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

D2.5 Learning resources 2

Deliverable Editor:	**Alexander Mikroyannidis (OU)**
Other contributors:	**Emily Vacher (ODI), David Tarrant (ODI), Ryan Goodman (ODI), Simon Bullmore (ODI), Huw Fryer (SOTON), Angelika Voss (Fraunhofer), Remi Brochenin (TU/e), Inna Novalija (JSI)**
Deliverable Reviewers:	**David Tarrant (ODI), Jean-Louis Lievin (IdeXlab)**
Deliverable due date:	**31/01/2017**
Submission date:	**22/02/2017 (as agreed with project PO)**
Distribution level:	**P**
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	08/12/2015	Alexander Mikroyannidis	Outline and responsibilities of contributors.
0.2	15/12/2016	Simon Bullmore	Added in new ODI learning modules.
0.3	07/01/2017	Alexander Mikroyannidis	Version for internal review.
0.4	14/02/2017	Alexander Mikroyannidis	Revised draft for approval by the scientific director.
0.5	16/02/2017	Alexander Mikroyannidis	Final version approved by the scientific director.
0.6	16/02/2017	Elena Simperl	Scientific approval
1.0	22/02/2017	Aneta Tumilowicz	Final QA

Table of Contents

Change Log 2
Table of Contents 3
List of Tables 4
List of Figures 4
1. Executive Summary 5
2. Introduction 6
3. Data Visualisation and Storytelling 9
3.1 Module overview 9
3.1.1 Introduction to data storytelling 9
3.1.2 The four step data storytelling process 10
3.1.3 Ensuring that data is correctly licensed for usage 10
3.1.4 Gathering data 10
3.1.5 Organising data 11
3.1.6 How to clean data for use 11
3.1.7 Filtering and pivot tables 11
3.1.8 Data visualisation formats 12
3.1.9 Data visualisation best practice 12
3.1.10 Visual deception 13
3.1.11 Narrating your story 13
3.1.12 Reason for original development 13
3.1.13 Further development plans 14
3.2 Learning materials & delivery methods 14
3.3 Relevance to curriculum 15
3.4 Relevance to demand analysis 16
4. Big Data Analytics 19
4.1 Module overview 19
4.1.1 Introduction 19
4.1.2 Collaborative filtering in the lambda architecture 19
4.1.3 Generating real-time recommendations 19
4.1.4 Big data analytics in Spark 20
4.1.5 Complex event processing with Proton 20
4.1.6 SQL operators for MapReduce with Teradata 20
4.1.7 In-memory processing 20
4.2 Learning materials & delivery methods 20
4.3 Relevance to curriculum 21

4.4 Relevance to demand analysis 21
5. Data Management and Curation 23
5.1 Module overview 23
5.2 Learning materials & delivery methods 23
5.3 Relevance to curriculum 24
5.4 Relevance to demand analysis 24
6. Statistical / Mathematical Foundations 25
6.1 Module overview 25
6.2 Learning materials & delivery methods 25
6.3 Relevance to curriculum 26
6.4 Relevance to demand analysis 26
7. Conclusions and next steps 27

List of Tables

Table 1: The revised core EDSA Curriculum (D2.2). ---------- 6
Table 2: Recommendations from the Study Evaluation Report (D1.4). ---------- 7
Table 3: Relevance of Data Visualisation and Storytelling to the EDSA syllabus. ---------- 16
Table 4: Relevance of Big Data Analytics to the demand analysis recommendations. ---------- 22
Table 5: Relevance of Statistical / Mathematical Foundations to the demand analysis recommendations. ---------- 26

List of Figures

Figure 1: Video excerpt from lesson 11 of the Data Visualisation and Storytelling module. ---------- 15
Figure 2: Video excerpt from the Big Data Analytics module. ---------- 21

1. Executive Summary

This deliverable presents the modules that have been added to the EDSA courses portfolio during the second year (Y2) of the project. According to the revised EDSA curriculum presented in D2.2 (M18), the following 4 modules were scheduled for release during Y2:

- Data Visualisation and Storytelling
- Big Data Analytics
- Data Management and Curation
- Statistical / Mathematical Foundations

Following the outcomes of the M18 project review, the focus of the project has now been shifted towards addressing the supply of training materials in order to bridge the data science skills gap. As a result, the EDSA courses portfolio is being extended to include a wider range of courses offered by renowned educational institutions both inside and outside the project consortium. These courses are selected based on their relevance to the EDSA curriculum and the EDSA demand analysis.

This deliverable thus presents both the learning resources that have been produced by the project in Y2, as well as the external learning resources that have been incorporated into the EDSA courses portfolio in order to cover the topics of the EDSA curriculum and address the current demand as identified by the EDSA demand analysis.

2. Introduction

As presented in D2.4, the EDSA courses portfolio includes a variety of data science courses that adopt different pedagogical models, in order to address different learning contexts and audiences. The main delivery channels of these courses are the EDSA courses portal[1] and the EDSA eBook.[2]

The EDSA courses portfolio is currently being expanded by incorporating a wider range of high quality learning resources, either offered by project partners or by third parties. This shift of focus aims at closing the gap between the demand of data science skills across Europe and the supply of learning materials suited for offering the required skills to job seekers.

The EDSA courses portfolio is thus being extended to include additional courses offered by renowned institutions both inside and outside the project consortium. These courses are available as:

- Massive Open Online Courses (MOOCs)
- Face-to-face courses
- Online courses
- Blended courses (delivered face-to-face and online)

The main criteria for the selection of both internal and external courses for inclusion in the EDSA courses portfolio are the *EDSA curriculum* and the *EDSA demand analysis*. Courses are selected based on their potential of addressing the EDSA curriculum topics as well as the training needs of data scientists as identified by the EDSA demand analysis.

As reported in D2.2, the EDSA curriculum has been revised to address feedback received by the community after having released the first version of the curriculum, in addition to further studying of the data science landscape in Europe. The revised curriculum is shown in Table 1. According to the revised curriculum, the modules scheduled to be released in Y2 of the project, should cover the topics of:

1. Data Visualisation and Storytelling
2. Big Data Analytics
3. Data Management and Curation
4. Statistical / Mathematical Foundations

The modules presented in this deliverable are therefore centred around the above 4 topics.

Table 1: The revised core EDSA Curriculum (D2.2).

Module	Topic	Stage	Status as of D2.2
1	Foundations of Data Science	Foundations	Released and revised
2	Foundations of Big Data	Foundations	Released
3	Statistical / Mathematical Foundations	Foundations	Newly Released
4	Programming / Computational Thinking (R and Python)	Foundations	To be released M30

[1] http://courses.edsa-project.eu

[2] http://courses.edsa-project.eu/mod/page/view.php?id=299

5	Data Management and Curation	Storage and Processing	Newly Released
6	Big Data Architecture	Storage and Processing	Released
7	Distributed Computing	Storage and Processing	Released and revised
8	Stream Processing	Storage and Processing	To be released M30
9	Linked Data and the Semantic Web	Storage and Processing	Released
10	Machine Learning, Data Mining and Basic Analytics	Analysis	Released and Revised
11	Big Data Analytics	Analysis	Newly Released
12	Process Mining	Analysis	Released
13	Social Media Analytics	Analysis	To be released M30
14	Data Visualisation and Storytelling	Interpretation and Use	Newly Released
15	Data Exploitation including data markets and licensing	Interpretation and Use	To be released M30

The EDSA demand analysis has produced the Study Evaluation Report (D1.4), which highlights the following 7 recommendations that should guide the development of the EDSA curriculum and the EDSA courses portfolio. These recommendations are summarised in Table 2. Internal and external courses are prioritised for inclusion into the EDSA courses portfolio based on the extent that they address these recommendations, as documented in the modules presented in this deliverable.

Table 2: Recommendations from the Study Evaluation Report (D1.4).

Recommendation	**Intervention level**	**Summary description**
1. Holistic training approach	General training approach	Refine EDSA's training approach and curriculum cycle to strengthen data science skills for data science teams and data literacy across various units of each organisation.

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

Additional criteria for the selection of courses are based on the *EDSA curriculum design and delivery values*.[3] According to these values, priority is given to the applicability of the learning materials on real-life settings as these are dictated by the data science industry. Learning materials that employ real-life case studies, tools and datasets are thus prioritised. As these values are in line with the aforementioned recommendations of the demand analysis (especially the recommendation regarding blended training), they are not explicitly referred to as selection criteria for the modules presented in this deliverable.

The rest of this deliverable presents the modules that have been incorporated into the EDSA courses portfolio in Y2, based on their relevance to the EDSA curriculum and the EDSA demand analysis. For each module, we present an overview of its objectives and learning outcomes, we identify the types of learning materials used and how these are delivered, and we discuss how each module is related to the EDSA curriculum and the demand analysis. Finally, this deliverable is concluded and the next steps of this work are outlined.

[3] http://edsa-project.eu/overview/edsa-values/

3. Data Visualisation and Storytelling

3.1 Module overview

When effectively analysed and presented in a clear and compelling way, data has the potential to create impact. Whether that's changing perceptions, offering counterintuitive insights or prompting action, impact happens when data acts as the catalyst for change. And at the heart of driving change is the skill of finding and telling stories using, where relevant, compelling visualisations.

No field is more experienced at finding and telling stories than journalism, and no field better at using data than data science. This set of 11 online lessons, offered by ODI, looks at what these fields can learn from each other, in order to find and tell compelling stories with data. Although the data needs to have a story, it is also necessary that each part of the story is an effective visualisation. The course also looks in detail at what makes an effective visualisation, and what best practices are for ensuring clarity and correctly conveying the results which are intended.

The 11 lessons are:

1. Introductions to data storytelling
2. The four step process
3. Understanding your rights to use data
4. Gathering data
5. Organising data
6. Cleaning data
7. Filtering and pivot tables
8. Data visualisation formats
9. Data visualisation best practice
10. Visual deception
11. Narrating your story

The objective of this course is to introduce the key role communication plays in data science work. Regardless of if the learner is already a competent statistician or programmer the key focus of this course in towards communication and how to tell good stories using data.

For data scientists who are already competent in many other areas, such as programming and visualisation, the theory aspects of this course are likely to lead them to investigating more in the area of the cognitive theories of infographics and not just learning on how to create them. For others, their next path may well be learning other data science skills such as data gathering, statistical methods of advanced computing. Each lesson on the course can thus be taken alone with the links at the end of each allowing users to build their own learning journey if they so wish.

Data scientists not only have to develop insights from data but also to communicate their findings to the organisation and world. The following sections outline why each lesson has been included and its relevance to data scientists.

3.1.1 Introduction to data storytelling

The aim of this lesson is to enable learners to explain why storytelling needs data. A modern data scientist is expected to be a catalyst for change in an organisation. This requires not only performing complex data manipulation, but also communicating findings to senior level decision makers. A key part of this is being able to tell powerful stories. This lesson looks at how data has changed the way decisions are made, the way that stories are told and some significant examples.

Related courses and materials:

- School of data journalism (School of data)[4]
- Introduction to Data Visualization [Storytelling with Data] (Persontyle)[5]
- Visual Journalism (Akademie for Publizistik Hamburg)[6]

3.1.2 The four step data storytelling process

The aim of this lesson is to enable learners to plan a story that uses data. A modern data scientist is expected to undertake many, if not all, of stages from gathering the data to visualising it, to communicating the story. It is thus essential to understand each stage and the time required. A simple four step process taken from the field of data journalism can help ensure time is well planned and that findings create the desired impact.

Related courses and materials:

- School of data journalism (School of data)[7]
- Data journalism handbook[8]
- Data Journalism (Tilburg University)[9]

3.1.3 Ensuring that data is correctly licensed for usage

The aim of this lesson is to enable learners to understand their rights when using data as part of a story. A modern data scientist is expected to use data from many sources. Being able to navigate the legal frameworks surrounding data is an essential part of a data scientist's role. The first stage in the process is understanding the legal right to gather and use data as pertaining to the current context.

Related courses and materials:

- Data science (Athens University of Economics and Business)[10]
- Business Intelligence (OBS Business School)[11]
- Leadership Challenges With Big Data - Turning Data Into Business (Rotterdam School Of Management, Erasmus University)[12]

3.1.4 Gathering data

The aim of this lesson is to enable learners to find data to use in a story. A modern data scientist should be able to identify data from a number of sources and gather that data regardless of structure, format or mode of access.

Related courses and materials:

[4] https://journalism.columbia.edu/data#Classes

[5] http://www.persontyle.com/introduction-data-visualization/

[6] http://www.akademie-fuer-publizistik.de/lehrgaenge/visuelle-publizistik/

[7] https://journalism.columbia.edu/data#Classes

[8] http://datajournalismhandbook.org/

[9] https://www.educationfair.nl/program/show/master/data-journalism/tilburg-university/4941

[10] http://www.cs.aueb.gr/en/datascience/curriculum

[11] http://www.obs-edu.com/en/master-in-business-intelligence

[12] https://www.rsm.nl/fileadmin/Images_NEW/ExecEd/24870_Leadership_challenges_with_big_data_3.pdf

- Data fundamentals (School of data)[13]
- Data Science Training (Graspskills)[14]
- Business Intelligence (Dalarna University)[15]

3.1.5 Organising data

The aim of this lesson is to enable learners to organise data so it can be easily and quickly processed. Having gathered data from a number different sources, an essential skill of a data scientist is to organise that data, ready to be processed. A modern data scientist should design and implement a structure for a dataset that can be understood by their audience.

Having structured and organised the data, it is common to think the data is ready for processing or even visualising. This is not the case. It is first necessary to ensure that data is of high quality and accuracy.

Related courses and materials:

- Introduction into exploring data (School of data)[16]
- Open Data Science (Open Data Institute)[17]

3.1.6 How to clean data for use

The aim of this lesson is to enable learners to ensure a story is using high quality, accurate data. Poor quality data leads to poor quality stories and insights. Before any analysis can take place, a data scientist needs to be able to identify and remove errors from data, thus removing or limiting the risk to themselves or to the organisation.

Related courses and materials:

- A gentle introduction to data cleaning (School of data)[18]
- Fundamentals of Data Science (Octo Technology)[19]
- Exploratory data mining and data cleaning (Book by Tamraparni Dasu)[20]

3.1.7 Filtering and pivot tables

The aim of this lesson is to enable learners to analyse data to find a story. Once your data is clean, it is ready for analysis. This analysis does not have to use complex statistical methods. A modern data scientist should be comfortable using common analysis techniques such as pivot tables in their daily work.

At this point, there are two possible directions. Further analysis may be required to find a story; this may require further gathering, organising and cleaning. Application of further advanced statistical techniques may also be required (such as the ones outlined in the courses and exercises linked in the

[13] https://schoolofdata.org/courses/

[14] http://www.graspskills.com/information-technology/data-science-training-classroom/united-kingdom

[15] http://www.du.se/en/Study-at-DU/Programmes/Business-Intelligence-/What-Will-You-Study/BI-for-Managers/

[16] https://schoolofdata.org/handbook/courses/gentle-introduction-exploring-and-understanding-data/

[17] https://theodi.org/courses/open-data-science

[18] https://schoolofdata.org/handbook/courses/data-cleaning/

[19] http://www.octo.com/fr/publications/15-data-science-fondamentaux-et-etudes-de-cas

[20] http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471268518.html

lesson). Alternatively, the data might have helped reveal the story and be ready for visualisation. The next stage is choosing the most appropriate visualisation for the data.

Related courses and materials:

- Data and tool understanding - Ends and means (Big data institute)[21]
- Power Excel: Analysis Data to make Business Decisions (Learning Tree International)[22]
- Data Analytics (General Assembly)[23]

3.1.8 Data visualisation formats

The aim of this lesson is to enable learners to choose an appropriate visualisation for data. A modern data scientist should be comfortable working with many different types of visualisation from common graphs to advanced statistical plots to geographic representations. Each visualisation technology has its own costs and benefits, notably the financial cost or platform restrictions versus ease of use; and the flexibility involved versus the complexity required to create other designs.

Choosing the correct visualisation is the first step, but it does not guarantee that it will communicate a story effectively. The visualisation will require careful design in order to make the story clear to the audience.

Related courses and materials:

- Tableau: Presenting Analytic Visualisations (Learning Tree International)[24]
- Geoinformatics (Aalborg University)[25]
- Introduction to data visualisation using R (Royal Statistical Society)[26]

3.1.9 Data visualisation best practice

The aim of this lesson is to enable learners to design effective data visualisations for telling stories. Picking the right visualisation is only the first part of telling effective stories. Plotting the right data in the correct way, making good use of colour and other visual cues is essential to making the story "pop out" at the audience. By understanding how the brain perceives information, these techniques can be successfully used to efficiently communicate stories, for example to high level managers with limited time.

Although these best practices can make a visualisation more compelling, they don't guarantee engagement. Nevertheless, they can ensure that the message is clear, and where structured for the right audience, the most effective presentation of the story can be achieved.

Related courses and materials:

- Introductory Course to Data Science: Director Level - Learn and Empower (LARK)[27]

[21] http://bigdatainstitute.dk/undervisning/program/#data-og-vaerktojsforstaelse

[22] https://www.learningtree.co.uk/courses/195/power-excel-advanced-excel-training-to-analyse-data-and-make-business-decisions/

[23] https://generalassemb.ly/education/data-analytics

[24] https://www.learningtree.co.uk/courses/1256/tableau-for-big-data-visualisation-training/

[25] http://www.en.aau.dk/education/master/surveying-planning-land-management-msc-in-tech/specialisations/geoinformatics

[26] http://www.rss.org.uk/Images/PDF/pro-dev/2016/Data-visualisation-using-R-05-Sept-2016.doc

[27] http://media.wix.com/ugd/22ae06_43a076bbd13642c799b749510e26fa22.pdf

- Data visualisation: A one-day workshop (The Guardian)[28]
- Design and Make Infographics (Coursera)[29]

3.1.10 Visual deception

Whilst understanding best practices to ensure a clear presentation is necessary, it is not always sufficient. On occasion, the audience may be uninterested, or lack the requisite knowledge to appreciate the story being presented without additional context or narrative.

The aim of this lesson is to enable learners to recognise when and how visual deception techniques can be used to engage their audience. As well as using visual cues, such as colour, an audience can also be engaged through creating intrigue. The use of visual deception is one such method by which this can be achieved, whilst at the same time it can be used to mislead.

A modern data scientist should be able to choose deception techniques to engage and inform, and avoid those which intentionally mislead. This could be likened to a box of chocolates: not as appealing without the packaging and unclear which chocolate is which flavour without the explanatory label.

Related courses and materials:

- Deception: Perspectives from Science, Technology and Art (Stanford)[30]
- Visual Journalism (Akademie for Publizistik Hamburg)[31]

3.1.11 Narrating your story

The aim of this lesson is to enable learners to add structure and narrative to a story. A modern data scientist is expected to be a catalyst for change in an organisation. This requires not only performing complex data manipulation, but also being able to present the outcomes to senior level decision makers in your organisation. This means creating a memorable story which not only contains an engaging visualisation but a compelling narrative.

Related courses and materials:

- School of Data Journalism (School of data)[32]
- Shaping the data-driven company (EIT Digital Professional School, iMinds)[33]
- Data Visualization Efficient Communication of Data (Inspiri)[34]
- Transmedia Storytelling: Narrative worlds, emerging technologies, and global audiences (The University of New South Wales - Australia)[35]

3.1.12 Reason for original development

Both the dashboard and demand analysis identify data analysis and visualisation as key skills needed by industry. Insights gained from surveys and interviews also make it clear that data scientists need to

[28] https://www.theguardian.com/guardian-masterclasses/2015/aug/07/data-visualisation-a-one-day-workshop-tobias-sturt-adam-frost-digital-course

[29] https://www.coursera.org/learn/infographic-design

[30] http://web.stanford.edu/class/sts121b/syllabus.html

[31] http://visuelle-publizistik.de/

[32] https://schoolofdata.org/school-of-data-journalism-international-journalism-festival-perugia/

[33] https://professionalschool.eitdigital.eu/professional-training-courses/shaping-the-data-driven-company/

[34] http://www.inspari.dk/kurser-og-events/kursus-i-datavisualisering/

[35] https://www.coursera.org/learn/transmedia-storytelling

improve how they engage with their business counterparts, and in particular how they communicate findings.

From a market analysis, it became clear that there are many options for data journalism courses, but none that provide an overlap with data science. Additionally, although there is a rich set of introductory data science courses available, many are lengthy or presented through an aging learning platform that fails to inspire learning and really practice what it is that is being taught.

David Venturi provides a great insight introductory level data science courses available and rates Data Science A-Z™: Real-Life Data Science Exercises Included (Kirill Eremenko/Udemy)[36] as the most comprehensive course in data science available. This course consists of 27 modules, the 26th of which is entitled "presenting for data science". If the course is followed in order then some learners may not even see this section if they can't get past "section 12: Building a robust geodemographic segmentation model". Although there is likely to be overlap, the Finding stories in data course has been designed to appeal to all learners without them needing to first learn to become a programmer or a statistician.

3.1.13 Further development plans

This course will be maintained and further developed in two key ways:

- *Increase the number of external links to beginner level and related data science courses*. As discovered by the demand analysis, well-structured learning for data science beginners is lacking. As and when more content is made available, applicable links will be added into the Data Visualisation and Storytelling course.
- *Add additional interactive visualisation exercises using European dataset.*: As part of other ODI activities, more interactive data visualisation exercises will be built and made available with minimum cost deployment to the EDSA project through the Data Visualisation and Storytelling course. As per the demand analysis, resources will be specifically designed to allow early career data scientists to bridge the gap towards more advanced training and qualification.

3.2 Learning materials & delivery methods

The entire course of 11 lessons is available entirely online[37] for anyone to engage with freely at any time without the need for enrolment. Although the lessons are numbered in order, learners can choose their own path if they choose. Each lesson is presented in the form of an interactive web page following the latest instructional design theory using an eLearning platform voted the most innovative 4 year in a row at the learnX awards.[38] The eLearning platform provides a completely cross platform responsive way of delivering interactive learning. Each lesson is backed with learning outcomes which are assessed through the use of interactive questioning at the end of each.

Each lesson features:

- Practical steps for accomplishing key tasks
- Case-studies
- Examples of good data-led storytelling and visualisations
- Knowledge-checks and quizzes

Some lessons also feature video content, for example: 'Narrating your story' (see Figure 1).

[36] http://click.linksynergy.com/fs-bin/click?id=SAyYsTvLiGQ&subid=&offerid=323058.1&type=10&tmpid=14538&RD_PARM1=https%3A%2F%2Fwww.udemy.com%2Fdatascience%2F%26u1%3Dcc-medium-career-guide-intro-to-data-science

[37] http://findingstories.learndata.info

[38] https://www.adaptlearning.org/index.php/about/

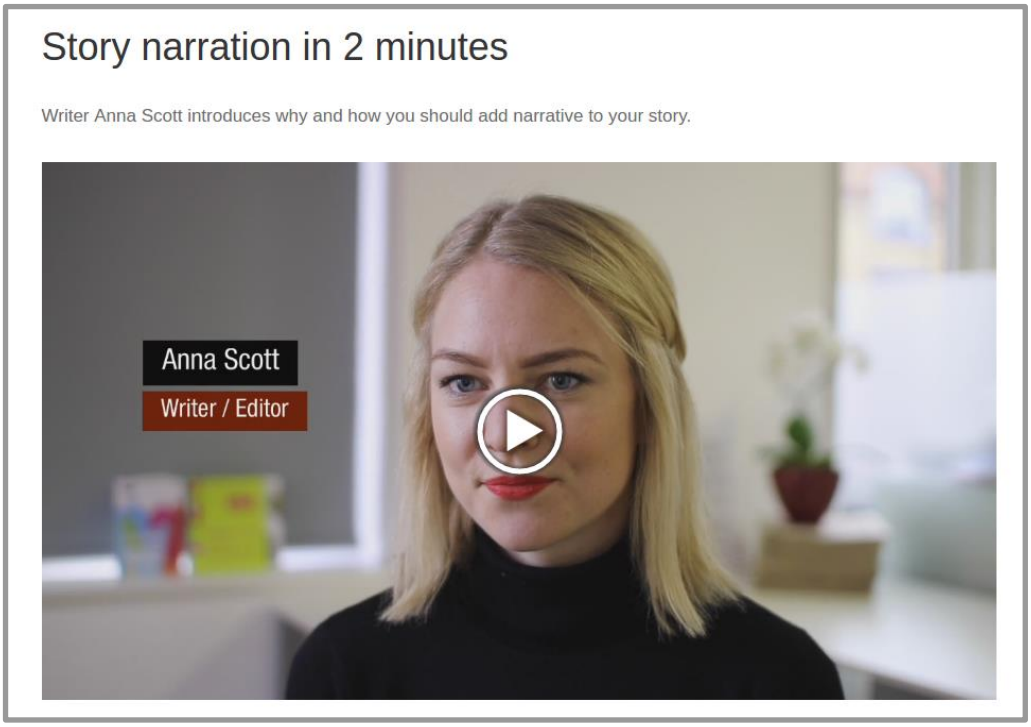

Figure 1: Video excerpt from lesson 11 of the Data Visualisation and Storytelling module.

3.3 Relevance to curriculum

In the second Data Science Curricula deliverable (D2.2), the Consortium outlined the main four stages of the syllabus:

- Fundamentals
- Planning
- Storytelling
- Producing visualisations

Having now completed the lesson outlines it was clear that there was a stage missing, that of gathering and processing data before it can be used to tell a story. This meant that the course assumed a lot of background knowledge about how to gather and process data. The EDSA demand analysis study evaluation report (D1.4) makes a number of recommendations that include the need to bridge the gap between newcomers to data science and the high level of existing knowledge that is required to take part in data science courses that are currently available. Adding a new focus on this missing stage of gathering and processing data helps ensure that all learner have equal opportunity to complete the course regardless of existing data literacy skills. Thus, the new areas of focus become:

- Fundamentals (planning)
- Data gathering and processing
- Data visualisation
- Telling a story

The 11 lessons match the syllabus as per Table 3.

Table 3: Relevance of Data Visualisation and Storytelling to the EDSA syllabus.

Fundamentals (planning)	● Introduction to data storytelling ● The four-step data storytelling process ● Ensuring that data is correctly licensed for usage
Data gathering & processing	● Gathering data ● Organising data ● Cleaning data ready for use ● Filtering and pivot tables
Data visualisation	● Data visualisation formats ● Data visualisation best practice ● Visual deception
Telling a story	● Narrating your story

3.4 Relevance to demand analysis

Both the dashboard and demand analysis identify data analysis and visualisation as key skills needed by industry. Insights gained from surveys and interviews also make it clear that data scientists need to improve how they engage with their business counterparts, and in particular how they communicate findings. This was a key driver for producing these lessons.

The demand analysis study evaluation report (D1.4) makes a number of recommendations for guiding the development of the EDSA curricular. Table 1 from this study makes seven recommendations, summarised as:

1. Holistic training approach
2. Open source based training
3. Soft skills training
4. Basic data literacy skills
5. Blended training
6. Data science skills framework
7. Navigation and guidance

The data visualisation and storytelling course focusses on many of these areas, in particular:

1. Holistic training approach

Most professional course providers offer very focused courses that go into depth about how particular tools, such as Hadoop, R, Python or SPSS, can be used by businesses.

- EDSA study evaluation report (D1.4, 4.2.1, p65)

The data visualisation and storytelling course focuses on key concepts that can be applied in many tools; for example, the exercises related to lessons 4, 5 and 6 of the course can be applied in any spreadsheet application (such as Excel, Libreoffice and Google Docs).

The other lessons focus more on the theory of creating an effective visualisation, rather than the pure practical. For example, lessons 9 looks at areas including how colour can help messages "pop out", while lesson 8 looks at how you can use deception to emphasise a point (if this ethically the right thing to be doing).

This holistic approach will help equip a modern data scientist with the theoretical knowledge and soft skills required to help communicate the output from their work effectively.

2. Open source based training

The majority of skills development in data science is achieved through these means [open source tools and resources].

- EDSA study evaluation report (D1.4)

The data visualisation and storytelling course will be offered as a free interactive online course, built itself from open source software. Additionally, all practical exercises available within the course (as well as the majority of those linked to externally) will be based upon freely available open source tools. This gives learners the maximum potential for self-study and learning without any financial or other barriers to using tools.

3. Soft skills training

Data scientists are often hired with high expectations regarding their abilities to transform business tactics and strategies; thus soft-skills such as these are seen as desirable and need greater focus in data science training.

- EDSA study evaluation report (D1.4)

The data visualisation and storytelling course goes beyond teaching learners how to make a chart. The focus of the course is on how visualisations, as well as other graphics, can help lead change. However, very few visualisations are effective without a narrative to help engage.

This however requires strong presentation and communication skills in order to influence senior management and other functional departments to make the right decisions based on data.

- EDSA study evaluation report (D1.4)

Creating a narrative is a key skill of journalists. The data visualisation and storytelling course borrows heavily from (and links to) existing data journalism courses in order to enhance the soft skills that data scientists need in order to influence and create change. The Open Data Institute's writer and editor Anna Scott, a journalist by training, introduces key concepts in the final lesson on narrating a story that help tie everything together.

4. Basic data literacy skills

Visualisation and communication of a story does not have to involve the use of advanced computing and programming tools.

Perhaps the most striking quantitative result is that advanced computing [...] is less in demand.

- EDSA study evaluation report (D1.4)

The EDSA study evaluation report identified 456 courses related to data science across Europe and noted the significant lack of "beginner" level courses that introduce the basics. The market appears flooded with courses that require a high degree of existing technical knowledge.

According to our assessment, none of the masters courses are targeted at beginners.

- EDSA study evaluation report (D1.4)

The data visualisation and communication course does not expect learners to have any prior knowledge in any particular area. Additionally, the practical examples use commonly available spreadsheet tools that do not require any programming or computing experience. As the course takes a holistic approach, links are offered to additional learning and exercises that suit all levels of learner. As per the market, currently there will be more links to advanced courses which it is hoped can be addressed in the future.

5. Navigation and guidance

Each lesson in the data visualisation and storytelling course has been designed to stand alone. The introductory content of each lesson outlines the skills that will be acquired and how they are relevant to a modern data scientist. Links will be made to other lessons where necessary. The main navigation page for the lessons will allow users to choose their pathway through the lessons. While it will be recommended that users follow the lessons in order, interlinking the lessons will mean this is not strictly necessary to access relevant content.

At the end of each lesson, learners will be guided not only to the next lesson but also to any external exercises, content and courses through which they can expand their knowledge.

4. Big Data Analytics

4.1 Module overview

This module is offered by Fraunhofer and provides an overview of approaches facilitating data analytics on huge datasets. Different strategies are presented including sampling to make classical analytics tools amenable for big datasets, analytics tools that can be applied in the batch or the speed layer of a lambda architecture, stream analytics, and commercial attempts to make big data manageable in massively distributed or in-memory databases. Learners will be able to realistically assess the application of big data analytics technologies for different usage scenarios and start with their own experiments.

The module is structured as follows:

1. Introduction
2. Collaborative filtering in the lambda architecture
3. Generating real-time recommendations
4. Big data analytics in Spark
5. Complex event processing with Proton
6. SQL operators for MapReduce with Teradata
7. In-memory processing

The following sections provide more details about each part of the module.

4.1.1 Introduction

Big data analytics solutions ask for skills from two different fields. First, information technology skills are needed to provide horizontally scalable systems for storing and processing data. Second, data analysis skills are needed and, in particular, tools and methods for the mathematical modelling of data. Collaborative filtering provides a running example for all aspects of big data analytics. The use-case is to provide recommendations for products based on user preferences. A concrete example is given by the task of music recommendation based on the last.fm dataset. The latter provides user/artist pairings based on listening events and can be used to generate artist recommendations. The task makes it necessary to define similarity of disparate items. This is based on utility matrices and the Jaccard similarity resp. distance. Once a distance measure is defined, data can be analysed by clustering. On the one hand, this can be achieved using sampling to fit big data sets into classical analytics tools. On the other hand, big data analytics solutions can be integrated in a lambda architecture.

4.1.2 Collaborative filtering in the lambda architecture

Lingual provides an ANSI SQL interface for Apache Hadoop. This is applied for data understanding in the collaborative filtering use-case. Building the utility matrix for collaborative filtering is a typical batch processing task. This can be modelled on top of Hadoop using Cascading. Classical data analytics software cannot handle huge amounts of data. A possible solution is given by sampling a subset of the dataset which can then be analysed in classical tools using RHadoop. The actual cluster model for the batch view is computed in R whereas the application of this model is again performed using Cascading. The size of the final model is small enough such that validation is possible in R via Lingual.

4.1.3 Generating real-time recommendations

Stream-processing can be used to update the results from collaborative filtering for new entities which have not been part of the batch processing, yet. For large streams, the technique of stream synopses allows to keep the data size manageable. In particular, count sketches can be used to efficiently approximate the distance computation necessary for collaborative filtering. Such a solution can be integrated in the lambda architecture using Apache Storm.

4.1.4 Big data analytics in Spark

Spark is one of the fastest developing platforms for big data analytics. It provides means to overcome the disc I/O overhead often seen in MapReduce based processing. Concepts like data frames and Spark SQL make it convenient to work with structured data inside Spark programs. In particular, it provides the Spark MLlib, a library that contains many distributed machine learning approaches suitable for big data analytics. PySpark allows the powerful combination of Spark with Python and thus to combine distributed processing and the wide range of libraries for data analytics and visualisation in Python. Linear regression, collaborative filtering with alternating least squares (ALS), K-Means clustering, and power iteration clustering provide the means to implement large scale collaborative filtering in Spark. Spark offers a framework for machine learning pipelines. These make it easier to combine multiple algorithms into a single workflow with a common interface to parameters. Linear regression is available in both Spark and Python. It may be beneficial to use either the one or the other, depending on data size.

4.1.5 Complex event processing with Proton

The basic principle of complex event processing is to derive complex events on the basis of a possibly large number of simple events using an event processing logic. Proton on Storm allows running an open source complex event processing engine in a distributed manner on multiple machines using the STORM infrastructure. Event processing networks provide a conceptual model describing the event processing flow execution. Such a network comprises a collection of event processing agents, event producers, and event consumers that are linked by channels. Fraud detection on call detail records is an illustrative use-case example showing how these concepts can be implemented.

4.1.6 SQL operators for MapReduce with Teradata

Database management system providers seek to enhance their traditional database and make them applicable to big data use-cases. A basic concept to achieve this is given by partitioning of tables, leading to massively parallel databases. Table operators allow making use of the partitioning for distributed algorithms using MapReduce. A selected commercial tool offering these approaches is the Teradata Aster solution.

4.1.7 In-memory processing

More and more main memory becomes available at a reasonable price. As access speed is reduced significantly once data outside of main memory is accessed, high performance applications focus on keeping as much data as possible in main memory. There is a wide variety of in-memory database systems available. Central performance and applicability measures to be kept in mind when choosing such a system comprise operating system compatibility, hardware requirements, license and support issues, runtime monitoring capabilities, memory utilisation, database interface standards, extensibility, portability, integration of open source big data technologies, local and distributed scaling and elasticity, available analytics functionality, persistence, availability, and security.

4.2 Learning materials & delivery methods

This module consists of learning materials supporting the Big Data Analytics course, which is offered by Fraunhofer and delivered face-to-face.[39] The learning materials for this module include an introductory webinar (see Figure 2), textual materials, as well as a quiz. The learning materials are available on the EDSA courses portal[40] and the EDSA eBook.[41]

[39] https://www.iais.fraunhofer.de/de/geschaeftsfelder/big-data-analytics/uebersicht/data-scientist-schulungen/big-data-analytics.html

[40] http://courses.edsa-project.eu/course/view.php?id=33

[41] http://courses.edsa-project.eu/mod/page/view.php?id=299

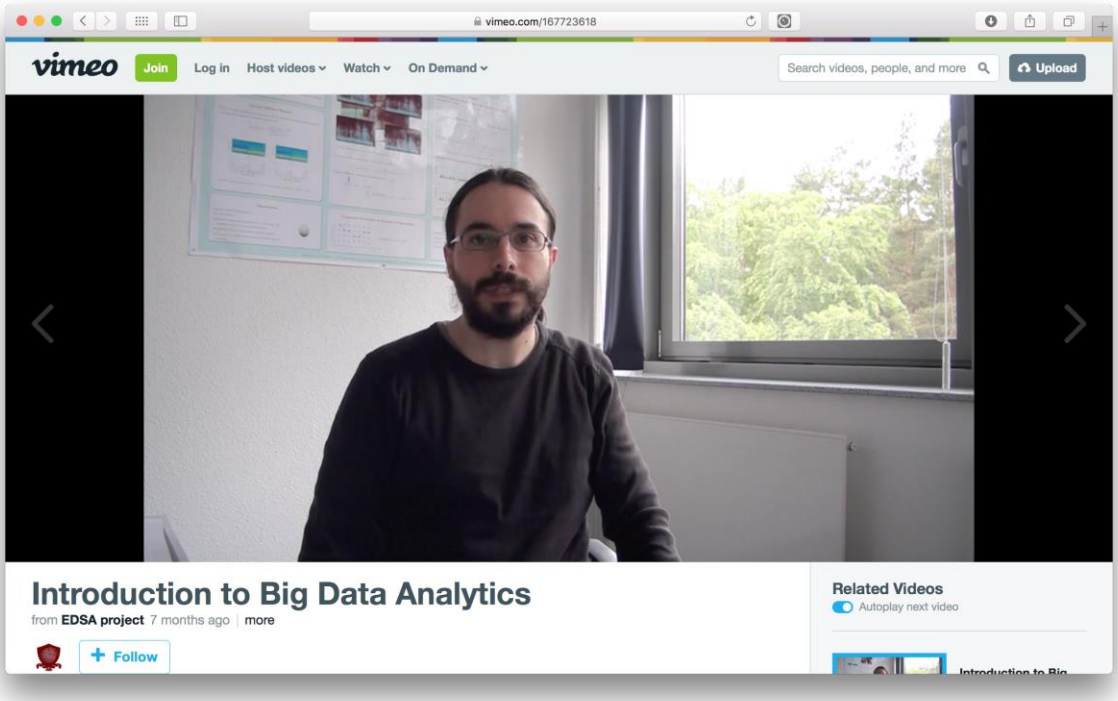

Figure 2: Video excerpt from the Big Data Analytics module.

4.3 Relevance to curriculum

This is module 11 "Big Data Analytics" in the EDSA curriculum in the introduction of this deliverable. It is based on modules 6 "Big Data Architecture" and 10 "Machine learning, data mining and basic analytics". Typical learning paths we find at Fraunhofer start with analytics (module 10), continue first with big data architecture (module 6) and then with big data analytics. These three modules are our most popular ones.

4.4 Relevance to demand analysis

This module applies concepts of data mining (job count in the dashboard: 11282) and machine learning (job count in the dashboard: 5116) to Big Data (job count in the dashboard: 9991). We can confirm the relevance expressed by these job counts from our own experience. Our corresponding face-to-face course is delivered five times per year and participants of our seminars regularly say that big data and big data analytics are the most relevant topics for their professional work and should be given the most time in our training.

Regarding the demand analysis, Table 4 summarises the relevance of the module to the recommendations of D1.4.

Table 4: Relevance of Big Data Analytics to the demand analysis recommendations.

Recommendation	Application to this module
Holistic training approach	According to the definition in the EDISON project, "Data science incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering." This module applies data analysis methods to big data and streaming data.
Open source based training	This module is open-source and the tools used are open source as well.
Soft skills training	No, the module is about the hardest skills in data science.
Basic data literacy skills	They are a precondition for this module.
Blended learning	Persons interested in our face-to-face course are directed to this module before they decide to enrol to it. After registration, they are invited to study this module as a preparation for the course.
Data science skills framework	In the EDISON curriculum for data science professionals, the module addresses the competence groups "data analysis" and "Data science engineering". Together, they are assigned 70 of 120 ECTS points.
Navigation and guidance	Most participants of this course have previously attended "big data architecture" and "data analytics". We strongly suggest such learning paths.

5. Data Management and Curation

5.1 Module overview

The objective of this module is to introduce and explain concepts and practices for the management of the whole lifecycle of data assets.

It includes the creation of a plan, organisation, file formats, and the creation of metadata. It also introduces knowledge required to manage the safe access to data, from storage to access rights and licensing, including notions about privacy issues.

The contents of this module are very abstract, and require from the students an understanding of the complexities of the topic, rather than an ability to apply a method matched to a single tool. Correspondingly, the courses teach an abstract method, using the description of real situations to let students feel the practical need for this knowledge. Hence there are no particular tools taught or practical exercises with tools, instead questionnaires ensure the engagement of students throughout the courses.

This module aims at covering parts of the following major books:

- 'Managing Research Data', Facet Publishing (2012)
- 'Principles of Data Management', Keith Gordon (2013)
- 'Preparing the Workforce for Digital Curation', National Academies Press (2015)

5.2 Learning materials & delivery methods

As no EDSA partner is offering a course on this topic, it has been decided to incorporate the following external courses into the EDSA courses portfolio and add them to the EDSA courses portal:

- **Introduction to Digital Curation 2016 (UCL / UCLextend, UK)**[42]

  Contents: This introductory course teaches the context of data curation, for which it provides a short theoretical basis. It also explains which further topics should be studied after the course.

  Method: This light course is mostly text-based, with some links to external videos. It provides tools for students to connect and discuss.

- **DC 101, digital curation 101 (DCC, UK)**[43]

  Contents: This intermediate to advanced course teaches data curation. It is designed for the UK, but its contents are useful for a wider audience. It covers all parts of data curation, which are conceptualized in a cycle. Topics covered include for instance: appraisal, preservation, and reuse.

  Method: This very comprehensive course is text-based. It was originally designed for people who would teach it in turn to another audience; hence its contents are originally supporting material for courses. Nevertheless, it provides a very comprehensive source of information for students who already have basic knowledge on the topic.

[42] https://extendstore.ucl.ac.uk/product?catalog=UCLXIDC25uDH16

[43] http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials

- **MANTRA Research Data Management Training (Univ. of Edinburgh, UK)**[44]
  Contents: This introductory course teaches data curation for research data. Besides generic data management and data organisation, it covers data transformation, data protection, data security, licensing, and metadata.

  Method: This medium-length course uses both text and video.

- **IEEEx's Storage101x**[45]
  Contents: This introductory to intermediate course teaches enterprise data storage and management. It covers basics of enterprise IT infrastructure, virtualization, storage devices, and insights in infrastructure management.

  Method: This comprehensive course relies on good quality videos and some text.

5.3 Relevance to curriculum

This module is entirely included in the "storage and processing" stage of the curriculum. It covers both the practical storage issues with IEEE's Storage101x, and the more abstract issues, from a data management perspective, with the three other courses. This module answers both questions of how to store and process, as well as what to store and process.

5.4 Relevance to demand analysis

The demand analysis does not address entirely the "Storage and processing" stage of the curriculum as its focus was more on the analysis part of data science, not data management or storage. It nevertheless found that among a defined set of abilities, "data collection and analysis" was the most "essential" one in the eyes of the poll sample. Data curation is closely related to data collection, and from this perspective this module seems essential.

Also, the authors of the analysis declare that their study shows the need to extend the skills of data scientists, but also non-data scientists. This is reflected in the 4th final recommendation "Basic data literacy skills". Both courses "MANTRA Research Data Management Training" and "Introduction to Digital Curation" definitely try to be accessible to non-data scientists, and hence fulfil this purpose.

[44] http://www.ed.ac.uk/institute-academic-development/postgraduate/doctoral/courses/online-courses/data-management

[45] https://courses.edx.org/courses/course-v1:IEEEx+Storage101x+2016_T2/info

6. Statistical / Mathematical Foundations

6.1 Module overview

The aim of the module is to provide the basics of statistics and mathematics for modern data scientist. Statistics and Mathematics Foundations module will help students to:

- Understand the basics of probability and statistical inference.
- State a statistical problem and select the most appropriate method for problem solving.
- Gain core math skills for data science.

6.2 Learning materials & delivery methods

No EDSA partner is offering a course on this topic, with the exception of the following videolectures offered by JSI:

- **Probability and statistics (JSI, Slovenia)**[46]

  Contents: Videolecture explains the basics of probability - definitions, laws, random variables and statistical inference - estimation, hypothesis testing.

  Method: Very comprehensive materials - contain videolecture and slides available for download.

As in the case of the Data Management and Curation module, the following external courses have been incorporated into the EDSA courses portfolio and have been added to the EDSA courses portal:

- **Data Science Math Skills (Duke University, USA at Coursera)**[47]

  Contents: This course is designed to teach learners the basic math they will need in order to be successful in almost any data science math course and was created for learners who have basic math skills but may not have taken algebra or pre-calculus. Data Science Math Skills introduces the core math that data science is built upon, with no extra complexity, introducing unfamiliar ideas and math symbols one-at-a-time.

  Method: MOOC for beginners.

- **Introduction to Probability and Data (Duke University, USA at Coursera)**[48]

  Contents: This course introduces learners to sampling and exploring data, as well as basic probability theory and Bayes' rule. Learners examine various types of sampling methods, and discuss how such methods can impact the scope of inference. A variety of exploratory data analysis techniques are covered, including numeric summary statistics and basic data visualization.

  Method: MOOC for beginners.

- **Basic Statistics (University of Amsterdam, The Netherlands at Coursera)**[49]

  Contents: Understanding statistics is essential to understand research in the social and behavioural sciences. This course introduces the basics of statistics; not just how to calculate them, but also how to evaluate them. In the first part of the course, methods of descriptive statistics are discussed, e.g. what cases and variables are and how measures of central tendency can be computed (mean, median and mode) and dispersion (standard deviation and variance). Next, the assessment of relationships between variables is discussed, and the concepts

[46] http://videolectures.net/eswc2016_rupnik_probability_statistics/?q=jan

[47] https://www.coursera.org/learn/datasciencemathskills

[48] https://www.coursera.org/learn/probability-intro

[49] https://www.coursera.org/learn/basic-statistics

correlation and regression are introduced. The second part of the course is concerned with the basics of probability: calculating probabilities, probability distributions and sampling distributions. The third part of the course consists of an introduction to methods of inferential statistics. Confidence intervals and significance tests are discussed and learners are trained to calculate and generate these statistics using freely available statistical software.

Method: MOOC for beginners.

6.3 Relevance to curriculum

Statistics and Mathematics Foundations module is one of the core modules for EDSA curriculum. We expect this module to be one of the starting modules for individual learning pathway. EDSA modules related to data mining, machine learning, programming etc. are linked to the methods described in the Statistics and Mathematics Foundations module.

6.4 Relevance to demand analysis

The results of the demand analysis show that statistical skills are in high demand all over Europe. Table 5 summarises the relevance of the module to the recommendations of D1.4.

Table 5: Relevance of Statistical / Mathematical Foundations to the demand analysis recommendations.

Recommendation	**Application to this module**
Holistic training approach	This module addresses the basic skills required to proceed with data science training.
Open source based training	Materials for this module include publicly available videolectures and set of MOOCs.
Soft skills training	The module does not address soft skills.
Basic data literacy skills	The materials presented in the module are targeted at beginners.
Blended learning	People interested in face-to-face consultations can contact the module provider by email.
Data science skills framework	ACM Classification (2012) includes Mathematics of Computing facet (that covers Probability and Statistics) related to Data Science.
Navigation and guidance	The module is expected to be taken at the beginning of the individual learning pathway.

7. Conclusions and next steps

This deliverable has presented the modules that have been added to the EDSA courses portfolio in the second year of the project, in accordance with the EDSA curriculum and the EDSA demand analysis.

Following the M18 review, the EDSA courses portfolio will focus on incorporating a wider range of learning resources, either offered by project partners or by third parties. The EDSA courses portfolio is therefore being extended to include internal and external courses that are selected based on their relevance to the EDSA curriculum and the demand analysis.

During the third and final year of the project (Y3), the EDSA curriculum and courses portfolio will be further developed and finalised as follows:

- Throughout Y3, we will be monitoring on a monthly basis the demand trends for skills, as these are captured by the EDSA dashboard.
- We will be tailoring the curriculum and the courses portfolio accordingly in order to fit the current needs of the market.
- The EDSA curriculum will be finalised in M30 and published in D2.3.
- Based on the list of topics of the finalised curriculum, as well as the latest demand trends, we will produce a list of courses to be added to the project's portfolio. This list will consist of both internal and external courses.
- Each selected internal and external course will be evaluated against the demand analysis recommendations.
- The finalised set of courses will be published at the end of the project in D2.6 (M36), together with a report presenting the lessons learned and best practices (D2.7).

As a result of these planned activities, WP2 will continue collaborating with WP1, so that the demand analysis outputs dictate the shaping of the EDSA curriculum and the further development of the EDSA courses portfolio. As the project is progressing towards its final stages, WP2 will also be more strongly connected to WP5, in order to transform the courses portal into a unique offering of the EDSI beyond the duration of the project.

As part of the project's initiative for closing the gap between demand and supply, the courses portal will complement the EDSA dashboard, thus enabling users to not only explore the current market demand, but also find learning materials and training relevant to the skills they would need to secure a specific job position. Users will also be supported in building personalised learning pathways, consisting of courses and learning materials that will help them reach their learning goals.