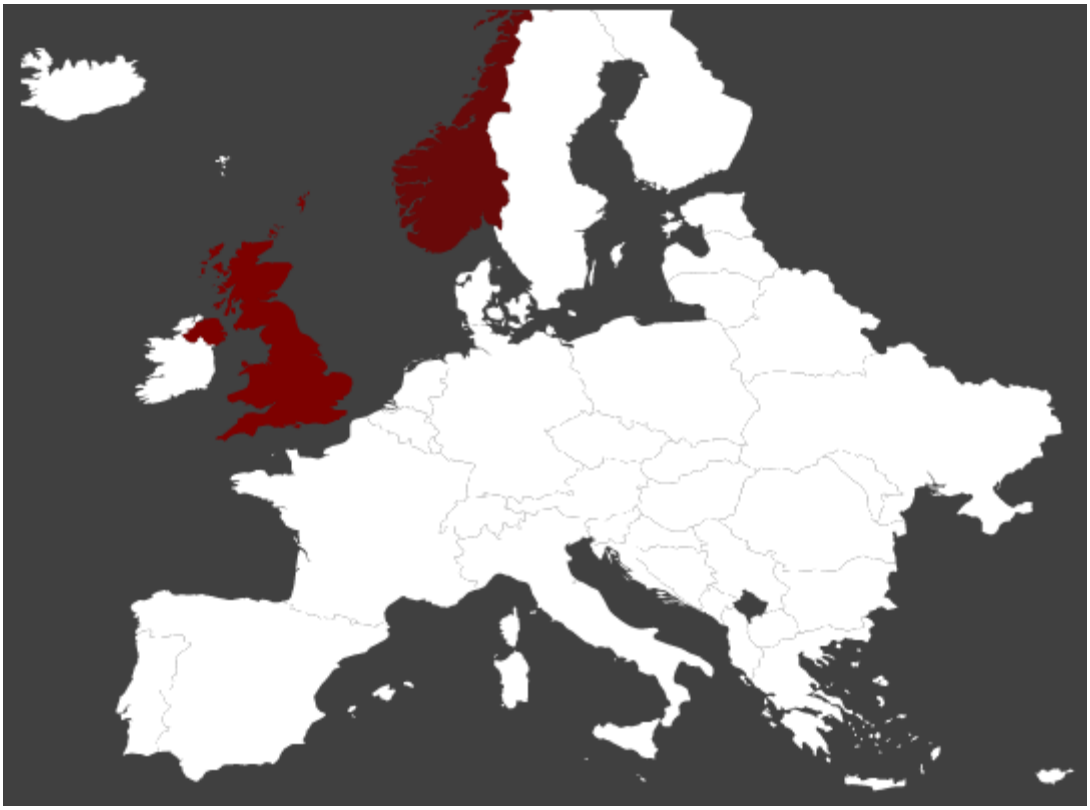


Figure 8 - Initial design of interactive map

Once a country or region is selected a number of different visualisations will present data relating to that country or region including:

- Skills analysis radar diagram coving the main eight topics outlined in Section 2.2
- A trend tracker displaying the automatically collected data.
- A topic map of key points from qualitative interviews (currently in design phase)
- A link to training courses available in each area (supplied as part of WP2)

In addition to aggregating the data at the country level, where possible the user will be able to view the same data for specific sectors. The number of sectors will depend on the breadth of data collected.

6.3 Data science capability radar diagrams

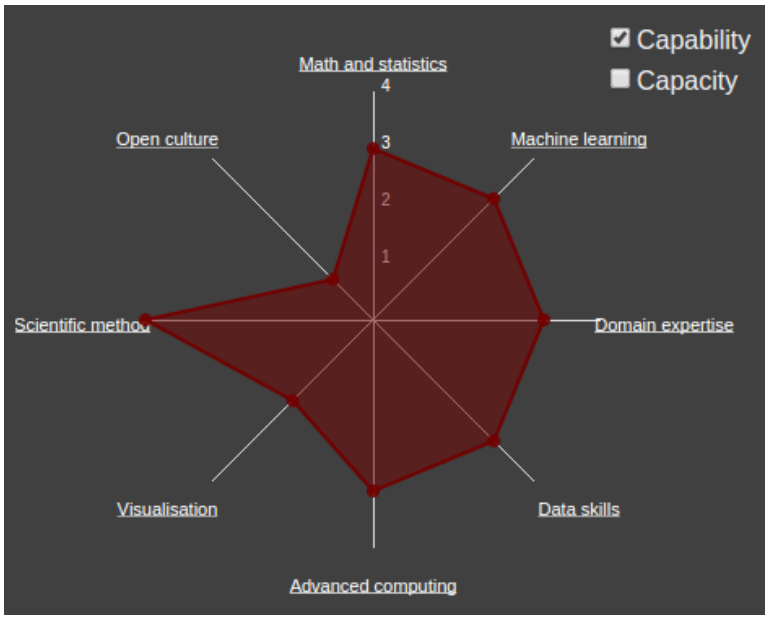


Figure 9 - Example data science capability radar diagram

Collected from a combination of focus groups, interviews and automated analysis, the radar graph (shown in Figure 9) will allow the comparison of countries and show opportunities for skills development.

The radar diagram therefore represents the current capability in a country or sector. Each of the eight axes will also have a clickable title which will bring up the trend tracker for that area.

6.4 Trend tracker

The trend tracker (shown in Figure 10) will show the changing nature of jobs and job advertisements in each country and, where possible, per sector for each of the eight axis of the radar graph. Figure 10 shows early data for the "Math and statistics" axis with associated job roles (as matched in Appendix 1). This will give a clear indication on the relationship between capability, as shown on the radar diagram, and demand, shown by the trend graph.

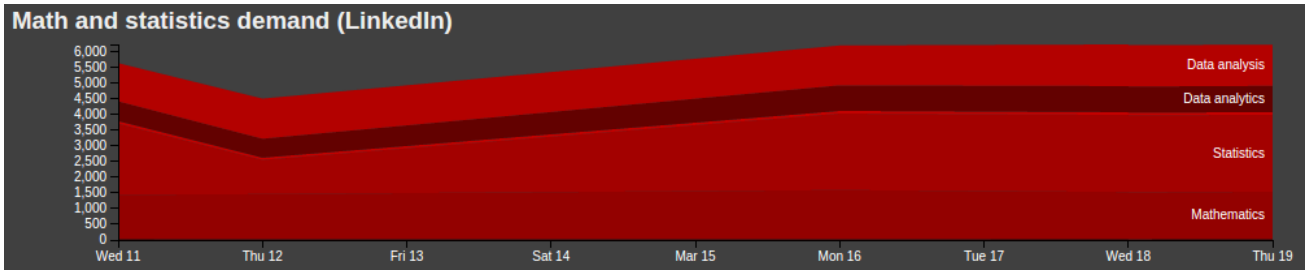


Figure 10 - Jobs demand in the area of math and statistics for UK (March 2015)

The biggest potential for skills development will be where there is a growth in advertised jobs but a low capability.

Further statistics and metrics for comparing countries will be added as required during the initial analysis of results

7. Demand analysis deliverables

There are two key deliverables in this work package for the demand analysis. The details and objective of each are outlined in this section.

D1.2 (M6): Initial study evaluation report

The focus of the initial study evaluation report will be on the launch of the European Data Science Academy dashboard containing results of surveys from partner counties (targets as specified in section 5.4). A breakdown of work for each deliverable can be found below:

- Initial qualitative surveys, target of 3 each. (All)
- Transcriptions of recorded interviews (external associates)
- One focus group and analysis (ODI)
- Testing of online survey including data collection (ODI, Southampton, OU, Persontyle)
- Deployment of data scientist expert platform (ideXlab)
- Deployment of automated data collection platform for trend tracker (ODI)
- Data analysis (Lead ODI, contributor: Southampton)
- The EDSA dashboard (ODI, OU, Southampton)

D1.4 (M18): Final study evaluation report

- Complete qualitative surveys (All and external associates)
- Transcriptions of all recorded interviews (external associates)
- Promotion of online survey and EDSA dashboard (All)
- Three further focus groups (ODI and partners)
- Data analysis of all collected data (ODI, Southampton)
- Final version of EDSA dashboard (ODI, OU, Southampton)
- Executive summary of results and recommendations for curriculum development (ODI, Southampton)

7.1 Use of subcontracting.

It is intended that subcontractors will be used, as outlined in the proposal for certain elements of Task 1.1. It was expressed by a number of partners that they did not have sufficient expertise in survey techniques to perform interviews, and other considerations such as quality of data collection and capability to accurately transcribe the interview were raised. Additionally, using an independent organisation to carry out interviews ensures fairness and protects the results against opinionated or biased questions or responses being collected. For this reason the ODI has decided to subcontract some of this work, where necessary.

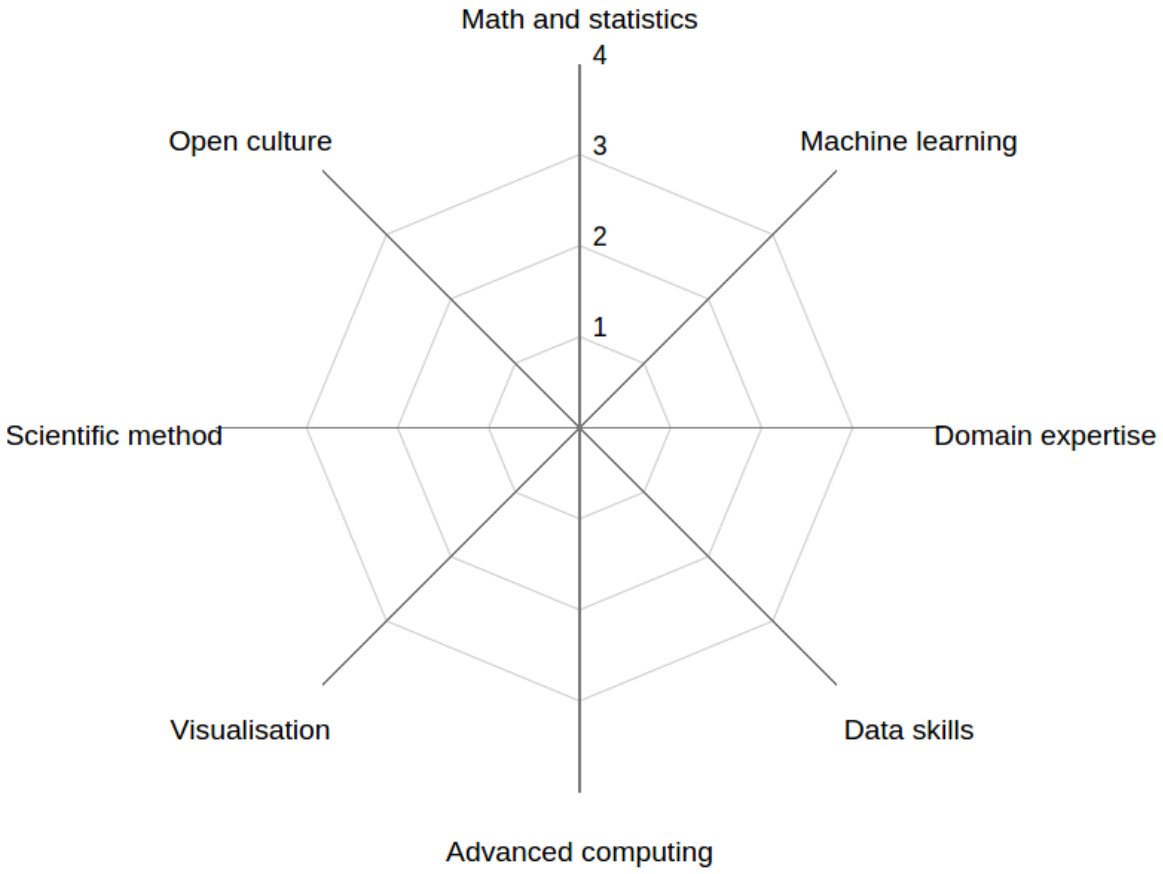
8. Appendices

Appendix 1 – Mapping of topics to specific skills

Math and Statistics • Linear Algebra and calculus • Statistics and probability • RStudio • Data analytics	Machine Learning • Machine learning • Social network analysis • Inference and reasoning • Process mining
Domain Expertise • Enterprise process • Business intelligence • Data anonymisation • Semantics and schemas • Data Licensing	Data skills • Databases • Data management • Data mining • Data formats and linked data • Information extraction • Stream processing • Large scale data processing
Advanced computing • Programming • Computational systems • Python and R • Cloud scale computing	Visualisation • Visualisation • Infographics • Interaction • Data Mapping • Data stories • Data journalism • D3js, Tableau
Scientific method • Research methodologies • Significance and reproducibility • Scientific publication • Process and data sharing techniques	

Appendix 2 – Exercise – Your data science capability

Exercise 1: On the graph below plot your personal capability in data science. Note that you will need to fill in the domain you feel your data science skills are most often applied.



Exercise 2: With a different colour pen plot a new line that applies to a specific data science project you have done.

- Do the two lines match up?
- Are you exploiting your ability on data science projects
- Which areas need improvement when applied to projects
- Are these the same as your personal gaps?

Exercise 3: Identify one area and come up with an action plan on how to improve on this axis

Data Skills 1. Just let me have a spreadsheet. 2. Databases and tables for me. 3. Big data is easy 4. JSON, GIS, RDF… no problem!	**Advanced computing** 1. I know how to create pivot tables 2. Excel is just one of the tools I use 3. R is amazing 4. Happy to program my own solution
Scientific Method 1. I publish results 2. I rigorously test hypothesis 3. I test for significance and look for related research. 4. My result is reproducible by someone else outside of my office.	**Visualisation** 1. Charts in excel 2. Happy to hack with existing solutions 3. Ggplot and tableau are brilliant 4. D3 is for me
Math and statistics 1. I know what an average is 2. I know four or more types of averages 3. The Gaussian curve is my favourite 4. Allow me to derive the GLS estimator	**Open culture** 1. Challenge accepted! 2. Not perfect, but it bloody works! 3. Didn't think it could do that, did you? 4. I think I just invented something, now I'll give it away.
Machine learning 1. I built a model once 2. I know more than three classifiers 3. If you torture the data, it will confess 4. I know the maths behind Random Forests.	**Domain expertise** 1. I apply my expertise to a domain 2. I understand the impact of decisions 3. I am the domain expert 4. I define the domain