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

D2.4 Learning resources 1

Deliverable Editor:	**Alexander Mikroyannidis (OU)**
Other contributors:	**Chris Phethean (SOTON), Inna Novalija (JSI), Angelika Voss (Fraunhofer), Rolf Bardeli (Fraunhofer), Mihhail Matskin (KTH), Shatha Jaradat (KTH), Ali Syed (Persontyle), Patrick Mukala (TU/e)**
Deliverable Reviewers:	**Elena Simperl (SOTON), Inna Novalija (JSI)**
Deliverable due date:	**31/01/2016**
Submission date:	**27/01/2016**
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	01/12/2015	Alexander Mikroyannidis	Outline and responsibilities of contributors
0.2	18/12/2015	Alexander Mikroyannidis	Updated draft with contributions from partners
0.3	07/01/2015	Alexander Mikroyannidis	Draft for internal review
0.4	21/01/2016	Alexander Mikroyannidis	Revised draft
0.5	27/01/2016	Alexander Mikroyannidis	Final version
1.0	27/01/2016	Aneta Tumilowicz	Final QA

Table of Contents

List of Tables

List of Figures

1. Executive Summary

This deliverable presents the learning resources of the modules developed by the project so far. According to the core EDSA curriculum presented in D2.1, the following 6 modules were scheduled for development during the first year of the project:

1. Foundations of Data Science
2. Foundations of Big Data
3. Big Data Architecture
4. Distributed Computing
5. Process Mining
6. Essentials of Data Analytics and Machine Learning (previously entitled "Machine Learning, Data Mining and Basic Analytics")

This deliverable also presents the courses offered by EDSA partners and associate EDSA partners beyond the core EDSA curriculum. Additionally, we detail the pedagogical models, channels and formats we have employed for the development and delivery of the EDSA learning resources to the Data Science community.

2. Introduction

The EDSA curriculum targets the following 4 themes, which provide the core framework for the development of the EDSA courses:

- Foundations of Data Science
- Data Storage and Processing
- Data Analysis
- Data Interpretation and Use

WP2 is tasked to develop a number of modules in order to cover all aspects of the above 4 themes. These modules are complemented by the videolectures and webinars produced by WP3.

The core EDSA curriculum has been shaped based on the demand analysis conducted in WP1. As this analysis will continue throughout the project's duration, the core curriculum is expected to be further refined and updated to reflect the latest market trends. The first version of the EDSA curriculum has been presented in D2.1 and consists of the modules and delivery dates listed in Table 1.

The remainder of this deliverable is structured as follows. First we introduce the EDSA courses portfolio in terms of the employed pedagogical models, as well as the different delivery channels and formats. We then proceed with presenting the learning resources of the core curriculum modules developed by the project so far. We distinguish between the courses developed for self-study and other types of courses, i.e. MOOCs, blended and face-to-face courses. Finally, the deliverable is concluded and the next steps of this work are discussed.

Table 1: The first version of the core EDSA curriculum.

Topic	Schedule
Foundations of Data Science	M6
Foundations of Big Data	M6
Big Data Architecture	M6
Distributed Computing	M6
Machine Learning, Data Mining and Basic Analytics	M6
Process Mining	M6
Statistical / Mathematical Foundations	M18
Data Management and Curation	M18
Big Data Analytics	M18
Data Visualisation	M18
Finding Stories in Open Data	M18
Programming / Computational Thinking (R and Python)	M30
Stream Processing	M30
Visual Analytics	M30
Data Exploitation including data markets and licensing	M30

3. The EDSA courses portfolio

The EDSA courses portfolio includes a wide range of Data Science courses that adopt a variety of pedagogical models, as well as employ different delivery channels and formats in order to address different learning contexts and audiences. The following sections detail the types of courses offered by EDSA, as well as the main delivery channels and formats.

3.1 Types of courses

Based on the pedagogical models used for the design, development and deployment of a course, we define the following categories of EDSA courses:

- *Self-study courses:* These courses consist of self-study learning materials available as Open Educational Resources (OERs). Learners can study them at their own pace, as there is no predetermined start or end date.
- *MOOCs*: These Massive Open Online Courses (MOOCs) are available on external MOOC platforms, such as Coursera and FutureLearn.
- *Blended courses:* These courses are taught in a blended way (face-to-face and online) by EDSA partners and associate EDSA partners.
- *Face-to-face courses:* These courses are taught face-to-face by EDSA partners and associate EDSA partners.

As it can be seen from the above list, the EDSA courses cover all types of learning contexts, from the traditional face-to-face pedagogical model, to the more recent trends in online education (MOOCs and OERs).

3.2 Delivery channels and formats

The EDSA courses employ different delivery channels and formats in order to maximise the impact of the EDSA learning materials on the community and bring them closer to as many people as possible. In particular, the EDSA courses are available:

- Via the Moodle Learning Management System (HTML format)
- As an eBook (iBooks and ePUB formats)

The HTML format can be accessed by any web browser, while the eBook is available to download and use even without an internet connection on iPads and iPhones (iBooks format), as well as other tablets and smartphones (ePUB format). The following sections provide more details about the employed delivery channels and formats.

3.2.1 The EDSA online courses portal

The EDSA online courses portal[1] is based on the Moodle Learning Management System. A Learning Management System (LMS) is an online software application offering facilities for student registration, enrolment into courses, delivery of learning materials to students, student assessment and progress monitoring. Moodle[2] is an open-source learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalised learning environments. Moodle has been adopted by numerous educational institutions worldwide, including the Open University[3]. Moodle currently has more than 79 million users across the academic and enterprise

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

[2] https://moodle.org

[3] https://www3.open.ac.uk/media/fullstory.aspx?id=7354

sectors. These figures make it the world's most widely used learning platform. Additionally, as it is open source it has attracted a sizeable community of developers, which offers a wide range of free and open plugins that extend and enrich the functionalities provided by Moodle.

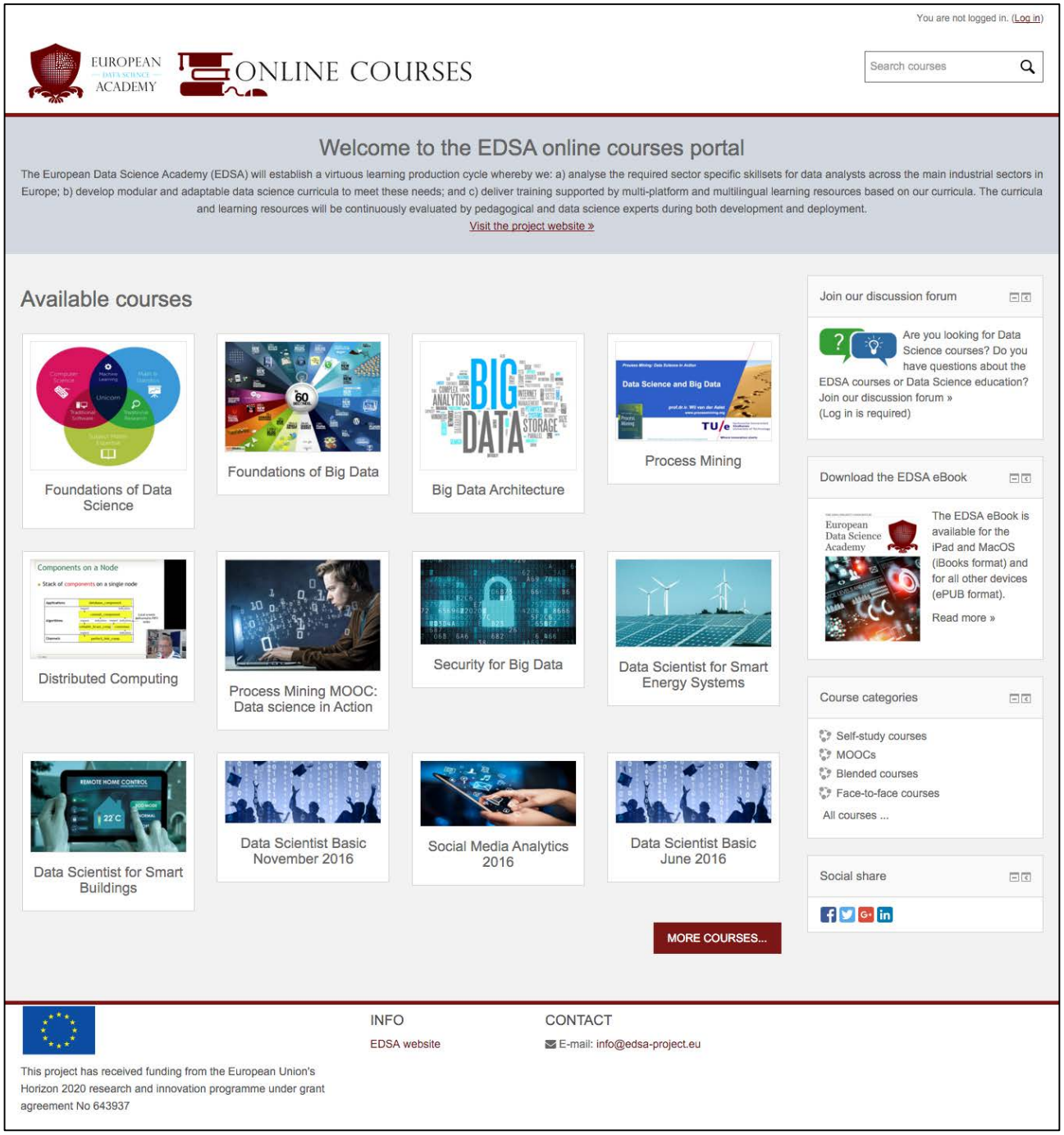

Figure 1: The EDSA online courses portal.

Figure 1 shows a screenshot of the homepage of the EDSA online courses portal. The homepage features a list of the available courses and their categories. The portal hosts the full learning materials (presentations, webinars, text, quizzes, etc.) for the self-study courses, to which learners can enrol and study at their own pace. Learners are not required to register in the portal in order to enrol to a course, but they can login using an existing social media account, such as Google, Facebook or LinkedIn. The portal also lists the other types of EDSA courses, namely MOOCs, blended and face-to-face courses. For

each of these courses, we provide a brief overview and a link to the dedicated course page on the website of the EDSA partner or associate EDSA partner that offers the course.

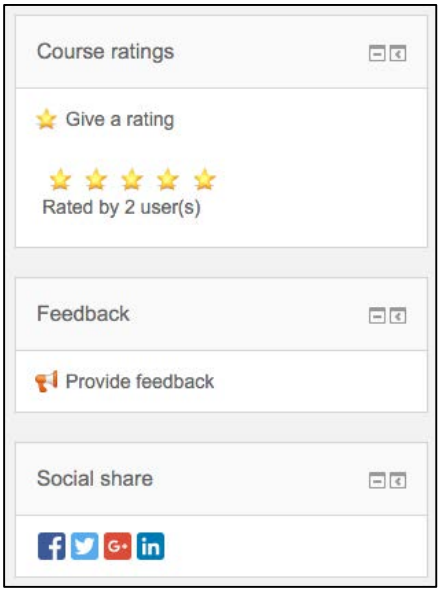

Figure 2: Feedback and sharing mechanisms available on the sidebar of an EDSA course.

A number of feedback and sharing mechanisms are offered to learners, as shown in Figure 2. Learners who enrol in a self-study course can rate it using a five-star rating system. These ratings are displayed on the sidebar of each self-study course. Enrolled learners also have the opportunity to offer more detailed feedback about a particular course by answering a questionnaire (see Appendix 7.1). The questionnaire is comprised of a set of statements to which the learner is requested to record his/her level of agreement via a 5-Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). These questions are focused on the ease of use and learning effectiveness of the EDSA learning materials. Additionally, open-ended questions ask about the things that the student has enjoyed most about the EDSA learning materials, possible improvements, as well as any additional comments. More details about this feedback instrument will be provided in D3.3 Report on the evaluation of course content and delivery (M18). Finally, learners can share any course or learning resource in their preferred social media.

Users of the portal (learners, educators, trainers, etc.) have the ability to join the EDSA community by participating in different discussion forums available throughout the portal. The homepage of the portal features a general discussion forum[4], where participants can post general questions about courses they are looking for, or questions about the offered EDSA courses. Additionally, each self-study course contains a dedicated discussion forum intended to provide peer-support to the learners that have enrolled in a particular course.

As mentioned before, Moodle offers a wide range of plugins that we can use to customise the platform based on our learners' needs. One of these plugins allows us to employ Learning Analytics technologies in order to track the activities of learners throughout the portal[5]. Details about these technologies and the ways we use them to monitor and analyse the behaviour of learners will be available in D3.3 Report

[4] http://courses.edsa-project.eu/mod/forum/view.php?id=284

[5] https://moodle.org/plugins/logstore_xapi

on the evaluation of course content and delivery (M18). The use of additional plugins, such as badges[6], will be explored throughout the duration of the project by studying the behaviour of learners and analysing their feedback. Finally, a Glossary will be developed in order to gather all the key data science terms and concepts described in the EDSA courses and provide their definitions.

3.2.2 The EDSA eBook

The EDSA eBook[7] is an additional delivery medium for the project's courses, targeting primarily tablet devices and mobile phones. An electronic book or eBook is a book-length publication in electronic form, consisting of text, images, and (depending on the format used to publish it) videos and other interactive elements. eBooks can be read on computers or other electronic devices, such as tablets and smartphones.

Figure 3: Screenshots from the EDSA eBook.

One of the most common formats for publishing an eBook, supported by the majority of electronic devices, is ePUB. ePUB is a free and open eBook standard developed by the International Digital Publishing Forum (IDPF)[8]. ePUB is essentially a ZIP archive that contains the files comprising the eBook's content, including HTML files, images, CSS style sheets and other assets, plus metadata that make the content able to be reliably consumed by any application or device compatible with the ePUB

[6] https://docs.moodle.org/30/en/Badges

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

[8] http://www.idpf.org

specification. In its latest version (3.0[9]), ePUB supports HTML5, allowing interactive elements like widgets[10] to be embedded in the eBook.

On the other hand, Apple has developed a proprietary format for eBooks specifically for the iPad and MacOS, called iBooks. The iBooks format allows the embedding of interactive elements inside an eBook, such as HTML5 widgets[11], thus enabling it to feature dynamic content offered by external online sources. The authoring software that we use for this purpose is the iBooks Author, an application for MacOS available to download for free[12]. The application contains templates for galleries, quizzes, Keynote presentations, interactive images, and more. Alternatively, users can create and embed their own HTML5 widgets.

In order to widen the audiences reached via different platforms, the EDSA eBook is available both in the iBooks and ePUB format. Figure 3 shows screenshots from the current version of the EDSA eBook. The eBook contains the textual and image/video learning resources of the EDSA self-study courses. As not all self-study courses are comprised of textual learning resources and are thus not suitable to be published in the form of a book, the eBook contains a subset of the courses available on the EDSA online courses portal. The eBook is a different type of medium intended for a different type of use compared to the portal, hence this variation in the content and presentation of the courses between the eBook and the portal.

Throughout the duration of the project, the eBook will grow as more learning resources will be added. Its content will also become more mature as it will be updated based on the feedback received from the Data Science community. We will also consider producing more than one eBooks, in order to target different levels of expertise and different training purposes.

[9] http://www.idpf.org/epub/30/spec/epub30-overview.html

[10] http://www.idpf.org/epub/widgets

[11] http://support.apple.com/kb/HT5068

[12] http://www.apple.com/uk/ibooks-author

4. Self-study courses

The following sections describe the learning resources comprising the modules of the core EDSA curriculum that have been developed in the first year of the project. These modules are offered in the form of self-study courses via the EDSA online courses portal. The modules with the appropriate format (i.e. text and images) are also included in the EDSA eBook. It should be noted that the overlaps that exist between some of these modules are intentional, as they address different levels of expertise or comprise different learning pathways.

In the following sections, we provide for each module an overview of the covered topics and learning objectives, we list the types of learning materials available for this module, as well as the future development plans for the module.

4.1 Foundations of Data Science

4.1.1 Module overview

This module has been developed by the University of Southampton and covers a range of key topics of Data Science. By the end of this module, learners will:

- Understand the foundations of the data science process
- Be able to evaluate data science tools and techniques based on their suitability for particular tasks
- Have gained hands-on experience using R to analyse data

The key areas addressed by this module are the following:

- *Core Data Science terminology*: This will cover key data science concepts that any data science student would need to understand, and will lay the foundations for further data science knowledge to be acquired.
- *Technology pipeline and methods*: In order to gain an understanding of the data science process, this topic will examine the key stages of the data science process and the technologies that are required and can be adopted at each point.
- *Data Science application scenarios and state of the art*: Helping to develop an understanding of the potential impact of data science, we will cover certain application scenarios where data science is already making huge changes, reflecting the state of the art developments in the field and how organisations are implementing these.
- *Data collection techniques (sampling and crawling, brief intro to QA and curation methods):* Before covering this later in the curriculum, we will cover an introduction to data collection methods in the context of the overall data science process that will be necessary in order to make sure that this process is understood. Additionally this will also show the different methods and approaches that could be adopted at this stage of the process.
- *Data analytics – basic statistical modelling, basic concepts, experiment design, pitfalls):* This will also cover a basic introduction to various analytical approaches that could be used at this stage of the data science process.
- *Introduction to R for data science:* A brief introduction and guide to using R for carrying out some of the basic analytics discussed in the previous topic. This will be essential in order for these skills to be developed further in later stages of the overall curriculum.
- *Data integration – linked data and Google/OpenRefine:* Key concepts around linked data, the star ratings etc. This would also introduce learners to Open Refine and would teach them the skills necessary to use this software.
- *Data interpretation and use (basic visualisation techniques):* This topic will cover some of the key ideas behind visualising data in a clear and coherent way. It will also provide a list of online resources for visualising data such as CartoDB.
- *High performance computing (MapReduce, Hadoop, NoSQL, stream processing solutions):* An introduction to the technical foundations of data science and how these technologies are evolving and facilitate many of the concepts required for data science to happen.

4.1.2 Learning materials

The learning materials for this module consist of presentations covering all the above topics of the course (see Figure 4). In addition, the module includes a set of tutorials on MongoDB, NodeJS, D3 and Linked Data.

Figure 4: Snapshot of the learning materials and structure of the "Foundations of Data Science" module.

4.1.3 Further development plans

Further materials will be added to the module to provide a narrative around the existing content. We will revisit the material to ensure that there is a consistent theme and message throughout, and to verify that it can be broken down in order to be used in various learning pathways for different groups of learners once these have been defined. We will investigate opportunities to make interactive exercises for some of the content - particularly those areas with tutorials - that can then be used within an eBook or MOOC as necessary.

4.2 Foundations of Big Data

4.2.1 Module overview

This module has been developed by the Jožef Stefan Institute (JSI) and introduces the core topics related to Big Data. By the end of this module, learners will:

- Understand the foundations of big data
- Get knowledge on the tools that operate with Big Data and Big Data applications
- Get experience on data analytics with QMiner tool

The key areas addressed by this module are the following:

- *Introduction to Big Data:* This topic describes the notion of Big Data, paying a special attention to interesting facts about Big Data. The topic provides a view on Big Data in numbers.
- *Big Data Definitions, Motivation and State of Market:* This topic shows the definitions of Big Data, the motivation behind operating with Big Data. The characterization of Big Data by volume, velocity, variety (V3) is provided as well as Big Data popularity on the Web, Big Data hype cycles, Big Data value chain and view on the Big Data market.
- *Techniques Overview:* This topic is dedicated to the analytic techniques that are used for operation with Big Data. The specific analytical operators for Big Data are discussed.
- *Tools Overview:* This topic provides a view on types of tools that are used for Big Data. Distributed infrastructure and Distributed processing are discussed. A particular attention is given to MapReduce, NoSQL databases. Open source Big Data tools are listed.
- *Applications:* This topic discusses a number of Big Data applications, such as recommendation, social networks, as well as media monitoring
- Mining Massive Datasets: This topic provides algorithms for extracting models and other information from very large amounts of data. The emphasis is on techniques that are efficient and that scale well.
- *Data analytics with QMiner:* This topic provides practical insights on data analytics using QMiner. QMiner implements a comprehensive set of techniques for supervised, unsupervised and active learning on streams of data.

4.2.2 Learning materials

The learning materials for this module consist of presentations, text and images (see Figure 5). At the end of the course, learners are provided with resources for further reading, consisting of related courses, books and web resources.

Figure 5: Snapshot of the learning materials and structure of the "Foundations of Big Data" module.

4.2.3 Further development plans

The module Foundations of Big Data Architecture, currently available in the form of presentations, text and images, will be extended within the first term of 2016 by a set of examples, exercises and relevant videos from Videolectures.net. We are planning to use Tonic (https://tonicdev.com/) as a platform for QMiner demonstration online.

4.3 Big Data Architecture

4.3.1 Module overview

This module has been developed by Fraunhofer and provides an overview of architectural designs and technical components related to Big Data. Based on computing concepts like "MapReduce", theoretical insights like the CAP theorem and non-functional requirements like processing in real time Big Data products are presented and rated. Learners will be able to realistically assess the application of big data technologies for different usage scenarios and start with their own experiments.

The key areas addressed by this module are the following:

- *Introduction:* The components used by internet giants, e.g. Google, Facebook or Amazon, to build big data applications are provided, and often extended in various ways, by open source communities. The technological building blocks for leveraging big data are freely available and also substantially present in portfolios of large systems providers. Still, there are some major obstacles to overcome before implementing big data applications. Big data components often provide less functionality than usually expected of operating systems, relational databases or BI systems.
- *Lambda Architecture:* A constructively usable pattern for concept and design of a big data application is the "lambda architecture", as published by Nathan Marz and James Warren. The main point is to differentiate between a batch layer for large volumes of data and a speed layer for real time processing of data streams. Both layers create analytical results and store them in scalable databases. An application service combines the results of both layers and presents them to the user. The modular design of the lambda architecture maps well to common requirements of big data applications and systematizes them. The architectural approach is helpful for discovery and evaluation of technical and non-functional requirements. This is independent of the way and extend the modules are used as technical components of the application. Required services are identified and a suitable selection of components can be made.
- *Batch Processing:* The horizontal scalability of big data systems allows processing large volumes of data in a short amount of time. For better utilization of the individual components of a distributed system, the interfaces of big data frameworks are restricted. The limitations are designed to discourage behaviour that works well on a single computer but are not suited for processing on multiple machines. Available open source solutions differ in focus and therefore in their interfaces. It is essential to select system components and compose them in a way that they fulfil the requirements of the application. The easiest way of implementing these steps is a concept of dataflow driven building blocks. We consider the popular disturbed processing scheme map reduce and how this scheme is made available by the open source software Hadoop. Further, we demonstrate the use of Cascading as a workflow engine to assemble advances processing schemes.
- *Speed Processing:* Within the Lambda-Architecture the input data is processed by multiple layers, i.e. the batch and speed layer. The transport is provided by a messaging system, in our example kafka. We describe the general publish subscribe mechanism of kafka and demonstrate the transport to a distributed processing engine we choose for the speed layer, Apache Strom. We give an overview on the organization of processing tasks in the Storm system and elaborate on the involved components, i.e. topologies, spouts, bolts and streams. In contrast to the views of the batch layer, views for the speed layer require high write performance. As an example of a system capable of providing such capability, we select the NoSQL database Cassandra. We

analyse the concepts used for data storage organization, for instance columnar storage organization and merkle trees.

- *Exercises:* We show the construction of batch and speed layer of an application by way of an example. The goal of the application is to monitor posts of an online board for technical terms and emotions, such as joy, concern or anger. The results are aggregated and presented for interactive exploration to the analyst via a web interface. The batch layer processes large amounts of historical data, while the speed layer analyses new posts in a timely manner. For the batch and speed functionality "Hadoop" with "Cascading" and "Storm", respectively, are used. The messaging bus "Kafka" distributes incoming information to the Hadoop File System (HDFS) and the data entry points of "Storm". Exercises consist in building and adapting workflows at the batch and speed layers.

4.3.2 Learning materials

The learning materials for this module consist of text and images addressing all topics of this module, as shown in Figure 6.

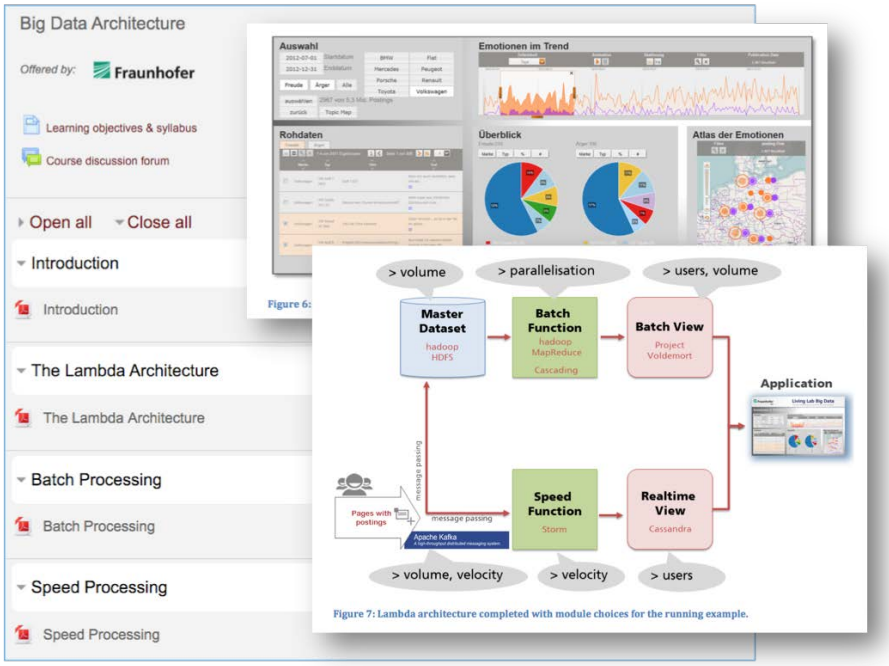

Figure 6: Snapshot of the learning materials and structure of the "Big Data Architecture" module.

4.3.3 Further development plans

The module Big Data Architecture, currently available in the form of text with figures, will be extended within the first term of 2016 by a set of quizzes and videos introducing the central points of each chapter, i.e. introduction, lambda architecture, batch processing and speed processing.

4.4 Distributed Computing

4.4.1 Module overview

This module has been developed by the Royal Institute of Technology (Kungliga Tekniska Högskolan – KTH) and provides learners with the basic concepts and principles of large-scale dynamic distributed systems and distributed algorithms to be applied to Big Data. By the end of this module, learners will:

- Understand and apply the main concepts and principles from large-scale dynamic decentralized systems.
- Implement and evaluate peer-to-peer algorithms in a simulation environment.
- Understand the main concepts and principles from cloud computing when building a distributed system.

The key areas addressed by this module are the following:

- Models of distributed algorithms
- Fault Tolerance Abstractions and Failure Detectors
- Reliable Broadcast and Causal Broadcast
- Replicated Shared Stores and Consistency Models
- Single Value Consensus, and Paxos
- Sequence Consensus, and Multi-Paxos
- Atomic Broadcast
- Replicated State Machine
- Reconfiguration
- Clocks in Distributed Systems

4.4.2 Learning materials

The learning materials for this module consist presentations, webinars, programming tutorials and exercises as shown in Figure 7. The programming tutorials and exercises offer learners the opportunity to develop their programming skills in distributed computing using the Kompics component model and programming framework.

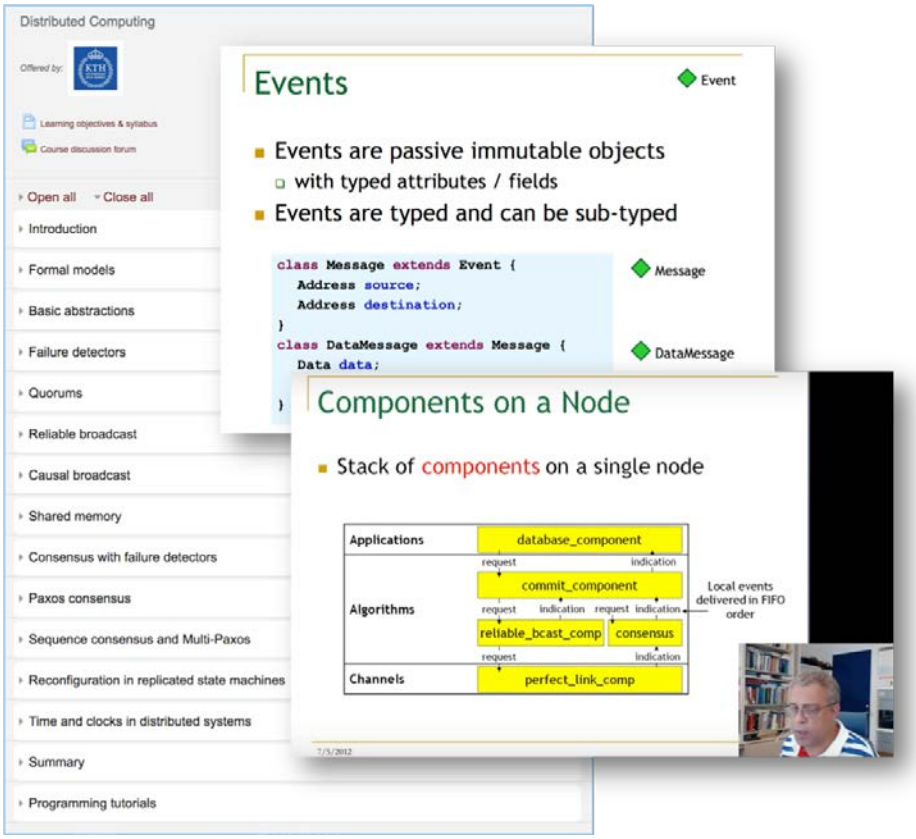

Figure 7: Snapshot of the learning materials and structure of the "Distributed Computing" module.

4.4.3 Further development plans

The learning materials of this module will be further refined and updated. In addition, two follow-up courses related to the "Distributed Computing" category will be developed as self-study courses, namely:

- Distributed Artificial Intelligence and Intelligent Agents
- Programming Web Services

4.5 Process Mining

4.5.1 Module overview

This module has been developed by the Technische Universiteit Eindhoven (TU/e) and provides data science knowledge that can be applied directly to analyse and improve processes in a variety of domains. By the end of this module, learners will:

- Have a good understanding of Business Process Intelligence techniques (in particular process mining),
- Understand the role of Big Data in today's society,
- Be able to relate process mining techniques to other analysis techniques such as simulation, business intelligence, data mining, machine learning, and verification,
- Be able to apply basic process discovery techniques to learn a process model from an event log (both manually and using tools),
- Be able to apply basic conformance checking techniques to compare event logs and process models (both manually and using tools),
- Be able to extend a process model with information extracted from the event log (e.g., show bottlenecks),
- Have a good understanding of the data needed to start a process mining project,
- Be able to characterize the questions that can be answered based on such event data,
- Explain how process mining can also be used for operational support (prediction and recommendation), and
- Be able to conduct process mining projects in a structured manner.

The key areas addressed by this module are the following:

- *Introduction, Process Modelling and Analysis:* Process mining is introduced and key concepts explored. In this topic students learn about event logs, the input for process mining, and about Petri nets, the process modelling notation used to explain foundational concepts. This topic also provides the theoretical foundations of process modelling and process discovery. In the next three topics, students will use these concepts in a more applied setting.
- *From Event Logs to Process Models:* In this topic, a practical aspect of process mining is introduced. Students learn basic discovery algorithms (i.e. alpha-algorithm) to discover models from event logs. Students are introduced to the process of turning data (from various data sources) into proper event logs, needed for process mining. Moreover, challenges encountered with event logs such as noise and incompleteness are discussed.
- *Advanced Process Discovery Techniques:* In this topic students learn even more process discovery algorithms. In the first part, the main focus is on conformance checking, i.e., aligning observed behaviour with modelled behaviour. This can be used for a wide variety of compliance questions: Where and why do people, machines, and organisations deviate? Students learn different ways of evaluating the conformance between a process model and the event log. The second part of lectures during this topic explores different perspectives that can also be mined from event logs. Techniques for social network analysis, resource behaviour and decision point are discussed.
- *Putting Process Mining to Work:* In this last topic, the emphasis in on the application of process mining to real life use cases. We demonstrate how to conduct a process mining project from start to finish. We also discuss ProM and other available process mining tools that can be used to perform experiments.

4.5.2 Learning materials

The learning materials for this module consist of presentations and quizzes, as shown in Figure 8. The quizzes are placed in various topics and act as self-assessment exercises, allowing learners to quickly put to the test what they have learned in different stages of the module.

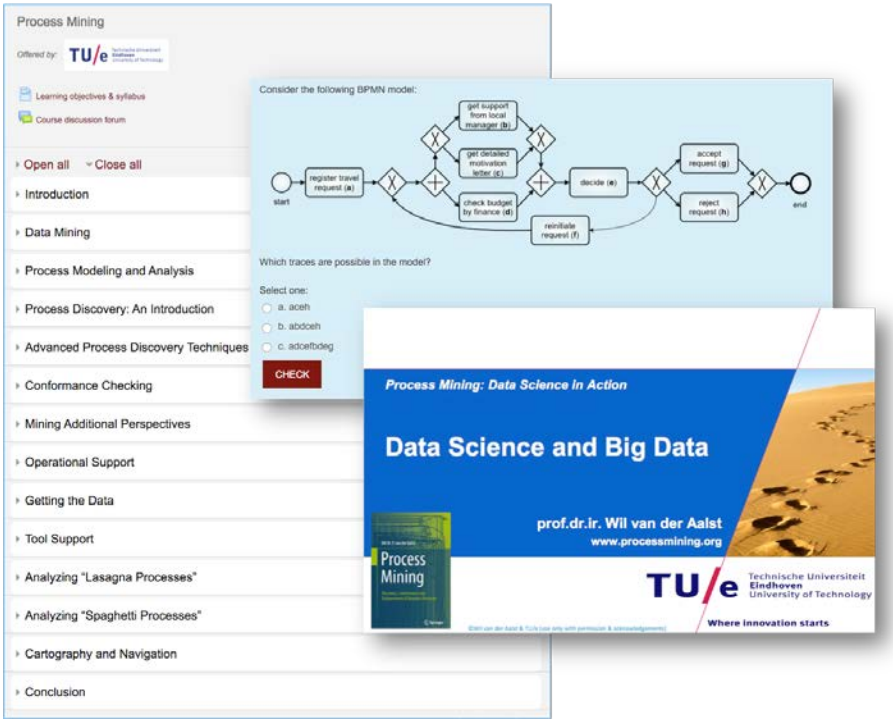

Figure 8: Snapshot of the learning materials and structure of the "Process Mining" module.

4.5.3 Further development plans

The contents of the Process Mining MOOC, upon which this module is based, will be moved to Coursera's on-demand platform in the coming weeks. TU/e is also preparing some follow up modules with a practical aspect: Introduction to Process Mining with ProM, Process Mining in Healthcare and Process Mining for Smart Environments. As of now, the MOOC on the Introduction to Process Mining with ProM is under production and is expected to be available on FutureLearn in the next couple of months (April-May 2016).

4.6 Essentials of Data Analytics and Machine Learning

4.6.1 Module overview

This module has been developed by Persontyle and introduces learners to the tools and concepts associated with Data Analytics and Machine Learning. By the end of this module, learners will acquire knowledge of:

- What Machine Learning and Data Mining entails and why it is important.
- The different types of Learning.
- Be able to use R to apply a number of the most common and powerful statistical machine learning and data mining techniques.
- Know how to implement such techniques in principle and therefore be able to apply their knowledge within paradigms outside R.
- Be able to appreciate the trade-offs involved in choosing particular techniques for particular problems.
- Be able to utilize rigorous methods of model selection.

- Understand the mathematical ideas behind, and relationships between, the various methods.
- Have a greater confidence in their knowledge and standing as a data scientist.
- How to use these algorithms in a variety of benchmark datasets.
- How to fine-tune these algorithms for better performance.

The key areas addressed by this module are the following:

- *R Refresher:* We begin with a refresher of R that ensures you have correctly installed the R studio IDE, examines how this IDE work and shows how exercise packages can be loaded. We then look at the basic functions we will be using in early exercises.
- *Basic Regression and Model Selection Techniques:* Understand the idea of supervised learning, as well as the form and applications of regression type statistical models. Be able to implement and apply linear, quadratic and polynomial regression, and have experience doing so on actual data. Understand the role of basis projection in polynomial regression and PCA. Be able to evaluate model performance using mean squared error. Be able to use hold-out validation and cross validation for model selection, and understand the relationship between model complexity and performance. Be able to implement and apply principle component analysis (PCA) and use PCA for feature selection, information compression and regression. Understand and be able to apply feature shrinkage and subset selection techniques within the context of simple regression models. Understand the idea of degrees of freedom for measuring model complexity. Be able to model regression error using error functions.
- Basic Classification Techniques: Understand the form and applications of classification type statistical models. Be able to implement and apply linear and quadratic discriminate analysis (LDA, QDA), and perceptron classification, and have experience doing so on real data. Be able to apply logistic regression, and have experience doing so on real data. Be able to evaluate model performance using misclassification error. Understand the idea behind Bayesian Methods in statistics. Be able to use noisy-or and Dirichlet-categorical distributions to encode expert knowledge with count and pseudo-count parameters, and have experience doing so.
- *Cluster Analysis:* Understand the idea of unsupervised learning, as well as the form and application of cluster analysis. Be able to implement and apply K-Means, K-Mediod and hierarchical clustering algorithms for cluster analysis, and have experience doing so with real data. Be able to use dendrograms to represent the results of hierarchical clustering algorithms.
- *Local Methods:* Understand the form and application of local methods, as well as their very distinctive strengths and weaknesses. Be able to implement and apply the K-Nearest-Neighbours, local regression and kernel density estimation algorithms, and have experience doing so with actual data. Understand and be able to work with kernel functions.
- *Trees and Boosting:* Be able to implement and apply regression/classification trees. Be able to implement the adaboost algorithm.
- *Advanced Techniques - Support Vector Machines and Neural Networks:* Understand how support vector machines (SVMs) and neural networks (NNs) work and the reasons for their success. What support vectors, optimal hyperplanes and support vector classifiers are and their relationship to SVMs. What back-propagation is and how it is used to train NNs. Understand the links between SVMs and NNs and the simpler statistical models from earlier modules. The use of kernels and implicit basis projection in SVMs. The role of adaptive basis projection in NNs. The role of linear and logistic regression in NNs. The relationship between radial basis networks and kernel basis functions and smoothing splines. The relationship between weight decay in NNs and ridge regression.
- *Advanced Model Selection:* Be aware of a number of additional statistical and information theoretic model selection and validation techniques and be able to apply them to real life problems. Understand the advantages and disadvantages of the different methods.

4.6.2 Learning materials

The learning materials for this module consist of text and images addressing the topics of this module.

4.6.3 Further development plans

The learning materials of this module will be further refined and updated to include Neural Networks methods, new data exercises and Azure Machine Learning Studio Using R. Also this module will be developed as a MOOC.

5. MOOCs, blended and face-to-face courses

As mentioned before, the EDSA courses portfolio spans beyond self-study courses. EDSA also offers a variety of Massive Open Online Courses (MOOCs), blended and face-to-face courses. These courses are all listed in the EDSA courses portal and are updated regularly. For this purpose, we have devised an online form, which is used internally by the EDSA partners in order to report new courses. At the time of writing this deliverable (December 2015), the EDSA partners offer the MOOCs, blended and face-to-face courses listed in the following sections.

5.1 MOOCs

5.1.1 Process Mining: Data science in Action

Offered by: TU/e

URL: https://www.coursera.org/course/procmin

Process mining is the missing link between model-based process analysis and data-oriented analysis techniques. Through concrete data sets and easy to use software the course provides data science knowledge that can be applied directly to analyse and improve processes in a variety of domains. This course is available as a Coursera MOOC.

5.2 Blended courses

5.2.1 Security for Big Data

Offered by: Fraunhofer

URL: http://www.professionalschool.eitdigital.eu/professional-training-courses/security-for-big-data/

Knowledge about increasing speed, mass and value is no longer sufficient for working with big data. There is a fundamental skills gap regarding security and data privacy in big data applications. At least basic knowledge in security is inevitable for a successful implementation of a big data product or service. This course sensitizes regarding security and privacy and provides the participants with basic knowledge to apply security solutions for big data environments.

This course will enable participants to understand the emerging trends and issues in security and privacy of big data systems and applications. Guided by practical examples, the participants will learn to identify the issues to be tackled, how to choose and apply appropriate approaches, technology and tools, and how to evaluate a solution.

5.2.2 Data Scientist for Smart Energy Systems

Offered by: Fraunhofer

URL: http://www.professionalschool.eitdigital.eu/professional-training-courses/data-scientist-for-smart-energy-systems/

The European power grid is facing massive challenges. The increasing amount of distributed renewable energy generation overthrows the known concepts of producers and consumers in the energy market - consumers turn to producers, traditional producers losing their key incoming sources and turning towards service providers. Volatile energy sources like wind and solar require the power grid to turn smart, in-order to keep the grid stability facing bottom-up energy and volatile flows. Hence, ICT is playing an increasingly important role steering energy flows and forming the power grid of the future.

The Data Scientist for Smart Energy Systems course will introduce the changes and challenges of the European power grid and discusses the roles of ICT regarding the arising issues. The course identifies the sources of data and presents software solutions to turn Big Data into Smart Data. Additional e-learning material is provided to deepen contents and own project challenges can be discussed in one-on-one consulting sessions with Smart Energy experts. Further, participants learn about Big Data

Technologies and get the chance to actively gain expertise using the Fraunhofer Big Data Framework GPI-Space.

5.2.3 Data Scientist for Smart Buildings

Offered by: Fraunhofer

URL: http://www.professionalschool.eitdigital.eu/professional-training-courses/data-scientist-for-smart-buildings/

The blended learning course Data Scientist for Smart Buildings deals with methods and software for intelligent energy management (monitoring, analysis, simulation, optimisation) in buildings or infrastructure based on measurement of data resources.

The participants will learn to identify and characterize problems within the intelligent energy management area, learn appropriate approaches and methods, see how to apply them, evaluate the outcomes, and communicate solutions developed. They will also learn and exercise the selection, application and evaluation of software tools and results. Techniques to communicate solutions are presented as well; contents include visualisation techniques for data, models, optimisation outcomes, and discussing alternatives.

5.3 Face-to-face courses

5.3.1 Data Scientist Basic

Offered by: Fraunhofer

URL: http://www.bigdata.fraunhofer.de/de/datascientist/seminare/zertifizierung.html

The participants acquire broad basic knowledge in data science, which enables them to collaborate in teams of data scientists. The Fraunhofer certificate “data scientist basic” can be obtained by passing a written exam about the content of the course, which is offered on the following day.

The following subjects are taught in this course:

- Business potentials
- Data engineering
- Data analytics
- Security and privacy

5.3.2 Social Media Analytics

Offered by: Fraunhofer

URL: http://www.iais.fraunhofer.de/socialmediaanalytics.html

The participants can distinguish different tasks in social media analytics, choose suitable methods and know how to compose a workflow for text analytics.

The following subjects are taught in this course:

- Scenarios, problems, tasks and state of the art
- Crawling and monitoring
- Repositories and preprocessing
- Analysis of complete social media posts
- Semantic similarity of concepts
- Recognition of names, products and companies
- Marketing in the automotive domain
- Marketing: emotions, tweeting soccer

5.3.3 Big Data Analytics

Offered by: Fraunhofer

URL: http://www.iais.fraunhofer.de/bigdataanalytics.html

The participants will know how to implement analytical and machine learning methods for a scalable big data architecture and have seen examples for batch and stream processing.

The following subjects are taught in this course:

- Sampling as an approach to the analysis of big data sets
- Analyzing big data sets in existing IT environments
- Developing statistical models in big data systems
- Analyzing big data streams
- K-means and clustering with spark
- Lineare regression with spark
- Data analysis exercises with spark
- Recognition of complex events for fraud detection with Proton
- Commercial big data systems

5.3.4 Basic Data Analytics

Offered by: Fraunhofer

URL: http://www.iais.fraunhofer.de/data-scientist-basicanalytics.html

After the course the participants can work on first analytics tasks and can assess the suitability of different machine learning methods.

The following subjects are taught in this course:

- The data analysis workflow according to CRISP-DM
- Data preparation and evaluation
- A simple workflow demonstrated with RapidMiner
- Selection of models and features
- Classification
- Regression
- Clustering
- Transformation
- Introduction to R
- A complete data analysis workflow with R

5.3.5 Multimedia Analytics

Offered by: Fraunhofer

URL: http://www.iais.fraunhofer.de/multimediaanalytics.html

The participants will understand the scope of different methods of multimedia analytics and assess the conditions for their application.

The following subjects are taught in this course:

- Automatic speech recognition
- Audio fingerprinting
- Optical recognition of icons in videos
- Recognition of logos
- Recognition of speakers
- Deep neural networks

5.3.6 Finding Stories in Open Data

Offered by: ODI

URL: http://theodi.org/courses/finding-stories-in-open-data

The course aims to help learners source relevant data, show how data can be analysed and used effectively and how to interpret findings. The course provides a workflow model for approaching data to find insights and shows what is possible with freely available tools.

By the end of the one-day course, learners will:

- Be armed with a list of great data sources
- Have selected a data set you can use for telling a story to others
- Understand new ways to find stories in data
- Know which free tools are available to help you clean, analyse and visualise data
- Be familiar using a spreadsheet and Open Refine
- Have created a visualisation of a dataset
- Know when they might need further help and what skills might be required

5.3.7 Open Data in Practice

Offered by: ODI

URL: http://theodi.org/courses/open-data-in-practice-3days

The essential course for those looking to take advantage of open data in their organisation and develop all the required skills. The Open Data in Practice course is for those who have experience with open data and have knowledge of the definitions of open data and best practices for publishing open data.

For individuals and teams who want to learn how to publish, consume and exploit open data to improve efficiency, create new services and to act on opportunities. We will cover the very best practice involved in using and publishing open data and the legal and policy requirements in order to remove any potential risks.

By the end of the course, learners will be able to:

- Apply the (six) steps of publishing open data to a dataset
- Choose the most appropriate methodologies for handling a dataset
- Know what job roles needs to be involved in an open data project
- Validate, clean and enrich a dataset
- Know the value of open data across an organisation
- Describe how companies are innovating with open data
- Explain the benefits of open data
- Analyse the risk in opening up a dataset
- Examine how law and licensing affects the publication and usage of open data around the world
- Summarise the opportunities and challenges involved in managing open data projects
- Describe the (five) steps in using open data
- Be able to critically evaluate, appraise visualisation techniques for data
- Create an interactive visualisation

5.3.8 Open Data Science

Offered by: ODI

URL: http://theodi.org/courses/open-data-science

This one-day course provides practical knowledge in open data science and the skills required in this area. After the end of the course, learners will be able to:

- Define open data science
- Describe a number of key data science stories

- Identify the characteristics of open data science projects
- Apply tools to unlock and transform data available on the web
- Perform a number of statistical experiments on data
- Create a simple visualisation for communicating the value of data
- Assess strategies for implementing open data science principles in projects

6. Conclusions and next steps

This deliverable has introduced the EDSA courses portfolio in terms of the employed pedagogical models, as well as the different delivery channels and formats. The deliverable has also presented the learning resources of the core curriculum modules developed during the first year of the project.

The EDSA courses consist of self-study learning resources, MOOCs, blended and face-to-face courses. They are delivered via different channels and formats, specifically via the Moodle Learning Management System (HTML format), as well as in the form of an eBook (iBooks and ePUB formats). The EDSA online courses portal has been developed in order to deliver the EDSA courses in the HTML format, which can be accessed by any web browser. The portal also enables learners, educators and trainers to connect with each other and exchange their views on the offered courses and submit their feedback. On the other hand, the EDSA eBook allows learners to download it and use it even without an internet connection on their iPads and iPhones (iBooks format), as well as other tablets and smartphones (ePUB format).

The next steps of this ongoing work will involve the development of new courses, as well as the continuous improvement of the existing ones. As the demand analysis conducted in WP1 will continue throughout the project's duration, the core curriculum is expected to be further refined and updated to reflect the latest market trends. As a result, the EDSA courses will be further updated, based also on the feedback received from the community. Additionally, courses offered outside of the EDSA project by external institutions will be collected and integrated with our curriculum. For this purpose, we have launched a call for interested parties who wish to join the project and potentially contribute to the EDSA curriculum[13]. The next version of the curriculum will be presented D2.2 (M18). The courses developed during the second year of the project will be presented in D2.5 (M24). Finally, the methods for collecting feedback from the community about the offered curriculum and courses, as well as the analysis of this feedback will be presented in D3.3 (M18) and D3.5 (M36).

Furthermore, we will be enriching new and existing courses with more metadata, so that they can be retrieved more easily and efficiently by learners, educators and trainers. In particular, we will be adding additional metadata about a course's level, prerequisite skills, as well as the acquired skills upon completion of the course. This will enable us to develop a new faceted search interface, allowing a more detailed filtering of courses based on their metadata. We will also be linking the course skills to the skills retrieved by the EDSA dashboard, so that someone who explores the current demand for Data Science skills across Europe, can also be pointed to relevant courses that can help him/her acquire these skills. In this way, we will be enabling the Data Science community of learners, educators and trainers to build their own personalised learning pathways towards acquiring the skills currently requested by employers for different Data Science jobs in different geographical regions of Europe.

[13] http://edsa-project.eu/members/join-us/

7. Appendices

7.1 Course feedback form

1. What is your gender?

Female

Male

Other

2. What is your age?

< 21

21 - 30

31 - 40

41 - 50

> 50

3. Please indicate your level of agreement with the following statements:

Strongly disagree		**Neutral**		**Strongly agree**
1	**2**	**3**	**4**	**5**

a. The content of this course was easy to understand.
b. The content of this course was appropriate to my knowledge, skills and abilities.
c. The content of this course was of good quality.
d. The learning outcomes of this course were clear to me.
e. The structure of this course was easy to follow.
f. The various elements of this course, i.e. text, videos, exercise(s), were linked to each other.
g. The exercise(s) helped me understand the subject of this course.
h. The exercise(s) helped me self-assess my progress during this course.
i. The exercise(s) contained adequate instructions.
j. It was easy for me to perform the exercise(s) of this course.
k. Performing the exercise(s) of this course improved my learning experience.
l. This course allowed me to control the rate, order and process of my learning.
m. Overall, I am satisfied with the quality of my learning experience in this course.

4. Open-ended questions:

a. Please list 3 things that you enjoyed about this course.
b. Please list 3 possible improvements about this course.
c. Do you have any additional comments about this course?