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

D2.2 Data Science Curricula 2

Deliverable Editor:	**Chris Phethean (SOTON)**
	Elena Simperl (SOTON)
Other contributors:	**Module curricula prepared by partners as noted in document: ODI, FRAUNHOFER, TU/E, PERSONTYLE, JSI, KTH**
Deliverable Reviewers:	**Shatha Jaradat (KTH), Angi Voss (Fraunhofer)**
Deliverable due date:	**31/07/2016**
Submission date:	**29/07/2016**
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	03/06/2016	Chris Phethean	Structure of document and initial content.
0.2	15/06/2016	Chris Phethean	Linked Data module added
0.3	16/06/2016	Chris Phethean	Revised v2 curriculum detailed
0.4	04/07/2016	Chris Phethean	New modules for M18 added and detailed using contributions from relevant partners.
0.5	05/07/2016	Chris Phethean	Curricula revisions to M6 modules made.
0.6	07/07/2016	Chris Phethean	Skills assignment added. Conclusions written.
0.7	07/07/2016	Chris Phethean	Changes and updates in advance of review. Final content added.
0.8	08/07/2016	Chris Phethean	Final version of skills classification added.
0.9	19/07/2016	Chris Phethean	Reviewer comments addressed
0.9a	19/07/2016	Elena Simperl	Scientific Review
1.0	27/07/2016	Aneta Tumilowicz	Final QA

Table of Contents

List of Tables

1. Executive Summary

This deliverable presents the second version of the EDSA curriculum, which comprises of 15 core data science topics that are being delivered by the 3-year project. Previously, we released the first version of the curriculum containing 6 of 15 modules, each of which had a specific curriculum containing learning objectives, topic descriptions and various resources and materials that been identified. Modules are separated into four stages: foundations, data storage and processing, data analysis and data interpretation and use. These categories remain in version 2 and provide structure to our courses. We present new modules in this deliverable, reflecting a number of changes to the curriculum's list of topics, along with revisions to the previously released modules based on learner feedback, teaching experience and increased amounts of demand analysis.

2. Introduction

Before presenting the new version of the EDSA curriculum, we first revisit the initial version released in July 2015 based on preliminary demand analysis. This section recaps the main curriculum, comprised of fifteen topics, and discusses the progress made to date regarding the implementation of the modules released in Year 1 through learning resources now available on the EDSA Courses portal

2.1 Recap of curricula version 1

In D2.1 (Data Science Curricula 1), we released the first version of the EDSA curriculum. 15 core topics were presented, with the intention to deliver these in groups over the course of the 3-year project. At the time of writing (Month 6), the first 6 modules were released, each provided by a different EDSA partner. The following section outlines an overview of the curriculum (2.1.1.) and then revisits the 6 modules that were released in version 1 (2.1.2).

2.1.1 The core EDSA curriculum

We presented the 15 topics that make up the core data science curriculum for EDSA, which are reproduced below:

Table 1: Core EDSA Curriculum, version 1

Module	Topic	Stage
1	Foundations of Data Science	Foundations
2	Foundations of Big Data	Foundations
3	Statistical / Mathematical Foundations	Foundations
4	Programming / Computational Thinking (R and Python)	Foundations
5	Data Management and Curation	Storage and Processing
6	Big Data Architecture	Storage and Processing
7	Distributed Computing	Storage and Processing
8	Stream Processing	Storage and Processing
9	Machine Learning, Data Mining and Basic Analytics	Analysis
10	Big Data Analytics	Analysis
11	Process Mining	Analysis
12	Data Visualisation	Interpretation and Use
13	Visual Analytics	Interpretation and Use
14	Finding Stories in Open Data	Interpretation and Use
15	Data Exploitation including data markets and licensing	Interpretation and Use

Modules were structured and classified according to the stage or theme in the general data science process that they were most suited to (foundations, storage and processing, analysis and interpretation and use). By covering these stages equally, we ensure a broad focus that allows different audiences and stakeholders with varying needs to access our materials and find relevance and value in them.

2.1.2 Modules released in Year 1

Table 2: Topic Schedule and Partner Allocation

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

Following the schedule outlined in Table 2, curricula for the first six modules were released in M6:

- Foundations of Data Science
- Foundations of Big Data
- Big Data Architecture
- Distributed Computing
- Machine Learning, Data Mining and Basic Analytics
- Process Mining

Each curriculum contains the following details:

- Learning Objectives
- Syllabus and Topic Descriptions
- Existing Courses
- Existing Materials
- Example Quizzes and Questions
- Description of Exercises
- Further Reading

Each module is released as a syllabus which can then be used to deliver various types of courseware, in various delivery formats. This means that partners can target different audiences, sectors or levels of experience by delivering different types of courses, for example a face-to-face training course, an online MOOC, or as a set of self-study learning resources.

2.2 Delivery of learning resources

Accompanying the release of the first version of the curriculum in M6, partners then produced learning resources to release for the first six modules in M12, as outlined in D2.4. We have made these courses available on the EDSA courses portal (http://courses.edsa-project.eu), where there are a range of delivery channels and formats available:

- Self-study courses, allowing learners to study open educational resources at their own pace.
- MOOCs, hosted on external platforms such as FutureLearn and Coursera, allowing social learning in large, online courses.
- Blended courses, combining face-to-face elements with online materials, which are offered by EDSA partners and associates.
- Face-to-face courses, which are courses delivered in person by EDSA partners and associates.

Each of the M6 modules has at least a self-study course option available on the site, while courses such as TU/e's 'Process Mining' module also have further options such as MOOCs. The courses portal also includes additional courses from the project partners that are related to the EDSA curriculum.

3. Insights from demand analysis

Following the demand analysis carried out in WP1 and reported upon in D1.4, a number of insights have been gained into the curriculum which this deliverable builds upon. As summaries in D1.4, it can be argued that the EDSA curriculum presented in M06 is already generally in line with industry demands across Europe. As such, the current focus on comprehensively training the technical and analytical aspects of data science should be maintained.

Where the demand analysis revealed there was scope for innovation was through integrating this technical and analytical training into a more holistic approach to skills development, expanding to cover supporting skills that can increase the impact of data science within organisations. The recommendations from the demand analysis, initially presented in D1.4 are reproduced in Table 3. The highlighted rows (recommendations 2-4 inclusive) represent those which influence the curriculum design, some of which are reflected in this deliverable while the others will guide our revisions over the next year before the final release in M30.

Table 3: Recommendations for EDSA curriculum development, originally presented in D1.4

Title	Intervention level	Summary description
Holistic training approach	General training approach	Refine the EDSA's training approach and curriculum cycle to strengthen skills along the full data exploitation chain.
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.
Soft skills training	Expansion of curriculum	Implement soft skill training to increase performance and organisation impact of data scientists / data science teams.
Basic data literacy training	Expansion of curriculum	Develop basic data literacy and data science training for non-data scientists to improve basic skills across organisations and facilitate uptake of data-driven decision making and operations.
Blended training	Course delivery	Develop blended training approaches including sector-specific exercises and examples to increase effectiveness of training delivery.
Data science skills framework	Training approach and delivery	Implement data science skills framework to structure skills requirements, assess skills of data scientists, and identify individual skills needs.
Navigation and guidance	Training market	Develop quality assessment of third party courses; provide navigation support to identify relevant trainings from EDSA and third parties.

4. New curricula released for version 2

The following sections discuss the revisions to the curriculum based on the feedback discussed in Section 3, and any changes that have been made to the individual modules following their initial release.

4.1 Overview of version 2 release

Building on our experiences of having released the first version of the curriculum, in addition to additional studying of the data science landscape in Europe, a number of modifications have been made to the curriculum.

Linked Data

After reviewing the curriculum and key technological developments around data science, it was observed that there was no content covering linked data or semantic web technologies. Having assessed the availability of learning resources in this area, it was decided that as well as improving our covering of data management skills, this topic could be used as an initial module to test out the FutureLearn platform for delivering online MOOC courses. Using the previous EU project 'EUCLID'[1], whose consortium included a number of the EDSA partners, a further module was added to the EDSA data science curriculum, adapted from the syllabus and learning resources created in EUCLID.

The linked data module - titled 'An Introduction to Linked Data and the Semantic Web' - both fills a gap in our curriculum and allowed us to trial the running of a course on FutureLearn based on existing, high-quality, open learning materials. This meant that we were able to quickly adapt the content form the EUCLID iBook to be released as a MOOC on FutureLearn in April 2016 and provide insights into the platform to the rest of the project before moving forward with using it for the other modules. Statistics regarding the number of learners who registered on the course are presented in D3.2 (Report on delivery of videolectures, webinars and face-to-face trainings 2).

Modifications to curriculum structure

According to the original schedule, the M18 curriculum would include the following new modules:

- Statistical / Mathematical Foundations
- Data Management and Curation
- Big Data Analytics
- Data Visualisation
- Finding Stories in Open Data

After reviewing the intended aims for the 'Data Visualisation' and 'Finding Stories in Open Data' modules, a significant overlap was identified which meant that many of the concepts would be repeated across the two modules in order to convey their narrative effectively. As such, the leaders of these modules (Southampton for data visualisation, and ODI for Finding stories), collaborated on designing a merged curricula for 'Data Visualisation and Storytelling'. Southampton and the ODI will release resources in collaboration as a self-study option for this course as planned in M24, and additionally will collaborate to provide either an online course or eBook covering the entire curricula.

In addition, the 'Visual Analytics' module due to be delivered by Fraunhofer in M30 was replaced with 'Social Media Analytics', in order to teach crucial expertise on how to extract value from the ever-growing amounts of social data. This broadens our curriculum to ensure that we cover a greater range

[1] http://euclid-project.eu/

of essential data science topics, rather than focusing intently on topics where overlap would have been unavoidable.

Revised core curriculum

Given the changes discussed above, this deliverable sees the release of version 2.0 of the core EDSA curriculum, which is presented below. As the data visualisation and finding stories modules have been merged into one topic, and because we have added the linked data topic, we remain with a curriculum of 15 modules.

Table 4: Core EDSA Curriculum, version 2.0

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
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

* Module based on EUCLID curriculum, and released as FutureLearn MOOC in April 2016.

4.2 Revised modules in version 2

This section outlines changes to the topics covered in modules that have been updated since their curricula were released in M6: 'Foundations of Data Science', 'Distributed Computing', and 'Machine Learning, Data Mining and Basic Analytics'.

4.2.1 Foundations of data science (SOTON)

The syllabus for foundations of data science was revised to incorporate a small change of topics following the experiences of delivering the course as a face-to-face module within Southampton's MSc Data Science programme. Topics were also re-organised to fit within four categories that correspond to stages of the typical data science pipeline:

- **The Foundations of Data Science**
  - Introduction
    - This section covers the introduction to the course, and details the learning outcomes as originally laid out in D2.1.
  - Data Science in a nutshell
  - Terminology
  - The data science process
  - A data science toolkit
  - Types of data
  - Example applications
- **Data Collection and Management**
  - Data collection as the first stage of data science
  - Where data comes from
  - Pre-processing
  - Cleaning and filling in missing data
    - Fixing data
    - Noisy data
  - Data integration
  - Using the cloud and principles of cloud computing
  - Document related storage
    - MongoDB
- **Fundamentals of Statistics and Data Analysis**
  - Introduction to data mining
    - Disciplinary components
    - Machine learning and statistics
    - Applications of data mining
      - Real-world examples
  - Introduction to statistics
    - Nature of statistics and introduction
    - Central tendencies and distributions
    - Variance
    - Distribution Properties and arithmetic
    - Samples/CLT
    - Linear regression
  - Using the R statistical package
  - Introduction to Machine Learning
    - What is machine learning? What can it do?
      - Example applications
    - Supervised v unsupervised learning
    - Prerequisites for Machine Learning
    - Regression problems
    - Classification problems
      - Linearly separable v non linearly separable
      - Naive Bayes
      - Support Vector Machine
- **Data Visualisation**
  - Introduction to data visualisation
  - Types of data visualisation
    - Exploratory
    - Explanatory
  - Data for visualisation
    - Data types
    - Data encodings
    - Retinal variables

        - Mapping variables to encodings
        - Visual encodings
    - Chartjunk and the beauty paradox
    - Technologies for visualisation
        - Web Technologies and their role in visualisation
            - Hypertext
            - CSS
            - Javascript
        - D3.js
            - D3 chain syntax
            - D3 Libraries
            - Alternatives to D3: Python, high-level tools etc.
    - Data-driven journalism and Storytelling

4.2.2 Essentials of Data Analytics and Machine Learning (Persontyle)

This module is a renamed and revised version of Persontyle's 'Machine Learning, Data Mining and Basic Analytics' module that launched in M6, based on the experiences of developing the learning materials and delivering the course. The list of topics and their associated descriptions have been modified as follows:

- **Lesson 1: R REFRESHER**
  - This module provides an overview of the R programming language. It is intended to be much more than a quick introduction or refresher in R. It is designed to function as a source that you can reference when working with R on the examples and exercises in the remainder of the guide. It contains not only a lot of information about the R programming language, but also general information that is important to know when programming in any language.
- **Lesson 2: FEATURE INITIALIZATION**
  - We begin our pre-processing modules by considering the initial steps that must be taken to transform the raw data, which has been collected to a set of features which are capable of being examined and transformed in the further pre-processing steps, and eventually of being used in statistical modelling.
- **Lesson 3: FEATURE SELECTION**
  - Once quantitative data has been generated we should examine the features that we have available. In machine learning tasks, we often have access to far more features than we can hope to use, and an important sub-problem is selecting an information rich subset of features to use to build our statistical models. It is important to note that more features is not necessarily better: features may or may not contain information about the target variable, but they will certainly contain noise. Further, working with additional variables tends to entail additional parameters in our models, which can increase the potential for overfitting.
- **Lesson 4: FEATURE TRANSFORMATION**
  - Feature transformation is the name given to replacing our original features with functions of these features. The functions can either be of individual features, or of groups of them. The result of these functions is itself a random variable – by definition the function of any random variable is itself a random variable. Utilizing feature transformations is equivalent to changing the bases of our feature space, and it is done for the same reason that we change bases in calculus: We hope that the new bases will be easier to work with.
- **Lesson 5: MISSING DATA**
  - This module contains many references to later modules. This is because imputing missing data generally requires the use of statistical models with which to do the imputation. It is included here as it is part of pre-processing, but readers may wish to return to this module after reading the statistical modeling modules.
- **Lesson 6: BASIC REGRESSION MODELS**
  - In this module we will look at some of the simplest regression models: What they are, how they work, how they can be built from data and how they can be evaluated.
- **Lesson 7: MODEL EVALUATION, SELECTION AND REGULARIZATION**
  - When creating, selecting and evaluating statistical models, you should think in terms of a three step process:
    - Training: Train models using dedicated training data.
    - Evaluation/Validation: Select the best performing model, either using dedicated validation data or statistical methods. We will call this step evaluation and reserve the term validation for the evaluation/testing technique discussed in the following sections.

        - Testing: Obtain unbiased estimates of the performance of the selected model using dedicated test data or statistical methods.
    - The aim of this module is to clarify why these three steps are required and provide you with the means of performing the second and third steps.

- **Lesson 8: BASIC CLASSIFICATION METHODS**
    - In this module we look at a number of basic classification methods. We first look at cases where the independent variables are real, and examine the following algorithms:
        - Linear and Quadratic Discriminant Analysis
        - The Perceptron Algorithm
        - Logistic
    - We then look at the case where both the independent and target variables are discrete, which we will term discrete-discrete classification.

- **Lesson 9: ANALYSIS OF CLASSIFICATION MODELS**
    - We introduced misclassification error in module 8, where we also discussed the use of mean squared error for probability estimates. These are the most common error scores used for fitting parameters. You should note that misclassification is the simple accuracy statistic discussed in the next section, and hence shares its problem.

- **Lesson 10: LOCAL/NON-PARAMETRIC METHODS**
    - Local methods are so-named because when estimating a new case they work with a subset of the training data found to be near to the new case, using some distance metric or spatial division.

- **Lesson 11: BASIC KERNEL METHODS**
    - The order of topics is as follows: We will provide a basic overview of the kernel trick. We then list some common kernels, followed by an example of an application of the kernel trick with kernel regression. We then provide background of the theory behind the kernel trick, which delves into the quite complex mathematics and which may be skipped by the intimidated reader. Finally, we examine kernel PCA.

- **Lesson 12: SPLINES**
    - Splines are a clever development of the more basic regression techniques that can be understood as implementing two ideas:
        - Instead of generating a single model for the entire feature space, we can divide it into different sections and model each of these with an individual model.
        - To combine these models elegantly and plausibly we will require that the regression lines meet with a degree of smoothness at the section boundaries.

- **Lesson 13: GAUSSIAN PROCESSES** (New lesson to be delivered in August 2016)
    - A stochastic process is a collection of random variables, often representing the indeterminate evolution of some system over time.
    - Stochastic processes are:
        - Infinite variate extensions of multivariate distributions
        - Models of stochastic evolution of systems over time
        - Distributions over functions

- **Lesson 14: TREES & ENSEMBLE METHODS** (New lesson to be delivered in August 2016)
  - In this module we introduce decision tree based methods and use them to examine a number of ensemble methods. This combination of topics is not accidental – tree models are almost always used as part of ensembles, and the most popular ensemble methods are tree based.
  - Tree based methods are sometimes described as radically non-linear. In fact, they perform a form of discretization, where the feature space is divided into regions and a value estimated for the target variable in each region.
- **Lesson 15: SUPPORT VECTOR MACHINES** (New lesson to be delivered in August 2016)
  - In this module we introduce another well-known and very high performing machine learning technique: support vector machines (SVMs). We will explain SVMs in terms of solving the optimization problem for linear classifiers known as finding the maximally separating hyperplane when classes are inseparable. We examine the basic case, known as support vector classifiers, before looking at how we can obtain a more powerful procedure by transforming the variables we work with using kernels.
- **Lesson 16: NEURAL NETWORKS AND DEEP LEARNING**
  - Artificial neural networks (ANNs or just NNs) are perhaps the most famous machine learning technique. They have ridden repeated waves of hype, and disappointment, since first being formulated in the 1940s. With the development of deep neural networks (DNNs) they have proven to perform very well on a large number of tasks and we are currently in another period of strong interest in neural network techniques. This interest can be negative – too many students and data scientists have been caught up in the current hype cycle and place undue interest in what is merely one tool among many. Nonetheless, they are a powerful tool and deserve a place in any data scientist's arsenal.

4.2.3 Distributed computing (KTH)

KTH revised the syllabus for distributed computing based on experiences of teaching courses based on the M6 version. The existing topic list remained the same, however a third part was added to cover new and further topics:

Part I: Distributed Services and Distributed Algorithms

- Unchanged from D2.1

Part II: Clouds, Peer-to-Peer and Big Data

- Unchanged from D2.1

Part III: Follow-up courses in the Distributed Computing module

A. Distributed Artificial Intelligence and Intelligent Agents

- **Introduction to Agents**
  - This part introduces the notion of intelligent agent and considers individual and group perspectives of agent technology
- **Negotiation**
  - In this part will consider negotiation principles, negotiation protocols (voting, monotonic concession, Clark tax, auctions) and negotiation mechanisms (organizational structures, meta-level information exchange, multi-agent planning)
- **Coordination**
  - This part introduces coordination principles and considers coordination mechanisms (organizational structures, meta-level information exchange, multi-agent planning, norms and lows and cooperative work)
- **Interoperability**
  - Here we consider approaches to software interoperation, speech acts and Agent Communication Languages (ACL), KQML, FIPA.
- **Multi-Agent Systems architectures**
  - In this part we discuss low-level architecture support, DAI testbeds, Middle-agent based architectures and agent marketplaces
- **Software Engineering of Multi-Agent Systems**
  - This part considers basic approaches to agent systems engineering, as wee as GAIA and agent-UML approaches
- **Agent theory**
  - This part introduces agent theory, basics of modal logic and BDI-architecture
- **Agent architectures**
  - Here we discuss deliberative, reactive and hybrid agent architecture
- **Agent mobility**
  - We consider requirements, implementation and security for mobile agents as well as environments for mobile agents.

B. Programming web Services

- **Introduction**
  - This considers historical perspective of service technology and basic concepts of Web services.
- **Markup languages**
  - Here basics of markup languages and XML are considered.
- **XML messaging**
  - This considers approaches to XML messaging as well as SOAP and JSON
- **Service description**

- Here functional and non-functional components of service description are considered together with Web service description language.

- **Service discovery**
  - This considers approaches to service discovery (registry, index and peer-to-peer) and specifications for service discovery
- **Services coordination**
  - Basic components of the service coordination environments are considered together with specifications. Atomic transaction and business activity are discussed.
- **Service composition**
  - Here basics and approaches to service composition are considered, language for service composition is presented
- **Services security**
  - Basic elements of Web service security are presented, main security specifications are discussed
- **Web services and stateful resources**
  - Approaches and specifications for description of stateful resources in Web services are considered
- **Semantic Web Services**
  - Here basics of semantic Web and its specification for semantic Web services are discussed

5. New modules in version 2

Following the revisions detailed above, we progressed with designing curricula for the 4 modules in M18 (with data visualisation and storytelling comprised of both the original Data Visualisation module and the Finding Stories in Open Data module, as discussed above). As in D2.1, we provide a summary of each module's content, and a list of the materials that already exist, in preparation for the learning resources to be developed and released for M24.

5.1 Statistical and mathematical foundations – JSI

5.1.1 Learning Objectives

By the end of this module, learners will:

- Be familiar with concepts, definitions and methods from the field of mathematics and statistics.
- Have knowledge on the use of the mathematical and statistical concepts in Data Science.
- Have gained experience on mathematical and statistical methods with R and QMiner tools.

5.1.2 Syllabus and Topic Descriptions

- **Introduction to Mathematics**
  - This topic describes the notion of Mathematics course, provides syllabus, aims for the course and recommendations for the learners.

- **Foundations of Mathematics**
  - This topic gives the basic definitions from the field of Mathematics, such as arithmetic and algebra concepts, linear equations, quadratic equations, functions, differentiation and integration.

- **Linear Algebra**
  - This topic provides a number of concepts from the field of Linear Algebra, such as vector spaces, orthogonality, dot product, norm, matrices, determinant, matrix decompositions.

- **Optimization**
  - This topic gives insights into the field of Optimization and definitions related to optimization, such as minimum, maximum, saddle points, convex functions, primal and dual problems, Lagrange multipliers.

- **Introduction to Statistics**
  - This topic describes the notion of Statistics course, provides syllabus, aims for the course and recommendations to the learners.

- **Foundations of Statistics**
  - This topic provides the basic definitions from the field of Statistics, such as probability, random experiments, definition of probability, conditional probability, independence, interpretation of probability, Bayes' theorem.
  - The learners will get familiar with random variables, probability density functions, expectation and expectation values.
  - The topic also covers probability distribution, discrete/continuous distributions, sampling and sampling distributions.

- **Exploratory Data Analysis**
  - This topic provides a view on Exploratory Data Analysis and its methods. In particular, the topic covers descriptive/summary statistics – central tendency (mean, median, mode), variability (standard deviation and variance), skew, kurtosis, histograms, frequency polygons, box-plots, quartiles, scatter plots, heat maps.

- **Regression and Inferential Statistics**
  - This topic describes the notion of Regression and such important concepts, as covariance and correlation.
  - The learners will learn about the null hypothesis significance tests (NHST) and confidence intervals.

- **Data analytics with R**
  - This topic will provide a set of practical examples from the area of Mathematics and Statistics using R - a programming language and software environment for statistical computing and graphics.

- **Data analytics with QMiner**
  - This topic provides some practical insights on mathematical implementations and statistical data analysis using QMiner. QMiner implements a comprehensive set of techniques for supervised, unsupervised and active learning on streams of data.

5.1.3 Existing Courses

Table 5: Existing courses that include a module similar to Statistical and Mathematical Foundations

Institution	Course	Module	URL	Target Audience	Languages and Data Types
JSI	Data Analytics with QMiner	Data analytics with QMiner	http://videolectures.net/sikdd2014_rei_large_scale/?q=qminer	Computer scientists, mathematicians, statisticians.	QMiner
Princeton University	Statistics One MOOC (Coursera)	Statistics One	https://www.coursera.org/course/stats1	Computer scientists, mathematicians, statisticians, other.	R

5.1.4 Existing Materials

A number of relevant materials have been identified on Videolectures:

- Statistical Methods http://videolectures.net/acai05_taylor_sm/?q=statistics
- Introduction to Statistics http://videolectures.net/cernstudentsummerschool2010_cowan_statistics/?q=statistics
- Lecture 15: Statistical Thinking http://videolectures.net/mit600SCs2011_guttag_lec15/?q=statistics
- Introduction to Statistics http://videolectures.net/cernstudentsummerschool09_cowan_is/?q=statistics

- On the Computational and Statistical Interface and "BIG DATA" http://videolectures.net/colt2014_jordan_bigdata/?q=foundations%20of%20statistics
- Probability and Mathematical Needs http://videolectures.net/bootcamp2010_anthoine_pmn/

Additionally, material has been identified in the following resources:

- James Ward and James Abdey. Foundation course: Mathematics and Statistics http://www.londoninternational.ac.uk/sites/default/files/programme_resources/ifp/fp0001_maths_and_stats_taster_2013.pdf
- Master Statistics with R https://www.coursera.org/specializations/statistics
- Basic Statistics https://www.coursera.org/learn/basic-statistics
- Inferential Statistics https://www.coursera.org/learn/inferential-statistics-intro
- Foundations of Data Analysis - Part 1: Statistics Using R https://www.edx.org/course/foundations-data-analysis-part-1-utaustinx-ut-7-10x

5.1.5 Descriptions of Exercises

Exercises will be designed to allow students to carry out data analytics with the QMiner tool: http://qminer.ijs.si/

5.1.6 Further Reading

- G. Cowan, Statistical Data Analysis, Clarendon, Oxford, 1998 see also www.pp.rhul.ac.uk/~cowan/sda
- R.J. Barlow, Statistics, A Guide to the Use of Statistical in the Physical Sciences, Wiley, 1989 see also hepwww.ph.man.ac.uk/~roger/book.html
- L. Lyons, Statistics for Nuclear and Particle Physics, CUP, 1986
- F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006; (W. Eadie et al., 1971).
- S. Brandt, Statistical and Computational Methods in Data Analysis, Springer, New York, 1998 (with program library on CD)
- W.-M. Yao et al. (Particle Data Group), Review of Particle Physics,
- J. Physics G 33 (2006) 1; see also pdg.lbl.gov sections on probability statistics, Monte Carlo

5.2 Data management and curation - TU/E

5.2.1 Learning Objectives

The objective of this module is to introduce and explain concepts and practices for the management of research data assets, including 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.

5.2.2 Syllabus and Topic Descriptions

- **Data and data management**
  - **Research data explained**

This unit aims at being aware of different types of research data, and understanding general ideas about big data useful for the purpose of this course.

  - **Data management plans**

This unit aims at explaining what data management is, and how a data management plan can be designed.

- **Organization and documentation**
  - **Organization of data**

This unit aims at introducing the concepts of data organization, as well as file versioning, naming conventions, and best practices.

  - **File formats and transformation**

This unit presents data formatting, compression, and other transformations.

  - **Documentation, metadata, and citation**

This unit explains the use of documenting data, as well as metadata.

- **Data and data management**
  - **Storage and security**

This unit introduces concepts of safety of storage and safety guidelines. It also explains data encryption, and destroying sensitive data.

  - **Data protection, rights, and access**

This unit aims at describing data protection, rights, and access, from privacy to intellectual property rights.

  - **Sharing, preservation, and licensing**

This unit presents the practical implications of sharing data, as well as licensing, in particular through different kinds of open data licences.

5.2.3 Existing Courses

- Introduction to Research Data Management and Sharing, free course on Coursera, by The University of North Carolina at Chapel Hill and The University of Edinburgh.
  - https://www.coursera.org/learn/data-management
- Data Management, distance learning course by The Open University.
  - http://www.open.ac.uk/postgraduate/modules/m816
- Digital Curation, tutorial with extensive learning materials online, by the Digital Curation Centre.
  - http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials
- New England Collaborative Data Management Curriculum
  - http://library.umassmed.edu/necdmc/modules

5.2.4 Existing Materials

- Research Data Management Training, free course called "MANTRA", by the University of Edinburgh.
  - http://datalib.edina.ac.uk/mantra/
- New England Collaborative Data Management Curriculum
  - http://library.umassmed.edu/necdmc/modules
- Digital Curation, tutorial with extensive learning materials online, by the Digital Curation Centre.
  - http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials

5.2.5 Example Quizzes and Questions

Quizzes can include questions that test the understanding of concepts, terminology, regulations, and methods.

5.2.6 Description of Exercises

- Provide rules that implement a sample policy.
- Provide a description of an infrastructure responding to a sample situation.
- Evaluate the opportunity to preserve some data, in a sample situation..

5.2.7 Further Reading

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

5.3 Big data analytics - Fraunhofer

This module was developed 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.

5.3.1 Learning Objectives

Learners with experience in programming, data analytics and the architecture of big data systems get to know methods and tools to analyze big data. They will be able to realistically assess the application of big data analytics technologies for different usage scenarios and start with their own experiments.

5.3.2 Syllabus and Topic Descriptions

- 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.
- 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.
- 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.
- 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.

- 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.
- 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.
- 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.

5.3.3 Existing Courses

- Big data analytics, a face-to-face training offered by Fraunhofer IAIS http://www.iais.fraunhofer.de/bigdataanalytics.html

5.3.4 Existing Materials

- Data Mining and Applications Graduate Certificate at Stanford Center for Professional Development, http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=1209602
- Machine Learning With Big Data at Coursera https://www.coursera.org/specializations/bigdata
- Mining Massive Data Sets, Stanford University http://web.stanford.edu/class/cs246/
- DATA SCIENCE AND BIG DATA ANALYTICS by EMC https://education.emc.com/guest/campaign/data_science.aspx
- Spark Fundamentals I by BigData University http://bigdatauniversity.com/courses/spark-fundamentals/

5.3.5 Example Quizzes and Questions

Multiple choice questions concerning

- The Jaccard distance to build the utility matrix for collaborative filtering
- Locating collaborative filtering in the lambda architecture
- The role of classical data analytics for big data
- The countdistinctSketch algorithm
- The alternating least squares algorithm
- The parallelizatioin of k-means clustering
- Big data analytics in Spark
- In-memory processing

5.3.6 Further Reading

- http://github.com/ishkin/Proton/tree/master/IBM%20Proactive%20Technology%20Online%20on%20STORM
- http://www.teradata.de/products-and-services/analytics-from-aster-overview
- http://en.wikipedia.org/wiki/List_of_in-memory_databases
- http://s.fhg.de/in-memory-systems

5.4 Data visualisation and storytelling - SOTON/ODI

As discussed above, this module is a combination of the 'Data visualisation' and 'Finding stories in open data' modules that were listed in the first version of the curriculum. We present here an overall curriculum for the new module that combines the topics of the original modules.

5.4.1 Learning Objectives

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

- Explain the role of data visualisation in data science
- Describe why particular visualisations are either good or bad
- Design visualisations that tell a coherent story
- Explain how the brain perceives visual information
- Create static and interactive visualisations

5.4.2 Syllabus and Topic Descriptions

The main syllabus is separated into four 'stages': fundamentals, planning, storytelling and producing visualisations.

Fundamentals

- Importance of visualisation within data science: this section discusses the current data situation and how the volume of data means there is a growing need for actionable insights that can be revealed through visualisation.
- Dashboards and infographics: we will review the rise of dashboards and infographics and how these are used to tell coherent stories about data
- Origin of graphs: this covers a brief history of visualisation, looking at why certain graphs or charts were introduced in order to make sense of growing amounts of data.
- European and local laws: this topic will cover how law is used to protect freedom of speech within journalism
- Ethics in communicating stories: what are the ethical considerations required when content can be shared rapidly to proliferate across the Web?

Planning

- Audience-Story-Action - this section will cover three key considerations that a designer must make before creating a visualisation
- Headline writing and story production - looking at how to scope a data journalism project, we will cover how to effectively communicate insight based on a thorough understanding of your audience
- 80:20 rule - how this applies in data journalism and how you can reduce time spent on early stages of the journalism process.

Storytelling

- Visual perception and information design: we will look at the cognitive process behind how the brain interprets visuals, including psychological theories such as: gestalt theory, pre-attentive processing, elementary perceptual tasks. We will show how an understanding of these allows us to notice how the brain can be tricked, so that bad design can be avoided.
- Colour and interaction to draw attention to data
- Chartjunk and bad design - we will explore Tufte's theories of minimalism and chartjunk as guidelines for producing effective graphics.

Producing Visualisations

- Obtaining data: this topic looks at how to obtain the data required to tell a compelling story, and what datasets are required to tell different types of story.
- Cleaning data: this looks at how to prepare data for analysis once it has been gathered, beginning with the essential process of data cleansing
    - Identifying outliers, fix typography, correct missing data

  - Enriching data
- Types of chart: what charts can be used for what data, and which stories can these tell?
- Building visualisations with D3: This topic looks at the practical ways to visualise data to tell a story. Technologies covered include HTML5, D3.js and DC.js to produce compelling online graphics.

5.4.3 Existing Courses

- Data Visualization and D3.js MOOC (Udacity) - Zipfian Academy
  - https://www.udacity.com/course/data-visualization-and-d3js--ud507
- Big Data: Data Visualisation MOOC (Futurelearn) - Queensland University of Technology
  - https://www.futurelearn.com/partners/queensland-university-of-technology
- Data visualisation one-day workshop (The Guardian)
  - https://www.theguardian.com/guardian-masterclasses/2015/aug/07/data-visualisation-a-one-day-workshop-tobias-sturt-adam-frost-digital-course
- Data visualisation (module on Master's degree on Data Science) - Barcelona Graduate School of Economics
  - http://www.barcelonagse.eu/sites/default/files/course-d007-data-visualization.pdf
- Data Analysis, Visualisation and Communication (MSc course) - University of Aberdeen
  - http://www.abdn.ac.uk/study/postgraduate-taught/degree-programmes/46/data-analysis-visualisation-and-communication/

5.4.4 Existing Materials

- Finding Stories in Data - Course Materials - http://training.theodi.org/FindingStories/

Additional resources will be developed by both SOTON and ODI.

5.4.5 Example Quizzes and Questions

Quizzes will be in the form of multiple-choice questions to test knowledge of:

- The importance of data visualisation and storytelling
- How should particular data be encoded into visual dimensions
- The stages of planning a visualisation
- Chartjunk and minimalism in visualisation design

5.4.6 Description of Exercises

- Visualising data - students will guided through how to create an online visualisation, with participants discovering the challenges involved in the process.

5.4.7 Further Reading

- Data Journalism Handbook - http://datajournalismhandbook.org/
- Cairo, A. The Functional Art
- Tufte, E. The visual display of quantitative information
- Few, S. Show me the numbers

5.5 Linked Data and the Semantic Web - SOTON

This module is a new addition to the curriculum and fulfils a gap in the previous version regarding the coverage of semantic technologies and data management.

5.5.1 Learning Objectives

The module is intended for anyone wanting to gain a basic understanding of linked data, before seeking out further study opportunities, or as part of their existing training. It will appeal to many computer science, web science, or data science graduates, students and practitioners wishing to expand their knowledge.

Linked data allows the structured publication on the Web which facilitates the exposure of and interlinking of datasets so that data may be exchanged, reused and integrated. Increasing numbers of organisations and initiatives are starting to appreciate the power of linked data, and it provides significant power and potential to the growing domain of data science and web science.

This module will expose you to the knowledge and skills around using linked data – skills that are in increasing demand as the Web evolves from a Web of Documents to a Web of Data. It will help you understand how you can use linked data technologies and how to write queries in SPARQL, allowing you to exploit these technologies in your own work.

By the end of this primer in Linked Data, learners will be able to:

- Describe various background technologies and standards behind linked data;
- Describe the role of linked data in data science applications
- Understand how linked data applications are designed and used
- Use SPARQL to process linked data.

5.5.2 Syllabus and Topic Descriptions

- **Week One – Linked Data fundamentals.**
  - By the end of this week you will be able to:
    - Reflect on why linked data is important for structuring the Web of Data
    - Summarise the background technologies behind linked data
    - Outline the various background standards behind linked data
    - Explain what linked data is
    - Summarise the principles behind linked data
    - Explain what 5-star linked open data means
    - Evaluate different linked data tools
  - This part of the course covers the basic topics required to understand linked data and the semantic web. It begins with reviewing the technologies and standards behind the Web itself and how these are evolving to facilitate linked data. Students will gain an understanding of what linked data is, before we cover a number of principles behind it and the 5-star model of rating data openness.

- **Week Two – SPARQL Queries.**
  - By the end of this week you will be able to:
    - Describe the basic concepts of SPARQL
    - Recall the definitions of SPARQL terminology
    - Apply knowledge of SPARQL to formulate SPARQL queries
    - Detect different types of SPARQL queries
    - Construct SPARQL queries
  - This section of the course introduces the learners to SPARQL, the language for querying the semantic web. We cover general principles such as how to structure a query, and provide definitions for the required terminology, before introducing basic queries, such as ASK and more predominantly SELECT.

- **Week Three – More SPARQL Queries.**
  - By the end of this week you will be able to:
    - Design more advanced SPARQL queries
    - Detect more types of advanced SPARQL query
    - Describe the practical applications for SPARQL and Linked Data

  - The third part of the course moves on to more advanced SPARQL queries, particularly DESCRIBE and CONSTRUCT. We cover the essential syntax and structure of these queries, building on previously gained knowledge from earlier in the course about querying linked data. Finally we cover the various applications of linked data and how this technology can be used to develop rich tools that provide increased value through the linking of data.

5.5.3 Existing Courses

- This course takes inspiration from the EUCLID project's iBook which is openly available and contains quizzes and exercises to test knowledge in a self-study format.

5.5.4 Existing Materials

- The EUCLID iBook is available to download from the Apple iBooks store, and from the EU project's website at http://euclid-project.eu/

5.5.5 Example Quizzes and Questions

Quizzes test the students' knowledge of various concepts such as:

- The semantic web technology stack
- RDF graphs
- SPARQL query results

5.5.6 Description of Exercises

Interactive exercises will provide students with hands-on experience of using linked data and SPARQL. Two variations are provided: the first allows the student to write their own RDF statements and then use SPARQL to verify that it is correctly formed, while the second use a live SPARQL endpoint to provide a more complex dataset on which the queries outlined in the course can be tested and experimented with.

5.5.7 Further Reading

- The EUCLID project iBook:
  - Simperl, E., Norton, B., Acosta, M., Maleshkova, M., Domingue, J., Mikroyannidis, A., Mulholland, P. and Power, R., 2013. Using linked data effectively.
- T. Berners-Lee, J. Hendler and O. Lassila (2001) "The Semantic Web". Scientific American vol. 284 number 5, pp 34-43. Available on-line at http://www.scientificamerican.com/article.cfm?id=the-semantic-web.
- P. Hitzler, M. Krötzsch, and S. Rudolph (2010). "Query Langauges: Foundations of Semantic Web Technologies". CRC Press

6. Data science skill groups

In Table 3 above, we present insights from the demand analysis carried out in WP1. One of these relates to the recommendation that EDSA should develop a data science skills framework in order to structure the skills requirements of particular jobs. Additionally, this would allow data scientists themselves to assess their skills profiles, compare them with those required for data jobs, and identify areas in which they need to seek further training. Linking with the WP1 demand dashboard would facilitate the automatic search of relevant training courses for a job that a user may find which they need to skill for.

In this section we present an initial matching of modules to skills, based on the curricula released in this and the previous deliverable. For the purposes of this skills assignment, we use a list of skills that has been used throughout the EDSA project. The assignment of skills to modules is shown in Table 6.

In addition, additional skills that were not in the current vocabulary and which related to particular tools or technologies were extracted from the curricula in order to enhance the accuracy of this mapping:

- Foundations of Big Data
    - QMiner
    - MapReduce
- Big Data Architecture
    - MapReduce
    - Apache Storm
    - Cassandra
    - Hadoop
- Distributed Computing
    - Hadoop
- Essentials of Data Analytics and Machine Learning
    - R
- Process Mining
    - ProM
- Statistical and Mathematical Foundations
    - R
    - QMiner
- Big Data Analytics
    - Hadoop
    - R
    - Apache Storm
    - Spark
    - Python
    - MapReduce

Table 6: Skills that each module curricula targets

(continued on next page)

		Foundations of Data Science	*Foundations of Big Data*	*Big Data Architecture*	*Distributed Computing*	*Essentials of Data Analytics and Machine Learning*	*Process Mining TU/E*	*Statistical and mathematical foundations*	*Data management and curation*	*Big data analytics*	*Data visualisation and storytelling*	*Linked Data and the Semantic Web*	*Total Coverage*
advanced computing	**python**												0
	advanced computing			x	X					X			3
	programming					X		X		X			3
	computational systems			x	X								2
	coding					X		X					2
	cloud computing			X	X								2
data skills	**databases**			X						X			2
	data management	X		X					X			X	4
	data engineering												0
	data mining		X				X			x			3
	data formats								X				1
	linked data											X	1
	information extraction						X				X		2
	stream processing			x						X			2
domain expertise	**enterprise process**												0
	business intelligence						X				X		2
	data anonymisation								X				1
	semantics												0
	schema												0
	data licensing								X				1
	data quality												0
	data governance								X				1
general	**data science**	X								x			2
	big data		X	X	X					X			4
	open data												0

machine learning	**machine learning**			x		X				X			3
	social network analysis												0
	inference											X	1
	reasoning											X	1
	process mining						X						1
maths & statistics	**linear algebra**					X		X		X			3
	calculus							X					1
	mathematics					X		X					2
	statistics	X				X		X					3
	probability							X					1
	Rstudio												0
	data analytics		X			X				X			3
	data analysis					X		X		X			3
visualisation	**data visualisation**	X									X		2
	infographics										X		1
	data mapping										X		1
	data stories										X		1
	data journalism										X		1
	d3js										X		1
	tableau												0

This mapping represents an initial categorisation of the modules in our curriculum based on the skills that they train. In the coming months, further refinement will be made – in particular the back-end work of the dashboard is also producing an updated list of skills based on their increasing occurrence in European job posts over time. Using this will provide a more granular and accurate assessment of the exact skills covered in each module.

Additionally, the mapping is currently at a curriculum level, indicating which topics are covered by the syllabus of each module. Further complexity arises when different instances or delivery mechanisms of each course are considered: for example a partner may produce an online course for their module and in doing so base it around a specific language or tool, whereas for a face-to-face offering they might use a combination of tools instead. This requires a classification not only at the curriculum level, but at the course level, too.

7. Conclusion

Following the release of the first version of the EDSA curriculum in M6, we have now since had time to develop and run a number of EDSA modules and receive both student and teacher feedback from these. Additionally, insights from our demand analysis are beginning to shape changes to the modules and curriculum structure itself, ensuring that the EDSA curriculum remains an integral resource for data scientists to discover new opportunities for training and work.

This deliverable has reviewed those changes made to date, as well as presenting the next group of modules from the curriculum to be released. At M30, we will present the final version of the curriculum which will include fewer brand new modules, and instead will focus on bringing all the insights from our training delivery, demand analysis and learning analytics together. Each module and associated course will be tagged with the skills which a student will learn. Courses released at M24 and M36 will therefore relate directly to these skills to join up the demand analysis work with the course delivery offerings. A perspective data science worker will therefore be able to locate jobs that interest them, identify their own skills shortages, and then locate training materials in order to help them meet the requirements of the job post.