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

D2.1 Data Science Curricula 1

Deliverable Editor:	**Christopher Phethean (University of Southampton)**
Other contributors:	**Elena Simperl, Gareth Beeston (University of Southampton) Curricula contributions from: Mihhail Matskin (KTH), Ali Syed (Persontyle), Patrick Mukala (TU/e), Inna Novalija (JSI), Angelika Voss (IAIS)**
Deliverable Reviewers:	**Stefan Rueping (IAIS) Ali Syed (Persontyle)**

Deliverable due date:	**31/07/2015**
Submission date:	**31/07/2015**
Distribution level:	**PUBLIC**
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	07/07/2015	C Phethean	Curricula from first round of modules included with outline of overall curriculum and schedule for delivery.
0.2	08/07/2015	C Phethean	Related training activities added.
0.3	09/07/2015	C Phethean	Amendments to content. Addition of market analysis methodology, delivery methods, introduction and conclusion.
0.4	10/07/2015	C Phethean	Final amendments to review version, added in summary of results from market analysis.
0.5	23/07/2015	C Phethean	Changes based on reviewer feedback
0.6	24/07/2015	C Phethean	Final changes
0.7	28/07/2015	C Phethean	Further changes based on technical review
1.0	31/07/2015	A Tumilowicz	Final QA

Table of Contents

List of Tables

1. Executive Summary

The purpose of this deliverable is to provide the first version of the EDSA curriculum, which is comprised of 15 core modules that will be delivered over the course of the 3-year project. At this point, in Month 6, we present the outline of the overall curriculum along with specific module curricula for 6 of these modules, with the remaining modules to be delivered over the remainder of the project – along with revisions to any existing content at that point. Modules are separated into four main themes: foundations, data storage and processing, data analysis, and data interpretation and use. These themes help to structure the curriculum and demonstrate the variety of levels at which we aim to provide training and expertise, which will assist in the provision of specific learning pathways for the variety of audiences that we will target – which are based around different job roles, technology specialisation, industry sector and previous experience. The 6 topics covered at this point are Foundations of Data Science, Foundations of Big Data, Big Data Architecture, Distributed Computing, Machine Learning and Process Mining. These are based on demand and market analysis results, which are described below. At this point, we have an initial curriculum based on these 6 topics, which – as they are developed further with learning resources, and additional modules – will lead to a number of learning pathways that different audiences can take. As such there is potential for some overlap in the modules, but this ensures that prerequisites are taken into account based on whether or not someone would have been expected to complete an earlier 'foundations' level module.

2. Introduction

Before presenting the initial curriculum, we go over a number of factors regarding the background of the course including target audience(s) and delivery methods before moving on to discussing the market analysis carried out to assess existing courses and alignment with the EDSA goal.

2.1 Target Audience and Prerequisite Knowledge

The target audience of the curriculum is varied, due to the broad range of professionals who are impacted by the deluge of data that is characteristic of the contemporary world. Data analysts, statisticians, data engineers and managers all have differing requirements in terms of knowledge in order to comfortably adopt Data Science and the wealth of benefits that it can offer. Therefore we have identified a number of audience groups that the EDSA curriculum will target, and as such defining exact levels of prerequisite knowledge is difficult. We therefore provide 'foundation' level modules in the curriculum, along with more advanced topics (described in Section 3) for:

- Statisticians (modeling, numerical insights, etc.)
- Analysts (more involved with data gathering and reporting)
- Managers, Product owners and CDOs
- Programmers, Developers and System Engineers
- Data managers (including security experts).

In addition, due to the huge breadth in what 'Data Science' could cover, we have identified a number of other dimensions through which to classify the EDSA content – this will help tailor specific learning pathways to particular requirements:

- Tools and programming languages

    - We have identified a number of popular languages and tools involved for Data Science from the market analysis discussed later. Popular languages include R, Python and D3, while tools such as Hadoop and Open Refine are similarly frequent choices for Data Science.)
- Type of data
    - Depending on the type of data that somebody possesses, the type of analysis they can carry out will differ. For example, we differentiate between structured and unstructured data, and categorise data based on it being geographical or temporal, or whether it is social data, sensor data or transactional data (for example).
- Industry sector
    - Based on the sector an individual is based in, there may be specific skills that are required, or a particular application of certain knowledge that is necessary. For example the Process Mining module outline below has specific variations for the Internet of Things and for Healthcare. We identify the following sectors:
        - aerospace and defence
        - agriculture
        - automotive
        - consumer services
        - construction and engineering
        - consultancy
        - data and information systems
        - energy
        - finance, insurance, real estate
        - government and public sector
        - health
        - manufacturing
        - media and advertising
        - mining
        - professional services
        - telecoms
        - transport
- Level: Basic, advanced or expert.
    - We will offer pathways based on what previous skill-level or experience a learner is arriving from. Foundational modules will ensure that basic-level students can cover the essential prerequisites before moving on to more advanced and specialised modules.

2.2 Delivery Methods

The curriculum discussed below is based on a number of modules for which we will develop high-quality learning materials that can be adopted for use in a number of formats. This will help us to meet the varying needs of data scientists in all of the job roles discussed above.

Course material will be disseminated online, via the EDSA project website. Materials will also be hosted on sites such as videolectures.net, SlideShare and SlideWiki. This will allow learners to undertake self-training as they can explore the material themselves and at their own pace, with no trainer support or facilitation. However some more structured delivery will also be available. MOOCs allow the delivery of particular pieces of content to be delivered as standalone courses online, where learners can engage in peer-supported and community driven learning with a class of other students. Expert facilitators can also monitor discussions and progress through these courses, responding to questions where necessary, and therefore they offer a more traditional experience albeit online, remote, and in class sizes of up to the tens of thousands.

Webinars also offer online lecturing, and allow for some interaction between learners and the tutor. This is particularly suited to presentations or lectures, and seminars that can be attended virtually. Interactive eBooks will provide the 'traditional' textbook in a digital form, containing interactive exercises and quizzes directly embedded within the book that allow the students to test out what they have learnt from the written content.

The materials may also be used for offline, face-to-face training, whereby each partner will deliver courses where appropriate to help disseminate the EDSA project and address skills shortages in particular organisations. This will allow close relationships to be developed between project partners and public or private sector organisations, which can help to provide more focused and constructive feedback on the suitability and performance of the learning materials in a real-world environment.

3. Curriculum Outline and Delivery Schedule

In this section, we introduce and present the initial EDSA curriculum, before discussing each of the constituent modules in more detail in Section 4. We also discuss the schedule by which individual module curricula will be developed, and when their subsequent learning resources will be created.

3.1 Overall Curriculum Structure

The overall curricula were formed around 4 main themes that are based around the general Data Science process:

1. Foundations of Data Science
2. Data Storage and Processing
3. Data Analysis
4. Data Interpretation and Use.

These stages provide a core framework to the curriculum in which modules for EDSA could be placed in order to assess their alignment with the demand analysis in WP1. By carrying out a market analysis of existing Data Science courses and other demand analysis from job postings that were available at this point in the project, we were able to extract 15 topics that make up the first version of the EDSA curriculum.

3.2 Market Analysis

A centralized search was carried out, initially in English, to identify what Data Science courses were currently available across the EU. Additionally, we planned to investigate what content these courses typically contained. This helped us to assess the demand for certain modules and technologies, but also allowed us to identify gaps in existing Data Science training. Throughout the project, partners will be encouraged to add to this data with details about courses in their own language and based in their country.

A quantitative approach was followed, with a range of search criteria used to collect details on courses. For the centralized search, search engines were employed to locate courses using a set of keywords related to Data Science. For each search term, the first 15 pages of search engine results were analysed in order to gather a large corpus of data about courses; these were complemented with similar searches on course databases from which all pages were analyzed. A systematic method of data collection was used, focusing on the following methods:

- Keyword search on search engines (Google and Bing)
- Keyword search on course databases e.g. Find a Masters
- Search on professional sites including LinkedIn, School of Data and Data Science Academy
- Keyword translated into alternative languages and submitted through search engines
- Data collected manually from screen, formatted and cleaned into standard structure

A number of keywords were chosen that represented both 'Data Science' courses in general, and courses that are conducted within this field – such as machine learning or data analytics. The following keywords are examples of those used to collect the data, which were elicited based on an assessment of the modules:

- Data Science
- Machine Learning
- Big Data
- Data Analytics
- Data Analysis
- Data Visualisation
- Data Mining
- Intelligent Systems
- Business Intelligence

Data was collected about each course, including the course length, location and modules/topics that are included. We found that 121 out of the 137 courses identified across the EU were in English, however as this was based on translation of keywords *from* English initially, it is possible that this misses a number of other courses. The top countries where courses were found are shown below:

Table 1: Market analysis of courses across the EU

Country	Number of Courses
United Kingdom	76
Ireland	11
France	11
The Netherlands	9
Spain	7
Germany	7

We also identified 186 courses from the USA, which were analysed to gain an understanding of the global provision of data science training. This work will be continued throughout the project, in particular with further emphasis on non-English language courses across the EU to enhance our understanding of the existing market.

Exploratory statistics were used to provide an overview of the courses already available, which allowed us to determine which modules were commonly offered. The majority of courses found included modules such as Data Mining, Programming, Big Data, Machine Learning, Data Analytics, Visualisation techniques and an introduction to Data Science. Courses frequently used Python, R, Perl and Java as programming languages used in the teaching of Data Science techniques, with technologies such as Hadoop used often to familiarize learners with specific tools. This helped to provide the EDSA consortium with an awareness of what the key languages and tools that were used on existing courses where, which we could combine with preliminary demand analysis to assess key components of what our curriculum should contain.

With the data from the market analysis organized into a single document[1], an analysis was carried out manually to match courses to the EDSA curriculum, by coding each course based on which module it overlapped with in EDSA (based on the core curriculum developed in Section 3.3). This developed a categorized list of courses based on what areas match the EDSA course. This information was available to partners, who were assigned modules to develop curricula for in M6 of the project following the design of the core curriculum discussed in Section 3.3. Each partner then recommended and added similar courses for the module based on their own in-depth expertise of the topic area (see the 'Existing Courses' section of each module curriculum in Section 4, below).

3.3 EDSA Core Curriculum

A list of modules was curated, using the market analysis discussed above. In addition, the consortium used the existing demand analysis carried out as part of WP1 to help identify what modules were necessary to include in the EDSA curriculum. The consortium held a meeting in London in March 2015 to agree upon a core selection of modules that would be essential to a Data Science course based on these two corpora of data, which together gave us indications about what was missing from existing courses, what was essential for any Data Science course, and what were seen as key skills we need to impart upon learners.. Where possible, we have sought to utilise as much of the consortium's existing expertise and to offer modules around what each partner excels at.

Table 2: Core EDSA Curriculum - Initial Version.

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

[1] The spreadsheet containing the data is available to download from https://drive.google.com/file/d/0B1kijuRKFg5NWHRmaTRNb3ViX3c/view?usp=sharing

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

The order of the modules and the structure of their classification around the 4 stages discussed above means that a learner following a direct path through the material will gradually be exposed to increasingly complex topics, having begun with a comprehensive overview of the fundamental foundation elements of Data Science knowledge. It is planned that given a number of different audience classifications (by job role, sector, level etc. – discussed in Section 2.1 we will provide different learning paths through the material, however, so that the pathway for a manager would be different for a statistician, for example. This also means that participants with existing knowledge of a particular area can quickly recap the content of the earlier foundations modules before working through the later topics to advance their knowledge. The modular curriculum affords the re-use and modification of all EDSA materials, so that training courses can be developed based on these topics but adjusted to reflect specific trainees' or learners' needs.

An example learning pathway for a basic level data analyst in the healthcare domain may look like:

Foundations of Data Science

Foundations of Big Data

Stats/Maths Foundations

Programming (R and Python)

Data Management and Curation

Machine Learning

Big Data Analytics

Process Mining (Healthcare pathway)

3.4 Curriculum Delivery Schedule

With three versions of the curriculum due throughout the EDSA project, the fifteen modules were split so that the overall curriculum will be built up and completed in three stages to match these deliverables (Table 3). In order to allow for revision and adaptation of each topic's content, the allocation is skewed towards the earlier deliverables so that more modules are completed early in the project: 6 modules will be produced in the first year, 5 in the second year and the remaining 4 will be produced in the final year of the project. This will allow time for any modifications that are required later in the project, which we will identify through continued demand analysis and engagement with the Data Science community.

Table 3: EDSA Core Curriculum Schedule

Topic	Schedule
Foundations of Data Science	M6
Foundations of Big Data	M6
Big Data Architecture	M6
Distributed Computing	M6
Machine Learning, Data Mining and Basic Analytics	M6
Process Mining	M6
Statistical / Mathematical Foundations	M18

Data Management and Curation	M18
Big Data Analytics	M18
Data Visualisation	M18
Finding Stories in Open Data	M18
Programming / Computational Thinking (R and Python)	M30
Stream Processing	M30
Visual Analytics	M30
Data Exploitation including data markets and licensing	M30

For each of the three stages, the associated learning materials will be created for each module in advance of the respective deliverables for the WP2 learning resources (D2.4, D2.5 and D2.6). Therefore, the topics with curricula produced in M6 will have learning resources for M12, curricula in M18 will have them in M24 and the final set of learning resources for the curricula in M30 will be produced by M36. As discussed above, there will also be revisions to existing material that will occur in the later stages to incorporate the recommendations and changes required to keep the curriculum up-to-date.

Following the schedule presented in Table 3, this document therefore presents the first versions of the curricula for the following modules:

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

3.5 Data and Methods for Curricula Design

Each module assigned for delivery in M6 was then allocated one of the project partners based on their areas of expertise, and therefore these partners will act as champions and leaders for these modules. To develop the individual module curricula discussed in Section 4, the following areas were investigated:

- Learning Objectives
  - To frame the module around a particular aim, the module leaders created a number of learning objectives based on what learners should have achieved by the time they complete the module. These are dependent on the syllabus content for the topic.
- Syllabus and Topic Descriptions
  - Using their expertise of the area, the awareness of gaps in training gained from market and demand analysis, and the list of associated courses relevant to their module, partners created an outline of a syllabus for their allocated topic. This acts as a high-level overview of the module contents, with descriptions of what each sub-topic will focus on.
- Existing Courses
  - This section of each module curriculum lists the courses that were found during the centralized market analysis that relate to the particular module in question. Each partner also provided their own set of relevant modules based on their expert knowledge of the area around the particular module, which helped build an essential resource for knowing about what key tools and topics are commonly taught for each of these modules.

- Existing Materials
    - Materials related to each module were searched for by the module leader – including on platforms such as SlideShare and YouTube, with the aim of developing a base of open educational resources that could inform the development of each module further. In some cases, partners have begun developing materials for the EDSA modules they have been assigned to, and these have been listed where relevant.
- Example Quizzes and Questions
    - Either based on existing material, or on the syllabus designed for the module (or both), each partner identified areas where they could test knowledge and learning using quizzes that could be re-used between different formats of the course, e.g. MOOCs, eBooks and traditional face-to-face training.
- Description of Exercises
    - In similarity with the quizzes above, each partner has also begun designing exercises that could be used within each module. Where possible, these will be developed as HTML5 widgets that can be embedded within iBooks and MOOCs to allow interactive exercises within the learning environment that the student chooses to use.
- Further Reading
    - As the partners assigned to modules have significant expertise and experience in the module area, they are ideally positioned to recommend essential reading for the topic. Therefore each partner has provided a reading list that helps to inform the module design, and that offers complementary material to a student studying the course.

3.5.1 Future Input for Module Design

In order to continue development of the modules, and iterate them throughout the project to ensure their ongoing relevance, we have identified a number of areas in which to do further market research regarding data science training offerings.

With a number of data science training initiatives and resources available, such as The Data Lab (Scotland), School of Data, Data Science Central and Data Science Academy, we are looking at how the EDSA offerings align to these initiatives, and where there are gaps in these compared to our market and demand analysis. This will help us refine the syllabuses for each module to ensure that we are addressing the skills gaps and demand that motivates this project.

4. Initial EDSA Module Curricula

Following the plan outlined above, in this section we now present the first module curricula for the overall EDSA curriculum. The following subsections describe each of the modules by providing a summary of their content and a list of materials that already exist.

4.1 Foundations of Data Science

4.1.1 Learning Objectives

By the end of this module, learners will:

- Understand the foundations of the Data Science process
- Be able to evaluate Data Science tools and techniques based on their suitability for particular tasks
- Have some experience of using Google Refine, R and other tools to analyse data.

4.1.2 Syllabus and Topic Descriptions

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

    - This topic will cover some of the key ideas behind visualising data in a clear and coherent way. It will also provide a list of online resources for visualising data such as CartoDB.
- **High performance computing** (MapReduce, Hadoop, NoSQL, stream processing solutions)
    - An introduction to the technical foundations of Data Science and how these technologies are evolving and facilitate many of the concepts required for Data Science to happen.

4.1.3 Existing Courses

Table 4: Existing courses that include a module similar to Foundations of Data Science

Institution	**Course**	**Module**	**URL**	**Target Audience**	**Languages and Data Types**
Lancaster University	MSc Data Science	Data Science Fundamentals	http://www.scc.lancs.ac.uk/masters/DataScience	Graduates – computing, statistics or environmental scientists	Hadoop, Spark, Mahout, Giraph, HDFS, Hbase
City University London	MSc Data Science	Introduction to Data Science	http://www.city.ac.uk/courses/postgraduate/data-science-msc	Graduates – computing/engineering/physics/maths or business/economics/psychology/health with demonstrable maths aptitude	Hadoop, Python, R, Java, GPU programming, Matlab, SPSS
University of Sheffield	MSc Data Science	Introduction to Data Science	http://www.sheffield.ac.uk/is/pgt/courses/data_science/msc_data_science	Graduates	R, SPSS, Weka, Tableau, Spotfire
New York University	MS Data Science	Introduction to Data Science	http://cds.nyu.edu/academics/ms-in-data-science/	Graduates – mathematics, comp sci, applied stats	
University of Washington	Introduction to Data Science MOOC (Coursera)	Introduction to Data Science	https://www.coursera.org/course/datasci	Intermediate programming experience, familiarity with databases	Python, SQL, R
Chinese University of Hong Kong	MSc Data Science and Business Statistics	Foundations of Data Science	http://www.sta.cuhk.edu.hk/Programmes/PostgraduateStudies/MScinDataScienceandBusinessStatistics.aspx	Executives, professionals (public and private sector), educators/practitioners in education and public health agencies	
John Hopkins University	Data Science MOOC (Coursera)	The Data Scientist's Toolbox	https://www.coursera.org/specialization/jhudatascience/1	Programming experience, working knowledge of mathematics up to algebra	R, RStudio

4.1.4 Existing Materials

This module will be developed over the course of the next 6 months, and will be evaluated with students in Southampton, so that feedback can be received on the course content and delivery before it is utilized for EDSA.

4.1.5 Example Quizzes and Questions

- Multiple-choice questions (MCQs) to test knowledge of key terminology and definitions, as well as the Data Science process
- MCQs to assess understanding of data collection and analysis techniques

4.1.6 Description of Exercises

- Linking and analysing data using Open Refine – joining datasets etc.
- Basic statistics on data using R.
- Visualising geographic data with tools such as CartoDB.

4.1.7 Further Reading

- 'Doing Data Science: Straight Talk from the Frontline' by Cathy O'Neil & Rachel Schutt (2013)
- 'Uncharted: big Data as a Lens on Human Culture' by Erez Aiden & Jean-Baptiste Michel (2014)

4.2 Foundations of Big Data

4.2.1 Learning Objectives

By the end of this module, learners will:

- Understand the foundations of big data
- Have gained knowledge on the tools that operate with Big Data and Big Data applications
- Have gained some basic experience with data analytics using the QMiner tool

4.2.2 Syllabus and Topic Descriptions

- **Introduction to Big Data**
    - This topic describes the notion of Big Data, paying a special attention to interesting facts about Big Data. The topic provides a view on Big Data in numbers.
- **Big Data Definitions, Motivation and State of Market**
    - This topic shows the definitions of Big Data, the motivation behind operating with Big Data. The characterization of Big Data by volume, velocity, variety (V3) is provided as well as Big Data popularity on the Web, Big Data hype cycles, Big Data value chain and view on the Big Data market.
- **Techniques Overview**
    - This topic is dedicated to the analytic techniques that are used for operation with Big Data. The specific analytical operators for Big Data are discussed.
- **Tools Overview**
    - This topic provides a view on types of tools that are used for Big Data. Distributed infrastructure and Distributed processing are discussed. A particular attention is given to MapReduce, NoSQL databases. Open source Big Data tools are listed.
- **Applications**
    - This topic discusses a number of Big Data applications, such as
        - Recommendation
        - Social Networks
        - Media Monitoring
- **Mining Massive Datasets**

- This topic provides algorithms for extracting models and other information from very large amounts of data. The emphasis is on techniques that are efficient and that scale well.

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

4.2.3 Existing Courses

Table 5: Existing courses that include a module similar to Foundations of Big Data

Institution	Course	Module	URL	Target Audience	Languages and Data Types
JSI	SlideShare Videolectures	Big Data tutorial	http://www.slideshare.net/markogrobelnik/big-datatutorial-grobelnikfortunamladenicsydneyiswc2013, http://www.complacs.org/lsoldm2014/	Computer scientists, mathematicians, statisticians	Hadoop, NoSQL
Stanford University	Mining Massive Datasets MOOC	Mining Massive Datasets	https://www.coursera.org/course/mmds	Computer scientists, mathematicians	MapReduce, graphs, recommender systems, clustering, SVM, decision trees
JSI	Videolectures	Data analytics with QMiner (to be updated)	http://videolectures.net/sikdd2014_rei_large_scale/?q=qminer	Computer scientists, mathematicians, statisticians	QMiner
n/a	Videolectures	Various resources on big data	http://videolectures.net/Top/Computer_Science/Big_Data/	Mixed	Mixed

4.2.4 Existing Materials

- 'Understanding Big Data' by Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Big+Data+University/page/FREE+ebook+-+Understanding+Big+Data
- 'Big Data glossary' by Pete Warden https://www.ebooks-it.net/ebook/big-data-glossary
- 'Mining of Massive Datasets' by Jure Leskovec, Anand Rajaraman, Jeff Ullman http://www.mmds.org/
- QMiner http://qminer.ijs.si/

4.2.5 Example Quizzes and Questions

- Quiz to test knowledge of key terminology and definitions.
- Possible quizzes and questions from Big Data in Numbers:

    - http://www.slideshare.net/markogrobelnik/big-datatutorial-grobelnikfortunamladenicsydneyiswc2013
        - $600 - to buy a disk drive that can store all of the world's music
        - 5 billion mobile phones in use in 2010
        - 30 billion pieces of content stared on Facebook every month
        - 40% projected growth in global data generated per year vs 5% growth in global IT spending
        - 235 terabytes data collected by the US Library of Congress by April 2011
        - 15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress etc.

4.2.6 Description of Exercises

- Data analytics with QMiner - http://qminer.ijs.si/
- MapReduce by examples (from slideshare.net): by Andrea Iacono, http://www.slideshare.net/andreaiacono/mapreduce-34478449

4.2.7 Further Reading

- ‘The little book of Big Data’ by Noreen Burlingame http://www.amazon.com/Little-Book-DATA-2012-Edition-ebook/dp/B0079Q8NR0
- ‘Scaling up machine learning’ by Ron Bekkerman, Mikhail Bilenko, John Langford http://www.amazon.com/Scaling-Machine-Learning-Distributed-Approaches/dp/0521192242
- ‘Big Data: The next frontier for innovation, competition and productivity’ by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
- ‘Hadoop in Action’ by Chuck Lam http://www.amazon.com/Hadoop-Action-Chuck-Lam/dp/1935182196, http://www.amazon.com/Hadoop-Action-Chuck-Lam/dp/1617291226/ref=dp_ob_title_bk
- ‘Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work’ by Harlan Harris, Sean Murphy, Marck Vaisman, Publisher: O'Reilly Media, Released: June 2013
- ‘An Introduction to Data’ by Jeffrey Stanton, Syracuse University School of Information Studies, Downloadable from: http://jsresearch.net/wiki/projects/teachdatascience, Released: February 2013
- ‘Data Science for Business: What you need to know about data mining and data-analytic thinking’ by Foster Provost and Tom Fawcett, Released: August 2013

4.3 Big Data Architecture

4.3.1 Learning Objectives

The participants get a sound overview of architectural designs and technical components. Based on computing concepts like “MapReduce”, theoretical insights like the CAP theorem and non-functional requirements like processing in real time big data products are presented and rated. The participants will be able to realistically assess the application of big data technologies for different usage scenarios and can start with their own experiments. This module builds on the basics and preliminary knowledge that learners would have gained in the 'Foundations of Big Data' module.

4.3.2 Syllabus and Topic Descriptions

- **Introduction**
  - The components used by internet giants, e.g. Google, Facebook or Amazon, to build big data applications are provided, and often extended in various ways, by open source communities. The technological building blocks for leveraging big data are freely available and also substantially present in portfolios of large systems providers. Still, there are some major obstacles to overcome before implementing big data applications. Big data components often provide less functionality than usually expected of operating systems, relational databases or BI systems. This topic covers Big Data fundamentals and concepts, components of big data applications, and CAP theorem and eventual consistency.
- **Lambda Architecture**
  - A constructively usable pattern for concept and design of a big data application is the "lambda architecture", as published by Nathan Marz and James Warren. The main point is to differentiate between a batch layer for large volumes of data and a speed layer for real time processing of data streams. Both layers create analytical results and store them in scalable databases. An application service combines the results of both layers and presents them to the user.

    The modular design of the lambda architecture maps well to common requirements of big data applications and systematizes them. The architectural approach is helpful for discovery and evaluation of technical and non-functional requirements. This is independent of the way and extend the modules are used as technical components of the application. Required services are identified and a suitable selection of components can be made.

- **Batch-Processing (Map Reduce, Batch Workflow Organization)**
  - The horizontal scalability of big data systems allows processing large volumes of data in a short amount of time. For better utilization of the individual components of a distributed system, the interfaces of big data frameworks are restricted. The limitations are designed to discourage behavior that works well on a single computer but are not suited for processing on multiple machines. Available open source solutions differ in focus and therefore in their interfaces. It is essential to select system components and compose them in a way that they fulfil the requirements of the application. The easiest way of implementing these steps is a concept of dataflow driven building blocks.

    We consider the popular disturbed processing scheme map reduce and how this scheme is made available by the open source software Hadoop. Further, we demonstrate the use of Cascading as a workflow engine to assemble advances processing schemes.
- **Speed Processing**
  - Within the Lambda-Architecture the input data is processed by multiple layers, i.e. the batch and speed layer. The transport is provided by a messaging system, in our

example kafka. We describe the general publish subscribe mechanism of kafka and demonstrate the transport to a distributed processing engine we choose for the speed layer, Apache Strom. We give an overview on the organization of processing tasks in the Storm system and elaborate on the involved components, i.e. topologies, spouts, bolts and streams. In contrast to the views of the batch layer, views for the speed layer require high write performance. As an example of a system capable of providing such capability, we select the NoSQL database Cassandra. We analyse the concepts used for data storage organization, for instance columnar storage organization and merkle trees.

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

Table 5 lists additional tools and techniques that will be covered in various topics of the Big Data Architecture module.

Table 6: Additional Syllabus Details for Big Data Architecture Module

Topic	Concepts and methods
Introduction • Big Data – Fundamentals and Concepts • Components of Big Data Applications • CAP Theorem, eventual consistency	Big Data Processing vs. Big Data Storage Systems Hadoop & Dynamo Overview NoSQL databases SPRAIN Complex event processing Horizontal scalability CAP theorem Eventual consistency Big data definition
Lambda Architecture • Batch- and speed processes	Lambda architecture Processing as application of a function to all data Immutable data, data de-normalization, Pre-computed views (batch views) Batch vs. Speed View Relationship to commercial Big Data Architectures
Batch-Processing • Map Reduce • Batch Workflow Organization	Distributed computing and horizontal scalability Map reduce concept (incl. Map reduce Example) Open source implementation: Hadoop Hadoop distributed file system (HDFS) HDFS scaling, fault tolerance, replication Hadoop distributed execution Map reduce (Hadoop v1.x) Hadoop2 Yet Another Resource Negotiator (YARN) Cascading workflows
• No-SQL Key-Value-Stores • Exercises	Key-Value Store Interface Distributed hash tables Quorum-based systems

	Conflict resolution with vector clocks Distributed key value storage (Voldemort)
Speed processing • Message Passing	Scalable messaging systems (Kafka) Publisher/Subscriber communication model Distributed fault tolerant configuration management (Zookeeper)
• Stream Processing	Distributed stream processing (Storm) Topologies: Spouts, bolts, tasks Stream groupings Storm vs. Hadoop
• No-SQL-Database Cassandra	Key value stores with columns (Cassandra) High write performance Speed layer
Big Data Technology Groups	Overview Big Data technology groups (mostly open source products) NoSQL databases Distributed batch processing Distributed stream processing Complex event processing Data extraction & transformation technologies Data ex/import Distributed search & indexing of documents Distributed data analysis techniques Hadoop ecosystem
Developing Cascading workflows	cascading workflow engine Cascading: taps. Tuples, pipes and operators Workflow Example Join processing Workflow & Lambda Architecture
Developing Storm Workflows	building Storm topologies: feeding data tuples: Storm spouts parallel processing using Storm bolts connect processing steps using Storm groupings

4.3.3 Existing Courses

- Fraunhofer ‘Big Data Architecture’ course: http://www.iais.fraunhofer.de/data-scientist-architektur.html
- Training courses from Cloudera: http://www.cloudera.com/content/cloudera/en/training.html

4.3.4 Existing Materials

Learning materials for each of the syllabus topics described above are being developed and will be made available online at http://slidewiki.org/deck/12141_big-data-architecture

4.3.5 Example Quizzes and Questions

Quizzes are currently in development for this module and are available through: http://slidewiki.org/deck/12141_big-data-architecture#tree-0-deck-12141-1-quest. Example questions include:

- Which statements about Hadoop are true?
- Which of the following are characteristics of NoSQL?
- What are the main components of Google’s MapReduce technology?

4.3.6 Description of Exercises

Learners will carry out an exercise developing workflows in Cascading and Storm.

4.3.7 Further Reading

- "Skalierbarkeit und Architektur von Big-Data-Anwendungen" by M. Mock, K.-H. Sylla, D. Hecker, OBJEKTspektrum, IT-Trends 2014, .de, http://www.sigs-datacom.de/fileadmin/user_upload/zeitschriften/os/2014/IT-Trends/Mock_Hecker_Sylla_BigData_14.pdf
- "Eventually Consistent" by Werner Vogels, Communications of the ACM 2009,|voL. 52, no. 1, Doi:10.1145/1435417.1435432
- "Towards Robust Distributed Systems" by E. Brewer, Proc. 19th Ann. ACM Symp.Principles of Distributed Computing (PODC 00), ACM, 2000, pp. 7-10;
- http://www.cs.berkeley.edu/~brewer/PODC2000.pdf
- http://www.bitkom.org/files/documents/BITKOM_Leitfaden_Big-Data-Technologien-Wissen_fuer_Entscheider_Febr_2014.pdf
- Jeffrey Dean und Sanjay Ghemawat (Google Inc.): MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004
- Nathan Marz und James Warren, "Big Data – Principles and best practices of scalable real-time data systems", Manning Publications, 2013
- http://www.iais.fraunhofer.de/data-scientist.html
- http://www.ferari-project.eu , FP7, Grant No. 619461
- http://www.insight-ict.eu , FP7, Grant. No. 318225
- http://www.motortalk.de
- http://www.cascading.org/
- http://kafka.apache.org/cascading
- https://storm.apache.org/
- http://www.project-voldemort.com/voldemort/
- http://cassandra.apache.org/

4.4 Distributed Computing

4.4.1 Learning Objectives

The learning goals of this course is to provide students with basic concepts and principles of large-scale dynamic distributed systems and distributed algorithms to be applied to Big Data.

General learning objectives:

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

Specific Learning Objectives.

The student should be able to:

- Understand and specifying distributed services,
- Design and analyse distributed algorithms for reliable and fault-tolerant implementations of the distributed services

- Explain the common concepts of P2P, e.g., DHT, gossip based algorithms and content distribution.
- Implement in a simulator environment some of P2P algorithms.
- Write a summary and present the basic ideas of some recent research papers in the field and give a critical view of the contribution and the cons and pros of the papers

4.4.2 Syllabus and Topic Descriptions

Part I: Distributed Services and Distributed Algorithms

- **Introduction to Distributed Algorithms**
    - This part will present introduction to distributed computing and to distributed algorithms in the context of big data
- **Formal Methods**
    - This part will provide insight into formal models of distributed systems considering asynchronous and synchronous systems, causality and the computation theorem and logical clocks and vector clocks
- **Basic Abstractions**
    - This will consider basic abstractions of the distributed systems including event-based component model, specification of services, safety and liveness, node failure models, fair loss links, stubborn links, perfect links and timing assumptions
- **Failure Detectors**
    - This part considers different detectors of failure including classes of detectors, perfect failure detectors, leader election, reductions and relations between failure detectors
- **Broadcast**
    - Here the following topics are considered: broadcast abstractions, lazy reliable broadcast, eager reliable broadcast, uniform reliable broadcast, fail-silent broadcast, causal broadcast, no waiting algorithm, vector-clock algorithms and orderings broadcast
- **Shared Memory**
    - This part considers algorithms for memory utilization including replicated shared memory, regular register algorithms, linearizable registers and multiple writers algorithm
- **Protocols for solving consensus**
    - Here a consensus is introduced and different algorithms are considered including regular consensus fail-stop algorithm, Paxos algorithm, sequence consensus and multi-Paxos

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

- **P2P Introduction**
    - This part will provide an introduction into the concept of P2P and consider P2P overlay types (centralized, unstructured, super-peer and structured)

- **Small World Networks**
    - This will consider concept of Small World Networks, in particular, it will discuss Small Worlds vs. Random Graphs, Navigation in Small-Worlds (Kleinberg's model), Small-World based Structured Overlays and Non-uniform Structured Overlays
- **Examples of P2P systems**
    - Before going into details of the subject this part considers examples of the P2P systems, these systems include Chord and Kademia
- **Epidemic algorithms**
    - This part introduce and consider in depth Epidemical algorithms, in particular, it discusses Aggregation protocols, Membership Management, Topology management and Connectivity problems
- **BitTorrent**
    - This will consider a protocol for distribution of large amount of data via the Internet, in particular, it discusses Peer selection and Piece selection, Choke Algorithm, Bit torrent extensions, Application of BitTorrent and Future of BitTorrent
- **P2P Live Streaming**
    - This part will include discussion of P2P media streaming, Classification of P2P streaming systems, Security in P2P streaming systems and Sepidar/GLive – two P2P streaming systems
- **System Issues in P2P**
    - This will consider issues of building P2P systems including Node Heterogeneity, Overcoming Limited Direct Connectivity in IP, Congestion Control for P2P Systems and Secure Gossiping
- **Introduction to Hadoop**
    - This part will provide a big Data issues in the context of distributed computing, this will include discussion of what is Big Data, Historical Overview of Data Systems, HDFS: Hadoop Architecture, Processing Big Data (MapReduce), Spark: Resilient Distributed Datasets and Apache Flink and Hadoop on the Cloud. While there are some overlaps here with the Big Data Architecture module, this helps to ensure that this material is covered to allow for different learning paths through the various modules.
- **Cluster Management**
    - This part discusses different cluster management issues including Architectures for large cluster resource scheduling, Examples of scheduling policies (Capacity scheduler, Fair Scheduler, Reservation-based scheduler) and Scheduler architectures
- **Introduction to Cloud Computing**
    - This will consider cloud computing issues related to big data, it includes discussion of what is Cloud computing, Supporting Technologies, Public Clouds, Private Clouds, Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS)

4.4.3 Existing Courses

There are several related courses developed at KTH as parts of “Software Engineering of Distributed Systems”, “Distributed Systems” and “Cloud Computing and Services” Master programs. Some examples of existing courses are detailed below:

Table 7: Existing courses that include a module similar to Distributed Computing

Institution	Module	URL
Carnegie Mellon University	Distributed systems	http://www.cs.cmu.edu/~dga/15-440/F12/
MITp://oc w.mit	Distributed Computer Systems Engineering	htt.edu/courses/electrical-engineering-and-computer-science/6-824-distributed-computer-systems-engineering-spring-2006/index.htm
Columbia University	Distributed systems	http://www.cs.columbia.edu/~roxana/teaching/DistributedSystemsF12/index.html
Washington University	Distributed Systems	http://courses.cs.washington.edu/courses/csep552/13sp/
John Hopkins University	Distributed Systems	http://www.cnds.jhu.edu/courses/cs437/
Stanford University	Distributed Systems	http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId =11778
New York University	Distributed Systems	http://www.news.cs.nyu.edu/~jinyang/fa10/
The University of Edinburgh	Distributed Systems	http://www.inf.ed.ac.uk/teaching/courses/ds/
Umeå University	Distributed Systems	http://www.umu.se/english/education/courses-and-programmes/course?code=5DV147
KTH	Distributed systems, basic course	http://www.kth.se/student/kurser/kurs/ID2201?l=en
Uppsala University	Distributed Systems	http://user.it.uu.se/~justin/Teaching/DS/
KTH	Distributed Systems, advance course	http://www.ict.kth.se/courses/ID2203/

4.4.4 Existing Materials

There are several related courses developed at KTH as parts of the “Software Engineering of Distributed Systems”, “Distributed Systems” and “Cloud Computing and Services” Master programs.

4.4.5 Example Quizzes and Questions

In the quizzes we will ask questions based on the lectures notes and the corresponding papers. The quizzes will test understanding of the main terminology as well as algorithms and system solutions.

4.4.6 Description of Exercises

For Part I, the assignments will be carried out in the Kompics programming framework and cover the main concepts of the presented material including Broadcast, Shared memory and Consensus.

For the Part II assignments, a home assignment titled SWIM Through the NATs will be done. The goal of this assignment is to implement and evaluate a decentralized membership service that provides information about all the participating processes.

4.4.7 Further Reading

- Introduction to Reliable and Secure Distributed Programming (Links to an external site.), Christian Cachin, Rachid Guerraoui, and Luis Rodrigues, Springer, 2011, ISBN 978-3-642-15259-7
- The Paxos Made Simple (Links to an external site.) paper is a good and comprehensible explanation of Paxos
- Gossip-based Peer Sampling http://www.ict.kth.se/courses/ID2210/papers/peer%20sampling.pdf
- SWIM: *S*calable *W*eakly-consistent *I*nfection-style Process Group *M*embership Protocol. https://www.cs.cornell.edu/~asdas/research/dsn02-swim.pdf
- LEDBAT: the new BitTorrent congestion control protocol https://www.kth.se/social/files/5527dbf2f276543890e0c409/LEDBAT.pdf
- Apache Hadoop YARN: Yet Another Resource Negotiator https://54e57bc8-a-62cb3a1a-s-sites.googlegroups.com/site/2013socc/home/program/a5-vavilapalli.pdf?attachauth=ANoY7cp3OJ5QW0OazGrf2H5DyGeZALuXnjy4iNtwdhbAryCx0HI4g2Nq5-4avGACbRyj4a5Ozj9VSUK0cEmzTJ6x0gj-xf0PWYUM2bQZJp0JD1kUHUn-cIykBZvF2317Ute52MxE3baD0VIoWlTN64tHnMQQqKeZbg2LjH-ikNMFZPG9rSOeVAyW032JyaNcCIaHdm6Zthuw1HcBUacMEBcYnWeTHR7k6mVGSgS0eNd1-hWQeBSpJOY%3D&attredirects=1
- Sparrow: Distributed, Low Latency Scheduling http://people.csail.mit.edu/matei/papers/2013/sosp_sparrow.pdf

4.5 Machine Learning, Data Mining and Basic Analytics

4.5.1 Learning Objectives

By completing this module you will learn;

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

- Have a greater confidence in their knowledge and standing as a data scientist.
- How to use these algorithms in a variety of benchmark datasets
- How to fine-tune these algorithms for better performance

4.5.2 Syllabus and Topic Descriptions

Lesson 1: R Refresher

We begin with a refresher of R (a basic introduction to R will be covered in the foundations of Data Science module) that ensures you have correctly installed the R studio IDE, examines how this IDE work and shows how exercise packages can be loaded. We then look at the basic functions we will be using in early exercises.

Lesson 2: Basic Regression and Model Selection Techniques

- Understand the idea of supervised learning, as well as the form and applications of regression type statistical models.
- Be able to implement and apply linear, quadratic and polynomial regression, and have experience doing so on actual data.
- Understand the role of basis projection in polynomial regression and PCA.
- Be able to evaluate model performance using mean squared error.
- Be able to use hold-out validation and cross validation for model selection, and understand the relationship between model complexity and performance.
- Be able to implement and apply principle component analysis (PCA) and use PCA for feature selection, information compression and regression.
- Understand and be able to apply feature shrinkage and subset selection techniques within the context of simple regression models.
- Understand the idea of degrees of freedom for measuring model complexity.
- Be able to model regression error using error functions.

Lesson 3: Basic Classification Techniques

- Understand the form and applications of classification type statistical models.
- Be able to implement and apply linear and quadratic discriminate analysis (LDA, QDA), and perceptron classification, and have experience doing so on real data.
- Be able to apply logistic regression, and have experience doing so on real data.
- Be able to evaluate model performance using misclassification error.
- Understand the idea behind Bayesian Methods in statistics.
- Be able to use noisy-or and Dirichlet-categorical distributions to encode expert knowledge with count and pseudo-count parameters, and have experience doing so.

Lesson 4: Cluster Analysis

- Understand the idea of unsupervised learning, as well as the form and application of cluster analysis.

- Be able to implement and apply K-Means, K-Medoid and hierarchical clustering algorithms for cluster analysis, and have experience doing so with real data.
- Be able to use dendrograms to represent the results of hierarchical clustering algorithms.

Lesson 5: Local Methods

- Understand the form and application of local methods, as well as their very distinctive strengths and weaknesses.
- Be able to implement and apply the K-Nearest-Neighbours, local regression and kernel density estimation algorithms, and have experience doing so with actual data.
- Understand and be able to work with kernel functions.

Lesson 6: Trees and Boosting

- Be able to implement and apply regression/classification trees.
- Be able to implement the adaboost algorithm.

Lesson 7: Advanced Techniques - Support Vector Machines and Neural Networks

- Understand how support vector machines (SVMs) and neural networks (NNs) work and the reasons for their success.
    - What support vectors, optimal hyperplanes and support vector classifiers are and their relationship to SVMs.
    - What back-propagation is and how it is used to train NNs.
- Understand the links between SVMs and NNs and the simpler statistical models from earlier modules.
    - The use of kernels and implicit basis projection in SVMs.
    - The role of adaptive basis projection in NNs.
    - The role of linear and logistic regression in NNs.
    - The relationship between radial basis networks and kernel basis functions and smoothing splines.
    - The relationship between weight decay in NNs and ridge regression.

Lesson 8: Advanced Model Selection

- Be aware of a number of additional statistical and information theoretic model selection and validation techniques and be able to apply them to real life problems.
- Understand the advantages and disadvantages of the different methods.

4.5.3 Existing Courses

- Introduction to Artificial Intelligence by Sebastian Thrun and Peter Norvig.
- Machine Learning by Andrew Ng. Again,

- Statistical Learning by Trevor Hastie and Rob Tibshirani.
- Neural Networks for Machine Learning by Geoffrey Hinton.
- Learning From Data by Yaser Abu-Mostafa.
- Machine Learning 1 - Supervised Learning by Charles Isbell from Georgia Introduction to Data Science by Bill Howe.
- Mining Massive Datasets by Jure Leskovec, Anand Rajaraman and Jeff Ullman from Stanford.
- *Supervised Learning, Unsupervised Learning, Reinforcement Learning 3-course Machine Learning Series and is offered at Georgia Tech via Udacity. https://www.udacity.com/wiki/ml*

4.5.4 Existing Materials

Materials will be developed for this module throughout the project.

4.5.5 Example Quizzes and Questions

Quizzes will be developed to target students' understanding of various topics, targeting the R language where code-based queries are required.

4.5.6 Description of Exercises

Exercises for this module will be planned developed in R as part of the course development work.

4.5.7 Further Reading

- Machine Learning, Neural and Statistical Classification http://www1.maths.leeds.ac.uk/~charles/statlog/ by D. Michie, D.J. Spiegelhalter, C.C. Taylor (eds)
- Bayesian Reasoning and Machine Learning http://www.cs.ucl.ac.uk/staff/d.barber/brml/ by David Barber
- Introduction to Machine Learning http://arxiv.org/pdf/0904.3664.pdf by Amnon Shashua
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction http://statweb.stanford.edu/~tibs/ElemStatLearn/ by Trevor Hastie, Robert Tibshirani, Jerome Fried
- An Introduction to Statistical Learning with Applications in R http://www-bcf.usc.edu/%7Egareth/ISL/ by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

4.6 Process Mining

4.6.1 Learning Objectives

After taking this course, the learner will:

- have a good understanding of Business Process Intelligence techniques (in particular process mining),
- be able to relate process mining techniques to other analysis techniques such as simulation, business intelligence, data mining, machine learning, and verification,
- be able to apply basic process discovery techniques to learn a process model from an event log (both manually and using tools),

- be able to apply basic conformance checking techniques to compare event logs and process models (both manually and using tools),
- be able to extend a process model with information extracted from the event log (e.g., show bottlenecks),
- have a good understanding of the data needed to start a process mining project,
- be able to characterize the questions that can be answered based on such event data,
- explain how process mining can also be used for operational support (prediction and recommendation), and
- be able to conduct process mining projects in a structured manner.

4.6.2 Syllabus and Topic Descriptions

They syllabus for this module is based on a MOOC that would run for 4 weeks. Students are expected to work on average 4 to 6 hours per week. Provisionally, the weekly main topics to be covered include:

- **Week 1 : Introduction, Process Modeling and Analysis (Chapters 1- 2)**
  - During this week, process mining is introduced and key concepts explored. In this week students learn about event logs, the input for process mining, and about Petri nets, the process modelling notation used to explain foundational concepts. This week also provides the theoretical foundations of process modeling and process discovery. In the next three weeks, students will use these concepts in a more applied setting.
- **Week 2 : From Event Logs to Process Models (Chapters 4-5)**
  - In this second week, a practical aspect of process mining is introduced. Students learn basic discovery algorithms (i.e. alpha-algorithm) to discover models from event logs. Students are introduced to the process of turning data (from various data sources) into proper event logs, needed for process mining. Moreover, challenges encountered with event logs such as noise and incompleteness are discussed.
- **Week 3 : Advanced Process Discovery Techniques (Chapters 6 - 8)**
  - In this week students learn even more process discovery algorithms. In the first part, the main focus is on conformance checking, i.e., aligning observed behavior with modeled behavior. This can be used for a wide variety of compliance questions: Where and why do people, machines, and organizations deviate? Students learn different ways of evaluating the conformance between a process model and the event log. The second part of lectures during this week explores different perspectives that can also be mined from event logs. Techniques for social network analysis, resource behavior and decision point are discussed.
- **Week 4 : Putting Process Mining to Work (Chapters 10 - 12)**
  - In this last week, the emphasis in on the application of process mining to real life use cases. We demonstrate how to conduct a process mining project from start to finish. We also discuss ProM and other available process mining tools that can be used to perform experiments.

4.6.3 Existing Courses

Table 8: Existing courses that include a module related to Process Mining

Institution	**Course**	**Module**	**URL**	**Target Audience**	**Languages and Data Types**
Eindhoven University of Technology	MOOC	Process Mining: Data Science in Action	https://class.coursera.org/procmin-001/wiki/General_Information	Graduates, Data Science Practitioners and Enthusiasts.	xes, mxml, csv
Eindhoven University of Technology	Preparation Master BIS	Business Process Intelligence	http://venus.tue.nl/owinfo-cgi/owi_0695.opl?vakcode=2IIE0&studiejaar=2014	Bachelor Students, basic computer science skills	Rapid Miner, Java ??
Eindhoven University of Technology	Master BIS	Advanced process mining	http://venus.tue.nl/owinfo-cgi/owi_0695.opl?vakcode=2II66&studiejaar=2014	Graduates, BIS Master students	-
Eindhoven University of Technology	MS Data Science (EIT ICT Labs Data Science)	Introduction to process mining	https://venus.tue.nl/owinfo-cgi/owi_0695.opl?vakcode=2IMI35&studiejaar=2015	Graduates – computer science, mathematics or industrial engineering	-
Universitat Politècnica de Catalunya	MS Statistics and Operations Research (Summer School)	Process Oriented Data Science	https://mesioupcub.masters.upc.edu/en/summer-school-2015/courses/process-oriented-data-science	Scientists and professionals in general.	-

4.6.4 Existing Materials

Related materials are already available on Coursera (https://www.coursera.org/course/procmin), and these can be adapted and modified for use in EDSA.

4.6.5 Example Quizzes and Questions

Provisionally, there will be

- Weekly quizzes to test knowledge of key process mining terminology and definitions
- Questions to assess understanding of Transition Systems and Petri Net Properties
- Questions to assess understanding of Alpha-algorithm and its limitations
- Questions on understanding alternative Process Discovery Techniques
- Questions on understanding of conformance checking (testing Alignment of Observed and Modeled Behavior)
- Questions on understanding of process mining techniques for different mining perspectives such as social network and decision point analyses.

4.6.6 Description of Exercises

- Applying process mining techniques to derive process models from an example of event log.
- Analyzing and visualizing process behavior using ProM or Disco. Apply the L* life cycle to carry out a process mining project from inception to completion.

4.6.7 Further Reading

- The textbook "W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011" (http://springer.com/978-3-642-19344-6) is advised as background information. It can be ordered via http://springer.com/978-3-642-19344-6, http://amzn.com/3642193447, or http://www.bol.com/nl/p/process-mining/1001004011199272/.
- W.M.P. van der Aalst. Process Mining. Communications of the ACM, 55(8):76-83, 2012.
- W.M.P. van der Aalst and C. Stahl. Modeling Business Processes: A Petri Net Oriented Approach. MIT press, Cambridge, MA, 2011.

4.6.8 Domain-Specific Curricula for Process Mining

With the process mining module, we have two additional curricula that form domain-specific versions of the course to show the flexibility of training that EDSA will offer.

Process Mining for the Internet of Things

Learning Objectives

After taking this course, the learner will:

- have a good understanding of the Internet of Things (IoT) and its Applications
- be able to describe the categories of IoT products (smart home, smart city, etc.)
- be able to describe the trends and opportunities offered by IoT applications
- have a good understanding of IoT-generated data within the Big Data Framework and related Challenges
- be able to identify and describe the role of process mining on event data generated from IoT
- be able to discuss the IoT processes
- be able to describe challenges associated with data access and analysis on IoT
- be able to conduct process mining projects for IoT using the L* life cycle model.

Syllabus and Topic Description

Provisionally, this MOOC will run for 4 weeks or less. Students are expected to work on average 4 to 6 hours per week. Provisionally, the weekly main topics to be covered include:

- **Week 1 : Introduction and Context**
  - In this week, students will be introduced to the main concepts of Internet of Things (IoT). We discuss current trends on IoT, opportunities and challenges encountered and/or foreseeable with IoT. We explore the research on IoT in light of the Big Data Framework.
- **Week 2 : IoT products: *smart wearable, smart home, smart city, smart environment, and smart enterprise***
  - In this second week, a generic description of IoT architecture is introduced. We also explore the main categories of IoT products and further discuss them. For each

category, we provide examples of applications and the type of data stored. We explore the analytics potentials on IoT data for each category. Moreover, we also discuss related potential security and privacy challenges related to data exploitation in these products.

- **Week 3 : Process Mining and IoT Processes**
    - In this week, a critical step towards analysing IoT generated data is discussed. First we review the basic elements of process mining and succinctly highlight relevant techniques useful for analysing IoT data. We then discuss the possibility of exploring and analysing IoT data with process mining. We give examples of IoT processes and the value-added opportunities offered by process mining. We further discuss challenges related to data access and collection and outline possible solutions.
- **Week 4 : Applying Process Mining to IoT Data : Use Cases**
    - In this last week, the emphasis is on the application of process mining to real IoT use cases. We explore selections of events on the IoT products (i.e. smart city, smart environment etc.) and demonstrate how process mining can be used to analyse this data. We then discuss any foreseeable data quality issues.

Existing Courses

No specific courses exist for process mining for IoT. However we list a number of relevant courses:

Table 9: Relevant courses for Process Mining in IoT

Institution	Course	Module	URL	Target Audience	Languages and Data Types
Eindhoven University of Technology	MOOC	Process Mining: Data Science in Action	https://class.coursera.org/procmin-001/wiki/General_Information	Graduates, Data Science Practitioners and Enthusiasts.	xes, mxml, csv
Queen Mary University of London	MSc Big Data	Sensors and the Internet of Things	http://www.qmul.ac.uk/postgraduate/coursefinder/courses/121386.html	Candidates with second-class degree or above (good 2:1 minimum for Industrial Experience option) in electronic engineering, computer science, mathematics, or a related discipline.	-
Futuretext.com	Certificate	Data Science for Internet of Things(IoT)	http://www.futuretext.com/datascienceiot2/	A typical student is a developer who has skills in programming environments like Java, Ruby, Python, Oracle etc. and wants to learn Data Science within the context of Internet of Things. Thus, there are broadly three learning domains: Computer Science, Mathematics and IoT	-

Quizzes and Questions

Provisionally, there will be:

- Weekly quizzes to test understanding of process mining for IoT concepts.
- Questions to assess understanding of IoT applications and research opportunities
- Questions to evaluate the understanding of privacy and security challenges posed by IoT infrastructures

- Questions on assessing the understanding of the IoT processes
- Questions to assess understanding of application of process mining to IoT event data for smart city, smart home, smart environment etc.

Description of Exercises

- Step by step exercise on identifying processes on IoT data.
- Applying process mining techniques to derive process models from an example of event log (IoT process).
- Analyzing and visualizing process behavior using ProM or Disco. Apply the L* life cycle to carry out a process mining project from inception to completion.

Further Reading

- The Textbook: "Vermesan, O., & Friess, P. (Eds.). (2013). *Internet of things: converging technologies for smart environments and integrated ecosystems*. River Publishers."
- The Textbook:" Vermesan, O., & Friess, P. (Eds.). (2014). Internet of Things-From Research and Innovation to Market Deployment (pp. 74-75). River Publishers."
- The textbook "W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011" (http://springer.com/978-3-642-19344-6) is advised as background information. It can be ordered via http://springer.com/978-3-642-19344-6, http://amzn.com/3642193447, or http://www.bol.com/nl/p/process-mining/1001004011199272/.

Process Mining in Healthcare

Learning Objectives

After taking this course, the learner will:

- have a good understanding of Challenges in Healthcare
- have a good understanding of Healthcare processes
- be able to describe the opportunity and need to apply process mining to healthcare processes
- be able to describe the healthcare reference model
- be able to apply process mining techniques to healthcare data in light of the healthcare reference model
- have a good understanding of the data needed to start a process mining project in the context of healthcare data,
- be able to characterize the questions that can be answered based on such event data,
- be able to describe possible data quality issues that can be encountered in healthcare event data
- be able to conduct process mining projects in healthcare using the L* life cycle model.

Syllabus and Topic Description

The MOOC will run for 4 weeks. Students are expected to work on average 4 to 6 hours per week. Provisionally, the weekly main topics to be covered include:

- **Week 1 : Introduction and Context**
    - Healthcare costs have increased dramatically and the demand for high quality care will only grow in our aging society. At the same time, more event data are being collected about care processes. During this first week, students are introduced to the opportunities and challenges from Healthcare that can be exploited using process mining. The landscape for the reminder of the course is set by briefly describing healthcare processes, process mining and its application to these processes.
- **Week 2 : Process Mining and Healthcare Processes**
    - In this second week, a generic description of key process mining elements is presented first. We review the concept of Event data and process models, the types of process mining and relevant techniques, the process mining spectrum. We also discuss the tools available for using process mining. Then, Healthcare processes are discussed with a view on different levels of care, classification of healthcare processes as well as 4 main types of questions that can be answered with healthcare event data.
- **Week 3 : Healthcare Reference Model**
    - In this week, a critical step towards analysing healthcare data is discussed. Although the Hospital Information System (HIS) contains all the data needed for analysis, locating and extracting this data is either complex or too costly for the analyst. Therefore, we discuss a healthcare reference model whose goal is to locate event data easily and to support data extraction. The healthcare reference model was developed based on an analysis of the available data in several Dutch hospitals. We detail the different classes of the model and also investigate whether it is representative for data present in other hospitals.
- **Week 4 : Applications of Process Mining and Data Quality Issues**
    - In this last week, the emphasis is on the application of process mining to real healthcare use cases. We explore selections of events on healthcare data, identification and quantification of deviations and bottlenecks on healthcare processes. We also demonstrate how to compare healthcare processes. Moreover, we discuss some of the issues related to data quality that can be encountered on healthcare Event data. We classify them and explore guidelines for improving data.

Existing Courses

While there are no specific courses on process mining in healthcare to refer to, we list a number of broad courses on process mining:

Table 10: Relevant courses for Process Mining in Healthcare

Institution	Course	Module	URL	Target Audience	Languages and Data Types
Eindhoven University of Technology	MOOC	Process Mining: Data Science in Action	https://class.coursera.org/procmin-001/wiki/General_Information	Graduates, Data Science Practitioners and Enthusiasts.	xes, mxml, csv
University of California, Davis	Healthcare Analytics Certificate Program	Data Mining for Healthcare Analytics	https://extension.ucdavis.edu/section/data-mining-healthcare-analytics	Candidates with prior professional experience in a healthcare setting and familiarity with statistics	
Georgia Institute of Technology	Healthcare Informatics & Data Analytics Specialization (Coursera MOOCs)	Big Data Analytics for Healthcare	https://www.coursera.org/specialization/medicaltech/30/overview	Students with general knowledge of computer science and at least moderate programming skills	??
The College of St. Scholastica	MOOC	Health Data Analytics	https://www.css.edu/graduate/non-degree/massive-open-online-courses/health-data-analytics-mooc.html	N/A	Rattle and R
Union Graduate College	Online MS in Healthcare Data Analytics	Hospital Analytics	http://www.uniongraduatecollege.edu/management/MS-Healthcare.aspx	Individuals working with large amounts of healthcare such as Healthcare Consultancy Analyst, Application Developer for a health insurance company.	R, JMP, RapidMiner.

Quizzes and Questions

Provisionally, there will be:

- Weekly quizzes to test knowledge of key challenges and opportunities for Big Data in Healthcare data
- Questions to assess understanding of application of process mining to Healthcare data
- Questions to assess understanding of Healthcare processes
- Questions to evaluate the understanding of different levels of care in healthcare processes, classification of these processes as well as the types of questions on healthcare processes
- Questions on describing the healthcare reference model
- Questions on assessing the process of validation for the healthcare reference model
- Question on assessing selections of events on Healthcare event data
- Questions on identifying and quantifying deviations and bottlenecks on healthcare event data.
- Questions on issues related to data quality that can be encountered on healthcare Event data. We classify them and explore guidelines for improving data of quality.

Description of Exercises

- Making use of the Healthcare Reference Model to construct event logs

- Applying process mining techniques to derive process models from an example of event log (healthcare process).
- Analyzing and visualizing process behavior using ProM or Disco. Apply the L* life cycle to carry out a process mining project from inception to completion.

Further Reading

- The Textbook: "Ronny S.. Mans, & van der Aalst, W. (2015). Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Springer."
- The textbook "W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011" (http://springer.com/978-3-642-19344-6) is advised as background information. It can be ordered via http://springer.com/978-3-642-19344-6, http://amzn.com/3642193447, or http://www.bol.com/nl/p/process-mining/1001004011199272/.
- W.M.P. van der Aalst. Process Mining. Communications of the ACM, 55(8):76-83, 2012.
- W.M.P. van der Aalst and C. Stahl. Modeling Business Processes: A Petri Net Oriented Approach. MIT press, Cambridge, MA, 2011.

5. Conclusions

As the proliferation of data increases, so too does the demand for skilled data scientists who are trained and ready to tackle many of the problems related to big data, as well as to drive new insights and value from this data across all sectors. EDSA aims to address this skills demand by providing modular and adaptable learning curricula that can be tailored to suit particular needs depending on sector, job role and past experience. Therefore the curriculum offered here by EDSA covers the topic of Data Science from a broad perspective, taking into account the fundamental foundations behind much of the subject matter, and then covering the vital components of the Data Science process including data storage and processing, data analysis and data visualization.

This deliverable initially provides an overview of this curriculum, focusing on the 15 topics that we have assigned to be the first version of our 'core' curriculum. These cover the stages described above, and therefore when followed sequentially can gradually build up a student's knowledge of the entire Data Science process. We then provide individual module curricula for the first 6 of these 15 modules: Foundations of Data Science, Foundations of Big Data, Big Data Architecture, Distributed Computing, Machine Learning and Process Mining.

In later versions of the curriculum, we will provide further modules to complete the overall 15-topic curriculum, following the schedule outlined above. By skewing modules towards the earlier deliverables, we allow time for revising those modules that have already been designed and developed at later stages to take into account any changes in demand, or based on community feedback and suggestions. In line with each of these three versions of the curriculum, learning resources will be created that offer hands-on interactive exercises alongside slides, webinars, videos and other materials. These will be integrated into a number of formats such as MOOCs and eBooks, and we will design learning paths through the material for different audience groups to traverse, ensuring that they reach the content that is most relevant and appropriate for them. This will involve producing sequences of material that may be from a more 'traditional' module for one topic, and then using something different such as a MOOC for another, therefore consideration will be made regarding how best to sequence these course elements.