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

D2.6 Learning resources 3

Deliverable Editor:	**Alexander Mikroyannidis (OU)**
Other contributors:	**Huw Fryer (SOTON), Shatha Jaradat (KTH), Mihhail Matskin (KTH), Angi Voss (Fraunhofer), Ryan Goodman (ODI), David Tarrant (ODI), Emily Vacher (ODI), Gillian Whitworth (ODI)**
Deliverable Reviewers:	**Inna Koval (JSI), Angi Voss (Fraunhofer)**
Deliverable due date:	**31/01/2018**
Submission date:	**24/01/2018**
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	14/11/2017	Alexander Mikroyannidis	Outline and responsibilities of contributors.
0.2	10/01/2018	Alexander Mikroyannidis	Version for internal review.
0.3	23/01/2018	Alexander Mikroyannidis	Revised version.
1.0	24/01/2018	Alexander Mikroyannidis	Final QA.

Table of Contents

List of Tables

List of Figures

1. Executive summary

This deliverable presents the modules that have been added to the EDSA courses portfolio during the third year (Y3) of the project. According to the final EDSA curriculum presented in D2.3 (M30), the following 4 modules were scheduled for release during Y3:

- Programming / Computational Thinking (R and Python)
- Data Intensive Computing
- Social Media Analytics
- Data Exploitation including data markets and licensing

The project has been focused on bridging the data science skills gap via a supply of training materials. As a result, the EDSA courses portfolio includes a wide range of courses offered by renowned educational institutions both inside and outside the project consortium. These courses are selected based on their relevance to the EDSA curriculum and the EDSA demand analysis.

This deliverable presents the learning resources that have been added to the project's courses portfolio in Y3, in order to cover the topics of the final EDSA curriculum and address the current demand as identified by the EDSA demand analysis.

2. Introduction

The EDSA courses portfolio incorporates a wide range of high quality learning resources, either offered by project partners or by third parties. These courses are available as:

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

The main criteria for the selection of courses for inclusion in the EDSA courses portfolio are the *EDSA curriculum* and the *EDSA demand analysis*. Courses are selected based on their potential of addressing the EDSA curriculum topics as well as the training needs of data scientists as identified by the EDSA demand analysis. With regards to compliance with the EDSA demand analysis, courses are evaluated against the recommendations of the Study Evaluation Report (D1.4). With regards to compliance with the EDSA curriculum, courses are evaluated against the latest version of the curriculum and the topics it addresses.

The final version of the EDSA curriculum, together with a discussion about its objectives and how these have been achieved, has been presented in D2.3 and is shown in Table 1. In this table, modules are grouped by the stage they belong to. The modules scheduled for release in Y3 are highlighted and are the following:

1. Programming / Computational Thinking (R and Python)
2. Data Intensive Computing
3. Social Media Analytics
4. Data Exploitation including data markets and licensing

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

Table 1: The final EDSA curriculum.

Topic	**Stage**	**Schedule**	**Allocated Partner**
Foundations of Data Science	Foundations	M6	SOTON
Foundations of Big Data	Foundations	M6	JSI
Statistical / Mathematical Foundations	Foundations	M18	JSI
Programming / Computational Thinking (R and Python)	**Foundations**	**M30**	**SOTON**
Big Data Architecture	Storage and Processing	M6	Fraunhofer
Distributed Computing	Storage and Processing	M6	KTH

Data Management and Curation	Storage and Processing	M18	TU/e
Linked Data and the Semantic Web	Storage and Processing	M18	SOTON
Data Intensive Computing	**Storage and Processing**	**M30**	**KTH**
Machine Learning, Data Mining and Basic Analytics	Analysis	M6	Persontyle
Process Mining	Analysis	M6	TU/e
Big Data Analytics	Analysis	M18	Fraunhofer
Data Visualisation and Storytelling	Interpretation and Use	M18	ODI
Social Media Analytics	**Interpretation and Use**	**M30**	**Fraunhofer**
Data Exploitation including data markets and licensing	**Interpretation and Use**	**M30**	**ODI**

The remainder of this deliverable presents the modules that have been incorporated into the EDSA courses portfolio in Y3, based on their relevance to the EDSA curriculum and the EDSA demand analysis. For each module, we present an overview of its objectives and learning outcomes, we identify the types of learning materials used and how these are delivered, and we discuss how each module is related to the EDSA curriculum and the demand analysis, as well as the plans in place for their further development.

3. Programming / Computational Thinking (R and Python)

3.1 Module overview

This is a course about getting people to think like computer scientists. Many people think computer science is only about writing code. Whilst this is a part of it, in this course we aim to show more of the processes behind writing the code, adding the code as an afterthought. We base the course on one of our modules at the University of Southampton, which we use to get people from backgrounds other than computer science up to speed on our Web Science programmes. We also use the work of Wing (2006)[1] about computational thinking.

In addition to being able to focus on thinking like a computer scientist, we also wish to introduce learners to the practical skills they need to put this thinking into practice. As such, each week we have included instructions about using the Python programming language to apply core computational and programming concepts.

This course is an exception when compared to the other courses in the EDSA curriculum, in that it is a foundational course which provides background to complete the remainder of the data science courses. As such, it is slightly further away from what might be considered a "core" of data science.

3.1.1 Learning Objectives

After completion of this course, participants will be able to:

1. Describe how computing is possible through a series of abstractions
2. Assess the most effective algorithm for a particular task
3. Design simple algorithms for common computing tasks
4. Retrieve data stored from a Web API, and store it in a relational database
5. Develop simple applications using the Python programming language
6. Approach everyday problems with a computational focus

3.1.2 Syllabus and Topic Descriptions

This is an introductory course, which introduces the concepts required to be able to work as a computer scientist. The layout of the course is as follows:

3.1.3 Algorithmic Thinking

In this section, students are introduced to the concept that there are different ways of performing tasks, some of which are better than others. In addition, with these different ways, the students will learn how an algorithm may be evaluated.

Algorithms:

- **Introduction** - Introduction to algorithms, and why they are useful. Compares a computational algorithm to a food recipe
- **Different Methods** - There are many ways of doing the same thing, each with advantages and disadvantages. Shows that a better algorithm can lead to a greater improvement in performance than faster computers
- **Evaluation** - Having established that more than one algorithm can perform the same task, we introduce some formal means of evaluating the average case, best case, and worst case scenarios.

[1] http://dl.acm.org/citation.cfm?id=1118215

- **Tractability** - Not all algorithms can complete in a reasonable time, sometimes an approximate solution is a better option

Data Structures:

- **Introduction** - Data structures have different forms, to optimise a particular type of function given certain constraints
- **Simple data structures** - Introduces the stack and queue as data structures, and how efficiently they perform certain tasks

3.1.4 Computing as an abstraction

In this section, the course looks at ways in which computers make use of black boxes and other abstractions to build upon what has been completed before:

- **Computer architecture** - A computer is a clear example of an abstraction, discusses how it gets from switches and gates to an operating system
    - Binary/machine code, logic gates, hex, assembly
- **Programming as an abstraction** - Computers are good at automation, here we see how the automation of some principles leads to
    - Compiled vs interpreted
    - Different types of programming, e.g., procedural, functional, object oriented, event based
    - Higher level languages
- **Object oriented programming** - A class is an abstraction of a thing, containing representations only of the things we want it to.
- **Data as an abstraction**
    - How data or information are abstracted from binary to represent text, statistics, or other media content

3.1.5 The Web and the Cloud

The World Wide Web (the Web) is central to modern data processes. In this section, we introduce the Web as an instance of a computer network. With the ubiquity of Internet connections, an inevitable extension was the concept of the Cloud, where online operators provide and manage services which would previously have been done on a business's own infrastructure.

- **What are computer networks?** An introduction to basic concepts of how data are transported across a network
- **The WWW** - The most widely known network
    - Client/server model - Explains the model which is used by the Web
    - HTTP - Explains the background to how the protocol handles request and response to connect to the Web, and the use of different REST verbs
    - Displaying pages on the Web
        - HTML & CSS - separation of concerns between content and presentation
- **The Cloud**
    - **Introduction to the cloud** - Services provided by cloud companies vary from provision of hardware to fully running services. This section explains the cloud business model.
    - **Types of cloud services** - This section introduces the different types of services provided by cloud companies, such as Iaas, PaaS, SaaS

3.1.6 Data Management

This section uses "data management" in a broad way, as to define all different means of manipulating and representing data.

- **Introduction to data management**
  - Store data somewhere, whilst being able to have access to it
  - Different trade-offs with different situations
  - Types of data: Unstructured, semi-structured, structured
  - Different scenarios for data management
- **Data Representation** - Provides examples of a variety of scenarios of how data may be represented, such that it can be manipulated for analysis in the future.
  - **Text files** - The simplest way of representing data. Introduces the following data formats: csv/tsv/json/xml, and good practices for log files
  - **Images, videos** - Following analysis of data it may be represented in an image or video; or these could themselves be used as data to be analysed. Discussion of common image and video encoding.
  - **Relational database** - The most common way of storing data, using SQL to manipulate and retrieve.
  - **NoSQL** - Introduces the CAP theorem, and discusses trade-offs between relational and non-relational data.
  - **APIs** - Data stored in some manner can be made available from an API. Builds on the discussion from HTTP and introduces REST APIs

3.1.7 Python Programming

This is done alongside the other sections, as a means of applying the theoretical concepts introduced in the other week. It will not give expertise in Python programming, but will provide a foundation to develop expertise in the future. The intention is not to teach how to program in Python so much, but rather to apply the principles in the remainder of the course.

Python is chosen above R for this curriculum, since it is more general purpose, and more relevant to the principles of programming generally as opposed to R which fits better as a language for the mathematical side of programming. In addition, there are other courses within this curriculum which additionally cover R, such as '*Statistical and Mathematical Foundations*'. Teaching Python will cover only the essentials sufficient to perform these tasks, with students referred to other courses should they wish to gain a more detailed understanding.

- Setting up a development environment
- Primitive types
- Control flow and loops
- Classes and functions

3.2 Learning materials & delivery methods

The course is offered by the University of Southampton and will be delivered via the FutureLearn MOOC platform. The materials are delivered through a combination of videos/slides, as well as text written in markdown in the platform itself. Participants get an opportunity to practice the skills they have learnt through exercises, and quizzes in the FutureLearn system.

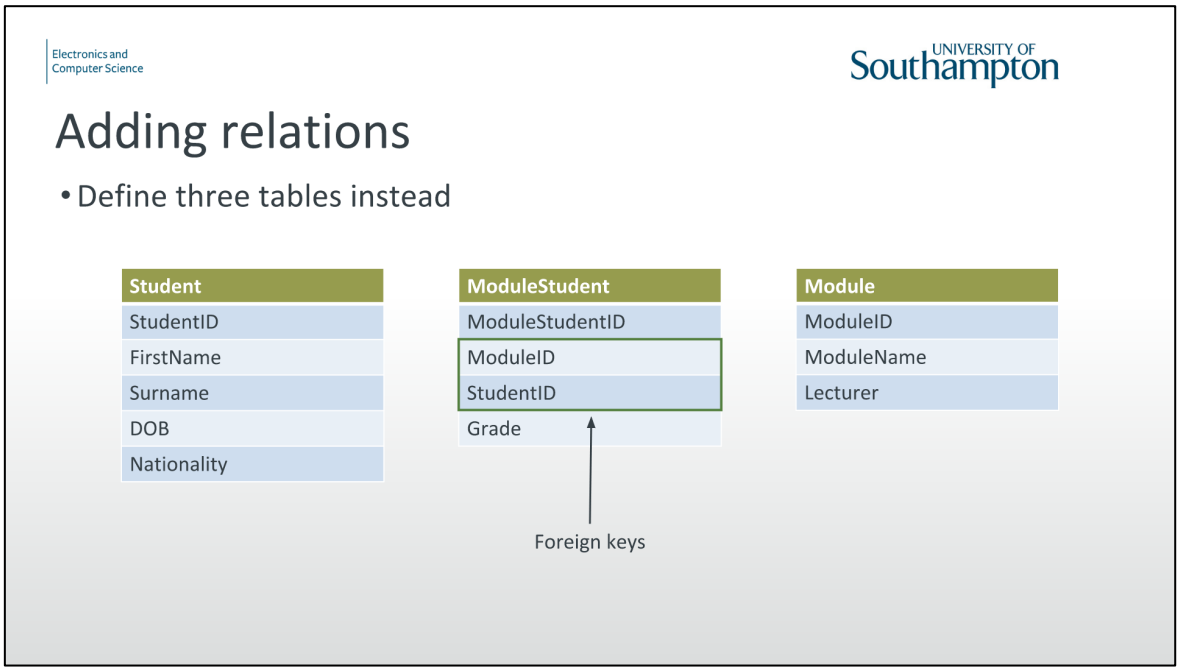

Figure 1: Sample of the learning materials of the Programming / Computational Thinking module.

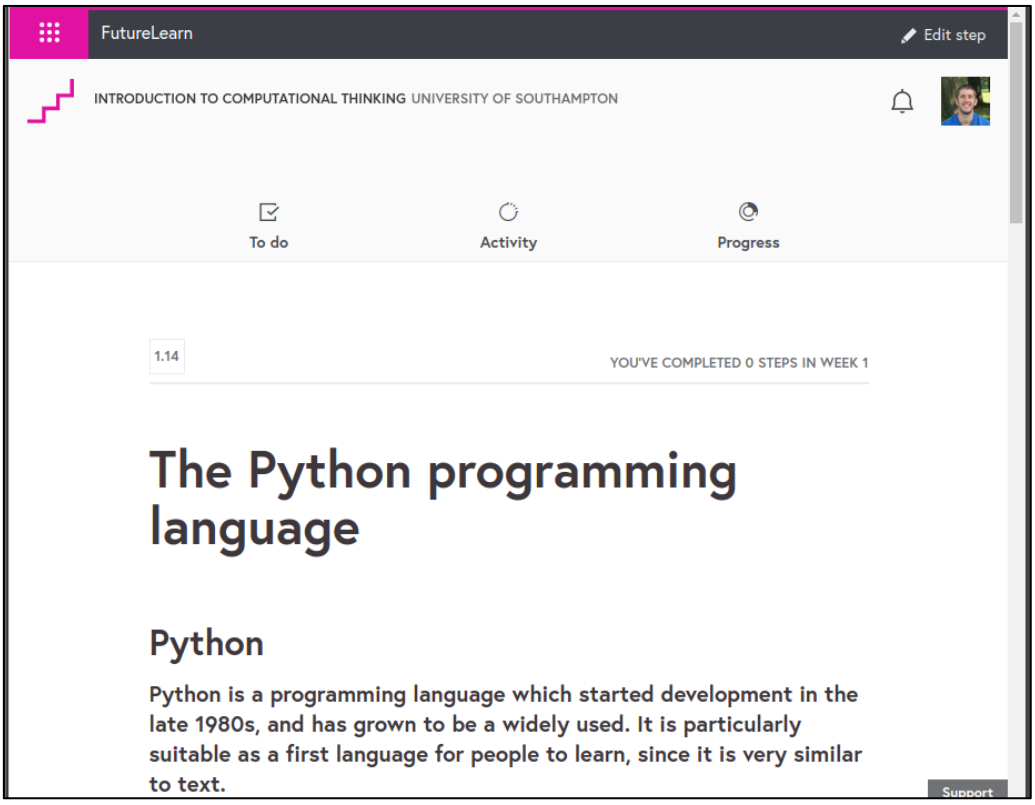

Figure 2: Screenshot of the Programming / Computational Thinking module hosted by the FutureLearn platform.

3.3 Relevance to curriculum

This course is a foundational module in the EDSA curriculum, scheduled to be released in M36. It addresses the "Computational thinking" module, as this was defined in the final EDSA curriculum presented in D2.3.

3.4 Relevance to demand analysis

In D1.4, it was recommended that EDSA should create a data science skills framework, which was done based on the EDISON skills areas. In addition, the demand for different skills based on the demand dashboard was used in order to identify relevant skills. As discussed in D2.3, there was a distinction between what should comprise core "data science", and skills which would be required to obtain a job in data science. This course fills the gap as a basic course, which provides the prerequisites for being able to perform the more advanced technical skills whilst allowing other courses to retain the focus on core data science skills.

3.5 Further development plans

This course will be released on the FutureLearn platform in early 2018. Following initial run, feedback will be sought with a view to developing based on this. In addition, extra areas will be investigated for possible inclusion in future versions of the course.

4. Data Intensive Computing

4.1 Module overview

The data intensive computing module is composed of three parts. **Part I** (Data-Intensive Computing) provides a solid foundation for understanding large scale distributed systems used for storing and processing massive data. A wide variety of advanced topics in data intensive computing are introduced. These topics include distributed file systems, NoSQL databases, processing data-at-rest (batch data) and data-in-motion (streaming data), graph processing, and resource management. **Part II** (Advanced Topics in Distributed Systems) complements the first part, by focusing on the concepts of graph theory which allow to explain the connectivity and dynamics of many real-world networks. The objective of **Part II** is to provide a deeper understanding and study of the behaviour of the networks arising within Distributed Systems. The course in **Part II** will cover the topics of Distributed Data Management, Large Graph Processing, Publish/Subscribe Systems and Navigable Small-World Overlays. **Part III** (Scalable Machine Learning and Deep Learning) provides the students with a solid foundation for understanding large-scale machine learning algorithms, in particular, Deep Learning, and their application areas. Techniques for efficient parallelization of machine learning algorithms are explained to show the intersection between distributed systems and machine learning fields.

4.1.1 Part I (Data-Intensive Computing) Syllabus Description

Detailed Content

Syllabus	Description
Introduction	Concepts and principles of cloud computing and data intensive computing. Cloud computing and Big Data (main trends, definitions and characteristics). Cloud Computing Models: IaaS, PaaS and SaaS. Cloud deployment models. Dimensions of Big Data: Volume, Velocity, Variety, Vacillation. Big Data Stack: Processing, Storage, Resource Management.
Storage	Distributed File Systems. Google File System (GFS). Master Operations. System Interations. Fault Tolerance. Flat Datacenter Storage (FDS). Databases and Database management. NoSQL. Dynamo. Data Consistency. Big Table. Cassandra
Parallel Processing	Programming Languages: Crash course in Scala. MapReduce. FlumeJava. Dryad. Spark and Spark SQL. Project Tungsten
Stream Processing	Introduction to stream processing. SPS programming model. Data Flow Composition and Manipulation. Parallelization. Fault Tolerance. Distributed messaging System. Kafka. Storm. SEEP. Naiad. Spark Streaming. Structured Streaming. Flink Stream. MillWheel. Google Cloud Dataflow
Graph Processing	Graph Algorithms Characteristics. Data-parallel vs. Graph-Parallel Computation. Pregel. GraphLab. PowerGraph (GrpahLab2). GraphX. X-Stream. Edge-Centric programming Model. Streaming partitions. Chaos. Storage and computation model
Machine	Data and Knowledge. Knowledge Discovery from Data. Data

learning with Mllib and Tensorflow	Mining Functionalities. Classification and Regression (Supervised Learning). Clustering (Unsupervised learning). MLlib. Tensorflow
Resource Management	Mesos. YARN

Existing Materials

- Data Intensive Computing, Royal Institute of technology, KTH
- Data-intensive Computing Systems, Duke University, https://www.cs.duke.edu/courses/spring15/compsci516/
- Data-Intensive Computing, Illinois Institute of Technology, http://www.cs.iit.edu/~iraicu/teaching/CS554-F13/
- Data-Intensive Scalable Computing, Brown University, http://cs.brown.edu/courses/csci2950-u/f11/
- Data-Intensive Computing, The City University of HongKong, https://www.cityu.edu.hk/ug/201415/course/CS4480.htm
- Data Intensive Computing and Clouds, University of Central Florida, http://www.eecs.ucf.edu/~jwang/Teaching/EEL6938-s12/
- Data Intensive Computing, University of Buffalo, http://www.cse.buffalo.edu/shared/course.php?e=CSE&n=587

Further Reading

- P. Mell, and T. Grance. The NIST Definition of Cloud Computing
- *800-145. National Institute of Standards and Technology (NIST), Gaithersburg, MD,* (*September 2011*)
- Armbrust, M.,Fox, A., Griffith, R. , Joseph, A. D. , Katz, R. H. and Konwinski, A., Lee, G., Patterson, D. A. , Rabkin, A., Stoica, I., and Zaharia, M. Above the Clouds: A Berkeley View of Cloud Computing. EECS Department, University of California, Berkeley, Technical Report No. UCB/EECS-2009-28, February 10, 2009
- S. Ghemawat, H.d Gobioff, and Shun-Tak Leung. The Google File System. 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003
- E. B. Nightingale, J. Elson, J. Dean and S. Ghemawat. Flat Datacenter Storage. USENIX Association 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI '12).
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W. Dynamo: Amazon's Highly Available Key-value Store. Proceedings of Twenty-first ACM SIGOPS Symposium on Operating Systems Principles, SOSP'07
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., Gruber, R. E. Bigtable: A Distributed Storage System for Structured Data. ACM Trans. Comput. Syst., 4:1--4:26, 2008
- Lakshman, and P. Malik. Cassandra: a decentralized structured storage system. *Operating Systems Review 44 (2): 35-40* (*2010*)
- J. Dean , S. Ghemawat. MapReduce: simplified data processing on large clusters. OSDI'04: PROCEEDINGS OF THE 6TH CONFERENCE ON SYMPOSIUM ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION

- Chambers, A. Raniwala, F. Perry, S. Adams, R. Henry, R. Bradshaw and Nathan. FlumeJava: Easy, Efficient Data-Parallel Pipelines. ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI), ACM New York, NY 2010, 2 Penn Plaza, Suite 701 New York, NY 10121-0701 (2010), pp. 363-375
- Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S, Stoica, I. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. 2012
- G. Cugola, Alessandro Margara. Processing flows of information: From data stream to complex event processing. ACM Comput. Surv. 44(3): 15:1-15:62 (2012)
- J.-H.Hwang, M. Balazinska, A. Rasin, U. Çetintemel, M. Stonebraker, S. B. Zdonik:
- High-Availability Algorithms for Distributed Stream Processing. ICDE 2005: 779-790
- R. C. Fernandez, M. Migliavacca, E. Kalyvianaki, and P. Pietzuch. 2013. Integrating scale out and fault tolerance in stream processing using operator state management. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (SIGMOD '13). ACM, New York, NY, USA, 725-736.
- M. Zaharia, T. Das, H. Li, S. Shenker, and I. Stoica. 2012. Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters. In *Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing* (HotCloud'12). USENIX Association, Berkeley, CA, USA, 10-10.
- T. Akidau, R. Bradshaw, C. Chambers, S. Chernyak, R. J. Fernández-Moctezuma, R. Lax, S. McVeety, D. Mills, F. Perry, E. Schmidt, and S. Whittle. 2015. The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proc. VLDB Endow.* 8, 12 (August 2015), 1792-1803.
- X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, DB Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M.Zaharia, and A. Talwalkar. 2016. MLlib: machine learning in apache spark. *J. Mach. Learn. Res.* 17, 1 (January 2016), 1235-1241.

4.1.2 Part II (Advanced Topics in Distributed Systems) Syllabus Description

Detailed Content

Syllabus	Description
Introduction	Networks: behaviour and dynamics. Analysis of Networks. Networks versus graphs. Complexity of networks.
Main Concepts	Basic definitions. Paths. Cycles. Connectivity. (Giant) Components. Distance.
Network models	G(n,m) model. Erdos-Renyi random graph. Random graphs and real world. Watts-Strogatz model.Small-Worlds model. Preferential attachment model. Random walks.
PageRank,	Convergence of Random Walk. Relation to Web search. Google

Graph Spectra	pagerank. Topic specific pagerank. Graph spectrum. Spectral graph partitioning.
Graph Exploration, Navigable Small-World Networks	How to explore big networks? Influence of node degree. Milgram's (Small-worlds) experiment. Implications for P2P systems. Navigation in Watts-Strogatz Small-Worlds. Kleinberg's model of Small-Worlds.
Navigable Structured Overlays, Gossiping Algorithms, Cyclon	Small-Worlds based P2P overlays. Approximation of Kleinberg's model. Traditional DHTs and Kleinberg's model. Basic navigation principles. Gossiping algorithms. Cyclon.
Topology construction by gossiping. Publish/Subscribe Systems	Proactive gossip framework. Topology creation. T-man. Publish/Subscribe systems. Topic based publish/subscribe. Content Topic based publish/subscribe. Filter based routing. Overlay-Per-Topic publish/subscribe. SpiderCast. Interest-Aware (greedy) Links. Tera. Rendezvous based publish/subscribe.
Hybrid Pub/Sub Systems Scribe	Decentralized publish/subscribe. Vitis. Building Navigable Structure.

Existing Materials

- Advanced Topics in Distributed Computing, at Royal Institute of Technology, KTH
- Network Analysis, University of Michigan, https://www.icpsr.umich.edu/icpsrweb/sumprog/courses/0131
- Social and Information Network Analysis, Stanford center for Professional Development, http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId=7932016
- Social Network Analysis, University of Michigan, Coursera, https://www.class-central.com/mooc/338/coursera-social-network-analysis
- Network Analysis and Modelling, Santa Fe Institute, http://tuvalu.santafe.edu/~aaronc/courses/5352/

Further Reading

- "Networks, Crowds, and Markets: Reasoning About a Highly Connected World" by David Easley and Jon Kleinberg
- "Networks: An Introduction" by Mark Newman
- “Foundations of Data Science” by J. Hopcroft and R.Kannan

4.1.3 Part III (Scalable Machine Learning and Deep Learning) Syllabus Description

Main Topics:

- Machine Learning (ML) Principles
- Using Scalable Data Analytics Frameworks to parallelize machine learning algorithms
- Distributed Linear Regression
- Distributed Logistic Regression
- Distributed Principal Component Analysis
- Linear Algebra, Probability Theory and Numerical Computation
- Feedforward Deep Networks
- Regularization in Deep Learning
- Optimization for Training Deep Models
- Convolutional Networks
- Sequence Modelling: Recurrent and Recursive Nets

Detailed Content

Syllabus	Description
Introduction	A brief history and application examples of deep learning, Large-Scale Machine Learning: at Google and in industry, ML background, brief overview of Deep Learning, understanding Deep Learning Systems, Linear Algebra review, Probability Theory review.
Distributed ML and Linear Regression	Supervised and Unsupervised learning, ML pipeline, Classification pipeline, Linear regression, Distributed ML, Computational Complexity
Gradient Descent and SparkML	Optimization theory review, gradient descent for Least Squares Regression, The Gradient, Large-Scale ML Pipelines, Feature Extraction, Feature Hashing, Apache Spark and Spark ML.

Logistic Regression and Classification	Probabilistic Interpretation, Multinomial Logistic Classification, Classification Example in Tensorflow, Quick Look in Tensorflow
Feedforward Neural Nets and Backprop	Numerical Stability, Neural Networks, Feedforward Neural Networks, Feedforward Phase, Backpropagation
Regularization and Debugging	A Flow of Deep Learning, Techniques for Training Deep Learning Nets, Regularization, Why does Deep Learning work?
Convolutional Neural Networks	How Convolutional Neural Nets Work, ConvNets with Depth, Memory Complexity, Case Studies, ConvNets for Everything
	Recurrent Neural networks, Sequence-to-Sequence Learning and Autoencoders, Long Short-term Memory (LSTM), Attention, Autoencoders: Unsupervised Feature Learning,
Deep Reinforcement Learning	Markov decision Processes, Overview of Reinforcement learning, Supervised Learning vs Reinforcement Learning, Deep RL, Q-learning, Deep Policy Networks, Distributed RL Architecture, Asynchronous RL.
Case Study	AlphaGo, When will Deep Learning become Intelligent?

Existing Material

- Scalable machine learning and Deep Learning, at Royal Institute of technology, KTH
- Scalable Machine Learning, edX, https://courses.edx.org/courses/BerkeleyX/CS190.1x/1T2015/info
- Distributed Machine Learning with Apache Spark https://www.edx.org/course/distributed-machine-learning-apache-uc-berkeleyx-cs120x
- Deep Learning Systems, University of Washington, http://dlsys.cs.washington.edu/
- Scalable Machine Learning, University of Berkeley, https://bcourses.berkeley.edu/courses/1413454/

Further Reading

- Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning, MIT Press
- SparkML Pipelines, http://spark.apache.org/docs/latest/ml-pipeline.html
- SparkML Overview, https://www.infoq.com/articles/apache-sparkml-data-pipelines
- SparkML under the Hood, http://spark.tc/machine-learning-in-apache-spark-2-0-under-the-hood-and-over-the-horizon-2/?cm_mc_uid=02948943737214702940997&cm_mc_sid_50200000=1472964414
- Tensorflow, https://www.tensorflow.org/versions/r0.11/tutorials/

4.2 Learning materials & delivery methods

The course is offered by KTH and each of its parts is delivered as follows:

Part I: Data-Intensive Computing

The course consists of thirteen face-to-face lectures in the form of presentations. Slides and video lectures are provided to students through an internal university portal. Students have to study a couple of research papers before each lecture. Four Exercises are provided as lab assignments, which are discussed with students with teacher assistants. Two reading assignments are to review papers and encourage the students to analyze papers in a critical way.

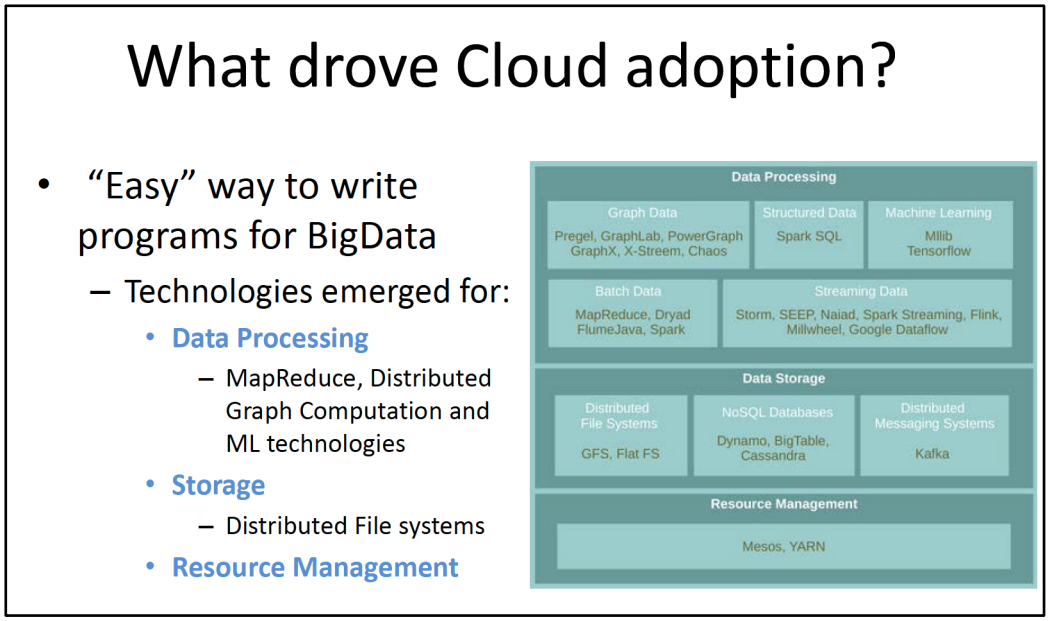

Figure 3: Screenshot from the Introduction presentation of Part I (Data-Intensive Computing).

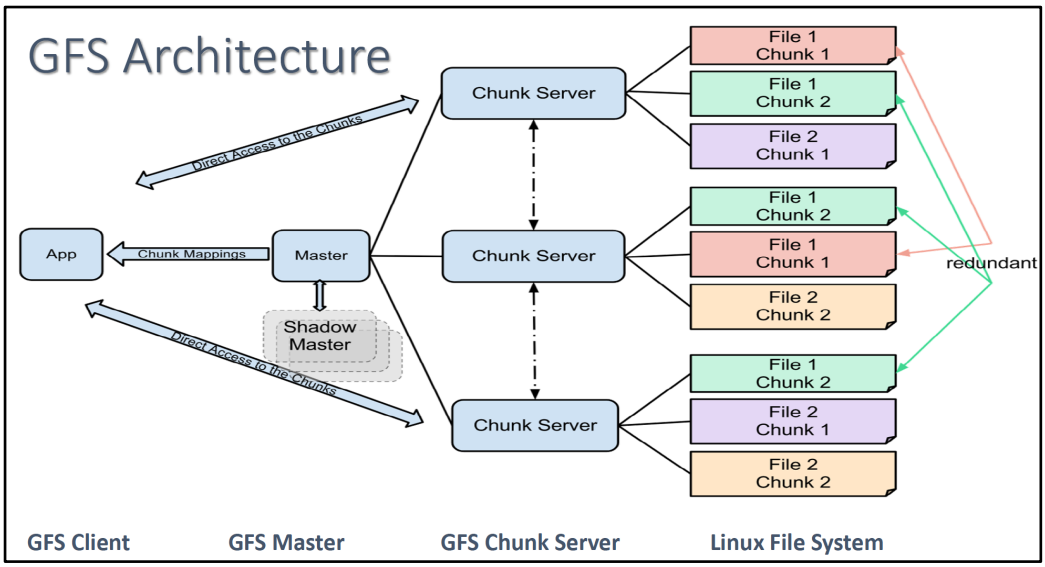

Figure 4: Screenshot from the Big Data Storage presentation of Part I (Data-Intensive Computing).

Part II: Advanced Topics in Distributed Systems

The course consists of eight face-to-face lectures in the form of presentations. Slides and video lectures are provided to students. Students have to study a couple of research papers before each lecture. Two reading assignments are to review papers and encourage the students to analyze papers in a critical way. Students are required to give an oral presentation related to the papers they review. Below are some screenshots from the course learning materials.

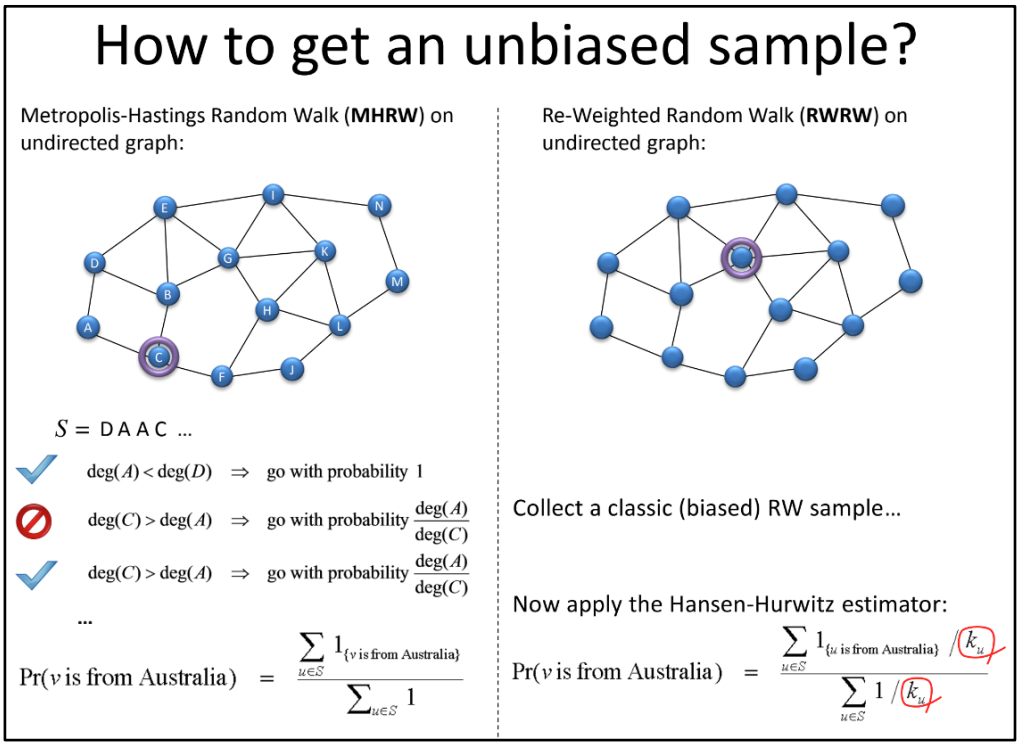

Figure 5: Screenshot from the learning materials of Part II (Advanced Topics in Distributed Systems).

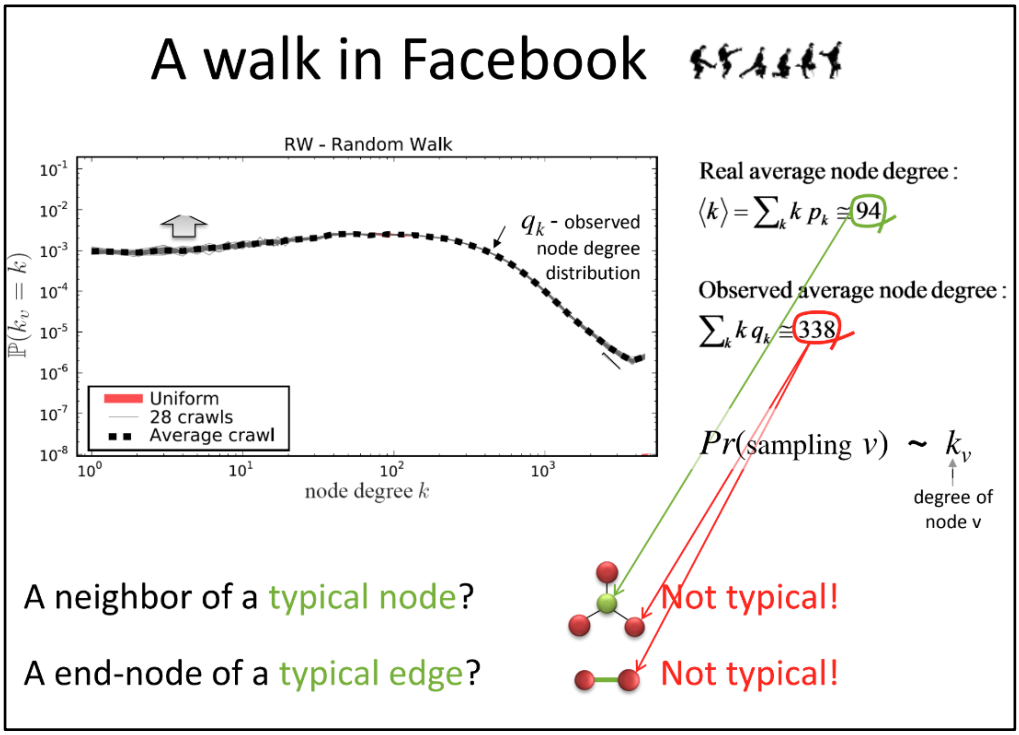

Figure 6: Screenshot from the learning materials of Part II (Advanced Topics in Distributed Systems).

Part III: Scalable Machine Learning and Deep Learning

The course consists of eleven face-to-face lectures in the form of presentations. Students are required to work on programming assignments using Scala/python programming languages. They are also required to succeed in a project which idea they choose from the context of the course and using deep learning. Below are some screenshots from the course learning materials.

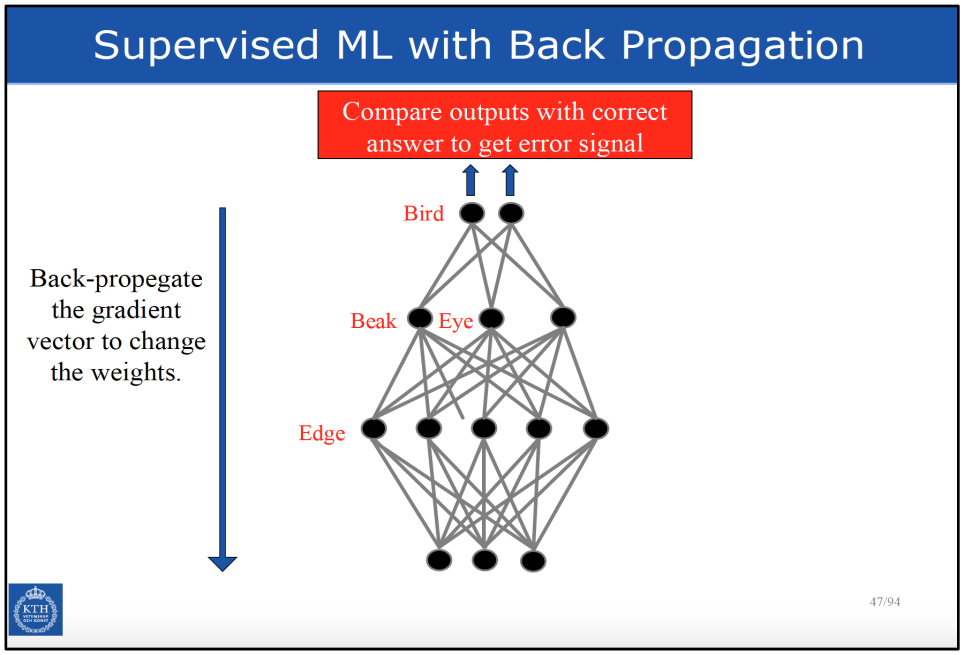

Figure 7: Screenshot from the learning materials of Part III (Scalable Machine Learning and Deep Learning).

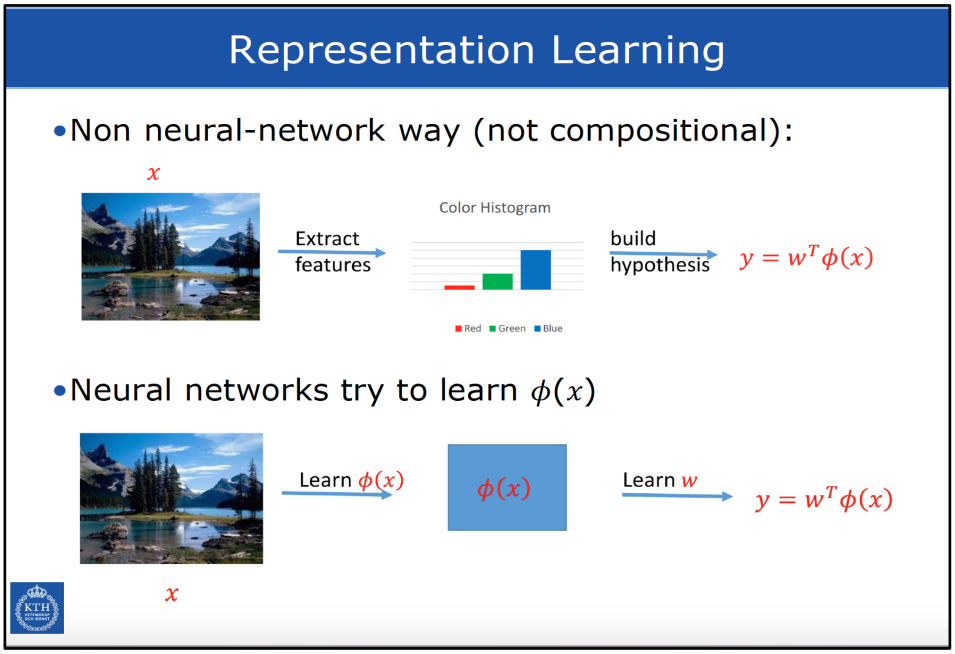

Figure 8: Screenshot from the learning materials of Part III (Scalable Machine Learning and Deep Learning).

4.3 Relevance to curriculum

The course corresponds to the "Data-Intensive Computing" module of the EDSA curriculum, as this was presented in D2.3.

4.4 Relevance to demand analysis

The courses provide a solid foundation for students in machine learning and data mining combined with distributed computing. Two of these courses are given in a blended format, and the third one is a face-to-face course. The courses are rich in top soft skills required currently in the data science field. Students are encouraged to do peer-reviews discussions, oral presentations and paper reviews in these courses, which improve their communication skills. All courses in this module use open-source tools and programming libraries. There are plans to make some of these courses reachable across different organizations.

4.5 Further development plans

The contents of all courses of this module are updated according to recent papers and improvements in the field. Lab assignments are also enhanced each time the course is given. Students provide evaluation at the end of course about the programming languages, and platforms that are used. Courses are adjusted according to reviews that are relevant. There are plans to provide some of these courses as MOOC courses.

5. Social Media Analytics

5.1 Module overview

The participants get a sound overview of social media analysis problems and analysis techniques. Target is the extraction and summarization of user statements from social media sources. The relevant text data mining concepts and deep learning approaches will be introduced. The underlying model assumptions are discussed as far as they are important for the practical application and interpretation of results. Special emphasis is put on the performance evaluation of procedures. To be able to process comprehensive collections specific Big Data implementations will be utilized. Most approaches are demonstrated by stepping through small Python scripts to show the necessary computations. The participants will be able to realistically assess the application of social media analysis technologies for different usage scenarios and can start with their own experiments using different analysis packages with Python.

Syllabus	**Concepts and methods**
Introduction • **Social Media** • **Media Analytics** • **Recent success stories**	Types of social media User base of social media Relevance of social media analysis Analysis by machine learning Main machine learning paradigms Different types of analysis questions Important applications
Download and preparation • **Document formats** • **Corpora** • **Google, Twitter, Facebook** • **Preparation of data**	Search engines for document collection crawling packages parsing of different document formats language detection sentence splitting and tokenization part of speech tagging Example scripts
Classify social media posts • **Common approaches** • **Performance evaluation** • **Application**	Definition Different types: logistic regression, support vector machine Overfitting test and performance measures feature selection Big Data approaches: Spark, TensorFlow Example Scripts

Word similarity in posts **• Group words in posts** **• Detecting topics in posts**	Unsupervised learning of word properties topic models: Assuming underlying topics word2vec: predicting words in the neighborhood stochastic optimization negative sampling Big Data implementations: Spark, Tensorflow Example scripts
Clustering social media posts **• Similarity of post** **• visualization**	Defining similarity measures using topic models and word embeddings Clustering approaches Visualization and specialization Interpretation of results Example scripts
Detecting names, products **• Detecting entities** **• Long range dependencies**	Predict properties of words Probabilistic: Conditional Random Field recurrent neural network automatic feature selection word models bidirectional recurrent neural network test and performance measures practical training and evaluation Example scripts
Opinions in social media **• Predict user sentiment** **• opinions and aspects** **• phrases and their relations**	Attaching labels to phrases and sentences Types of opinion mining Aspect based opinions Evaluation Example scripts
Practical social media analytics **• Lessons learnt** **• selecting algorithms** **• combining approaches**	Description of a practical evaluation Combining different approaches and algorithms Evaluation and performance assessment visualization

Existing training:

- Face-to-face training at Fraunhofer: http://www.iais.fraunhofer.de/socialmediaanalytics.html
- https://www.diygenius.com/10-free-online-courses-in-social-media-and-inbound-marketing/

- https://apps.ep.jhu.edu/course-homepages/3523-605.433-social-media-analytics-piorkowski-mcculloh
- https://www.coursera.org/specializations/social-media-marketing

Exercises: Starting with the example scripts the participants will be given new datasets. They have to select an analysis technique, perform the analysis and report and interpret the evaluation and visualizations.

Sources for further reading:

- G.F. Khan Seven Layers of Social Media Analytics (2015): Mining Business Insights from Social Media Text, Actions, Networks, Hyperlinks, Apps, Search Engine, and Location Data. Kindle
- T.Schreck, D.Keim (2013), Visual Analysis of Social Media Data,
- C.C. Aggarwal, C.X. Zhai (2012) Mining Text Data. Springer
- CD Manning, P Raghavan, H Schütze (2008): Introduction to Information Retrieval
- C. Biemann, A. Mehler (2015): Text Mining: From Ontology Learning to Automated Text Processing Applications
- http://www.iais.fraunhofer.de/data-scientist.html
- https://www.tensorflow.org/
- http://scikit-learn.org/stable/

Description of topics:

1. *Introduction*: Different types of social media (Facebook, Twitter, Instagram, etc.) are considered and the utilization by users is discussed. Social media analysis is relevant as many actual decision of users, e.g. buying goods or voting in elections are based on social media. This course aims at extracting and summarizing the statements of users from social media to give a realistic impression of the public view on relevant questions. Different types of analysis questions will be discussed. The relevant machine learning concepts will be introduced. Finally a number of application highlights will be presented.
2. *Download and Preparation:* In this lecture techniques for downloading, crawling and monitoring social media will be presented. As a second step the different format – pdf, text, html, ePUB – have to be parsed to extract the underlying text. This text usually is pre-processed to extract words and sentences. Often only specific parts of the documents are required, e.g. html navigation or advertisements have to be excluded. Part-of-speech tagging may be used to exclude specific types of words – e.g. function words. The approaches are demonstrated by Python sample scripts.
3. *Classify Social Media Posts*: Often we want to categorize the documents according to their contents, e.g. soccer, politics, recipes, etc. This can be done with high reliability by text classification procedures. These approaches require a training set with manually classified examples. Logistic regression and support vector machines are introduced as relevant techniques. We discuss the training process and the phenomenon of overfitting, which may be controlled by regularization. Corresponding performance measures based on an independent test set are required. We use up to date libraries like scikit-learn as well as Spark and TensorFlow, which are ready for Big Data applications. The approaches are demonstrated by Python sample scripts.
4. *Clustering Words and Social Media Posts*. Most interesting in social media analytics is to show the spectrum of issues which are addressed in posts. One relevant analysis approach investigates the similarity of words by putting words into a high-dimensional semantic space. Topic models assume that there are a number of underlying themes in a social media source. These topics are extracted without any user input (unsupervised) and each word in a document can be

represented as a mixture of topics. The different word embedding approach (word2vec) directly represents each word as a vector in a high-dimensional semantic space. Both techniques are introduced, and their training in the Big Data context is discussed. Both techniques may be extended to define the content of a document as well as the similarity of complete documents. The approaches are demonstrated by Python sample scripts.

5. *Detecting Names and Products*. A specific requirement in social media analytics is the detection of entities, i.e. phrases of a specific type. These have highly varying form and can only detected from their context. Hence, we introduce two types of word sequence analysis procedures using pre-labelled training data: Conditional Random Fields and Recurrent Neural Networks. We discuss the underlying assumptions and training procedures and assess their performance by appropriate evaluation measures. The approaches are demonstrated by Python sample scripts.
6. *Opinion Mining in Social Media*. Opinion mining (or sentiment analysis) has the aim to extract subjective user statements on specific issues from a social media post. There are different approaches ranging from assigning an opinion score to a whole document to detecting opinions specified for a specific aspect. We discuss a range of analysis techniques including classification and word sequence analysis. The approaches are demonstrated by Python sample scripts.
7. *Practical Social Media Analysis*. Social media content analysis requires the knowledge of a large toolbox of techniques with specific underlying models and limitations. From our own project experience, we give practical hints on the selection of techniques, the succession of analysis steps, and their evaluation. Most important is the visualization of results, especially for clustering approaches like topic modelling or word embedding. We demonstrate this using t-SNE and Gephi. Finally, users may discuss their own questions with the instructors.

5.2 Learning materials & delivery methods

The course is offered by Fraunhofer and consists of two face-to-face days, full of presentations with compact information. Scripts are shown and demonstrated, but there is no time for exercises. Below are some screenshots from the first presentation.

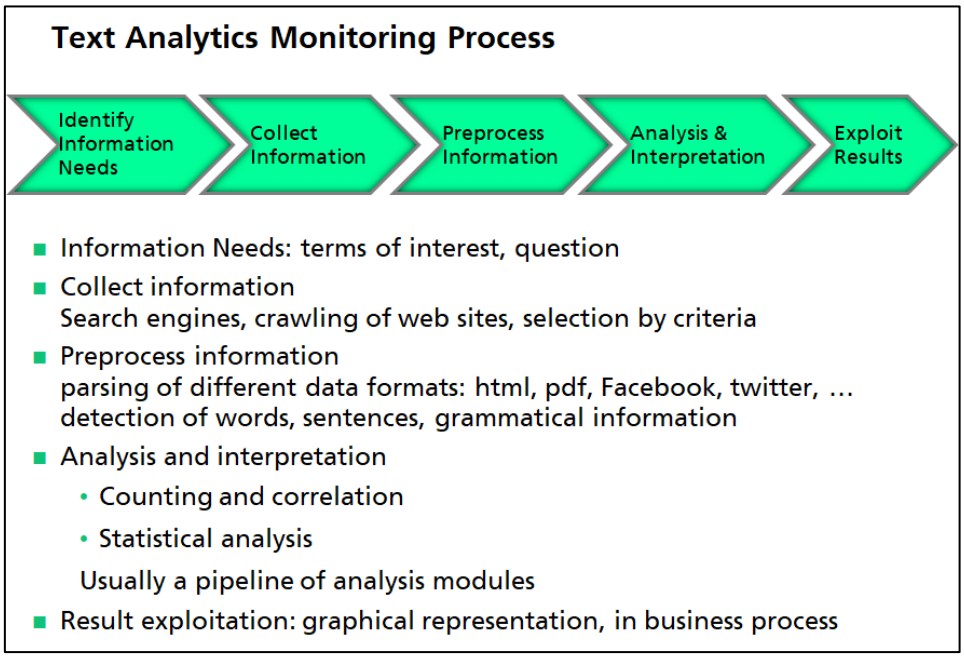

Figure 9: Screenshot from the learning materials of the Social Media Analytics module.

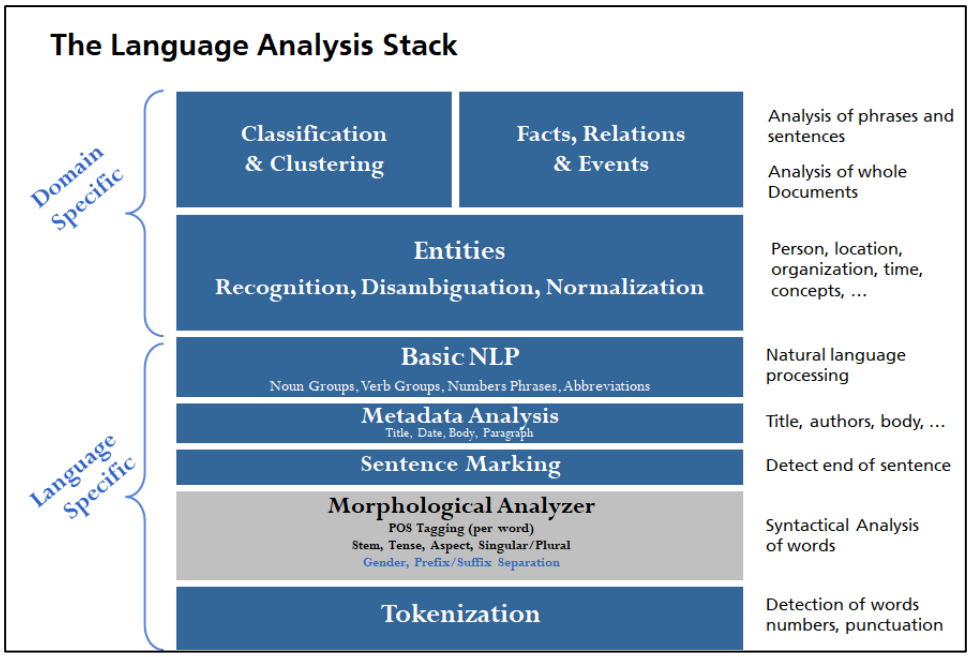

Figure 10: Screenshot from the learning materials of the Social Media Analytics module.

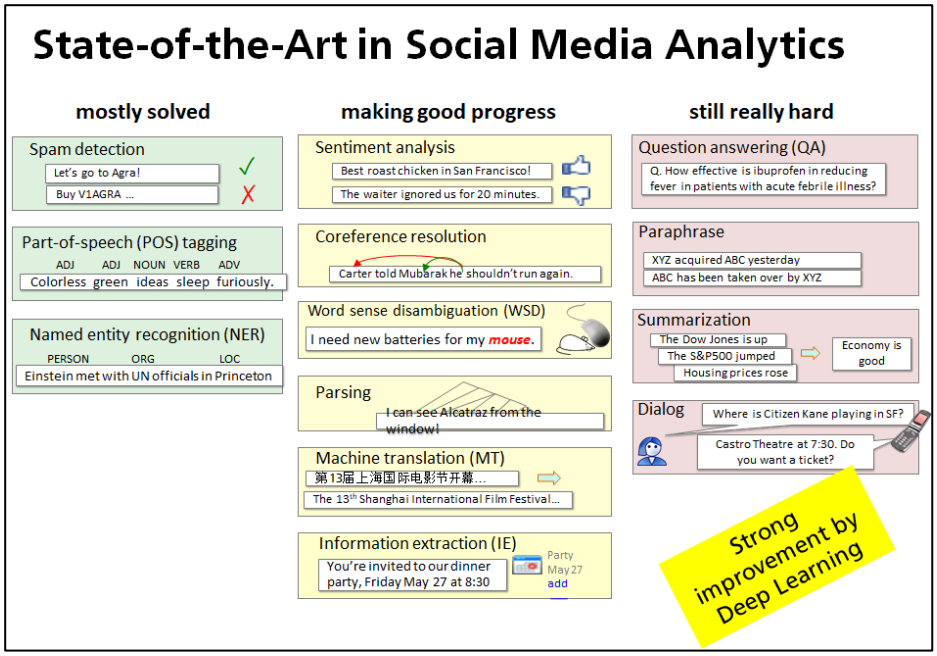

Figure 11: Screenshot from the learning materials of the Social Media Analytics module.

5.3 Relevance to curriculum

The course corresponds to module 13 of the final EDSA curriculum, as this was presented in D2.3. It addresses the analytic stage at an advanced level. It concentrates on textual data specifically from social media. Thus, it is an attempt to target a specific sector.

5.4 Relevance to demand analysis

The course concentrates on open source tools and libraries. But it is rather special and does not try to reach participants from different units. Also, it does not include blending learning formats, but this is going to change.

5.5 Further development plans

Fraunhofer has a three-level certification program for data scientists. In 2018, a new certificate "Data Scientist Specialized in Machine Learning" will be added. It includes an examination which tests for theoretical knowledge and practical skills. This examination is prepared by a group of blended courses. All courses start face-to-face and end with a practical phase, where the participants do exercises with big data sets in a distributed machine learning platform. The mandatory course is about deep learning and other current methods of machine learning. Among the optional, more special courses is a course on deep text analytics[2], first to be delivered in May 2018. This blended course will replace the former purely face-to-face course on "Social Media Analytics".

[2] https://www.bigdata.fraunhofer.de/de/datascientist/seminare/machine_learning_zertifizierung/textverstehen.html

6. Data Exploitation including data markets and licensing

6.1 Module overview

Data science is fundamentally about generating impact from data. Impact is achieved when data acts as a catalyst for change either for social, environmental or economic gain.

At the heart of driving change is finding and telling stories in data, as covered in the "Finding Stories in Data"[3] course. The combination of this along with strong communication and leadership skills can help lead to change in business practice in either profitability, market positioning and/or efficiency. Data science can inform and bring evidence for new and changing business models and effective data exploitation can also unlock new and unexpected markets.

This set of 9 online lessons, offered by ODI, looks at how exploiting a world of data can lead to unexpected benefits for businesses and citizens alike. The 9 lessons are divided into 3 sequential modules that will guide learners through three key topic areas in data exploitation, data markets and licensing.

- Course 1: Exploiting data science to create value
- Aim: Enable you to recognise the opportunities to exploit data science and analyse the risks of doing so.
    - Lesson 1: Exploiting data science
    - Lesson 2: Managing risk and reward
    - Lesson 3: Open innovation and data science
- Course 2: Data science means business
- Aim: Enable you to analyse a number of different business models to exploit data science.
    - Lesson 1: Business models in data science
    - Lesson 2: Pitching for data science
    - Lesson 3: Data markets and licensing
- Course 3: Data science ecosystems
- Aim: Enable you to exploit data science in different ecosystems.
    - Lesson 1: Embedding data science within a business
    - Lesson 2: Exploring the startup ecosystem
    - Lesson 3: Exploiting data science within a sector

Curated content in the syllabus will be accompanied with narration to help guide a learner in their interpretation of the resources. Additionally, supporting materials and guidance from those in the community will help explain how data science has opened up new markets.

6.1.1 Course 1: Exploiting data science to create value

This set of lessons look broadly at the ways that data science can create value, how to identify the right opportunities and how to manage risk. By the end of this module learners should be able to make the right go/no-go decisions for exploiting data science in their own businesses.

Exploiting data science

Across the world, people are exploiting data science to inform decisions, build new businesses and form new collaborations. The aim of this lesson is to help learners understand how data science creates value and introduce them to a number of case studies. For example, we look at how Transport for London used WiFi access points to track movement of passengers through the network in their first ever short study able to collection 400,000 points of data for analysis.

[3] http://courses.edsa-project.eu/course/view.php?id=52

Managing risk and reward

In order to exploit data science successfully, it is important to be able to identify the potential risks and consider solutions to manage problems that may arise. One of the main concerns is that data science is blurring "disciplinary boundaries and field with long histories of ethical review are being inundated with work from fields with no history of review and indeed active movements against such requirements."[4] As well as ethics, this lesson looks at opportunity and legal risk and introduces case studies that found themselves in hot water. The module concludes with some tools and guides that will help mitigate risks when exploiting data science.

Open innovation and data science

Data science blurs disciplinary boundaries and involves experts from many fields. Often one company will not have all the required skills in house to rapidly exploit data science. Open innovation is a key enabler to fast innovation. Open innovation is a collaborative approach to creating innovative products and services. Being collaborative with the community can also help improve product reputation or help identify risks earlier. The lesson looks at how open innovation plays a key role in data exploitation along with key case studies.

6.1.2 Course 2: Data science means business

Building from the first course, this course looks at ways to embed innovation from data science within an existing (or new) business model. A key part of Business Intelligence is how to use data to drive forward a business model. This set of lessons looks at where data science fits in different business models and how to pitch for data science against a business model. Further this course looks at how licensing is a core part of the data market and how different models can be exploited both for commercial gain and for innovation.

Business models in data science

Many businesses are now using data science as part of their business model. In order to successfully exploit data science, organisations need to understand how data is creating value for their business. This lessons looks at how to create a value proposition and the link between this and various business models. The aim of this lesson is to enable learners to realise how to exploit data science deeply within a business model.

Pitching for data science

Once a potential new or innovative value proposition and business model have been built, the next key stage is to communicate this effectively either within an organisation or to external funders. Having a good pitch is essential for gaining interest from organisation leaders, investors and potential partners. This helps establish and grow data science businesses. This lesson looks at how to create an effective pitch for data science led innovation.

Data markets and licensing

Data is a new raw material. A key role of a data scientist is to refine this material to create something more valuable. At the same time raw materials at various stages of refinement are traded or sold to others. This lesson looks at some of the markets surrounding data and how these related to a spectrum of opportunities and business models.

6.1.3 Course 3: Data science ecosystems

The aim of this course is to enable participants to exploit data science in different ecosystems. Within the field of data science, there are many different levels of ecosystem. These range from a internal

[4] https://www.forbes.com/sites/kalevleetaru/2017/10/16/is-it-too-late-for-big-data-ethics/#aa20de23a6d1

business ecosystem that is made up of many different departments and teams, to all the actors and stakeholders within a sector or industry.

Embedding data science within a business

Modern data scientists often face the challenge of how to embed data science within a business. For a business to fully embrace data science, it must be at the heart of all business activities. In this module, we explore the importance of a data science strategy and how to develop a data science strategy, as well as how to reduce friction and increase buy-in. This will ensure that data science sits at the heart of business activities.

Exploring the startup ecosystem

A data scientist might work within an organisation or they might have a new and novel idea that they wish to bring to market. For data scientists entering the startup ecosystem, it can be a very different experience from the ecosystem of an individual business. In this module we explore the different stages of startups that exist within the ecosystem, the roles of different funding types and sources, and the additional support that is available to those navigating the data science startup ecosystem.

Exploiting data science in a sector

A modern data scientist is expected to combine domain expertise with more traditional data science skills. Knowing how to exploit data science within a sector is an essential skill for a modern data scientist. In this lesson, we explore data science ecosystems within sectors, drawing on examples from the entertainment, transport and physical activity sectors, as well as the difference between data science in a closed, shared or open sector ecosystem.

6.2 Learning materials & delivery methods

The entire 3 course curriculum encompassing 9 lessons is offered by ODI and is available entirely online for anyone to engage with freely at any time without the need for enrolment. Although the lessons are numbered in order, learners can choose their own path if they choose. Each lesson is presented in the form of an interactive web page following the latest instructional design theory using an eLearning platform voted the most innovative 4 year in a row at the learnX awards[5]. The eLearning platform provides a completely cross platform responsive way of delivering interactive learning. Each lesson is backed with learning outcomes which are assessed through the use of interactive questioning at the end of each.

Each lesson features:

- Practical steps for accomplishing key tasks
- Case-studies
- Knowledge-checks and quizzes

6.3 Relevance to curriculum

So far, the EDSA curriculum has provided educational resources of skills and methodologies which cover many skills areas in depth. For example, there are at least 5 courses in the areas of “Big data technologies and systems” and “Computing methodologies”. However, currently there are no courses on “Business analysis organisation and management” and only one in the areas of “Business analysis and enterprise organisation” and “Business process management”. By focussing specifically on Business Intelligence and exploiting data science in businesses this course has been specifically designed to cover such areas.

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

6.4 Relevance to demand analysis

The key outcomes of the demand analysis (D1.2) provided a clear indication that there was a need for broad introductory level training in data science. 8 key curriculum areas were identified of which 6 focus on "hard skills":

1. Big Data
2. Machine learning and prediction
3. Data collection and analysis
4. Maths and statistics
5. Interpretation and visualisation
6. Advanced computing and programming

This leaves two areas of the curriculum which are more "soft skills" based:

1. Business intelligence and domain expertise
2. Open source tools and concepts

This course has been designed focus on how data science can unlock value for new and existing businesses. Since the start of the European Data Science Academy project in 2015 data science has been growing in interest, while related areas such as Big Data, AI and Machine learning continue to also show a constant focus.

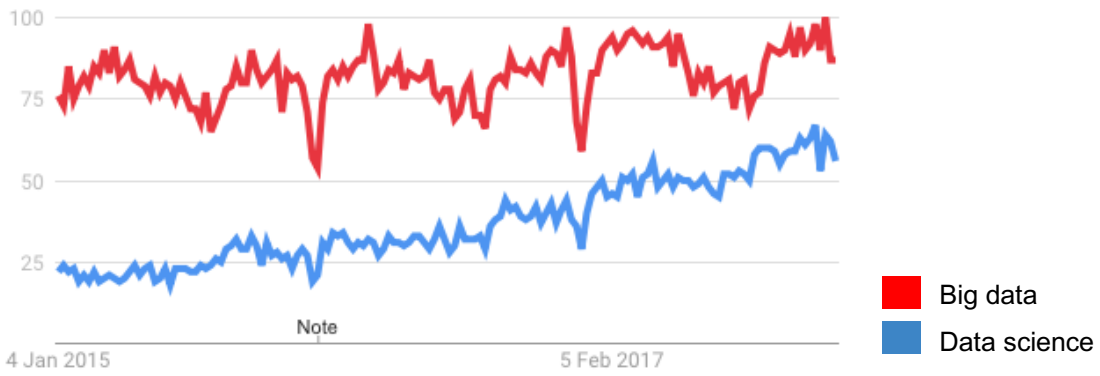

Figure 12: Google trend analysis for Big data and Data Science.

Importantly this shows that data science, as a discipline is gaining in importance. Data science is a discipline that brings important scientific and business approaches to tools and techniques like machine learning and it is important to thus also focus on the soft skills on what applying data science means to businesses and how managers can make the most of data science in existing businesses.

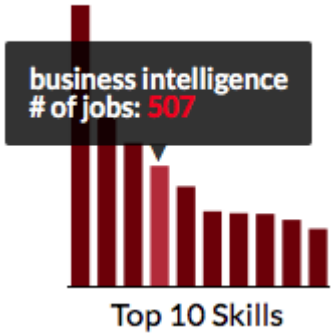

Figure 13: The EDSA dashboard skills chart.

The current EDSA dashboard also shows the growing interest in exploiting open data in businesses. The 4th most required skill in jobs is currently business intelligence.

Business intelligence is fundamentally about the strategies and technologies that allow business to build insight from data. In order to exploit BI requires knowledge of both the hard skills (technologies) and the soft skills (strategies) to understand the implications of such data analysis. Thus, data scientists require some level of strategic knowledge while existing managers need to understand the opportunities and risks of applying data science and other BI techniques.

This requirement was also emphasised in the demand analysis (D1.2). D1.2 pointed out a report by McKinsey on Big Data and the lack of managers willing it take risks on employing data scientists[6].

D1.2 recommended that curricular needs to address the challenges of data science being embedded within business at all levels such that potential is maximised. This module has been specifically designed to help bridge the gap between the hard data science skills and their implications on businesses. By the end of the module learners will have a better business intelligence of the opportunity and risks of exploiting data science, both within existing organisations and to build new ones.

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

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

The data exploitation including data markets and licensing course focuses on many of these areas, in particular.

1. Holistic training approach

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

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

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

The data exploitation including data markets and licensing course focuses on soft skills that are easily transferable into different domains. The course introduces a variety of modelling tools that can be used in many domains and sectors; for example, the lesson "Business models in data science" introduces the business model canvas which is a widely used tool for identifying great business opportunities.

This holistic approach will help equip a modern data scientist with the theoretical knowledge and soft skills required to help integrate data science into many different markets, sectors and business.

2. Open source based training

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

- EDSA study evaluation report (D1.4)

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

3. Soft skills training

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

- EDSA study evaluation report (D1.4)

The data Exploitation including data markets and licensing course is entirely focussed on the aspects of how to transform their applications into business tactics and strategies. The three courses all take a key business focus and look not just at how data science fits within different business models but also how this can help change a sector and what strategies to take.

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

- EDSA study evaluation report (D1.4)

In the second part of the course on "Data science means business" there are lessons specifically on how to create and pitch a value proposition to business leaders in order to bring about change. These skills are essential for leading data science innovators.

4. Basic data literacy skills

One key aspect of data literacy is understanding when data should not be used to inform a decision. Specifically, in the first part of the Data Exploitation including data markets and licensing course there is a lesson entitled "Managing risk and reward". This looks specifically about the risks of data science blurring boundaries between disciplines that results in a loss of expertise and rigor of results that can create a negative impact.

5. Navigation and guidance

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

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

6.5 Further development plans

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

1. *Increase the number of external links to beginner level and related data science courses*: As discovered by the demand analysis, well-structured learning for data science beginners is lacking. The ODI will continue to maintain any relevant links in the learning resources to ensure relevance to learners.
2. *Repackage as a commercial product:* The course will be offered as a commercial product by ODI, as described in the ODI exploitation plan presented in the Project Exploitation Report (D5.4).

7. Conclusion

The EDSA courses portfolio has incorporated a wide range of learning resources, either offered by project partners or by third parties. This deliverable has presented the modules that have been added to the EDSA courses portfolio in the third year of the project, in order to cover the following 4 topics of the final EDSA curriculum.:

- Programming / Computational Thinking (R and Python)
- Data Intensive Computing
- Social Media Analytics
- Data Exploitation including data markets and licensing

In particular, we have presented an overview of the objectives and learning outcomes of the new modules, we have identified the types of learning materials used and how these are delivered, and we have discussed how the new modules are related to the EDSA curriculum and the demand analysis, as well as the plans in place for their further development.