EDISON

Education for Data Intensive Science to Open New science frontiers

Project no. 675419
Coordination and Support Action
Funded by the Horizon 2020 Framework Programme of the European Union

Call identifier: **H2020-ICT-2015-1**
Topic: **INFRASUPP-4-2015 - New professions and skills for e-infrastructures**
Start date of project: **1 September 2015 (24 months duration)**

Deliverable D2.2

Existing educational and training resources inventory and analysis

Due date: 31/05/2016

Submission date: 31/05/2016

Deliverable leader: UiS

Dissemination Level

☒	PU:	Public
☐	PP:	Restricted to other programme participants (including the Commission Services)
☐	RE:	Restricted to a group specified by the consortium (including the Commission Services)
☐	CO:	Confidential, only for members of the consortium (including the Commission Services)

EDISON project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675419

List of Contributors

Participant	Short Name	Contributor
University of Amsterdam	UvA	Yuri Demchenko, Adam Belloum, Spiros Koulouzis
Engineering - Ingegneria Informatica s.p.a.	ENG	Andrea Manieri, Giulia Patucca
University of Stavanger	UiS	Tomasz Wiktorski, Aleksandra Królak, Anousheh Shirazi
Research Institute for Telecommunication and Cooperation	FTK	Holger Brocks
University of Southampton	SOUTHAMPTON	Steve Brewer

Change history

Version	Date	Partners	Description/Comments
0.1	15/02/2016	UiS	Initial draft, incorporated information from several separated sources
0.2	8/03/2016	UiS	Documentation of Inventory and standardization work
0.3	1/04/2016	UiS	Results from inventory analysis, initial input on taxonomy
0.4	15/04/2016	UiS	Major updates to inventory analysis
0.5	30/04/2016	UvA, UiS	Automated taxonomy analysis
0.6	9/05/2016	UiS	Learning Outcomes and mapping to taxonomy
0.7	10/05/2016	UvA	Updated taxonomy analysis and extraction from the text, updated mapping of professions
0.8	11/05/2016	UiS, UvA, FTK, SOUTHAMPTON	Various improvements to all sections
0.9	13/05/2016	UiS	Final draft for internal review
1	31/05/2016	UiS, UvA	Final version

Executive summary

This document contains an overview and analysis of available educational and training resources covering Data Science and related subjects. It is important to understand the current state of Data Science education to properly plan what areas should be addressed in later work on Model Currricula, Community Portal and other EDISON tasks. The inventory of resources that was created in this task has a value of its own for the community and it paves the way for a common interchange format about programs and courses in Data Science with external partners, such as RDA, Elixir, CODATA, etc.

The primary focus of our analysis were academic programs, but academic and industrial courses, books, and other training materials where included when relevant. The initial iteration was gathered by WP2 partners in collaboration with partner universities (including Champion universities) and also with contributions from the ELG and RDA IG-ETRD. The resulting inventory was then made publicly available which lead to its extension both in terms of the amount of entries and also their quality. As a result, the EDISON Inventory is already one of the most comprehensive catalogue of information on available educational resources in Data Science and related subjects. The inventory analysis so far suggests that existing programs are poorly balanced w.r.t. Data Science competence groups and one of the reasons might be the lack of cross-department collaboration in creating the programs. Learning outcomes are seldom specified and even when they are present, educational theory (e.g. Bloom's taxonomy) is usually not considered when defining them. The approach of the analysis and detailed results are described in Section 2, Section 3 and Appendices A-C.

We further extended work (started in D2.1) on the taxonomy of Data Science based on Inventory analysis, Data Science Competence Framework CF-DS, (D2.1) providing enumerated list of Data Science competences by competence groups, an early version of the Data Science Body of Knowledge (T2.3), and analysis of skills in job advertisements. The further definition of the Data Science professional profiles is complemented with the definition of the corresponding competences according to CF-DS competence groups. The extension also reflects differences in the related professions like Data Scientist, Data Analyst, Data Engineer, Data Steward, Scientific Data or e-Infrastructure manager, etc. In relation to CF-DS we defined corresponding learning outcomes specific to Data Science and mapped them to the taxonomy. This work is described in Section 4.

Finally, we performed a gap analysis between the current state of Data Science educational offerings as expressed by Inventory and requirements originating from CF-DS and learning outcomes. Results suggest that the Data Science Model Curriculum should be competence-based and flexible in terms of specific technologies and courses. It is necessary to include courses that connect competence from all three CF-DS competence groups early in the education process. There should be a focus on assessment methods used to achieve a higher level of knowledge necessary for Data Scientists (especially on graduate level). Programs should be a result of cross-department collaboration. Computing and programming competences together with domain knowledge should be given proper coverage, an aspect missing in majority of existing programs.

TABLE OF CONTENTS

List of figures

List of tables

1 Introduction

This deliverable contains an overview and analysis of available educational and training resources on Data Science and related subjects. It also contains a high-level taxonomy of Data Science, including a mapping to learning outcomes.

The deliverable and related task descriptions define, in the context of earlier developments in EDISON, the following goals for the work described in this document:

- to gather data on existing programs, academic and industrial courses, books and also other training resources (with primary focus on European offerings, but also covering representative inputs from North America, Asia, and other regions);
- to ensure the relevance of inventory work in collaboration with external parties, in particular ELG and RDA;
- to analyse Inventory data to identify common patterns and important gaps based on project developments and educational theory in order to provide a basis for Model Curricula, Body of Knowledge, and other efforts in the project;
- to extend the work on the taxonomy of Data Science by considering existing taxonomies beyond ACM, a family of Data Science professions, available education programs, required skills as expressed in job advertisement and related domain taxonomies;
- to work within the context of education theory by mapping CF-DS and taxonomy to learning outcomes.

The deliverable is organized as follows. In Section 2 we describe the work on creating and populating the EDISON Inventory. We start by explaining the organization of the Inventory and methods for populating it with content. Furthermore, we detail EDISON's initiative to standardize the exchange format for Data Science educational offerings, performed in collaboration with RDA, CODATA, Elixir, and including inputs from the APARSEN and FOSTER projects. We also describe the ongoing effort of providing the Inventory as a community service. The Inventory is hosted online but we also include a snapshot in Appendices A-C for illustration purposes.

In Section 3 we describe the analysis of the inventory. We perform quantitative analysis of several aspects including: origin, source, coverage of domain knowledge, naming of programs, etc. We also perform qualitative analysis of selected degree-giving programs and other resources w.r.t. CF-DS, Bloom's taxonomy and constructive alignment.

In section 4 we describe on-going work on Data Science taxonomy. In particular, we discuss how different professions from the Data Science family of professions could be addressed in the taxonomy. We also create learning outcomes for Data Science programs and courses based on CF-DS and taxonomy.

In section 5 we summarize findings from Inventory work and identify gaps in the context of requirements resulting from earlier-defined learning outcomes and CF-DS. It leads to a set of recommendations for further development of Data Science curricula.

2 Inventory

To best support Data Science education in the future, we first have to fully understand the current landscape of existing programs, academic and industrial courses (subjects), and books. There exist several lists of programs and courses, some of which we mention later, but they usually only list the name of the program and institution. There also seems to be little quality control over inclusion in these lists and there is no detailed information on how the programs or courses are actually constructed. These shortcomings make it impossible to understand current state of Data Science education. For instance, how well various programs cover necessary competences. In this section, we describe our work aimed at filling this gap.

Please note that we use the word *course* to mean a single subject, and the word *program* to mean a set of courses usually leading to a degree or a certificate. It does not mean that this is the only correct way that we suggest to the community. Use of these words has geographical connotations and we settle on these definitions only for the clarity in EDISON documents.

2.1 Organization of inventory

The EDISON inventory resides primarily online to allow for frequent updates. We also present a snapshot per submission date of deliverable in appendices of this document for reference.

The EDISON inventory contains information about:

- academic programs
- academic courses
- industrial courses
- books

We include MOOCs in our analysis as a part of academic courses, which they usually are. There are no full academic programs offered as MOOCs, but we notice early attempts at providing short, focused programs.

Industrial course differ from academic by usually being offered by non-academic institutions. They also strictly focus on developing technical skills, related to a relatively narrow set of technologies. In contrast to academic courses that aim at more general skill development.

The following data elements were collected (when available) for each program:

1. Name of program
2. University
3. Country
4. Unit (such as faculty or department)
5. Language of instruction
6. Level (such as bachelor, master, or doctoral)
7. Title awarded (if any)
8. Link to program website
9. Abstract (short description of the program as provided by university)

All these data are made available publically on the EDISON project website and everybody interested is allowed to submit request for updates as is further described in section 2.4. In addition, we have collected the following data that is not published but was also used for analysis:

10. Contact person (name, email)
11. Degree of coverage of competence groups (domain knowledge, data analysis, computer engineering)
12. Learning Outcomes (if specified)

Additional data are a basis for taxonomy work and inventory analysis, described in later sections of this report.

For academic and industrial courses we have collected:

1. Organization (university and unit for academic courses)
2. Course title
3. Level
4. Language (academic only)
5. Link

For books we have collected:

1. Title
2. Authors
3. Main topics and technologies
4. ISBN
5. Link

We have collected information about more than 300 programs (divided roughly equally between European and non-European), and about more than 100 industrial and academic courses (excluding usually those from programs). The Inventory is open for new inputs beyond the duration of this task, through EDISON website.

2.2 Methods for population

Populating the EDISON Inventory is a continuous process, in which we aim to engage the Data Science community in a way that is independent of the interests of any particular organization. Nevertheless, it is important to provide an initial critical mass of content, on the one hand to support immediate project needs, and on the other hand to position the EDISON Inventory on the forefront of similar resources.

The core of the population process was performed through a web search based on a set of keyword terms with relation to data science. These terms included, but we're not limited to: data science, machine learning, data analysis/analytics, data mining, business intelligence, business analysis/analytics. Each entry was analysed w.r.t. its contents to determine whether it should be included. At this stage, we focused on the goals of a particular program originating from a description and marketing of the program.

Further, the Inventory was extended through a network of partners with knowledge about specifics of educational system in various, especially European, countries. Due to language difference such offerings might be underrepresented in a general search. Inventory was also presented to the EDISON Liaison Group in order to elicit additional contributions. It makes it a useful reference for the community and a solid basis in the context of the project to support gap analysis and creation of model curricula. Any missing offerings should not qualitatively change the outcomes of the analysis.

We excluded many generic programs in computer science or information science with only minor elements of data analysis or domain knowledge. Such programs are unfortunately common in other non-curated lists. While such programs might with time develop toward Data Science direction, they do not a provide meaningful basis for analysis of actually existing Data Science programs. Finally, we excluded many programs not granting a degree. We note that they usually are simply ad-hoc offerings with limited importance from a perspective of proper curriculum development. However, a significant amount of such offerings provides yet another important signal about the growing importance of Data Science education.

While the breadth of the coverage was important, we simultaneously focused on depth of each entry, in particular, in analysis of content of each program w.r.t Data Science competence groups and the detailed definition of intended learning outcomes (sometimes also called objectives or goals).

The depth and quality of coverage stands in contrast to other existing lists, such as the „Colleges with Data Science Degrees“[1], which is also the most comprehensive. That list offers greater breadth than ours since it has been compiled an over longer period of time. At the same time, it allows for non-curated inclusion of the programs, what results in large percentage of programs very remotely related to data science. Moreover, the coverage of degrees from outside United States (and partially UK) is severely limited. Due to the origin of the EDISON project, European offerings were given high priority in our work.

Available lists of other types of resources e.g. courses or books, were not nearly as comprehensive as program lists. We have created inventory of these resources following similar methodology as applied for programs. We are not aware of any other offering comparable to ours.

2.3 Common interchange format

The goal of this format is to simplify the gathering and publishing of information about courses and programs in data science and related domains. It is purposefully very generic to accommodate for a wide variety of courses, both regular and one-off.

We included the fields that we consider to be important both for general informational purposes but also from the perspective of education theory and alignment with EDISON's Data Science Competence Framework (CF-DS), Body of Knowledge (CF-BoK), RDA EU's Training Specification, CODATA and Elixir. In case the content for some of the fields might be difficult to obtain for all courses and programs, we suggest keeping them in the format but making them optional.

The tables for Courses and Programs are identical except for "Name of Presenter(s)" and "Related Program(s)" that are only present in the Courses table, and "Track Name" and "Course List" that are only present in the Program list.

Cross-organization and cross-project agreement on the contents of the interchange format can enable a qualitative improvement in sharing information about available programs and courses. In order to further facilitate this development we plan to provide an extended technical specification with a sample implementation as a part of efforts in WP3.

We identified and reviewed four existing standards to determine to what extent they could cover some of identified needs. iCalendar was reviewed based on RFC5545[2]. Schema.org/event was reviewed based on schema.org website[3]. XCRI-CAP was reviewed based on a summary spreadsheet of PG XCRI-CAP[4]. LRMI was reviewed based on dublincore.net website[5].

A review of the existing standards demonstrates that no single standard would fully cover all the requirements specified by project partners and organizations included during the consultation process. In some cases, there is no explicit coverage for a particular field, but there are some closely related fields, which is reflected in the presented tables.

It is important to represent relevant information with the iCalendar standard. Such approach would facilitate an easy import into various calendar applications common today.

Schema.org/event does not offer advantages over iCalendar. Considering the iCalendar adoption in calendar applications, schema.org/event does not seem to be useful for our purpose.

XCRI-CAP covers the majority of requested fields. However, after an initial review, it seems to be a fairly complicated standard. Despite the fact most fields we care about are covered, this coverage is often indirect or requires additional structures and information which do not seem to be necessary for our purpose. The complexity of XCRI-CAP might be a hurdle in adoption, especially for educational and training purposes.

LRMI does not cover as many fields as XCRI-CAP; however, it seems to be more straight-forward to use due to its structure and also its relation to widely accepted developments in Dublin Core.

In Table 1 we present a list of fields with description for information exchange about courses, in Table 2 the list for programs is given. First column specifies field name, second whether a particular field is mandatory, recommended, or only optional. In further four columns we indicate if information carried by the field is already covered in major related standards: LRMI, XCRI-CAP, Schema.org/event, and iCal.In the last column a short description is provided.

We recommend that LRMI be extended further, including some form of integration with iCalendar simultaneously. As for the semantic specification of such approach, an expert opinion should be sought. It is recommended as a part of development of community portal in WP3.

Table 1 Fields for information exchange about courses (subjects)

Field Name	Mandatory Recommended Optional	iCal	Schema.org/event	XCRI-CAP	LRMI	Description
Title	Mandatory	+	+	+	+	A meaningful short title
Name of Presenter(s)	Optional	-	+	-	+	A person of a list of people delivering the course, with their affiliations
Organizer	Mandatory	-	+	+	+/-	Institution, company, project organizing the course
Type of Course	Mandatory	-	-	+	+	Webinar, academic course, …
Related Program	Recommended	-	-	+/-	-	URI(s) to programs(s) this course is a part of
Location	Mandatory	+	+	+	-	A country and city (or full address) where course takes place, unless online
Start Date and Time	Mandatory	+	+	+	-	The start date and time of the item (in ISO 8601 date format[6], preferably in UTC with time offset to local time zone).
End Date and Time	Mandatory	+	+	+	-	The end date and time of the item (in ISO 8601 date format[6], preferably in UTC with time offset to local time zone).
URL	Mandatory	+	+	+	+	Link to further information
Contact	Mandatory	+	+/-	+/-	-	A person/email that should be used for contacting
Language	Mandatory	-	+	+	+	Language of instruction
Level	Optional	-	-	+/-	-	Which level of studies following either Bologna[7] or US approach
Credit	Recommended	-	-	+	-	Recommended for academic courses, including grading system
Prerequisites	Recommended	-	-	+	+/-	Required prior knowledge, preferably based on a EDISON Body of Knowledge or Taxonomy
Target Audience	Optional	-	-	-	-	E.g. “social scientists”, “biologists”, “data managers”, “policy makers in the UK”, or other
Knowledge Areas	Recommended	-	-	+	+/-	Knowledge areas covered by the course, preferably based on a EDISON Body of Knowledge or Taxonomy
Learning Outcomes	Recommended	-	-	+	+/-	Including objectives, preferably based on a EDISON Competence Framework
Description	Recommended	+	+	+	+	E.g. The course will provide a strong basis in administrative, programing, and algorithm design aspects of data intensive systems.
Registration Deadline	Optional	-	-	+/-	-	The date and time of the item (in ISO 8601 date format, preferably in UTC with time offset to local time zone).
Payment	Optional	-	-	+	-	Use three letter currency symbols (in ISO 4217 format[8]) and payment methods

Table 2 Fields for information exchange about programs

Field Name	Mandatory Recommended Optional	iCal	Schema.org/event	XCRI-CAP	LRMI	Description
Title	Mandatory	+	+	+	+	A meaningful short title
Track Name	Optional	-	-	-	-	Name of the track within the program
Course List	Recommended	-	-	+/-	+/-	URI to courses being part of the program, limited to the track if specified
Organizer	Obligatory	-	+	+	+/-	Institution, company, project organizing the course
Type of Program	Mandatory	-	-	+	+	Summer school, academic program, …
Location	Mandatory	+	+	+	-	A country and city (or full address) where the course takes place, unless online
Start Date and Time	Mandatory	+	+	+	-	The start date and time of the item (in ISO 8601 date format[6], preferably in UTC with time offset to local time zone).
End Date and Time	Mandatory	+	+	+	-	The end date and time of the item (in ISO 8601 date format[6], preferably in UTC with time offset to local time zone).
URL	Mandatory	+	+	+	+	Link to further information
Contact	Mandatory	+	+/-	+/-	-	Contact information of the responsible party (name, email or phone number)
Language	Mandatory	-	+	+	+	Language of instruction
Level	Optional	-	-	+/-	-	The level of studies following either Bologna[7] or US approach
Credit	Recommended	-	-	+	-	Recommended for academic courses, including grading system
Prerequisites	Recommended	-	-	+	+/-	Required prior knowledge, preferably based on a BoK or taxonomy
Target Audience	Optional	-	-	-	-	E.g. “social scientists”, “biologists”, “data managers”, “policy makers in the UK”, or other
Knowledge Areas	Recommended	-	-	+	+/-	Knowledge areas covered by the course, preferably based on the EDISON Body of Knowledge or Taxonomy
Learning Outcomes	Recommended	-	-	+	+/-	Including objectives, preferably based on the EDISON Competence Framework
Description	Recommended	+	+	+	+	E.g. The course will provide a strong basis in administrative, programing, and algorithm design aspects of data intensive systems.
Registration Deadline	Optional	-	-	+/-	-	The date and time of the item (in ISO 8601 date format, preferably in UTC with time offset to local time zone).
Payment	Optional	-	-	+	-	Use three letter currency symbols (in ISO 4217 format[8]) and payment methods

2.4 Inventory as a community service

The initial version of the EDISON Inventory was constructed by EDISON partners as an internal tool. However, to extend the inventory's impact and improve its quality it was advantageous to publish it online in an interactive version. This way the Data Science community can not only get a better overview of existing resources, but also directly contribute to this shared asset.

The EDISON website displays information about University and other programs that have been captured in this task, as is presented in Figure 1. Programs can currently be found under the top-level menu: Library/Discussion documents as a University program list. This may change over time as the website evolves but it will be easy to locate.

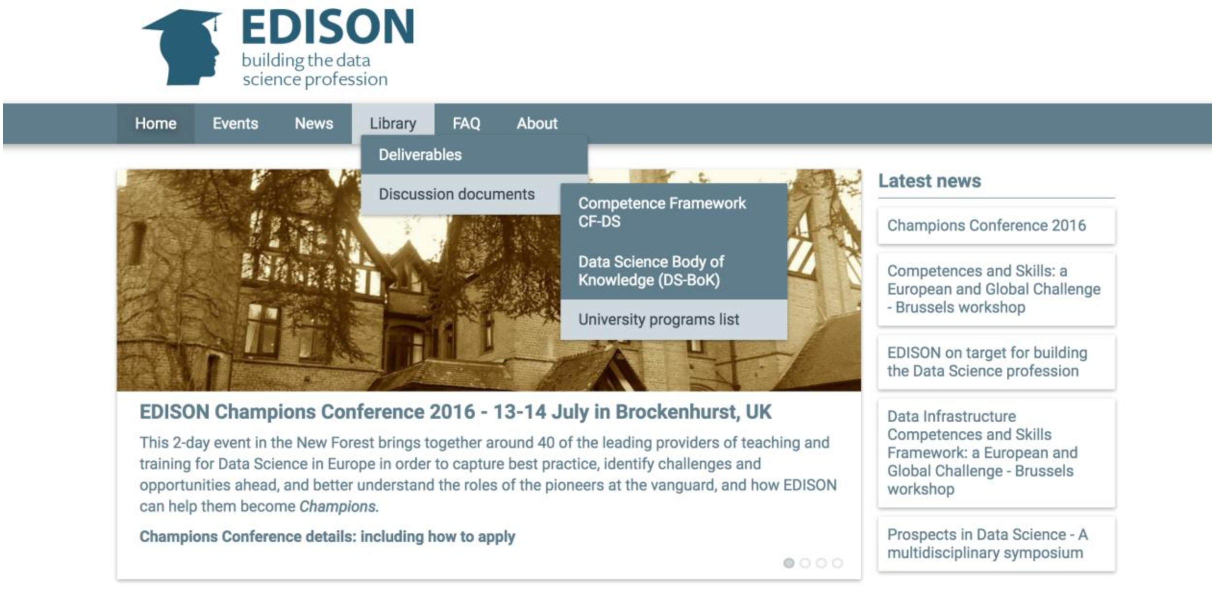


Figure 1 Location of university program list

By default, the screen currently displays the first page of a list of all courses that have been recorded in the system. This list can be paged through using the navigational buttons on the screen. The list can be filtered using two filters: country and language, where language refers to the language in which the course is delivered, presented in Figure 2. Users can also perform a search using the title field.

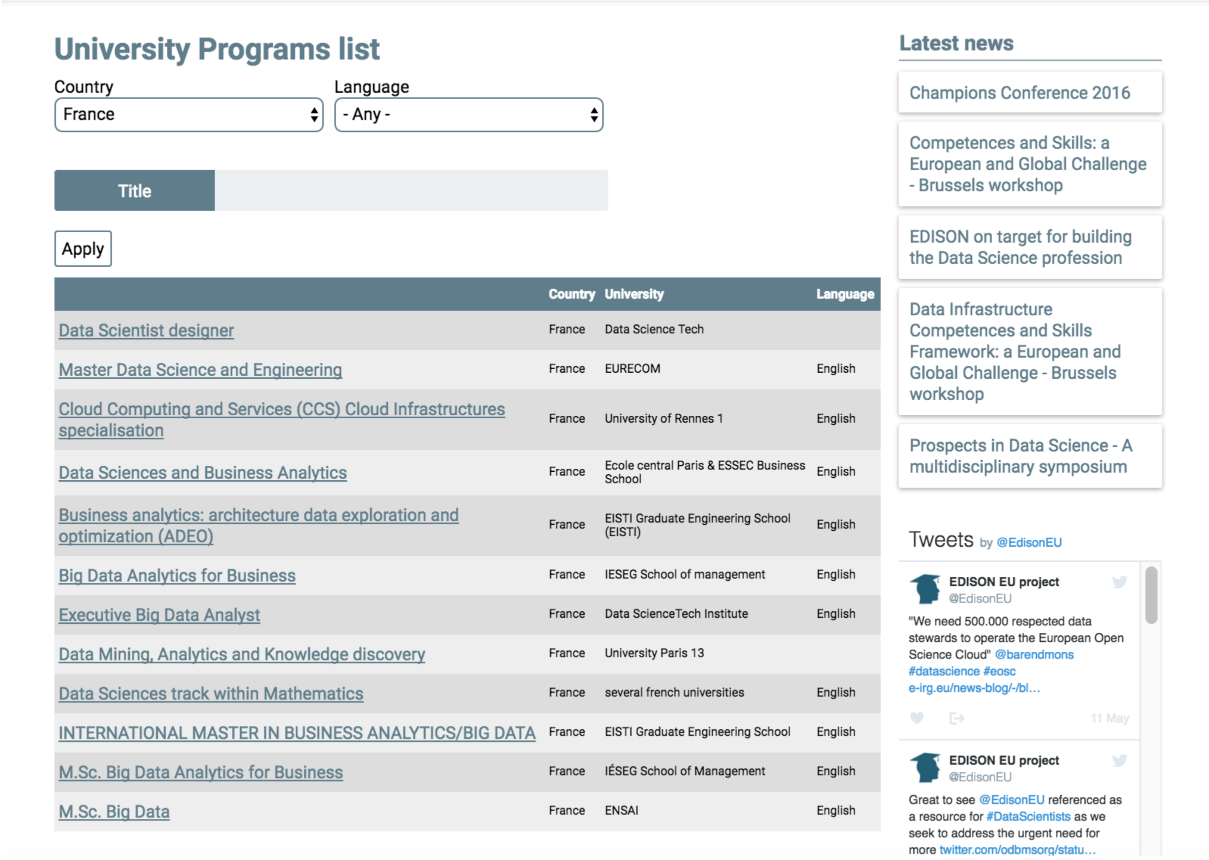
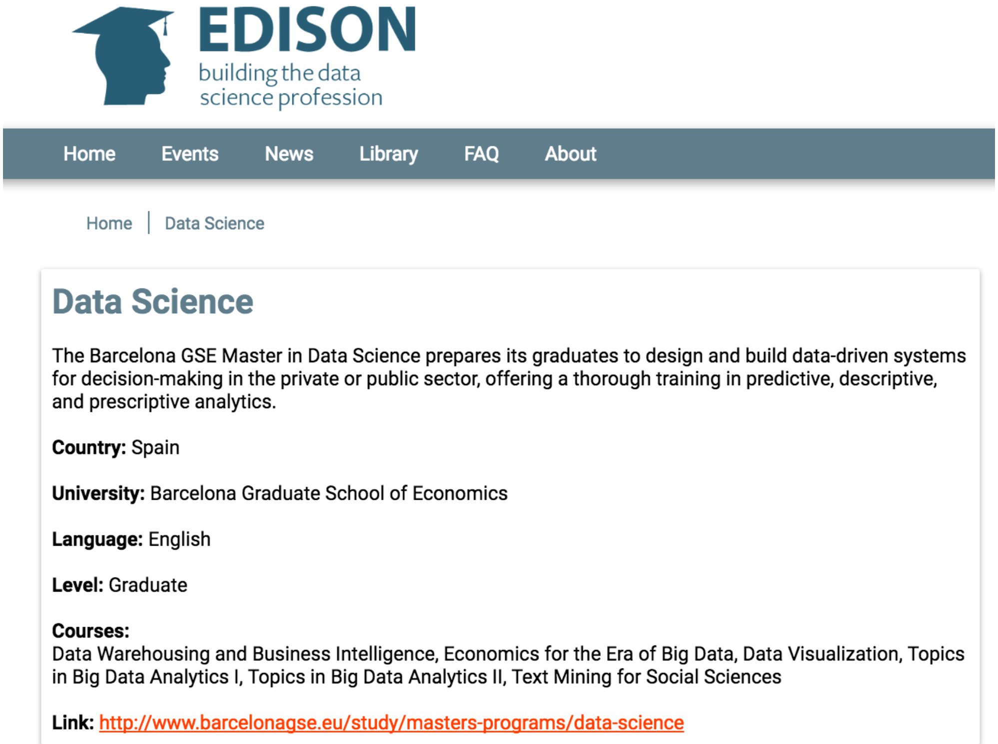


Figure 2 University programs list with filtering options

Clicking on the title of a particular item in the list brings up the details of the course, presented in Figure 3. For editors of the system, the list can be updated by adding more content of the type “University programs”. Users with the appropriate rights can create this content type and fill in the fields. To edit or remove existing material, editors need to select content and filter the list of all content in the system to view only “University Programs”, presented in Figure 4. Material can also be filtered by content status. For non-admin/editor users of the site there will be a request to email us with suggestions for changes and additions.



EDISON
building the data science profession

Home Events News Library FAQ About

Home | Data Science

Data Science

The Barcelona GSE Master in Data Science prepares its graduates to design and build data-driven systems for decision-making in the private or public sector, offering a thorough training in predictive, descriptive, and prescriptive analytics.

Country: Spain

University: Barcelona Graduate School of Economics

Language: English

Level: Graduate

Courses:
Data Warehousing and Business Intelligence, Economics for the Era of Big Data, Data Visualization, Topics in Big Data Analytics I, Topics in Big Data Analytics II, Text Mining for Social Sciences

Link: http://www.barcelonagse.eu/study/masters-programs/data-science

Figure 3 Details of a program in the Inventory

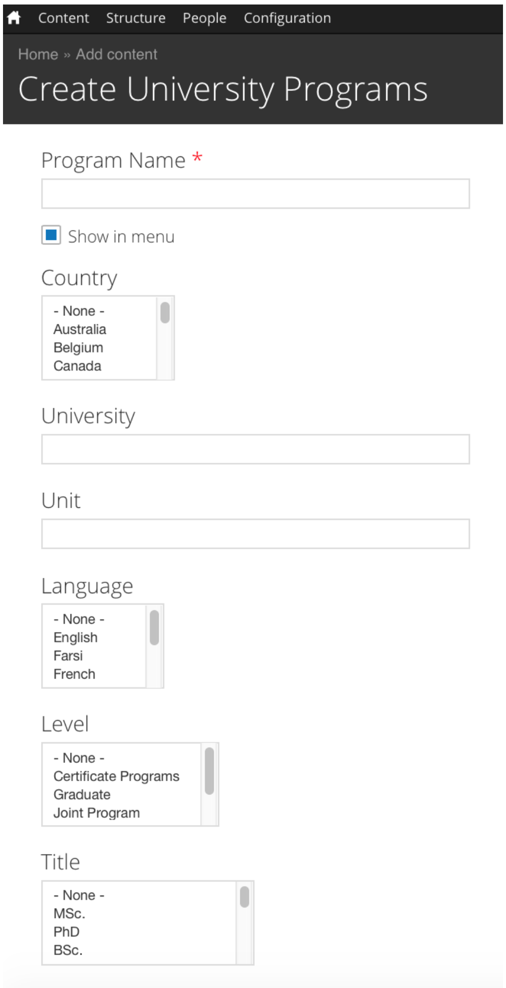


Figure 4 Adding a new program to Inventory

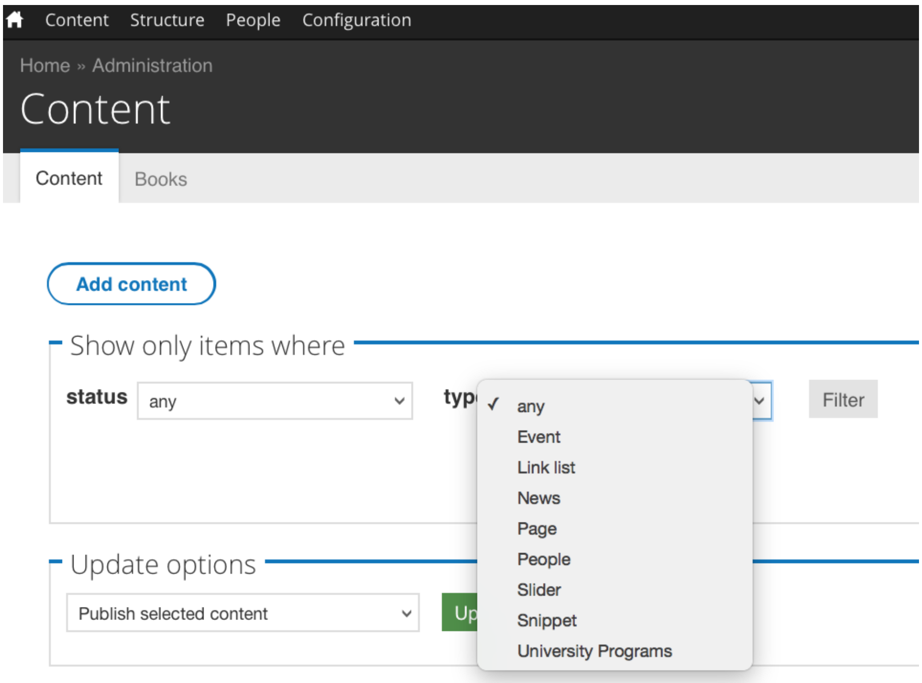


Figure 5 Filtering university programs for editing

2.5 Summary

In this section we described the process of creating EDISON Inventory of Data Science education resources. Main focus of Inventory was on academic programs because analysis of existing programs is an important component for designing Model Curricula. We also included other resources such as academic and industrial courses, which become useful in further work in WP3.

At the same time, we started work on an exchange format for information about Data Science education and training, with partners including RDA, CODATA and Elixir. An agreement was reached regarding necessary fields. Generic standards for educational information exchange were identified and reviewed to determine to what extent they could cover our needs. Based on this analysis we recommend extending LRMI (Learning Resource Metadata Initiative) standard in further work in WP3.

Finally, Inventory of programs was published as a service to Data Science community. It is also open for correction and inclusion of new entries.

3 Inventory Analysis

This section describes common conceptual elements and gaps among the present educational offerings in the EDISON Inventory and compares it with requirements given by the Data Science Competence Framework (CF-DS) from T2.1. Identified gaps are subsequently analysed and reformulated using Bloom's Taxonomy of Learning to ensure complete coverage of cognitive domains – from remembering to creating. We also suggest appropriate forms of teaching-learning activities and examination forms for various competences based on the theory of Constructive Alignment.

The analysis presented in this chapter is based on the initial population of the Inventory. Thanks to the interchange format and community portal to be developed in WP3, the contents of the Inventory will be extended and the analysis can be updated. The update can serve two purposes. First, it will provide an even deeper picture of the Data Science domain; second, by comparison it can show developments in the domain, measuring the impact of EDISON and other initiatives.

3.1 Competence framework and other basis for analysis

3.1.1 Data Science Competence Framework

The basis for quantitative analysis of entries in the EDISON Inventory is the Data Science Competence Framework (CF-DS) originating in the EDISON Deliverable 2.1 "Data Scientist Competences and Skills Framework (CF-DS) and BoK definition (first version)". In particular, we have related the contents of programs in the Inventory with the Data Science Competence Groups as defined in Section 4.4 of the aforementioned deliverable and visualized in Figure 8(a) there. We reproduce that visualization for reference in Figure 6.

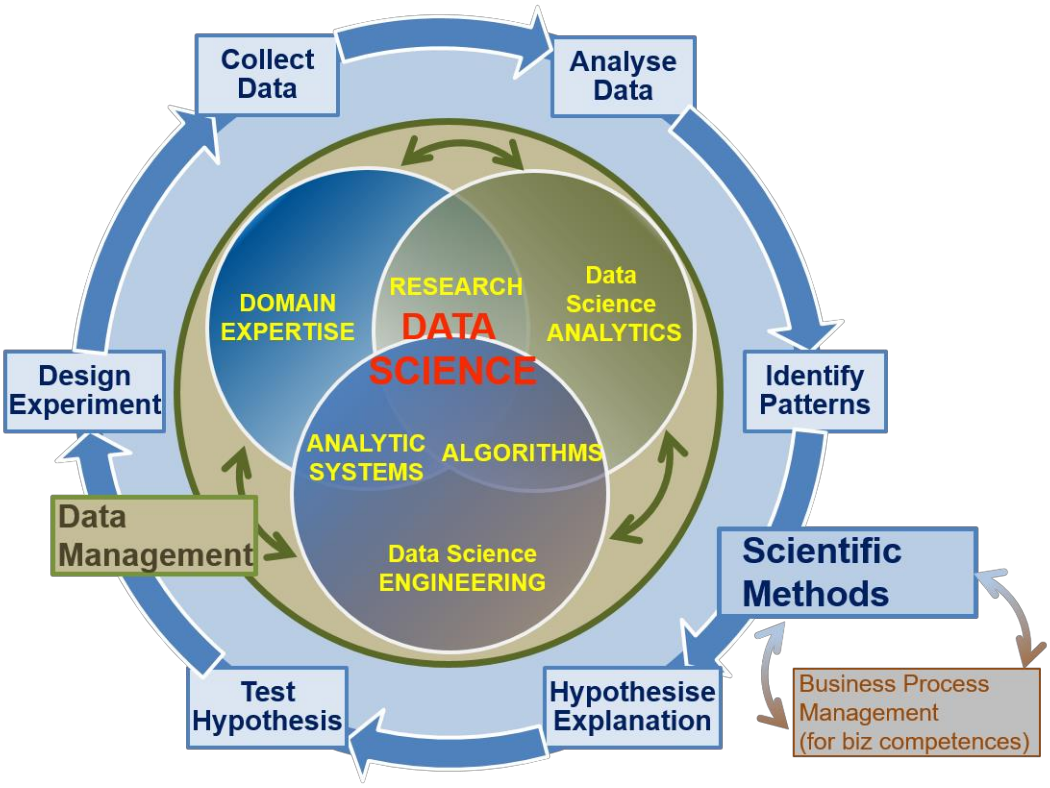


Figure 6 Data Science competence groups for general or research oriented profiles

We analysed the curriculum of each program in the inventory, including: definition of the program, list of courses, and definition of courses where available. Outputs were mapped to the three main DS competence groups: Data Science data analytics (mostly related to applied statistics), Data Science engineering competence (relating mostly to computer and software engineering), and and domain expertise. Each course might at the same time cover more than one domain to a certain extent and that was also taken into account. Available data did not allow more detailed classification, especially, regarding scientific methods and data management. Most of the programs and courses, unfortunately do not contain specific information on competences or learning outcomes.

In principle, we should expect roughly equal coverage of each competence group. Balance in covering competence groups is a key to educating successful data scientist. Small differences in coverage are natural. We observed that the difference between the most and least covered competence group cannot exceed 20 pp. (percent point) in order for the whole program to still be able to well cover the whole Data Science spectrum. This difference should preferably be even lower, but we thought that a stricter criterion would be misleading at this early stage of Data Science curriculum design. Between 20 pp. and 30 pp. we classified programs as having a small imbalance. If the difference exceeds 30 pp. it means usually that one of the competence groups is not covered at all or to a minimal extent, while another exceeds 60%. We classified such programs as having significant imbalance.

Considering the infrequent explicit definition of competence and learning outcomes in current programs, the analysis as presented here is an approximation. At the same time, given the large amount of programs analyzed and and our classification into three simple competence groups, the analysis can be considered meaningful as long as one is careful about what type of conclusions they drawn from it.

All the results are presented as a 2 digit percentage due to convenance. However, quantitative differences of just a few percent points should not be over-interpreted. The focus should be on qualitative differences. The analysis presented in the following subsections follows this recommendation. In addition to curriculum aspects we also investigated the source of programs, their naming and types of offered degrees.

3.1.2 Bloom's Taxonomy

Bloom's taxonomy provides a conceptual framework to organize levels of learning of a topic or subject, and assignes action verbs to each level that help to understand activites related with particular level of learning. For instance students start at the *knowledge* level when they can *name* and *identify* relevant technologies. The further move to *comprehension* level when they can *explain* how technologies work. They can then move to *application* level when they can *choose* right technology to *solve* a problem. Further they can progress to *analysis*, *synthesis*, and finally *evaluation* levels. Below example shows typical attributes of the different level of learning and example questions testing this level, levels are organized in Figure 7.

Knowledge

Exhibit memory of previously learned materials by recalling facts, terms, basic concepts and answers
Knowledge of specifics - terminology, specific facts
Knowledge of ways and means of dealing with specifics - conventions, trends and sequences, classifications and categories, criteria, methodology
Knowledge of the universals and abstractions in a field - principles and generalizations, theories and structures
Questions like: What are the main benefits of implementing Big Data and data analytics methods for organisation?

Comprehension

Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, describing, and stating the main ideas
Translation, Interpretation, Extrapolation
Questions like: Compare the business and operational models of private clouds and hybrid clouds.

Application

Using new knowledge. Solve problems in new situations by applying acquired knowledge, facts, techniques and rules in a different way
Questions like: What data analytics methods should be applied for specific data types analysis or for specific business processes and activities Which Big Data services architecture is best suited for medium size research organisation or company, and why?

Analysis

Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations

Analysis of elements, relationships, organizational principles

Questions like: What data analytics methids and services are required to support typical business processes of a web trading company? Give suggestions how these services can be implemented with the selected data analytics platform, including on-premises or outsoured to cloud. Provide references to support your statements.

Synthesis

Compile information together in a different way by combining elements in a new pattern or proposing alternative solutions

Production of a unique communication, a plan, or proposed set of operations, derivation of a set of abstract relations

Questions like: Describe the main steps and tasks for implementing data analytics and data managemen services for an example company or research organisation? What services and data analytics can be moved to clouds and which will remain at the enterprise premises and run by company's personel?

Evaluation

Present and defend opinions by making judgments about information, validity of ideas or quality of work based on a set of criteria

Judgments in terms of internal evidence or external criteria

Questions like: Do you think that implementing Agile Data Driven Enterprise model creates benefits for enterprises, short term and long term?

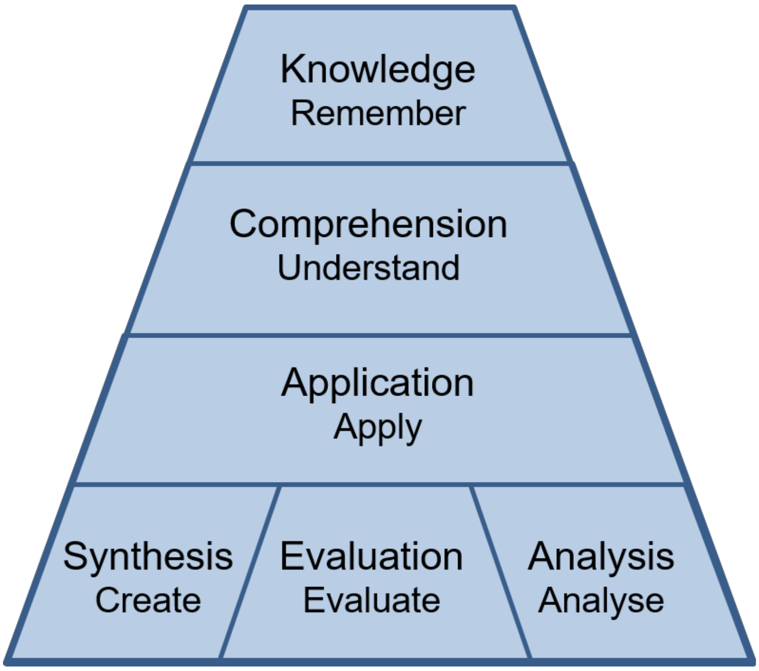


Figure 7 Simple Bloom's taxonomy

Figure 8 provides consolideated presentation of the Bloom's Taxonomy [13] structure, attributes and action verbs that can be effectively used for designing effective curricula and knowledge evaluation.

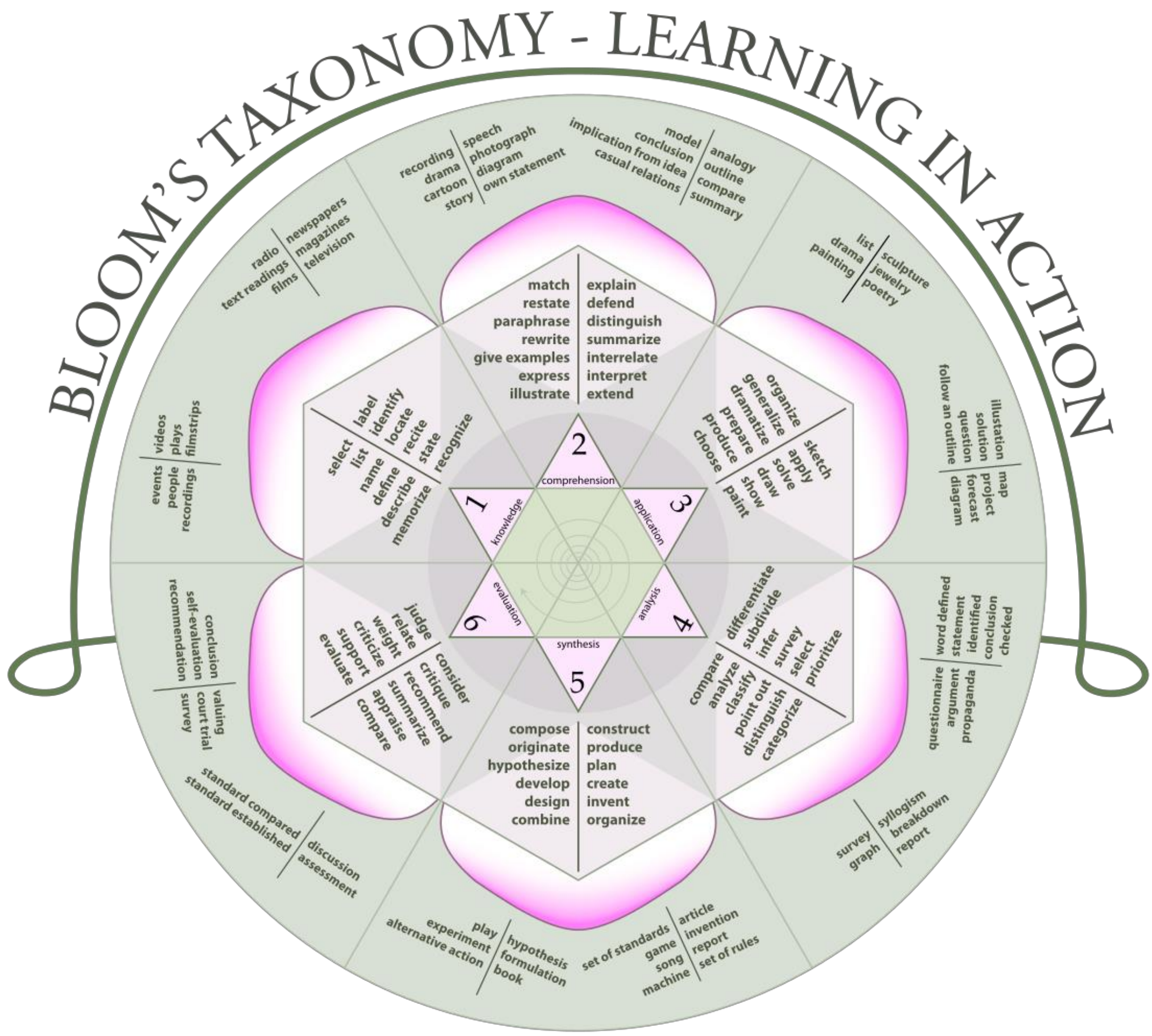


Figure 8 Extended Bloom's taxonomy[1]

When designing Learning Outcomes for a course or program it is essential to ensure that all levels will be adequately covered. Consideration of Bloom's taxonomy assists instructors both on the design phase of a course or program, and during grading process. It is a reliable and simple method to distinguish e.g. between familiarity with many concepts and actually being able to use them in a practical setting.

The traditional and still usual approach in science and engineering education is based on a behaviorist or objectivist epistemology, in which the student is passively imparted with knowledge by the teacher. Student's participation in the learning process is limited to memorizing schemes given by the instructor, which are assessed through instruments such as examinations and quizzes that measure the degree of conformance to a norm instead of actual competences [41]. In contrast, a constructivist epistemology puts the student in the center of the learning process as an active participant in constructing knowledge [14].

Problem Based Learning (PBL) [15, 16] is an alternative approach to instruction based on providing student with a non-trivial problem to solve, and guidance in obtaining the necessary competences. PBL is underlined by a constructivist epistemology that emphasizes active student participation in the construction of their knowledge from learning activities and motivating them through careful alignment of evaluation activities, leading to a concept called Constructive Alignment described by Biggs [18]. Ben-Ari [17] describes the applicability of constructivism to computer science education. Despite certain differences in epistemology between computer science and other sciences, constructivism is a useful approach to computer science education.

These education concepts provide guidance not only for analysis of the inventory, but also for further definition of Learning Outcomes and finally Model Curricula.

3.2 Quantitative analysis of degree-giving programs

3.2.1 Origin of programs

Figure 9 presents the distribution of programs in EDISON Inventory across the country of origin. It is important to see that lack or underrepresentation of certain countries might mean two different things. First, it might simply indicate that Data Science academic offerings in certain countries has not been yet developed. Alternatively, it might indicate that it was not included in the Inventory. This is of particular risk in Europe, where discovery of academic resources across borders is difficult due to language differences. It is impossible to distinguish between these two reasons at the current stage.

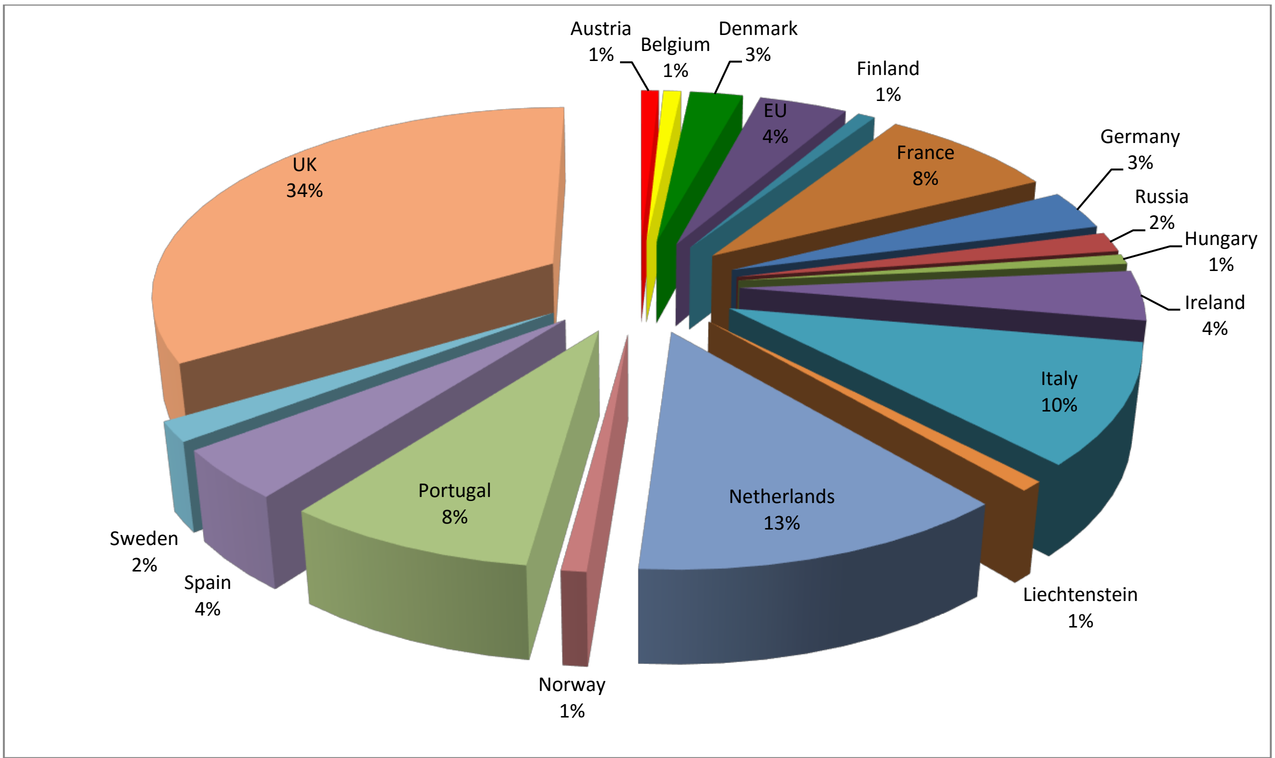


Figure 9 Origin of European Programs

As explained in Section 2.2 the Inventory is a result of combination of search results together with input from EDISON and ELG participants. Results from search give particual weight to programs conductred in English, which are naturally most common in UK. At the same time, many partners from e.g. Netherlands and Italy, result in good coverage of these countries.

3.2.2 Source of programs

Data Science programs can be created by different departments or units. Understanding where the program comes from can help to better understand what competences are well represented and what elements might require support.

In Figure 10 we present the distribution of the source of the programs among European institutions. The majority (38%) of the programs come from various types of Computer Science departments. Business and

Management departments are also an important source, with 27%. 14% of programs were created as an effort across several department or by a new specialized department.

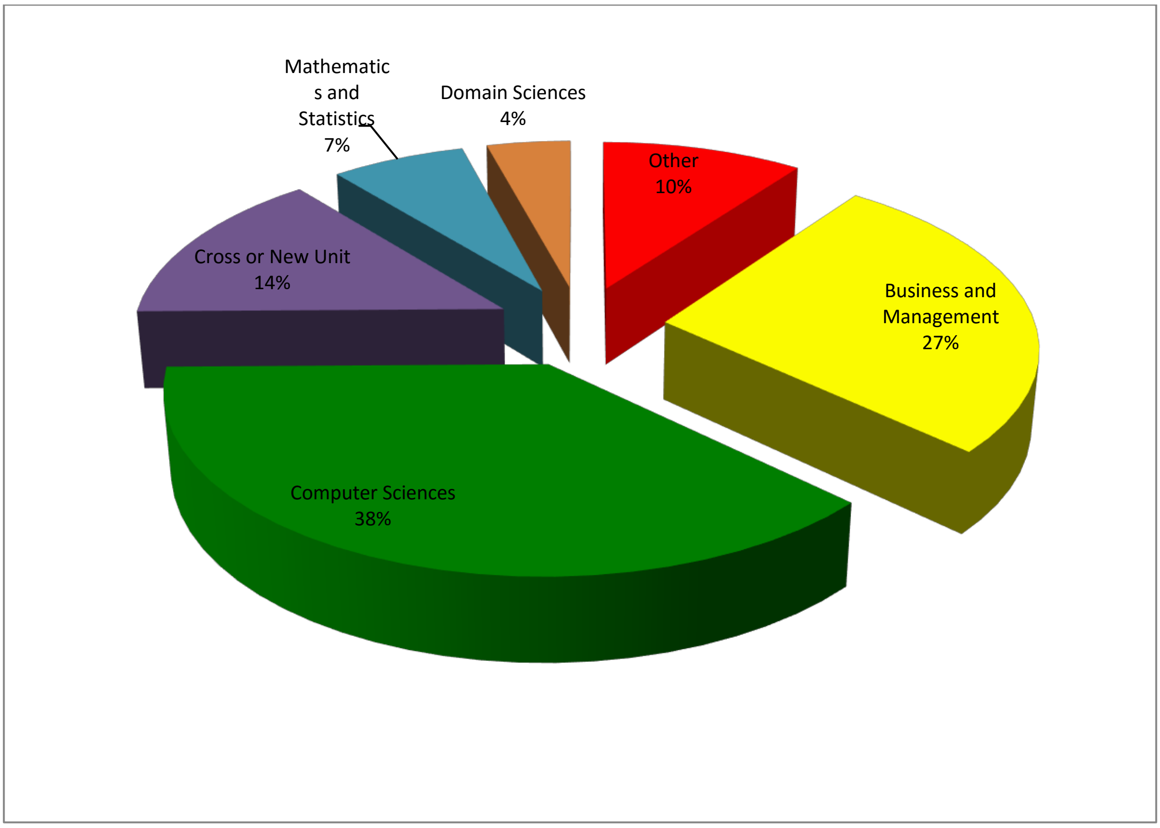


Figure 10 Source of European Programs

In Figure 11 we present the distribution of the source of the programs among Non-European institutions. The majority (37%) of programs come from Business and Management departments. Computer Sciences are a source of only 16% of programs.

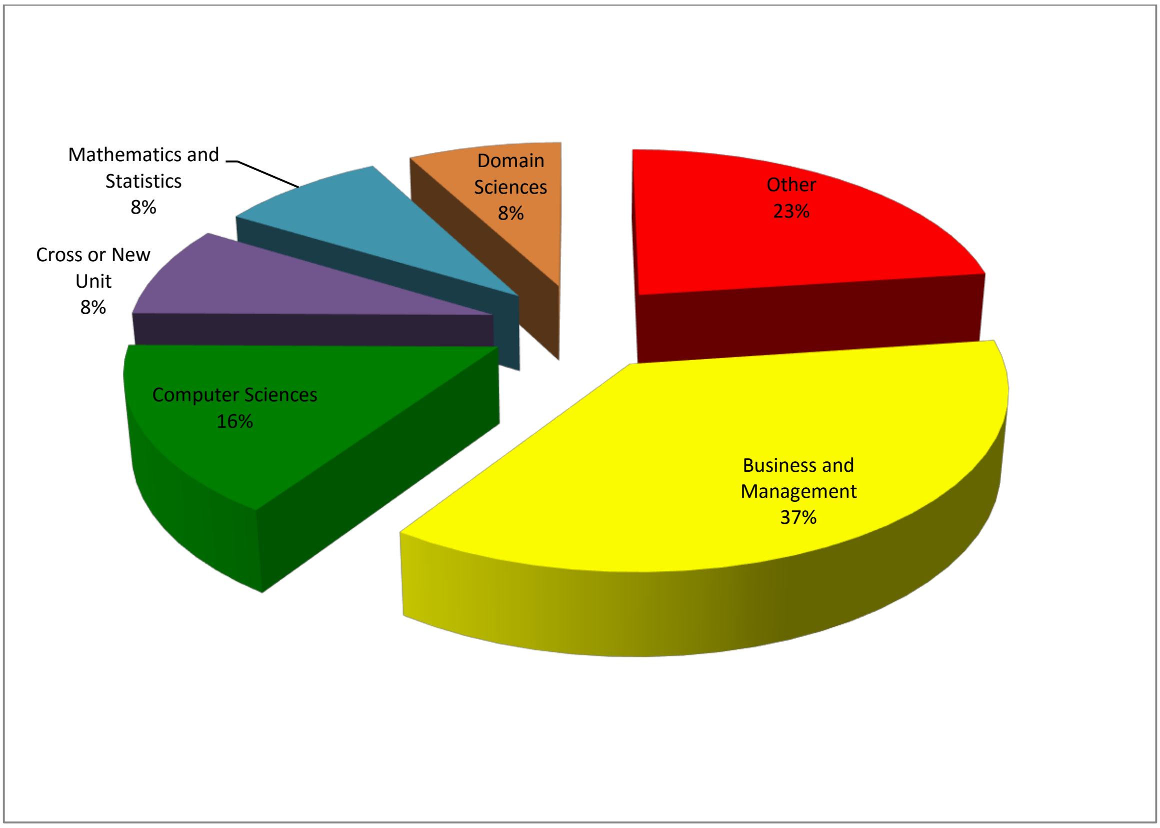


Figure 11 Source of Non-European Programs

We notice two major differences between European and Non-European programs (mostly influence by US institutions). First of all, while Compter Science departments are the main driver behind Data Science programs in Europe, outside Europe it is Business and Managment departments. Moreover, outside Europe, there are fewer (by 50%) programs coming from across several departments.

3.2.3 Coverage of domain knowledge

Each program in the inventory was analyzed in detail to determine to what extent courses in its curriculum cover competence groups. Some courses might naturally cover more than one group. In some cases, especially in the case of project courses (e.g. master thesis), they might provide coverage of all areas simultaneously. Such aspects were accounted for during our analysis.

In Figure 12 and Figure 13 we present the results of the analysis. 59% of European and 50% of Non-European programs are significantly imbalanced. This means that one of the competence groups is not covered properly or not at all. Additional 14% and 15% of programs respectively have smaller imbalances. Only 27% and 35% of the programs respectively could be considered balanced, despite the fact that the threshold we set was relatively low.

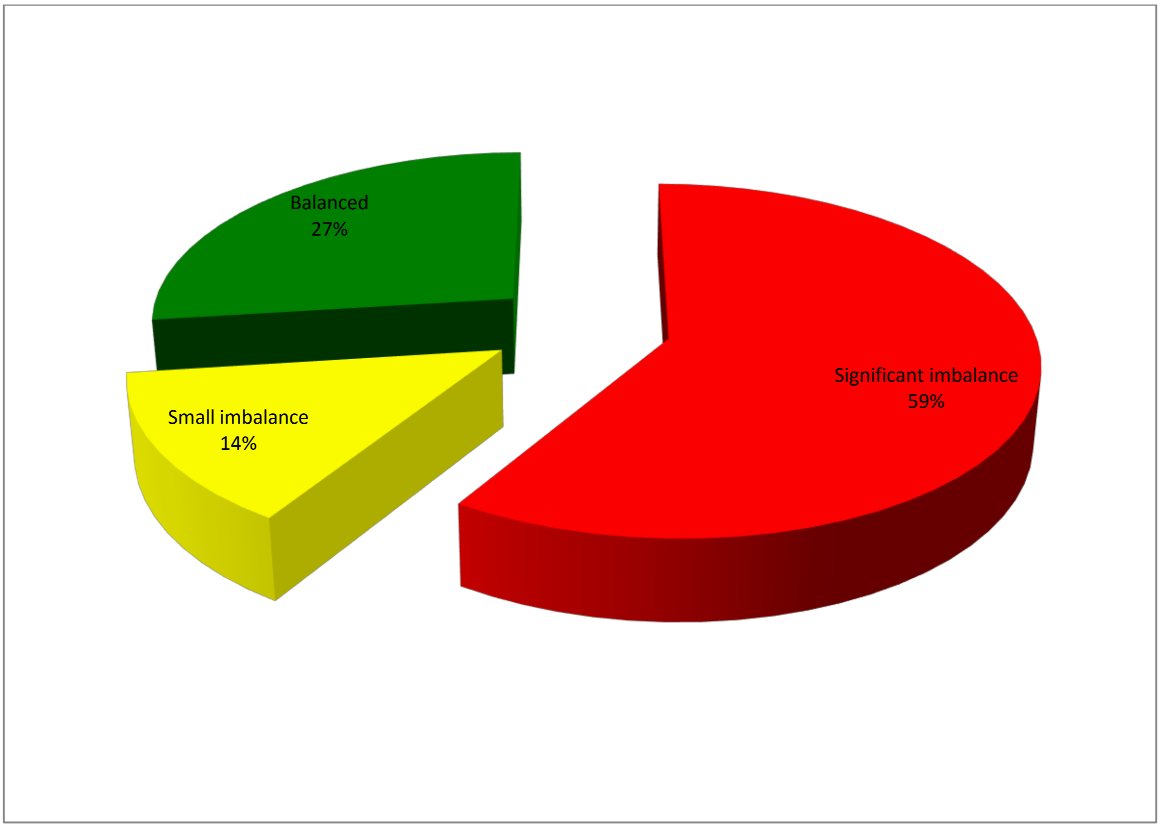


Figure 12 Balance of European Programs

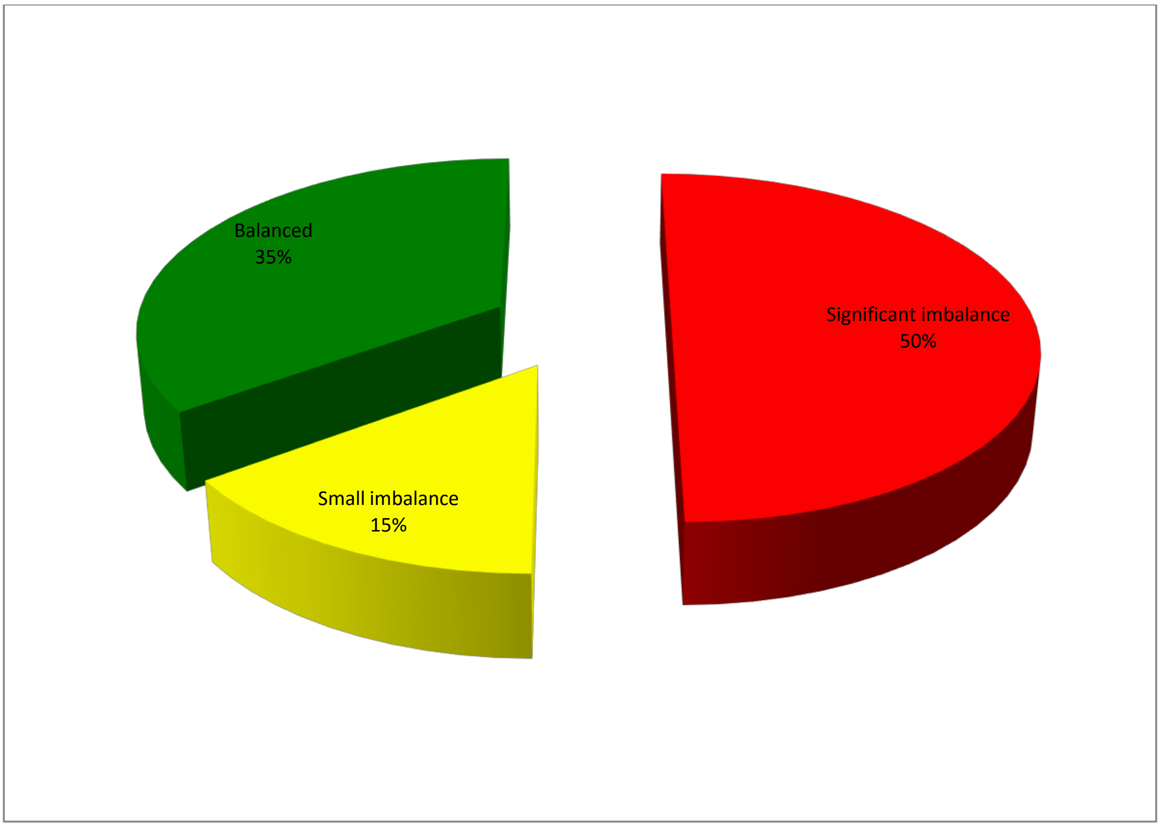


Figure 13 Balance of Non-European Programs

The distribution of the imbalance between competence groups is not equal. The Data analytics group is usually covered to a sufficient extent in almost all programs. On the other hand, (computer) engineering competences are often missing in programs not originating from computer science or computer engineering departments. At the same time, domain knowledge is often overlooked for programs from the aforementioned departments.

Another issue is uncontrolled flexibility of around 20% of the programs. The way their elective courses are structured might lead to imbalance for a particular student. Flexibility and electives should of course be encouraged, but they should be divided into competence groups and students should choose equally from each group.

In a large subset of programs, in which domain knowledge appears to be properly covered, deeper inspection reveals that offered courses overemphasize generic management and business skills. There is little conceptual connection between courses offered to cover domain knowledge and those covering other competence groups.

Such courses might be relevant to certain programs and business schools, but it seems they are used as a rushed solution, due to limited relation of these courses to the rest of the program, to superficially cover missing elements in the program. It is important to notice that we excluded from this argument specialist courses in economy, financial analysis or similar.

Many programs appear to place an equal sign between data scientist and business analyst. While business analysis might be considered a special case of Data Science, the opposite is certainly not correct.

Finally, in Figure 14 we look at balance in programs depending on what type of source they are coming from. We clearly see that for almost all cases, more than 50% are significantly imbalanced. The only exception are programs that come from cross-department collaboration, where more than 50% of programs are balanced. There are some minor differences between other sources but they should not be overinterpreted in the early stages of Data Science curricula development.

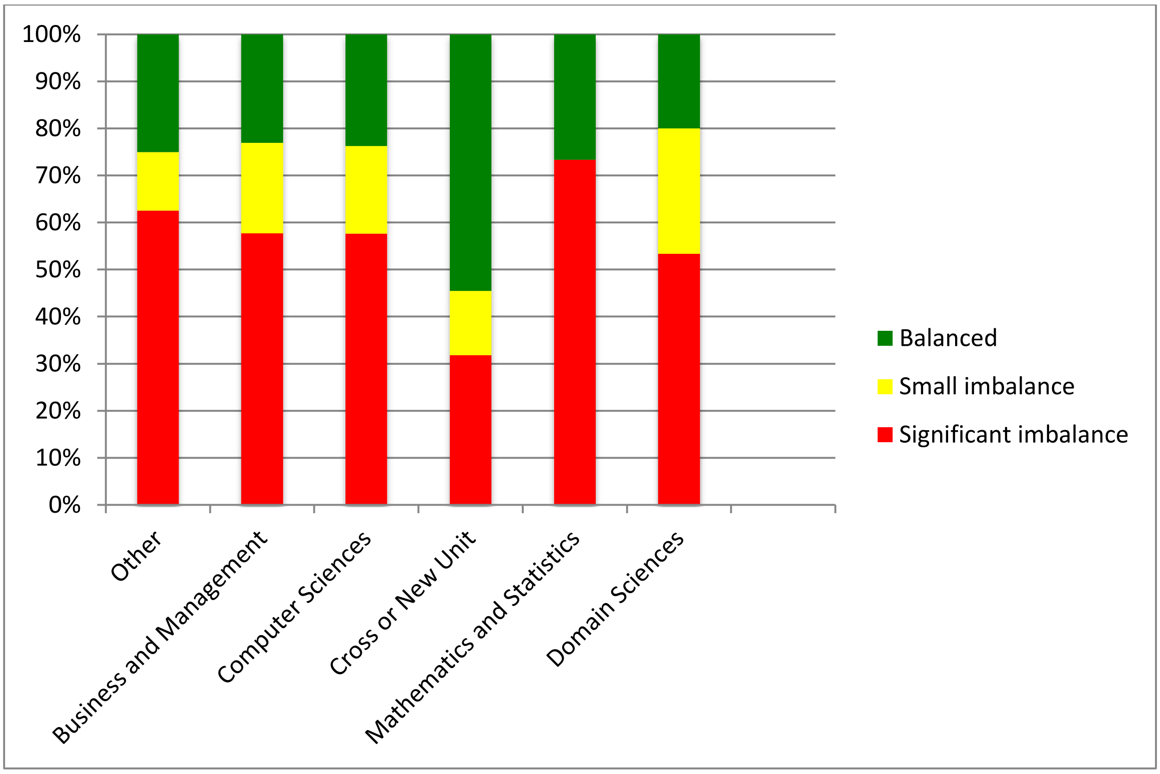


Figure 14 Balance of Programs w.r.t. Source Department

3.2.4 Naming of programs

Names used for programs in the Inventory are presented in Figure 15 for European offerings and Figure 16 for non-European. It is clear that the name Data Science is already in use, even if the contents of programs are not yet well structured. Otherwise, the distribution follows what we have already learnt from analyzing the source of programs. It means Computer Science related terms dominate in Europe and business related terms dominate outside Europe, mostly due to US influence.

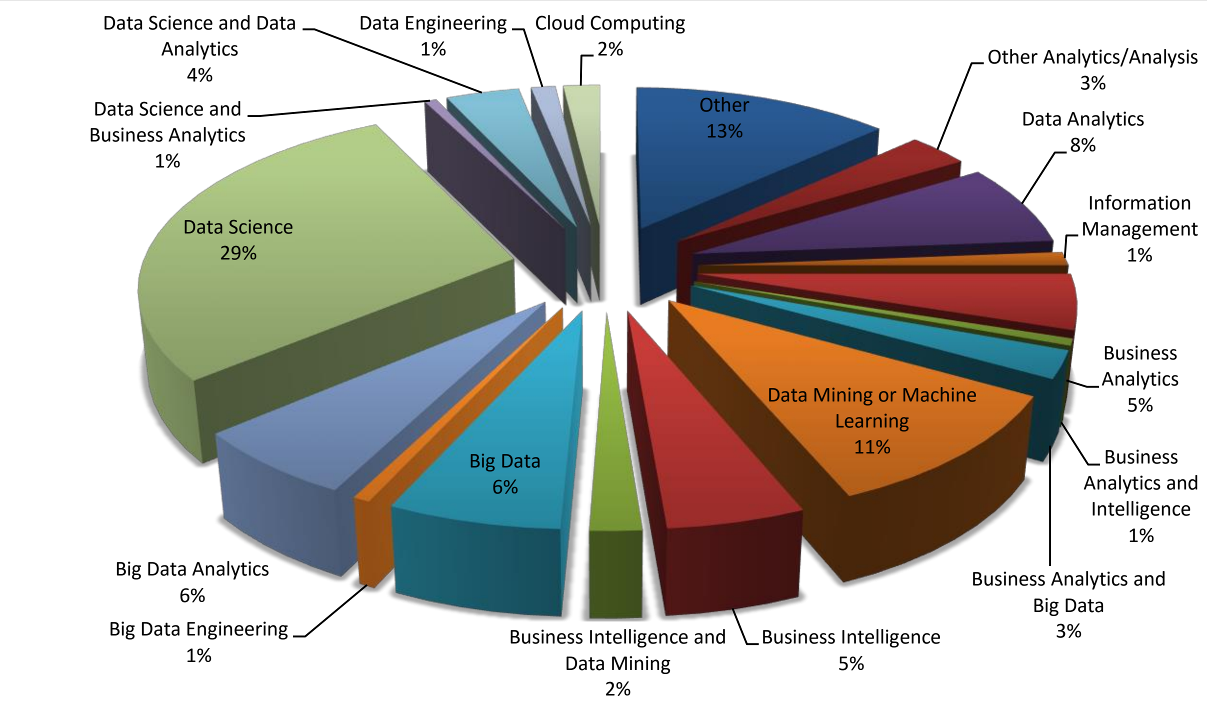


Figure 15 Naming of European Programs

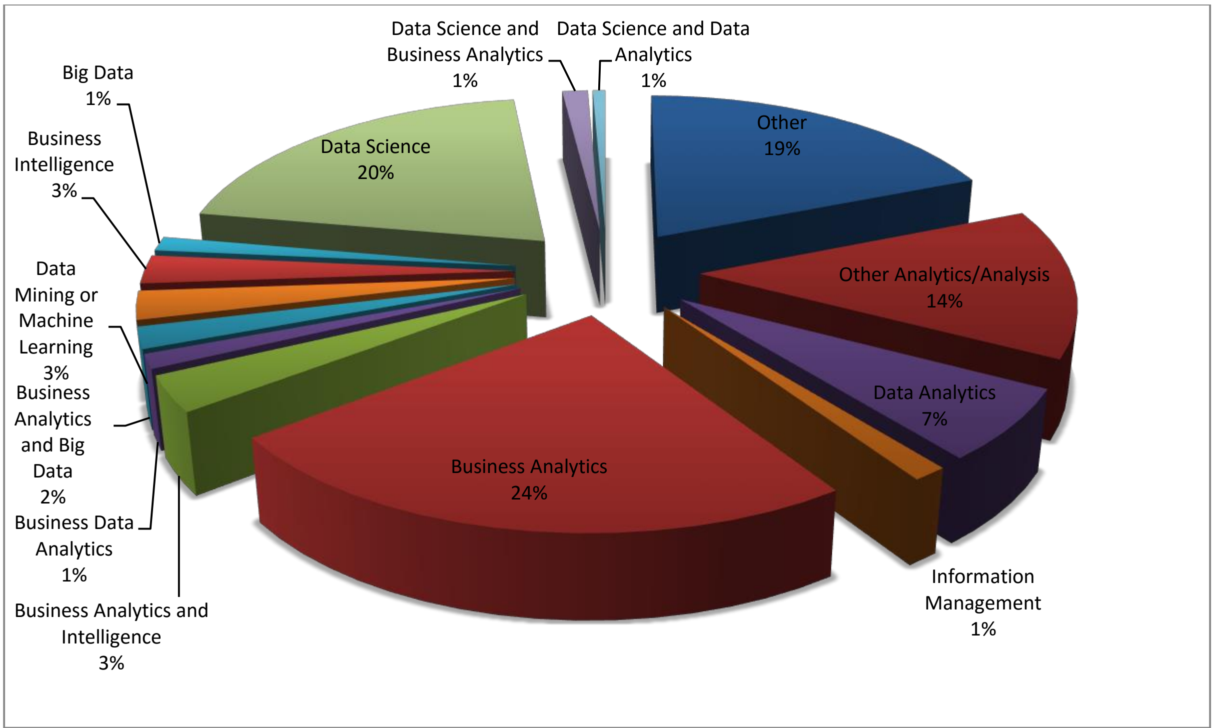


Figure 16 Naming of Non-European Programs

3.2.5 Degree and type of program

Figure 17 and Figure 18 present the distribution of the level of study for programs in Europe and outside Europe, expressed through a degree awarded. The majority of the programs are designed to be Master-level studies, (on average 9 out of 10), both in Europe and outside. One of the reasons, coming from analysis of these results in the ELG, might be that Bachelor programs are usually more regulated in term of contents and new programs require extra time for establishment.

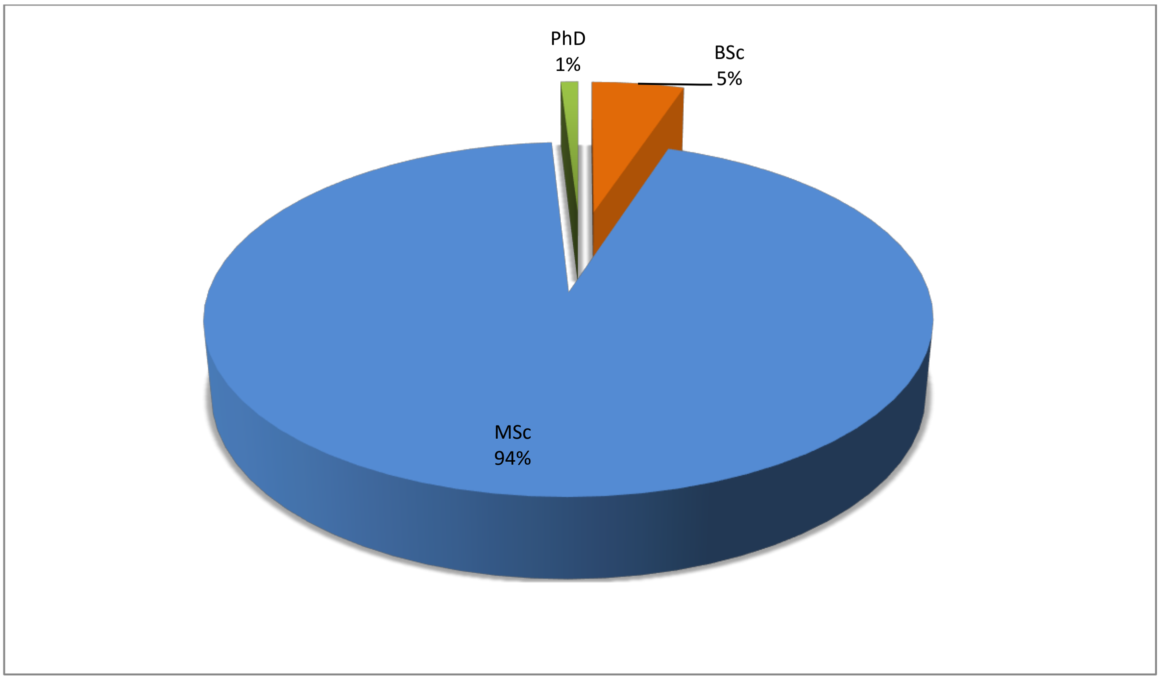


Figure 17 Degrees Awarded in European Programs

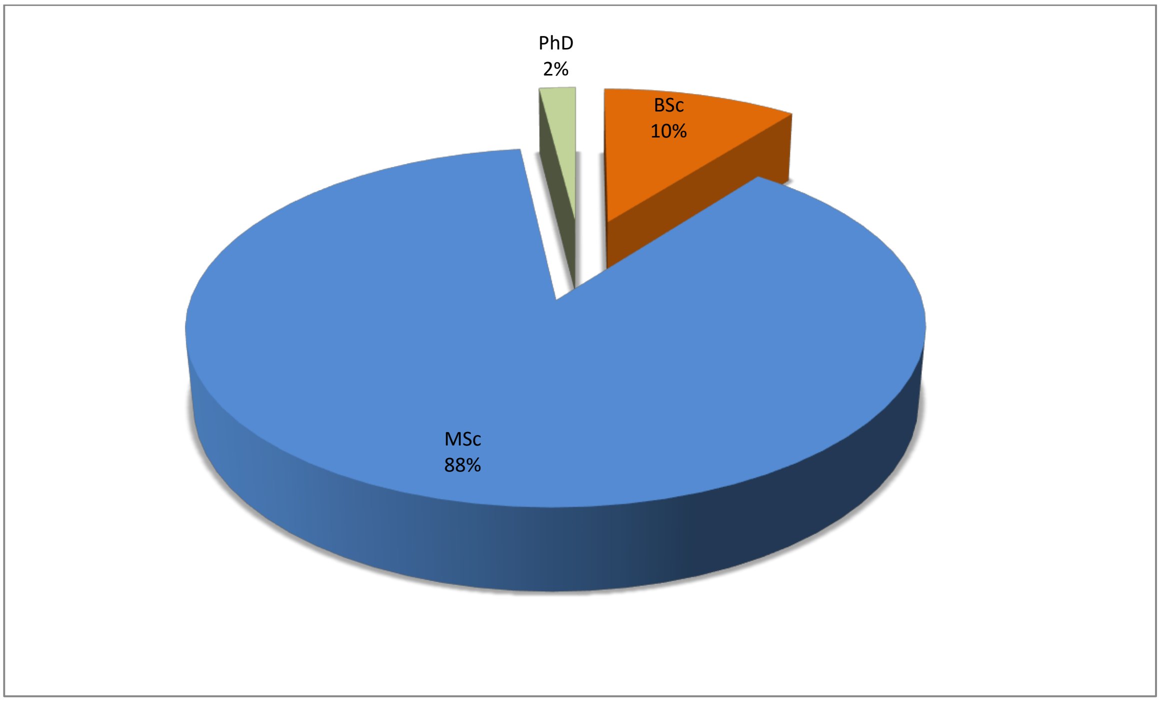


Figure 18 Degrees Awarded in Non-European Programs

Figure 19 and Figure 20 present type of programs offered, whether these are full Data Science related programs or special tracks in other existing programs. Both in Europe and outside most programs are standalone. Around 10% of programs are extensions in form of specialization, tracks, minors, etc. The difference in terminology reflects general difference between naming conventions in Europe and US, and is not specific to Data Science.

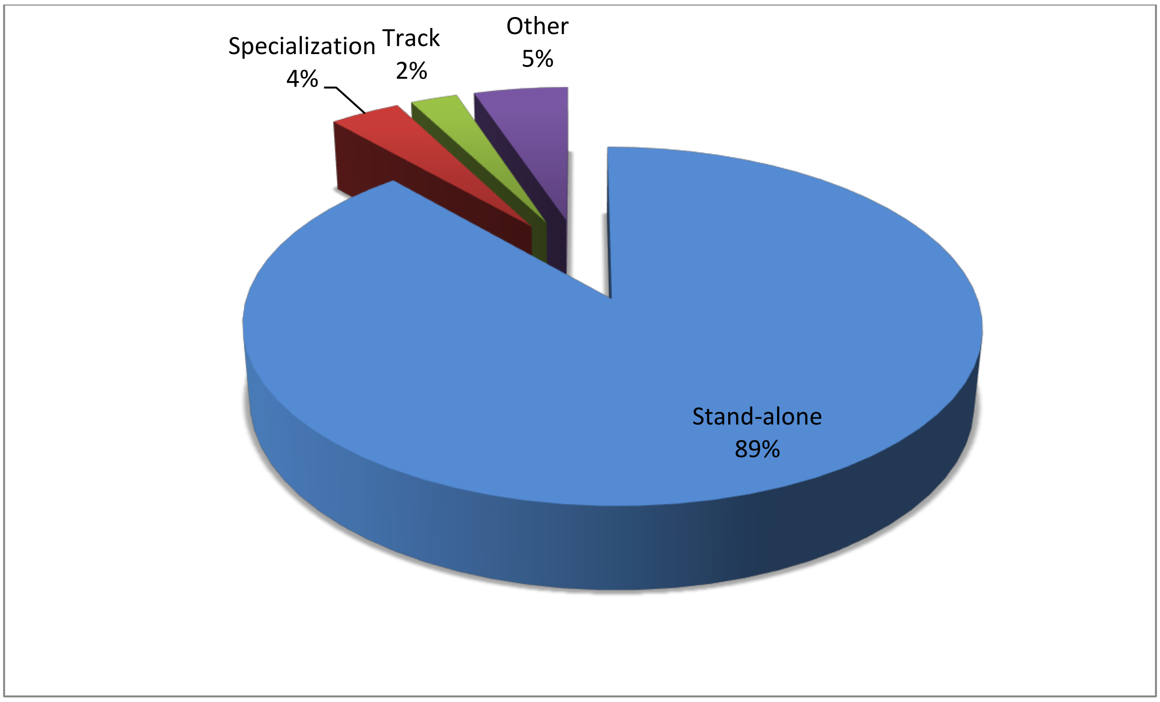


Figure 19 Program Type in European Programs

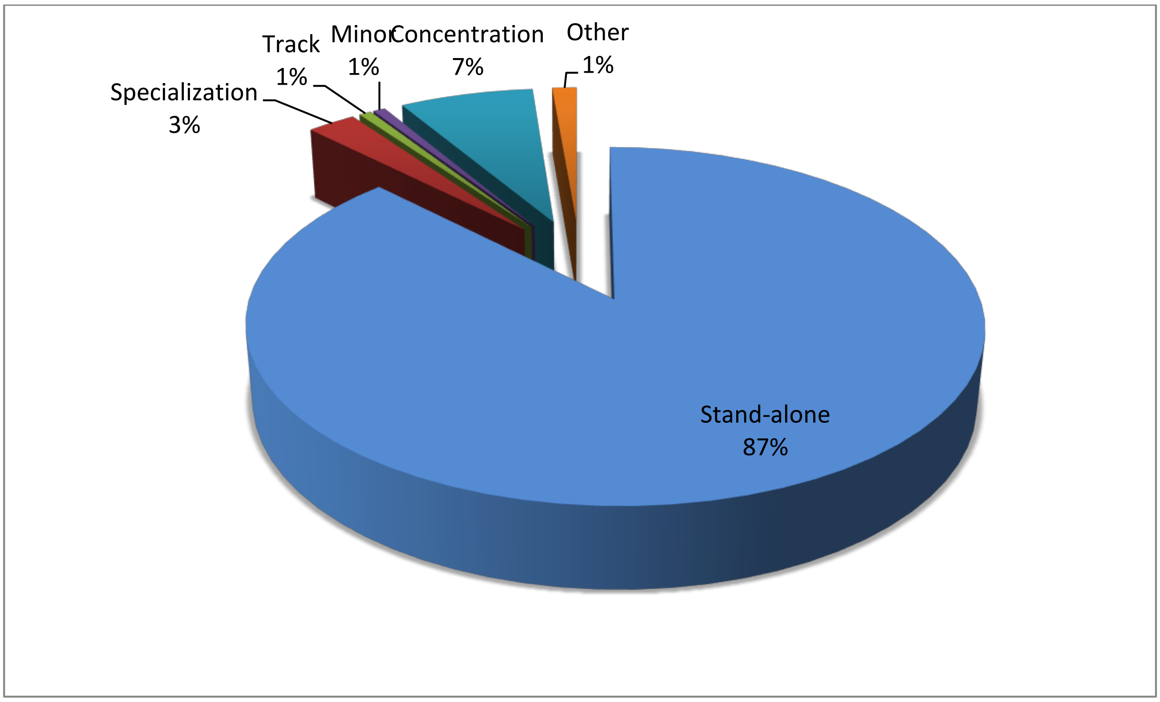


Figure 20 Program Type in Non-European Programs

3.3 Qualitative analysis of coverage of Data Science Competences in selected degree-giving programs

Only a small percentage of programs define some form of learning outcome, which can also include terms like goals, competences and objectives. Figure 21 shows that only 8% of European programs have such definitions Figure 22 shows that the corresponding number is 16% for non-European programs, mostly due to US influence. It is far fewer than expected, considering that all acadmiec programs should formalize learning outcomes. Due to limited data only general conclusions can be extracted.

Lack of formally specified learning outcomes might be a reason for the earlier discovered poor balance of existing programs. It is a reasonable proposition; however, there is not enough data yet to statistically confirm it.

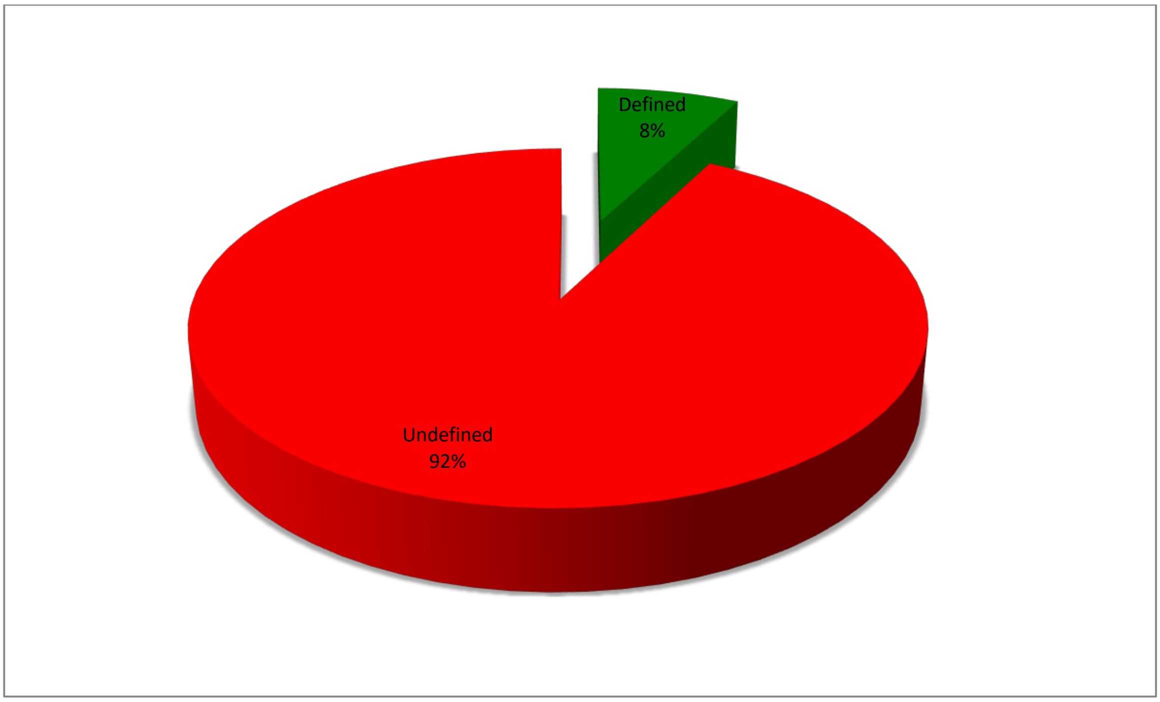


Figure 21 Learning Outcomes in European Programs

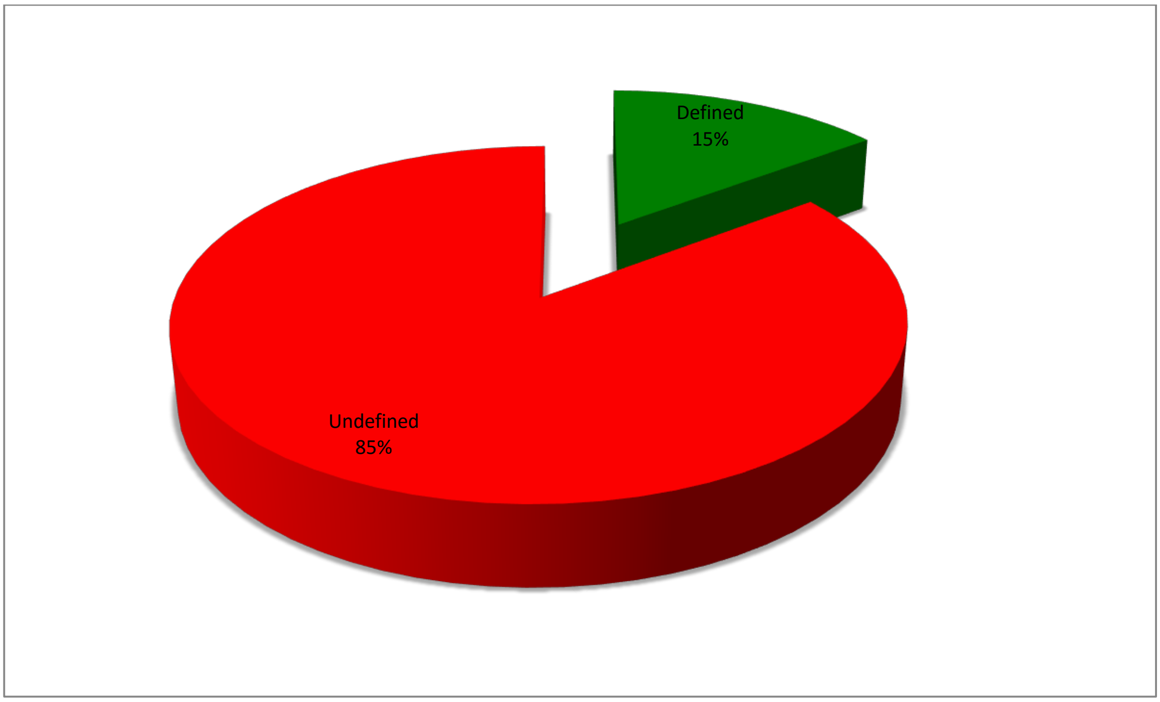


Figure 22 Learning Outcomes in Non-European Programs

When we evaluated the quality of learning outcomes w.r.t. Bloom's taxonomy the results were also worrying. Very few programs explicitly distribute learning outcomes across various learning levels. Usually, learning outcomes seemed very generic and offer little useful information.

3.4 Summary

In this section we analyzed programs in EDISON inventory. We noticed significant differences between Europe and outside Europe (mostly United States) in department from which Data Science programs originate. For Europe Computer Science departments are the main source, while it is Business Schools for programs outside Europe. In Europe we also mark more programs coming from cross-department initiatives. It is important, because was we also show, cross-department collaboration leads to better balance between Data Science competences in the program.

Naming of programs seems to follow the source departments. Data Science, Big Data, Machine Learning are most common names in Europe, while Business Analytics and Data Analytics outside Europe.

So far most of programs are offered on Master level. Learning Outcomes are usually not defined and even if defined they do not reflect established educational frameworks, such as Bloom's taxonomy.

4 Taxonomy of Data Science professions and learning outcomes

4.1 Methodology

4.1.1 General analysis

The taxonomy of Data Science as a multi-faceted discipline must be a combination of elements included in a Venn diagram. It is to be developed on the basis of available taxonomies and training resources for the following fields: Computer Engineering, Analytics, Statistics, Data Mining, Algorithms, Research Methods, and selected topics representing Certain Domain Knowledge. Not only existing taxonomies but also available syllabuses are being analyzed together with the Data Science Competence Framework, the Data Science Body of Knowledge and job advertisements, in order to develop a classification merging current offerings and required needs, academic disciplines and market sectors.

A number of professions can be defined in the field of Data Science, such as Data Scientist, Data Analyst, Data Engineer, Data Steward, Scientific Data or e-Infrastructure manager, etc. The proposed classification must reflect differences between these professions, and it should also be related to the Data Science competence groups defined in D2.1.

The proposed classification is based on the following sources:

- EDISON Data Science Competence Framework (CF-DS) and Data Science Body of Knowledge (DS-BoK);
- guiding principles described in the European e-Competences Framework (e-CF);
- ACM (2012) Computer Science Classification facets related to Data Science;
- descriptors defining levels in the European Qualifications Framework (EQF);
- American Mathematical Society (AMS) 2010 Mathematics Subject Classification topics related to Statistics required in Data Science profession; http://www.ams.org/msc/pdfs/classifications2010.pdf
- Core Subject Taxonomy for Mathematical Sciences Education defined by the Mathematical Association of America. http://www.maa.org/press/periodicals/loci/joma/subject-taxonomy
- Syllabi and taxonomies for: Data Mining, Analytics, Algorithms, Research Methods and Statistics.

4.1.2 Statistical approach to identification of skills

Skills and qualifications required from data scientists are reflected by job descriptions, scientific papers, blogs etc. Therefore, we can extract these skills from a set of documents related with that topic. Analyzing millions or even thousands of documents manually to identify skills is labor-intensive and time consuming. To automate this process we will employ natural language processing and text mining methods to construct a hierarchical taxonomy of the skills required by data scientists [21]. Automating the extraction of skills from job Ads and other documents is an important step to be able to follow trends in terms of demand and thus will help to define the appropriate curricula and keep them up to date with Job market demand. In this section we describe the approach used in Edison to develop tools to automate the analysis of a large corpus of documents. Preliminary results are presented in Appendix F.

4.1.2.1 Taxonomies

A taxonomy is a set of rules or conventions that describe the arrangement of things or concepts into ordered categories and it usually has a hierarchical structure. A taxonomy can be used to group related things based on a set of features and therefore can be seen as a knowledge map. A taxonomy may also be an empirical tool for building classifications to allow the ordering and retrieval of large amounts of data [22, 23]. To extract a taxonomy from unstructured text we have identified first the vocabulary used to describe Data science Skills and competencies and then discovered the relationships among the identified terms. We have thus followed a two step approach: (1.) term extraction, (2.) relation discovery.

4.1.2.2 Term extraction (identifying the constrained vocabulary)

Term or terminology extraction attempts to identify the body of terms used in a subject or content. A term may be single or multi-word expressions that have a particular meaning within in specific domain [24, 25]. There are three main approaches to term extraction: 1. statistical, 2. linguistic and 3. hybrid. Statistical approaches are mainly concerned with defining a degree of "termhood" for candidate terms. That is to find appropriate metric that can rank to what extent a terms belongs to a list of possible term. Linguistic approaches use syntactic rules aiming to identify specific syntactic term patterns. Hybrid approaches attempt to combine the two approaches.

4.1.2.3 Statistical Approaches

Statistical approaches for term extraction apply metrics to identify repeated sequences of lexical items. They also produce a ranked list of terms identifying the most important terms extracted from a text. Statistical approaches usually start by identifying all the unique words that appear in a text. Next, they construct all possible n-grams that can be identified. An n-gram is a contiguous sequence of n "items" (items for our purpose are words) from a given sequence of text. The next step in statistical approaches is to determine the "termhood" of each term and rank it accordingly. Term frequency (tf) is one of the most common and simple metrics used for statistical term extraction and it measures how frequently a term occurs in a document. Often tf is normalized by dividing the number of times a term appears in a document by the total number of terms in that document. However, with tf all terms are treated "equally" which means that terms repeated often but have little importance would be ranked higher than less repeated but important terms. This is treated by weighing down frequent terms while scaling up rare ones with a log function. This measure is called Inverse Document Frequency (idf) and in essence measures how important a term is. Combining the two provides the term frequency-inverse document frequency (tf-idf) which is a statistical measure used to evaluate how important a term is to a document in a corpus. Besides tf-idf, there are numerous other metrics that rank candidate terms such as T-score [26], C-value [27], Dice coefficient [28], Log-Likelihood ratio [29] etc.

4.1.2.4 Linguistic Approaches

Linguistic or contextual approaches attempt to identify syntactical patterns in a text in order to extract terms. Usually terms tend to have characteristic syntactic structures [30]. Contextual analysis usually starts by filtering out terms that are unlike to be terms based on their syntactical pattern. For example pronouns like she, few, many, are very unlikely to be part of a term. As a next step there is a contextual attempt to identify terms as combinations or sequences of nouns [31]. The construction of these syntactic patterns is usually done empirically. Part-of-speech taggers [32] are essential for these type of approaches as they are used to identify nouns, verbs, pronouns, etc. Therefore, linguistic approaches first tag the text using part-of-speech taggers, next filter out verbs, pronouns, etc. and then with the use of regular expressions extract sequences of words that follow certain patterns. Linguistic approaches are language-dependent and therefore are not flexible and adaptable to other languages [33].

4.1.2.5 Hybrid Approaches

Hybrid approaches use both statistical and linguistic information. For the most part these approaches depend more on statistics and use syntactic rules as a complementary method to filter the appropriate terms. Therefore, in these approaches a linguistic analysis is performed to exclude words like pronouns and verbs. This step may also be applied to identify patterns and sequences of part-of-speech and pass these on to statistical measures to rank possible terms. Other approaches include linguistic information in the ranking process [27]. The biggest challenge for any term extraction approach is validation. Judging the accuracy of any approach involves a human expert that needs to evaluate the results.

4.1.2.6 Relation Discovery

Relation discovery attempts to define or extract a set of rules that can be used to group together terms [34]. This process can be divided into two approaches: Non-hierarchical and hierarchical. Non-hierarchical methods depend heavily on clustering and classification techniques. The main idea behind identifying "clusters" of terms is that conceptually similar terms should use the same set of words to define them and therefore should be grouped together. Hierarchical approaches depend on hypernym-hyponym or holonym-meronym relations. An hypernym-hyponym relation represents an is-a or a superclass-subclass relation between terms and can be used to define a hierarchy. Similarly, a holonym-meronym or a container-member relation between terms define a has-a: is a relationship between them where a term "belongs to" another class or term [35, 36]. For both methods it is necessary to perform "semantization" for each term to obtain the meaning or definition for each of them. Online dictionaries such as WordNet [37] or Bablenet [38] or encyclopedias like Wikipedia can be used to semantize terms. However, each term, especially single-words or acronyms may have multiple meanings. Therefore, it is necessary to disambiguate terms to narrow down the meaning of each term to one. A simple and effective approach for disambiguation is to use n-grams where words that appear near the ambiguous term are used for giving context to that term. These “context” words are compared with the definitions obtained from the dictionary or encyclopedia and the term with the highest similarity is chosen as the “semantized” term.

A crucial step for both disambiguation and clustering is to be able to express terms as numerical vectors. With numerical vectors the application of measures such as Euclidean distance or cosine similarity is straightforward. The most common way for representing a term as a vector is to use one of the statistical methods mentioned earlier. As with term extraction the validation of relations between terms must be performed by an expert.

4.1.2.7 Taxonomy eXtraction from Text (TEXT)

Using the techniques and method described above we developed tools to build taxonomies from relevant corpus. Given a corpus, the first step to extract terms is “tokenization” where text is broken up into words. Next, it is necessary to filter out “stop words” like “the”, “or” etc. as they have little lexical content. To be able to apply any kind of measure from simple word count to more complex statistical measures words need to be “lemmatized”, which is the process of grouping together different forms of a word so they can be analyzed as the same (e.g. scientist and scientists should be considered as the same word). These processes allow us to build the term dictionary which is a list of all unique words used in the corpus. In the next step we use a set of hybrid term extraction methods to rank the relevant terms. During relation discovery we first build non-hierarchical relations and with the use of hypernym-hyponym relations we build hierarchical relations within each cluster. At every step it is necessary to perform a validation.

4.1.3 Overview of existing taxonomies

There are several attempts to create a taxonomy for Data Science. Most of them are based on an online article published in 2010 on the dataists.com portal. In “Taxonomy of Data Science” Hilary Mason and Chris Wiggins attempted to answer three questions: ‘Where to find a good Data Scientist?’, ‘What to learn to become a Data Scientist?’, and ‘What is Data Science?’. Finally Data Science was defined according to 5 steps:

(1) Obtaining data “O”
(2) Scrubbing data “S”
(3) Exploring data “E”
(4) Modeling data “M”
(5) iNterpreting data “N”

These 5 steps are described as the OSEMN model, which is pronounced as ‘*awesome*’. Elements of the model should be considered in an iterative and nonlinear manner since in practice one should move back and forth between them or perform multiple steps at the same time.

The steps are formulated according to the list of tasks a Data Scientist should be familiar with. The authors of the article pointed out that in real life Data Scientists have different levels of expertise with each of these five areas. The knowledge fields for each of the five areas and the goals defined for each step are summarized in Table 3.

Table 3 OSEMN model

Area	Goal	Skills
Obtain	Obtain data for given problem	Unix command line tools SQL in Databases APIs for ?? for what? Scripting languages (e.g. Python) for data retrieval (e.g. presented in JSON)
Scrub	Clean and refine messy data	Command line tools (sed, awk, grep) Scripting languages (Pearl, Python) Databases: syntax for representing data; querying databases
Explore	Get to know gathered data to	Command line tools

	define hypothesis	Data visualization techniques (histograms, pairwise histograms, scatter plots) Dimensionality reduction methods (MDS, SVL, PCA, PLS) Clustering (unsupervised ML techniques, Gaussian mixture modeling, K-means)
Model	Create statistical model of data	Clustering Classification Regression Dimensionality reduction Command line tools APIs
Interpret	Draw conclusions from data Evaluate results' meaning Communicate the result	Statistical tools Data visualization techniques

Another attempt to develop a classification of Data Science was presented in the online publication from 2013 posted at the learningbymarketing.com webpage. It is based on a Data Science Venn diagram and presents the hierarchy of knowledge areas:

(1) Big Data and Distributed Database Systems
(2) Data Mining/CRM
(3) Machine Learning and Statistics
(4) Business Intelligence and Descriptive Statistics

The goals of each level are as follows:

- make use of huge amounts of data and distributed database systems
- model building
- explain results after model building
- present results based on statistical calculations

Data Science Central is an online community for Data Science and Big Data practitioners. It offers social interaction between DS professionals, forum-based support and also provides information about the latest technologies, tools and trends. In November 2013 DSC published results of the study concerning the fields most frequently associated with Data Science. The study was based on LinkedIn data about endorsements of persons reporting themselves as Data Scientists. The author of the study considered top 5 skills listed by top 10 Data Scientists identified on LinkedIn. The analysis of gathered data resulted in listing the skills required from a Data Scientist with percentage weight. These coefficients and skills were composed into the "data science formula"(http://www.datasciencecentral.com/profiles/blogs/data-science-connected-fields-pioneers):

Data Science = 0.24*Data Mining+0.15*Machine learning+0.14*Analytics+0.11*Big Data+0.07*Predictive Analytics+0.06*Data Analysis+0.05*Predictive Modeling+0.03*Hadoop+0.03*Text Mining+0.03*Statistics+0.02*Natural Language processing+0.02*Start-Ups+0.02*Algorithms+0.01*Distributed Systems+0.01*Map Reduce+0.01*Data Warehousing+0.01*Business Intelligence+0.01*SQL&R+0.01*Scalability

Further analysis of data gathered in this study led to the development of "Taxonomy of Data Scientist" that concluded with identifying four leading skills for a Data Scientist: Data Mining, Machine Learning, Analytics and Big Data.

In each publication concerning Data Science classification or Data Science Taxonomy DS is presented as a diverse field which is a mixture of various specializations. Selected sources are summarized in the Table 4.

Table 4 Source for DS taxonomy

Source	Specializations
Taxonomy of Data Science **http://www.dataists.com/2010/09/a-taxonomy-of-data-science/**	"command line fu" for data procurement and preprocessing Machine Learning and Statistics to 'look at

	data'
Data Science Taxonomy: Who Cares About the Name http://www.learnbymarketing.com/49/data-science-taxonomy-who-cares/	Mathematics Probability Historical data analysis
What is data science http://datascience.nyu.edu/what-is-data-science/	Computer science Applied mathematics Statistics Data modeling Data visualization Domain knowledge: social sciences, economics, engineering, law, business, medicine, science
What is Data Science? https://datajobs.com/what-is-data-science	Mathematics Expertise Technology and Hacking Skills Business/Strategy Acumen
Data Science Ontology https://www.thoughtworks.com/insights/blog/data-science-ontology	Learning algorithms Model validation Model performance Data visualization Production Programming languages Data cleaning Data preparation Statistics

4.2 Defining taxonomy for the Data Science professions family

This section provides updates on the ongoing research to define the Data Science professions family that could be instrumental in defining education and training profiles for students and for practitioners. Deliverable D2.1 provided the initial definition of the Data Science occupations family as a proposed extension to the ESCO taxonomy of occupations by adding new occupation hierarchies (see [D2.1] and Appendix D).

The Data Scientist occupation groups are placed in the following top level ESCO hierarchies:

- Managers (for managerial roles);
- Professionals (for analytics applications developers and for infrastructure and datacenter engineers);
- Technicians and associate professionals (for operators and technicians)
- Optionally, some data management occupations can be also placed into the Clerical support workers group such as digital data archivist, digital librarians.

Correspondingly, the following new 3rd level occupation groups are proposed:

- Data Science/Big Data Infrastructure Managers
- Data Science Professionals
- Data Science technology professionals
- Data and information entry and access (this is a candidate group under Clerical support workers top level hierarchy)

It is proposed that the existing ESCO group "Database and network professionals" should be extended with new occupations (or professions) related to Big Data or cloud based databases: Large scale (cloud) database administrator/operator and Scientific database administrator/operator, however further identification of such occupations needs to be done.

A group of occupations related to digital librarians, data archives management, data stewardship and data curation are currently placed in the 3rd proposed group:

> *Professionals > Information and communications technology professionals > Data Science technology professionals > Data handling professionals not elsewhere classified,*

however potentially it can also be added in a new 2nd level group "*Clerical support workers > Data handling support workers (alternative)*". The motivation for this is a growing need for data support workers in all domains of human activities in the digital data driven economy.

To ensure a smooth Data Science professions acceptance by industry and employment bodies, the proposed profiles should be compatible with the relevant standards ESCO, eCFv3.0 [ecf] (future CEN standard EN 16324), CWA 16458 2012 ICT Profiles [cwa].

Table 5 provides an initial definition of the identified Data Science professional profiles collected from job advertisements, blogs and recent discussions at different forums, in particular, with the Research Data Alliance, and digital curation and data preservations communities.

Table 6 provides a mapping between professional profiles and Data Science competence groups which are identified in D2.1 as follows:

Data Analytics (DSDA)
Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations

Data Management (DSDM)
Develop and implement a data management strategy for data collection, storage, preservation, and availability for further processing.

Data Science Engineering (DSENG)
Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management

Scientific and Research Methods (DSRM) for research domain and Business Process Management (DSBP)
Create new understandings and capabilities by using the scientific method (hypothesis, test/artifact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organizational goals

Data Science Domain Knowledge (DSDK)
Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organizational roles and relations

The initial definition of the Data Science competence groups listed above was provided in D2.1.

Table 6 provides a ranking of different competence groups relevance for Data Science profiles where 1 is less relevant and 5 is highly relevant.

Table 5 Data Science professional profiles definition

Profile ID	Data Science Profile title	Data Science Profile Summary statement	*Alternative titles and legacy titles*
Managers			
DSP01	Data Science (group) Manager	Proposes, plans and manages functional and technical evolutions of the data science operations within the relevant domain (technical, research, business).	Data analytics department manager
DSP02	Data Science Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the big data infrastructure within the relevant domain (technical, research, business).	Big Data Infrastructure Manager
DSP03	Research Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the research infrastructure within the relevant scientific domain.	Research Infrastructure data storage facilites manager
Professionals			

DSP04	Data Scientist	Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data.	Data Analyst
DSP05	Data Science Researcher	Data Science Researcher applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to scientific problem, business process, or reveal hidden relations between multiple processes.	Data Analyst
DSP06	Data Science Architect	Designs and maintains the architecture of Data Science applications and facilities. Creates relevant data models and processes workflows.	System Architect, Applications architect
DSP07	Data Science (Application) Programmer/Engineer	Designs/develops/codes large data (science) analytics applications to support scientific or enterprise/business processes.	Scientific Programmer
DSP08	Data Analyst	Analyses large variety of data to extract information about system, service or organisation perfomance and present them in usable/actionable form	
DSP09	Business Analyst	Analyses large variety of data Information System for improving business performance.	Business Development Manager (Data science role)
Professional (data handling/management)			
DSP10	Data Stewards	Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers	
DSP11	Digital data curator	Finds, selects, organises, shares (exhibits) digital data collections, maintains their integrity, up-to-date status and fresheness, discoverability	Digital curator, digital archivist, digital librarian
DSP12	Digital Librarians	Selection, acquisition, organization, accessibility and preservation of digital information/library. Manages digital materials, takes a lead role in the creation, maintenance and stewardship of digital collections, including the digitization of special collections. Develops strategies for effective management and preservation of library digital assets.	Digital data curator
DSP13	Data Archivists	Maintain historically significant collections of datasets, documents and records, other electronic data, and seek out new items for archiving.	Digital Archivists
Professional (database)			
DSP14	Large scale (cloud) database designer	Designs/develops/codes large scale data bases and their use in domain/subject specific applictions according to the customer	Large scale (cloud) database developer

		needs.	
DSP15	Large scale (cloud) database administrator	Designs and implements, or monitors and maintains large scale cloud databases	
DSP16	Scientific database administrator	Designs and implements, or monitors and maintains large scale scientific databases	Large scale (cloud) database administrator
Technicians and associate professionals			
DSP17	Big Data facilities Operator	Manages daily operation of facilities, resources, and responds to customer requests. Includes all operations related to data management and data lifecycle	
DSP18	Large scale (cloud) data storage operator	Manages daily operation of cloud storage, Including related to data lifecycle, and responds to requests from storage users	
DSP19	Scientific database operator	Manages daily operation of scientific databases, Including related to data lifecycle, and responds to requests from database users	Large scale (cloud) data storage operators

Table 6 Mapping Data Science competence groups to the proposed profiles

Profile ID	Data Science Profile title	Data Science Competences Groups (relevance 1 - low, 5 – high)				
		Data Analytics	Data Management	Data Science Engineering	Research Methods, Business methods	DS Subject Domain
Managers						
DSP01	**Data Science (group) Manager**	3	4	3	3	2
DSP02	**Data Science Infrastructure Manager**	2	4	4	2	2
DSP03	**Research Infrastructure Manager**	2	4	4	3	2
Professionals						
DSP04	**Data Scientist**	5	3	4	5	3
DSP05	**Data Science Researcher**	4	3	2	5	4
DSP06	**Data Science Architect**	4	3	5	3	3
DSP07	**Data Science (Application) Programmer/Engineer**	4	2	5	3	4
DSP08	**Data Analyst**	5	3	3	3	4
DSP09	**Business Analyst**	5	3	3	4	5
Professional (data handling/ management)						
DSP10	**Data Stewards**	3	5	3	3	3
DSP11	**Digital data curator**	1	5	2	2	3
DSP12	**Digital Librarians**	2	5	2	2	3
DSP13	**Data Archivists**	1	5	1	1	3
Professional (database)						
DSP14	**Large scale (cloud) database designer**	2	4	4	3	3
DSP15	**Large scale (cloud) database administrator**	2	4	3	2	3
DSP16	**Scientific database administrator**	2	4	3	2	3
Technicians and associate professionals						
DSP17	**Big Data facilities Operator**	1	4	4	2	3
DSP18	**Large scale (cloud) data storage operator**	1	4	3	1	1
DSP19	**Scientific database operator**	1	4	3	2	3

4.3 Learning outcomes for CF-DS and relation to taxonomy

The data Science Competence Framework (CF-DS) provides a guidance on what competences a future data scientist should have. The definition of competence is a useful starting point to formally define Learning Outcomes (LOs) that should guide development of future programs and courses. A similar approach is taken e.g. in ACM curriculum guidelines.

In guidelines for the Information Technology curriculum [10], a competence-based learning model is used and it focuses on the extent that students learn given competencies (knowledge, skills, qualifications), instead of focusing on so called „seat time". A competence model for constructing curricula is based on defining measurable learning outcomes, of which 50 are defined, instead of a set of topics. Learning Outcomes are then grouped in technical competencies and workplace skills. Each learning outcome can be assessed on a three-tier system, which roughly follows Bloom's taxonomy.

In guidelines for the Computer Science curriculum [11], the competence approach is not considered directly. The BoK is constructed based on topical themes, it is presented in Appendix E. It consists of 18 Knowledge Areas (KA), each containing several Knowledge Units (KU). Each KU has several topics and Learning Outcomes defined. We also attempted to establish a relation between various Knowledge Areas, it is presented in Appendix G. LOs complement the definition of topics by adding verbs relating to level of mastery, which roughly follow Bloom's taxonomy. We analyze the relation to Bloom's taxonomy later in one of the following subsections.

The European e-Competence Framework (e-CF) [19], despite being a competence framework, resembles ACM/IEEE's KAs and KUs, especially in dimension 2. EDISON D2.1 Table 3.1 presents example competence definitions for each DS competence group. These are later reflected in proposed extensions to e-CF. The extensions are general in their character. Data Science skills are defined in Table 3.2 and also Table 3.3, and they correspond to a certain degree with topics in ACM definition and example use of the extension in Section 4.9 follows a ACM-like two-tier approach dividing competences into essential and additional.

In this section we first compare Master levels as used in the European Qualifications Framework (EQF) [20], e-CF, ACM/IEEE guidelines for Computer Science curriculum [10][11] and Bloom's taxonomy. It leads to the definition of mastery levels necessary to define Learning Outcomes in MC-DS. We then follow with definition of Learning Outcomes for CF-DS (in particular EDISON's extensions to e-CF).

4.3.1 Mastery levels

The European qualification framework defines eight levels of knowledge achieved through stages of education. Level 6 is considered to be achieved through a bachelor degree, level 7 through a masters degree and level 8 through a PhD degree. Levels 3-8 are mapped to 5 levels in e-CF dimension 3. The mapping and description is presented in Table 7. By comparing e-CF levels directly with education requirements from EQF we can notice a certain mismatch. It is impossible to achieve a desired e-CF level by simply following an education path based on EQF. It is not enough to get a masters degree to become a Lead Professional. Rather, education requirements should be interpreted as a necessary condition, but not sufficient.

Table 7 Description of EQF and e-CF levels

EQF level	EQF level description	e-CF level	e-CF level description
8	Knowledge at the most advanced frontier, the most advanced and specialized skills and techniques to solve critical problems in research and/or innovation, demonstrating substantial authority, innovation, autonomy, scholarly or professional integrity.	e-5	**Principal** Overall accountability and responsibility; recognized inside and outside the organization for innovative solutions and for shaping the future using outstanding leading edge thinking and knowledge.
7	Highly specialized knowledge, some of which is at	e-4	**Lead Professional/Senior Manager**

	the forefront of knowledge in a field of work or study, as the basis for original thinking, critical awareness of knowledge issues in a field and at the interface between different fields, specialized problem-solving skills in research and/or innovation to develop new knowledge and procedures and to integrate knowledge from different fields, managing and transforming work or study contexts that are complex, unpredictable and require new strategic approaches, taking responsibility for contributing to professional knowledge and practice and/or for reviewing the strategic performance of teams.		Extensive scope of responsibilities deploying specialized integration capability in complex environments; full responsibility for strategic development of staff working in unfamiliar and unpredictable situations.
6	Advanced knowledge of a field of work or study, involving a critical understanding of theories and principles, advanced skills, demonstrating mastery and innovation in solving complex and unpredictable problems in a specialized field of work or study, management of complex technical or professional activities or projects, taking responsibility for decision-making in unpredictable work or study contexts, for continuing personal and group professional development.	e-3	**Senior Professional/Manager** Respected for innovative methods and use of initiative in specific technical or business areas; providing leadership and taking responsibility for team performances and development in unpredictable environments.
5	Comprehensive, specialized, factual and theoretical knowledge within a field of work or study and an awareness of the boundaries of that knowledge, expertise in a comprehensive range of cognitive and practical skills in developing creative solutions to abstract problems, management and supervision in contexts where there is unpredictable change, reviewing and developing performance of self and others.	e-2	**Professional** Operates with capability and independence in specified boundaries and may supervise others in this environment; conceptual and abstract model building using creative thinking; uses theoretical knowledge and practical skills to solve complex problems within a predictable and sometimes unpredictable context.
4	Factual and theoretical knowledge in broad contexts within a field of work or study, expertise in a range of cognitive and practical skills in generating solutions to specific problems in a field of work or study, self-management not within the guidelines of work or study contexts that are usually predictable, but are subject to change, supervising the routine work of others, taking some responsibility for the evaluation and improvement of work or study activities.		
3	Knowledge of facts, principles, processes and general concepts, in a field of work or study, a range of cognitive and practical skills in accomplishing tasks. Problem solving with basic methods, tools, materials and information, responsibility for completion of tasks in work or study, adapting own behavior to circumstances in solving problems.	e-1	**Associate** Able to apply knowledge and skills to solve straight forward problems; responsible for own actions; operating in a stable environment.

EQF descriptions provide reference both to actual levels of knowledge, but also to additional skills related to knowledge application, analysis, synthesis and evaluation. It is quite similar to Bloom's approach. At the same time, levels in EQF do not only correspond to higher levels of conceptualization, but also to more specialized knowledge, experience and interpersonal skills related to people management, and professional integrity and responsibility. e-CF adds to its description of typical tasks regarding their complexity and autonomy. Therefore, higher levels of EQF and e-CF should not just be seen directly as the same higher levels in Bloom. At the same time, higher levels in Bloom's taxonomy are necessary to move up in e-CF and EQF. It follows the earlier argument about education requirements forming necessary but not sufficient conditions.

EQF has 8 levels, e-CF has 5 levels and Bloom's has 6 levels. Designing LOs of whole programs is a balance between precision and avoiding micromanagement of further definition of courses, especially when designing a guideline for programs instead of a specific program. It might be useful to limit the amount of levels on which LOs are considered. Such an approach is used in ACM/IEEE Computer Science and Information Technology curricula guidelines. Information Technology guidelines [10] define the three levels as: emerging, developed and highly developed. Computer Science guidelines [11] define the three levels as: familiarity, usage, and assessment. Bloom's taxonomy defines the six levels as: knowledge, comprehension, application, analysis, synthesis and evaluation.

The three levels as used in ACM/IEEE Computer Science guidelines are of particular importance because significant parts of a related taxonomy and BoK is used in the definition of CF-DS and BoK-DS in EDISON. A description of these three levels is presented in Table 8. The verb usage is not fully consistent with the original Bloom's taxonomy [12] or revised version [13], which is acknowledged in the document.

In principle, these levels are useful, though the synthesis level of Bloom's taxonomy seems to be somewhat omitted both in the naming of levels and also in their description. Furthermore, the analysis level of Bloom's taxonomy is sometimes mixed with the evaluation level. Deeper inspection suggests that ACM/IEEE's familiarity level maps to knowledge and comprehension levels in Bloom's taxonomy. Further, usage level in ACM/IEEE maps to analysis level in Bloom's taxonomy; and finally, assessment level in ACM/IEEE maps to analysis level in Bloom's taxonomy. As a result synthesis and evaluation levels from Bloom's taxonomy are to a large extent omitted. Such omission might be acceptable for undergraduate curricula that ACM and IEEE consider in these documents.

Table 8 ACM/IEEE CS curricula master levels

Level	Description
Familiarity	The student understands what a concept is or what it means. This level of mastery concerns a basic awareness of a concept as opposed to expecting real facility with its application. It provides an answer to the question "What do you know about this?"
Usage	The student is able to use or apply a concept in a concrete way. Using a concept may include, for example, appropriately using a specific concept in a program, using a particular proof technique, or performing a particular analysis. It provides an answer to the question "What do you know how to do?"
Assessment	The student is able to consider a concept from multiple viewpoints and/or justify the selection of a particular approach to solve a problem. This level of mastery implies more than using a concept; it involves the ability to select an appropriate approach from understood alternatives. It provides an answer to the question "Why would you do that?"

While not required in undergraduate curricula, the holistic definition covering all EQF, e-CF levels, requires also full coverage of levels in Bloom's taxonomy. At the same time, limitation to 3 levels should be maintained to preserve simplicity and compatibility. We suggest the following three levels: familiarity as understood by knowledge and comprehension in Bloom's taxonomy, usage as understood by application and analysis in Bloom's taxonomy, creation as understood by synthesis and evolution in Bloom's taxonomy. We present the three levels together with action verbs in Table 9. Action verbs were defined based on the original and revised Bloom's taxonomy with adjustments tailored to Data Science curricula.

Table 9 Knowledge levels for learning outcomes in Data Science model curricula (MC-DS)

Level	Action Verbs
Familiarity	Choose, Classify, Collect, Compare, Configure, Contrast, Define, Demonstrate, Describe, Execute, Explain, Find, Identify, Illustrate, Label, List, Match, Name, Omit, Operate, Outline, Recall, Rephrase, Show, Summarize, Tell, Translate
Usage	Apply, Analyze, Build, Construct, Develop,

	Examine, Experiment with, Identify, Infer, Inspect, Model, Motivate, Organize, Select, Simplify, Solve, Survey, Test for, Visualize
Creation	Adapt, Assess, Change, Combine, Compile, Compose, Conclude, Criticize, Create, Decide, Deduct, Defend, Design, Discuss, Determine, Disprove, Evaluate, Imagine, Improve, Influence, Invent, Judge, Justify, Optimize, Plan, Predict, Prioritize, Prove, Rate, Recommend, Solve

4.3.2 CF-DS and extensions to e-CF

In Table 10 we recall EDISON's extensions to e-CF as presented in Table 3.4 Section 4.6 of D2.1. These competences will be used as a basis for LOs definition in the following subsection. The remaining competences, which are not new to Data Science, will be considered in Model Curricula in WP3.

Table 10 Proposed e-CF3.0 extension with the Data Science related Competences

Competence group	Competences related to Data Science
A. PLAN (and Design)	A.10* Organizational workflow/processes model definition/formalization A.11* Data models and data structures
B. BUILD (Develop and Deploy/ Implement)	B.7* Apply data analytics methods (to organizational processes/data) B.8* Data analytics application development B.9* Data management applications and tools B.10* Data Science infrastructure deployment
C. RUN (Operate)	C.5* User/Usage data/statistics analysis C.6* Service delivery/quality data monitoring
D. ENABLE (Use/Utilise)	D10. Information and Knowledge Management (powered by DS) - refactored D.13* Data presentation/visualisation, actionable data extraction D.14* Support business processes/roles with data and insight (support to D.5, D.6, D.7, D.12) D.15* Data management/preservation/curation with data and insight
E. MANAGE	E.10* Support Management and Business Improvement with data and insight (support to E.5, E.6) E.11* Data analytics for (business) Risk Analysis/Management (support to E.3) E.12* ICT and Information security monitoring and analysis (support to E.8)

New competences were not assigned to e-CF levels in D2.1. The refactored D10 was originally placed between e-3 to e-5. Most of the original e-CF competences from groups A, D, E were placed between e-3 and e-4, and from groups B and C between e-2 to e-3.

Table 11 provides an example of competences definition for different groups that is improved after initially proposed in D2.1.

Table 11 Competences definition for different Data Science competence groups

Data Analytics (DSDA)	**Data Management/ Curation (DSDM)**	**DS Engineering (DSENG)**	**Scientific/ Research Methods (DSRM)**	**DS Domain Knowledge (for example, Business Apps) (DSDK)**

Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSDA01 **Use predictive analytics to analyse big data and discover new relations**	DSDM01 Develop and implement data strategy, in particular, in a form of Data Management Plan (DMP)	DSENG01 Use engineering principles to research, design, prototype, data analytics applications, or develop structures, instruments, machines, experiments, processes, systems	DSRM01 Create new understandings and capabilities by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods	DSDK01 Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
DSDA02 **Use appropriate statistical techniques on available data to deliver insights**	DSDM02 Develop and implement data models including metadata	DSENG02 Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	DSRM02 Direct systematic study toward a fuller knowledge or understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSDK02 Use data to improve existing services or develop new services
DSDA03 **Develop specialized analytics to enable agile decision making**	DSDM03 Collect and integrate different data source and provide them for further analysis	DSENG03 Develops specialized data analysis tools to support executive decision making	DSRM03 Undertakes creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	DSDK03 Participate strategically and tactically in financial decisions that impact management and organizations
DSDA04 **Research and analyze complex data sets, combine different sources and types of data to improve analysis.**	DSDM04 Visualise complex and variable data.	DSENG04 Design, build, operate relational non-relational databases	DSRM04 Apply ingenuity to complex problems, develop innovative ideas	DSDK04 Provides scientific, technical, and analytic support services to other organisational roles
DSDA05 **Use different data analytics protforms to**	DSDM05 Develop and maintain a historical data	DSENG05 Develop solutions for secure and reliable data access	DSRM05 Ability to translate strategies into action plans and follow through	DSDK05 Analyse multiple data sources for marketing purposes

process complex data	repository of analysis results		to completion.	
		DSENG06 Prototype new data analytics applications	DSRM06 Contribute to and influence the development of organizational objectives	DSDK06 Analyse customer data to identify/optimise customer relations actions

4.3.3 Defining learning outcomes for Data Science competences

In Table 12 we present ten guiding Learning Outcomes corresponding to data science competences defined in CF-DS using master levels defined earlier and action verbs compliant with Bloom's taxonomy. In parenthesis by each LO we indicate which competences are related to this particular LO (based on both CF-DS and extensions to e-CF). Names of Learning Outcomes groups follow the names of competence groups.

By *guiding* we mean that they are abstracted from specific technologies and particular algorithms. This way they can be adjusted to fit the sneeds of a particular program or course. In particular, they should be fine-tuned to each program and course using relevant subsets of Data Science taxonomy. Several example mappings between Learning Outcomes and Taxonomy are presented in Appendix H.

Table 12 Learning outcomes for CF-DS and extensions to e-CF

Learning Outcomes group	Familiarity	Usage	Creation
DSDA	LO.01 Choose and execute existing analysis (DSDA02, DSDA05, C.5)	LO.02 Examine available data, and infer and visualize data insights (DSDM04, D.13, D.14)	LO.03 Assess, adapte, and combine data sources to improve analytics (DSDA04)
DSDA/DSENG		LO.04 Apply and develop data analytic methods and applications (DSDA01, DSDA03, DSENG02, DSENG03, , B.7, B.8)	
DSENG	LO.05 Configure and operate exsisting applications and services (DSENG04, C.6)	LO.06 Inspect, identify and make use of required security monitoring (DSENG05, E.12)	LO.07 Design and evaluate new analytics applications (DSENG06)
DSENG/DSDK			LO.08 Assess, design and evaluate data infrastructures (B.10)
DSDK	LO.09 Outline and translate domain knowledge and problems into an abstract mathematical framework (DSDK01)	LO.10 Model and experiments with domain problems and processes (DSDK04, A10)	LO.11 Evaluate, improve, design processes for data, information and knowledge management (DSDM01, DSDK02, D.10, A.10)
DSDK/DSDA			LO.12 Assess, influence, and prioritize organization improvement and risk management with data (DSDA03, DSDK03, DSDK04, E.10, E.11)
DSDM	LO.13 Collect and describe data for further	LO.14 Identify, organize and develop processes	LO.16 Evaluate, improve and design data models

	analysis (DSDM03)	for data, information and knowledge management (DSDM05, D.10, A.10) LO.15 Build and organize data models and preservation processes (DSDM02, D.15, A.11)	and preservation processes (DSDM02, D.15, A.11) LO.17 Plan, recommend and design data management applications and tools (B.9)

There are no separate Learning Outcomes defined for Research Methods competence group. Related competences are covered indirectly through Learning Outcomes in other groups. CF-DS is to be updated in further work in WP2; at the same time Learning Outcomes will start to be verified through work in WP3. These two activites will leaded to an updated version presented in future deliverables.

Further more, we suggest the following adjustments in CF-DS definition:

1. DSDM04 “Visualise complex and variable data”, should be moved from Data Management to Data Analytic competences.
2. DSENG04 “Design, build, operate, relational non-relation databases”, should be abstracted from suggesting a particular technological choices.
3. DSDK05 “Analyse multiple data sources for marketing purposes” and DSDK06 “Analyse customer data to identify/optimise customer relations actions” should be abstracted from suggestion a particular application domain.

For main competence groups learning outcomes spread all levels from Familiarity to Creation. At the same time learning outcomes overlapping two competence groups can be usually found on Usage and Creation levels, which is not surprising. Related competences build on top of basic computing, analytic and domain competences. As a result, they correspond to higher conceptual levels.

4.4 Summary

In this section we presented statistical approaches to analysis of skills and competences; we also overviewed existing taxonomies related to Data Science.

We defined taxonomy for data science professions family, including 19 professions ranging from Scientific Database Operator to Data Science (group) Manager. Finally. we defined 17 learning outcomes covering Data Science competence groups spread over three mastery levels: familiarty, usage, creation). Professions family together with learning outcomes are a basis for creation of Model Curricula in WP3.

5 Summary and further steps

5.1 Summary of achievements

We created EDISON Inventory of Data Science education resources. Main focus of Inventory was on academic programs because analysis of existing programs is an important component for desiging Model Curricula. We also included other resources such as academic and industriul courses, which become useful in further work in WP3. Inventory of programs was published as a service to Data Science cummunity. It is also open for correction and inclusion of new entries.

We initiatied work on an exchange format for information about Data Science education and training, with parthners including RDA, CODATA and Elixir. An agreement was reached regarding necessary fields. Generic standards for educational information exchange were identified and reviewed to determine to what extent they could cover our needs.

Subsequently, analysed programs in EDISON inventory. We noticed significant differences between Europe and outside Europe (mostly United States) in department from which Data Science programs originate. For Europe Computer Science departments are the main source, while it is Business Schools for programs outside Europe. In Europe, we also mark more programs coming from cross-department initiatives. It is important, because as we also demonstrated, cross-department collaboration leads to better balance between Data Science competences in the program.

Naming of programs seems to follow the source departments. Data Science, Big Data, Machine Learning are most common names in Europe, while Business Analytics and Data Analytics outside Europe. So far, most of programs are offered on Master level. Learning Outcomes are usually not defined and even if defined they do not reflect established educational frameworks, such as Bloom's taxonomy.

Finally, we defined taxonomy for data science professions family, including 19 professions ranging from Scientific Database Operator to Data Science (group) Manager. We also defined 17 learning outcomes covering Data Science competence groups spread over three mastery levels: familiarity, usage, and creation. Professions family together with learning outcomes are a basis for creation of Model Curricula in WP3.

5.2 Gap analysis and further work

Identifying all, or at least a majority, of relevant programs in data science is currently a difficult task, especially in Europe, due to language differences and lack of standardization.

Based on our analysis we recommend extending LRMI (Learning Resource Metadata Initiative) standard in further work in WP3. The common interchange format can help to improve the understanding of a wider spectrum of programs and courses, especially with an explicit link to standardized competences and learning outcomes, which could help to overcome some of the language issues in analysis of programs.

Better balance in programs is a key issue for designing future Data Science programs. The data analysis competence group tends to be covered relatively well in the majority of the programs, but either programming (and general computing) or domain competences are often missing. Programming (and general computing) competences are not well connected with data analysis and domain knowledge. Right now, students often have to wait until thesis work to explore such connections.

There is a need for cross department collaboration to improve the balance of available and future programs. It is necessary to include courses that connect competence from all three CF-DS competence groups early in the education process.

There are many competences to cover in a Data Science program, but each course should target several competences at the same time. This is possible if courses are properly defined w.r.t. learning outcomes, what is usually missing right now. It could be achieved, for instance, by exposing students to non-trivial problems through project-based courses, already in early stages of education; first year in bachelor programs and first semester in master programs.

Curricula should be competence-based and flexible regarding specific technologies and courses. Competences specific for Data Science, are not tied to particular technologies and can be adjusted for different programs and courses.

There is little interest in assessment forms, which are important in achieving higher levels of knowledge. Especially that a majority of Data Science learning outcomes reside high on the scale of Bloom's taxonomy.

Assessment forms should be considered with greater care to improve students' achievements of intended learning outcomes. Therefore, assessment forms should become integral part of Model Curricula.

6 References

[1] College & University Data Science Degrees, http://datascience.community/colleges, Last visited on 30.11.2015

[2] Internet Calendaring and Scheduling Core Object Specification (iCalendar), https://tools.ietf.org/html/rfc5545

[3] Schema.org specification of Event object, http://schema.org/Event

[4] XCRI-CAP Data Definitions Summary - PG rollout, http://www.alanpaull.co.uk/xcri/web/files/pgXCRI/ExamplePGSummarySpreadsheet.xls

[5] LRMI Specification Version 1.1, http://dublincore.org/dcx/lrmi-terms/1.1/

[6] Date and time format, http://www.iso.org/iso/home/standards/iso8601.htm

[7] Bologna Process – European Higher Education Area, http://www.ehea.info/article-details.aspx?ArticleId=73

[8] Currency codes, http://www.iso.org/iso/home/standards/currency_codes.htm

[9] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[10] Information Technology Competency Model of Core Learning Outcomes and Assessment for Associate-Degree Curriculum(2014) http://www.capspace.org/uploads/ACMITCompetencyModel14October2014.pdf

[11] Computer Science 2013: Curriculum Guidelines for Undergraduate Programs in Computer Science http://www.acm.org/education/CS2013-final-report.pdf

[12] Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

[13] Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing, Abridged Edition. Boston, MA: Allyn and Bacon..

[14] D. A. Kolb, Experiential learning: experience as the source of learning and development. Prentice-Hall, 1984.

[15] P. C. Blumenfeld, E. Soloway, R. W. Marx, J. S. Krajcik, M. Guzdial, and A. Palincsar, "Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning," Educational Psychologist, vol. 26, no. 3–4, pp. 369–398, 1991.

[16] T. W. Malone and M. R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for learning," Aptitude, learning, and instruction, vol. 3, pp. 223–253, 1987.

[17] M. Ben-Ari, "Constructivism in computer science education," Journal of Computers in Mathematics and Science Teaching, vol. 20, no. 1, pp. 45–73, 2001.

[18] J. Biggs, "Enhancing teaching through constructive alignment," Higher education, vol. 32, no. 3, pp. 347–364, 1996.

[19] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[20] European Qualifications Framework (EQF) [online] https://ec.europa.eu/ploteus/content/descriptors-page

[21] Diana Maynard, Yaoyong Li, and Wim Peters. Nlp techniques for term extraction and ontology population. In Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge, pages 107–127, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[22] Christina Yip Chung, Raymond Lieu, Jinhui Liu, Alpha Luk, Jianchang Mao, and Prabhakar Raghavan. Thematic mapping - from unstructured documents to taxonomies. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02, pages 608–610, New York, NY, USA, 2002. ACM.

[23] Philip Rich. The organizational taxonomy: Definition and design. Academy of Management Review, 17(4):758–781, 1992.

[24] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: methods, evaluation and applications, volume 123. IOS press, 2005.

[25] Maria Teresa Pazienza. A domain-specific terminology-extraction system. Terminology, 5(2):183–201, 1998.

[26] Uri Zernik. Lexical acquisition: exploiting on-line resources to build a lexicon. Psychology Press, 1991.

[27] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima, Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.

[28] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. Comput. Linguist., 22(1):1–38, March 1996.

[29] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. Computational linguistics, 19(1):61–74, 1993.

[30] Lois L Earl. Experiments in automatic extracting and indexing. Information Storage and Retrieval, 6(4):313–330, 1970.

[31] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In Proceedings of the 14th conference on Computational linguistics-Volume 3, pages 977–981. Association for Computational Linguistics, 1992.

[32] Eric Brill. A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language, pages 112–116. Association for Computational Linguistics, 1992.

[33] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. The balancing act: Combining symbolic and statistical approaches to language, 1:49–66, 1996

[34] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: methods, evaluation and applications, volume 123. IOS press, 2005.

[35] Ronald J. Brachman. What is-a is and isn't: An analysis of taxonomic links in semantic networks. Computer;(United States), 10, 1983.

[36] Barbara Liskov. Keynote address - data abstraction and hierarchy. In Addendum to the Proceedings on Object-oriented Programming Systems, Languages and Applications (Addendum), OOPSLA '87, pages 17–34, New York, NY, USA, 1987. ACM.

[37] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995

[38] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 216–225. Association for Computational Linguistics, 2010.

[39] Filippo Menczer. Lexical and semantic clustering by web links. Journal of the American Society for Information Science and Technology, 55(14):1261–1269, 2004.

[40] Guido Vetere and Maurizio Lenzerini. Models for semantic interoperability in service-oriented architectures. IBM Systems Journal, 44(4):887–903, 2005.

[41] Wlodarczyk, Tomasz Wiktor, and Thomas J. Hacker. "Problem-Based Learning Approach to a Course in Data Intensive Systems." Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on. IEEE, 2014.APA

Appendix A – Inventory of programs

Name	Country	University	Unit	Language	Level
Data Studies	Austria	Danube University Krems	Arts, Culture and Building	German, English	graduate
Marketing analysis	Belgium	Ghent University	Department of Marketing	English	graduate
Data Science	Denmark	Danish Technical University	Department of Applied Mathematics and Computer Science	English	graduate
Data Engineering	Denmark	Aalborg University	Department of Computer Science	English	graduate
Economics and business Administration - Business Intelligence	Denmark	Aarhus	School of Business and Social Sciences	English	graduate
Data Mining and Knowlegde management (DMKM)	EU	Erasmus Mundus program (france:Pierre-and-Marie-Curie University, romania, italy, spain)		English	graduate
Cloud Computing and Services	EU	Universidad Politecnica de Madrid (TU/e Eindhoven; UNS Nice Sophia-Antipolis)	EIT Digital Master School	English	graduate
Data Science	EU	Universidad Politecnica de Madrid (TU/e Eindhoven; UNS Nice Sophia-Antipolis)	EIT Digital Master School	English	graduate
Computational Big Data Analytics	Finland	University of Tampere	School of information sciences	English	graduate
Data Scientist designer	France	Data Science Tech	Data ScienceTech Institute MSc Programmes	English	graduate
Data Sciences and Business Analytics	France	Ecole central Paris & ESSEC Business School		English	graduate
Big Data	France	ENSAI ;Ecole nationale de la statistique et de l'analyse de l'information	Statistic and Computer Science	English	graduate
Big Data Analytics for Business	France	IESEG School of management	School of Management	English	graduate
Master Data Science (DSC)	France	Universite Nice Sophia Antipolis		English	graduate
Executive Big Data Analyst	France	Data ScienceTech Institute	Nice Sophia Antipolis Campus	English	graduate
Data Mining, Analytics and Knowledge discovery	France	University Paris 13	Institute Galilee	English	graduate
Data Sciences track within Mathematics	France	several french universities	5 universities	English	graduate
Master Data Science and Engineering	France	EURECOM	Graduate School and Research Center in Communication Systems	English	graduate
Master Data Mining and Knowledge Management	France, Romania, Italy, Spain		composed of six universities in four countries	English	graduate
M.Sc. Machine Learning and Data Mining	France	Universities of Saint-Etienne (France) and Alicante (Spain)		English	graduate
Data Engineering	Germany	Jacobs University	Mobility	English	graduate

Big Data Systems	Russia	National research University Higher School of Economics	Faculty of Business and Management	English	graduate
Data Science	Germany	Technical University Dortmund	Faculty of statistics	English	graduate
Data Science (DSC),Design, Implementation, and Usage of Data Science Instruments Specialization	EU	TU Berlin	EIT ICT Labs Master School	English	graduate
IT4BI Master Programme	Germany	TU Berlin	Erasmus Mundus Joint Master Degree's Programme in Information Technologies for Business Intelligence (IT4BI)	English	graduate
Management and Data Science	Germany	University of Luneburg (Leuphana)	Institute of electronic business processes (IEG)	English	graduate
Analytical Business Intelligence	Hungary	Budapest University of Technology and Economics	Department of telecommunications and media informatics	English	graduate
Computing with focus on Data analytics	Ireland	Dublin City University (DCU), Dublin Institute of Technology	School of Computing	English	graduate
Computing with focus on Information and knowledge management	Ireland	Dublin Institute of Technology	School of Computing	English	graduate
Computing with focus on Data analytics	Ireland	Dublin Institute of Technology	School of Computing	English	graduate
Data Science and Analytics	Ireland	University College Cork	Science, Engineering and Food Science	English	graduate
Computer Science - Data Analytics	Ireland	NUI Galway	College of Engineering and Informatics	English	graduate
Big Data Analytics and Social Mining	Italy	University of Pisa	Department of Computer Science	English	Graduate
BABD - International master in Business Analytics and Big Data	Italy	Politecnico di Milano	MIP, Cefriel	English	Graduate
Data Scientist	Italy	University of Bologna	Bologna Business School	English	Graduate
Business Intelligence and Big Data Analytics	Italy	University of Milano-Bicocca	School of Ecnomics and Statistics, Department of Statistics and Quantitative Methods	English	Graduate
Big Data Analytics and Technologies for Management	Italy	University of Florence	Department of Science Economics and Business	English	Graduate
Data Science	Italy	University of Rome - Sapienza	Department of Informatics (DI), Department of Computer, Control and Management Engineering (DIAG), Department of Information Engineering, Electronics and Telecommunications (DIET), Department of Statistics (DSS)	English	Graduate
Big Data Management	Italy	Luiss Business School	Department of Economica and Finance	English	Graduate
Stochastics and Data Science	Italy	University of Torino	Department of Mathematics,	English	Graduate

			Department of ESOMAS, Department of Computer Science		
Marine science, Ocean physics and technology	Italy	University of Bologna, University of Naples - Partehenope	School of Sciences, Department of Physics and Astronomy	English	Graduate
Data Science	Italy	University of Rome - Tor Vergata	Department of Hitorical, Philosophical and Social, Cultural Heritage and Territory, Department of Enterprise, Governament, Philosophy and Civil Engineering and Computer Engieneering	English	Graduate
Customer Experience and Social Media Analytics	Italy	University of Rome - Tor Vergata	Department of Corporate Governance and Philosophy	English	Graduate
Computer Science and Engineering	Italy	University of Bologna	Department of Computer Science and Engineering	English	
Information Systems (specialization: Data Science)	Liechtenst ein	Universität Liechtenstein	Institute of Information Systems	English	graduate
track within Computer Science: Data Science and Technology track	Netherlan ds	Delft University of Technology	Computer science	English	graduate
Business Analytics and Quantitative Marketing (track within Master Econometrics and Management Science)	Netherlan ds	Erasmus University Rotterdam	Erasmus School of Economics	English	graduate
master/Business Information Management	Netherlan ds	Erasmus University Rotterdam	Rotterdam School of Management	English	graduate
Data Science (track of Computing Science)	Netherlan ds	Radboud university	Faculty of Science: Institute for Computing and Information Sciences	English	graduate
Data Science in Engineering	Netherlan ds	Technische Universiteit Eindhoven	Mathematics and computer science	English	graduate
Data Science: Business and Governance	Netherlan ds	Tilburg University	Economics and Management, Law, Social and behavioral, humanities	English	graduate
Data Science	Netherlan ds	Tilburg University & TU Eindhoven	Economics and Management, Law, Social and behavioral, humanities	English	under graduate
Data Science in Engineering	Netherlan ds	TU Eindhoven	Mathematics and computer science	English	graduate
MASTER'S PROGRAMME BUSINESS INFORMATION TECHNOLOGY(Business analytics specialization)	Netherlan ds	Universiteit Twente	Faculty of Electrical Engineering, Mathematics and Computer Science	English	graduate
MASTER'S PROGRAMME COMPUTER SCIENCE (Data Science and Smart Services specialization)	Netherlan ds	Universiteit Twente	Faculty of Electrical Engineering, Mathematics and Computer Science	English	graduate
MASTER'S PROGRAMME APPLIED MATHEMATICS (Operations Research specialization)	Netherlan ds	Universiteit Twente	Faculty of Electrical Engineering, Mathematics and Computer Science	English	graduate
MBA: Big Data & Business	Netherlan	University of Amsterdam	Amsterdam Business	English	graduate

Analytics	ds		School		
Econometrics: Big Data Business Analytics	Netherlands	University of Amsterdam	Amsterdam School of Economics	English	graduate
Artificial Intelligence (with Data Science specialization)	Netherlands	University of Amsterdam	Faculty of Science	English	graduate
Business analytics	Netherlands	Vrije Universiteit Amsterdam	Department of Mathematics	English	under graduate
Modelling and Data analysis	Norway	University of Olso	Department of Mathematics	English	graduate
ADVANCED ANALYTICS	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Information Analysis and Management	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Risk Analysis and Management	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Marketing Research and CRM	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Knowledge Management and Business Intelligence	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Information Systems and Technologies Management	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Specialization in Marketing Intelligence	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Geospatial Technologies (Erasmus Mundus)	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	English	graduate
Doctorate in Information Management	Portugal	NOVA Information Management School (|NOVA IMS)	NOVA IMS	Portuguese/English	graduate
Data Analytics and Decision Support Systems	Portugal	University of Porto	School of economics and management	English	graduate
Big Data Analytics	Russia	Novosibirsk State University	Science and Tech	English	graduate
Data Science	Spain	Barcelona Graduate School of Economics		English	graduate
Business Analytics and Big Data	Spain	IE School of social and behavioral sciences	Social and behavioral sciences	English	graduate
Big Data Analytics	Spain	UC3M; Universidad Carlos III de Madrid	Graduate School of engineering and basic sciences	English	graduate
Data Science	Spain	Universitat Autonoma de Barcelona	Graduate School of Economics	English	graduate
Master of Science in IT Strategic Management	Spain	Universitat Pompeu Fabra	Barcelona School of Management	English	graduate
Business Intelligence	Sweden	Dalarna University	-	English	graduate
Data Science	Sweden	University of Skovde		English	graduate
Business Intelligence	UK	Birmingham City University		English	graduate
Data Science and Analytics	UK	Brunel University London		English	graduate
Data Science MSc	UK	City University London	School of mathematics, computer science and engineering ; Department of Computer Science	English	graduate

Data Science and Computational Intelligence	UK	Coventry University	Faculty of Engineering and computing	English	graduate
Business Intelligence Systems and Data Mining	UK	De Montfort University		English	graduate
Data Science (new since sep 2014)	UK	Goldsmiths University of London	Department of Computing	English	graduate
Data Science	UK	Heriot Watt University	School of Mathematical and Computer Sciences	English	graduate
Advanced Computing	UK	Imperial College London	Data Science Institue	English	graduate
Business Analytics	UK	Imperial College London	Business School	English	graduate
Computing (Machine Learning)	UK	Imperial College London	Data Science Institue	English	graduate
Data Science(Computing Specialism, Statistical Inference Specialism, Environment Specialism)	UK	Lancaster University	School of Computing and Communications; Department of Mathematics and statistics; lancaster Environment Centre	English	graduate
Network Science	UK	Queen Mary University of London	School of Mathematical Sciences	English	graduate
Data Science and Analytics	UK	Royal holloway University of London	Computer Science	English	graduate
Machine Learning	UK	Royal holloway University of London	Computer Science	English	graduate
Machine Learning	UK	University College London	Engineering Sciences(Department of Computer Science)	English	graduate
Web Science and Big Data Analytics	UK	University College London	Engineering Sciences(Department of Computer Science)	English	graduate
Data Science	UK	University of Bedfordshire	Computing and Information Systems	English	undergra duate
MSc in Advanced Computing - Machine Learning, Data Mining and High Performance Computing	UK	University of Bristol	Faculty of Engineering, department of computer Science	English	graduate
Data Science	UK	University of Dundee	School of Computing	English	graduate
Knowledge Discovery and Data Mining	UK	University of East Anglia	Computing Science	English	graduate
Data Science	UK	University of Edinburgh	Informatics Centre for doctoral training in Data science	English	graduate
High Performance Computing with Data Science	UK	University of Edinburgh	College of Science (Astronomy and Physics)	English	graduate
Data Science	UK	University of Glasgow	School of Computing	English	graduate
Advanced Computer Science (Data Analytics)	UK	University of Leeds	Faculty of Engineering	English	graduate
Advanced Computer Science (Cloud Computing) MSc	UK	University of Leeds	Faculty of Engineering(School of Computing)	English	graduate
Advanced Computer Science (Intelligent Systems) MSc	UK	University of Leeds	Faculty of Engineering(School of Computing)	English	graduate
Big Data Management	UK	University of Liverpool	Management School	English	graduate
Master Big Data and High Performance Computing	UK	University of Liverpool	Department of computer science	English	graduate
Advanced Computer Science with Internet Economics	UK	University of Liverpool	Department of computer science	English	graduate
Advanced Computer Science	UK	University of Liverpool	Department of computer science	English	graduate

Data and Knowledge Management	UK	University of Manchester	School of Computer Science	English	graduate
Data Science	UK	University of Sheffield	Information School, School of Social Sciences	English	graduate
Big Data and Quantitative Methods	UK	University of Warwick	Politics and International studies & Centre for Interdisciplinary Methodologies	English	graduate
Big Data and the Digital Futures	UK	University of Warwick	Centre for Interdisciplinary Methodologies	English	graduate
MSc Data Analytics	UK	University of Warwick	Department of Computer Science	English	graduate
Data Science	UK	University of Warwick	Department of static and Department of Computer Science	English	undergraduate
Data Science	UK	Worcester Polytechnic Institute		English	graduate
Data Science	UK	university of southampton	Electronics and Computer Science (ECS)	English	graduate
Statistics	UK	University of Southampton	Mathematical Sciences	English	graduate
Statistics with Applications in Medicine.	UK	University of Southampton	Mathematical Sciences	English	graduate
Business Informatics	Lithuania	Mykolas Romeris University	BUSINESS AND MEDIA SCHOOL	English	graduate
Big Data Engineering	Italy	Polytechnic University Of Turin	Mathematical Sciences	Italian	graduate
MSc Degree in Information Systems Engineering with Focus on Data Mining and Business Intelligence	Israel	Ben-Gurion University Of The Negev	Department of Information Systems Engineering	English	graduate
Master of Science in Computing (Business Intelligence & Data Mining)	Ireland	Institute Of Technology Blanchardstown	INFORMATICS AND ENGINEERING	English	graduate
MSc in Data Business	Ireland	Irish Management Institute		English	graduate
MSc in Data Analytics	Ireland	National College Of Ireland	School of Computing	English	graduate
MSC (INFORMATION SYSTEMS MANAGEMENT)	Ireland	National University Of Ireland, Galway	College of Business, Public Policy, and Law	English	graduate
MSc Business Analytics	UK	Aston University	Aston Business School	English	graduate
MSc Applied Data Analytics	UK	Bournemouth University	Department of Computing & Informatics:	English	graduate
Business Intelligence and Social Media MSc	UK	Brunel University London	Brunel Business School	English	graduate
Data Science and Analytics MSc	UK	Brunel University London	Department of Computer Science	English	graduate
Data Science	UK	Goldsmiths, University of London	Department of Computing	English	graduate
Cloud Computing MSc	UK	Newcastle University	School of Computing Science	English	graduate
Data Science and Analytics (MSc)	UK	Royal Holloway, University Of London	Department of Computer Science	English	graduate
MSc Big Data Analytics	UK	Sheffield Hallam University	Department of Computing	English	graduate
Data and Knowledge Management MSc	UK	The University Of Manchester	School of Computer Science	English	graduate
Computational Statistics And Machine Learning	UK	University College London	DEPARTMENT OF COMPUTER SCIENCE	English	graduate

MSc Big Data Analytics	UK	University of Derby	Department of Computing and Mathematics	English	graduate
BSc (Hons) Data Science*	UK	University of Derby	Department of Computing and Mathematics	English	undergraduate
Analytics (Joint Honours)	UK	University of Derby	College of Engineering and Technology	English	undergraduate
MSc Data Science	UK	University Of East London	Architecture, Computing and Engineering	English	graduate
MSc Big Data and Text Analytics	UK	University Of Essex	School of Computer Science and Electronic Engineering	English	graduate
Big Data and Business Intelligence, MSc	UK	University Of Greenwich	FACULTY OF ARCHITECTURE, COMPUTING & HUMANITIES	English	graduate
Business Analytics - MSc	UK	University of Kent	KENT BUSINESS SCHOOL	English	graduate
Data Analysis for Business Intelligence	UK	University of Leicester	Computer Science, Mathematics	English	graduate
Data Science BSc	UK	University of Nottingham	School of Mathematical Sciences School of Computer Science	English	undergraduate
MSc Applied Statistics and Data Mining	UK	University Of St Andrews	School of Mathematics & Statistics	English	graduate
Business Analysis and Consulting	UK	University of Strathclyde	Strathclyde Business School	English	graduate
Business Analytics MSc	UK	University of Surrey	BUSINESS AND MANAGEMENT	English	graduate
Business Intelligence And Analytics	UK	University Of Westminster	Science and Technology	English	graduate
INTERNATIONAL MASTER IN BIG DATA, DATA ANALYTICS, DATA SCIENCE, DATA ARCHITECTURE	France	EISTI	GRADUATE SCHOOL IN COMPUTER SCIENCE AND MATHEMATICS ENGINEERING	English	graduate
Data Science Specialisation	France	ENSAE Paris Tech	Statistician Economist		undergraduate
Big Data at Telecom ParisTech - Data Scientist - Machine Learning	France	Telecom Paris Tech	the college of innovation through digital technology		
Master of Science Data Analysis and Pattern Classification	France	Telecom Sudparis	Institut National des Télécommunications, Département EPH	English	graduate
Master's Programme in Machine Learning and Data Mining	Finland	Aalto University School of Science	Department of Information and Computer Science	English	graduate
Algorithms, Data Analytics and Machine Learning	Finland	University Of Helsinki	Department of Computer Science	English	graduate
Data Mining and Knowledge Management	France	Université De Nantes	Département d'Informatique et Statistique	English	graduate
Master in Business Analytics and Big Data	Spain	Instituto de Empresa	School of Human Sciences and Technology	English	graduate
Machine Learning And Data Mining	Spain	Universities Of Alicante	Ministry of Education, Culture and Sports	English	graduate
Industrial Phd In Big Data Analysis	Denmark	Aarhus University	Department of Computer Science	English	graduate
MASTER'S DEGREE PROGRAMME IN ECONOMICS AND BUSINESS ADMINISTRATION -	Denmark	Aarhus University	SCHOOL OF BUSINESS AND SOCIAL SCIENCES	English	graduate

BUSINESS INTELLIGENCE (MSC)					
DIGITAL MEDIA ENGINEERING, Data Science focus	Denmark	Technical University of Denmark	DTU Compute	English	graduate
Data and Knowledge Engineering	Germany	Otto Von Guericke University Magdeburg	Faculty of computer science	English/ German	graduate
Information Engineering - Bachelor of Science	Germany	Universität Konstanz	Department of Computer & Information Science	English/ German	graduate
Big Data Analytics	Brazil	Mackenzie Presbyterian Institute	Information Technology		graduate
Master of Business Analytics	Australia	Deakin University	Faculty of Business and Law	English	graduate
BACHELOR OF ARTS WITH A MAJOR IN DATA SCIENCE	Australia	Macquarie University	Department of Statistics Faculty of Science and Engineering	English	graduate
Master of Data Science	Australia	University of South Australia	School of Information Technology & Mathematical Sciences	English	graduate
Bachelor of Science in Analytics	Australia	University of Technology Sydney	FACULTY OF SCIENCE	English	graduate
Master of Science in Computational Biology and Quantitative Genetics	USA	Harvard University		English	graduate
Master of Science in Data Science	USA	Columbia University in the City of New York		English	graduate
Information Management and Analytics	USA	Stanford University		English	graduate
Biomedical Informatics MS Degree	USA	Stanford University	school of medicine	English	graduate
MS in Statistics: Data Science	USA	Stanford University	Department of Statistics	English	graduate
Master of Science in Computational Analysis & Public Policy	USA	University of Chicago	Harris School of Public Policy and the Computer Science Department	English	graduate
Master of Science in Analytics	USA	University of Chicago	Graham School	English	graduate
MASTER OF SCIENCE IN ANALYTICS	USA	Northwestern University	McCormick School of Engineering	English	graduate
Online Master's in Predictive Analytics	USA	Northwestern University	school of professional study	English	graduate
Geographic Information Systems	USA	Johns Hopkins University	advanced academic programs	English	graduate
MS in Information Systems	USA	Johns Hopkins University	carey business school	English	graduate
Master's Study in Applied Statistics	USA	University of Notre Dame	Department of Applied and Computational Mathematics and Statistics	English	graduate
Master of Science in Business Analytics	USA	University of Notre Dame	MENDOZA COLLEGE OF BUSINESS	English	graduate
Computational and Data Sciences	USA	George Mason University	Department of Computational and Data Sciences (CDS)	English	Undergraduate
Marketing Analysis	USA	Bentley University	School of Business	English	graduate
MISM Business Intelligence and Data Analytics	USA	Carnagie Mellon University	School of Information Systems and Management	English	graduate
	USA	Carnegie Mellon University	School of Computer Science	English	graduate
Predictive Analytics	USA	DePaul University	College of Computing	English	graduate

			and Digital Media		
CS Specialization in Data Analytics	USA	Illinoise Institute of Technology	Department of Computer Science	English	graduate
Business Analytics	USA	Lousiana State University	Department of Information Systems & Decisions Sciences	English	graduate
Business Analytics	USA	Michigan State University	Department of Accouting & Information System	English	graduate
Analytics	USA	North Carolina State University	Master of Science in Analytics	English	graduate
Predictive Analitycs (E-learning)	USA	Northwestern University	School of Professional Studies	English	graduate
Business Analytics	USA	NewYork University	Center for Business Analytics	English	graduate
Business Intelligence and Analytics	USA	Stevens Institute of Technology	School of Business	English	graduate
Business Analytics	USA	University of Cincinnati	Lindner College of Business	English	graduate
Analytics	USA	University of San Francisco	College of Art and Science	English	graduate
Business Analytics Degree	USA	auburn university	Department of Aviation and Supply Chain Management	English	under graduate
MS in Marketing with a Specialization in Marketing Analytics	USA	the university of Alabama	Alabama Operations Management faculty	English	graduate
MASTER OF SCIENCE IN APPLIED STATISTICS	USA	the university of Alabama	online	English	graduate
Business Analytics	USA	the university of Alabama	online	English	graduate
Master of Science In Management Science - Business Analytics(MSMS-BA)	USA	The University of Alabama in Hunstville	College of Business Administration	English	graduate
Business Data Analytics	USA	Arkansas Tech University	Accounting & Economics Faculty	English	under graduate
master of science in business analytics	USA	Arizona State University	Department of Information Systems	English	graduate
Business Data Analytics	USA	Arizona State University	Department of Information Systems	English	under graduate
Online Master of Science in Business Analytics	USA	Arizona State University	Department of Information Systems campus online	English	graduate
Cross Disciplinary Studies Minor in Data Science	USA	California Polytechnic State University	Computer Science Department	English	joint program
Business Analytics: Master of Science In Business Administration	USA	California State University-East Bay	college of businuss and economics	English	graduate
MS in Information Systems and Decision Sciences concentration	USA	California State University-Fullerton	MBA and graduate programs	English	graduate
Doctorate in Computational and Data Sciences	USA	Chapman University	computational science faculty	English	graduate
Master's of Science in Data Science	USA	Galvanize U		English	graduate
Master of Science in Data Analytics	USA	National University	Computer Science, Information and Media Systems	English	graduate
Master of Science in Health & Life Science Analytics	USA	National University	School of Health and Human Services	English	graduate
MIMS Program	USA	University of California Hastings College of Law	school of information	English	graduate

B.S. in Data Science	USA	University of California-Irvine	Department of Statistics	English	under graduate
About the Master of Advanced Study in Data Science and Engineering	USA	University of California-San Diego	Departments of Computer Science & Engineering	English	graduate
Information and Data Science	USA	University of California, Berkeley	I School faculty	English	graduate
Data Science	USA	University Of San Francisco	Department of Mathematics and Statistics	English	under graduate
Master of Science in Computer Science with Specialization in Data Science	USA	University of Southern California	COMPUTER SCIENCE	English	graduate
Master of Science in Business Analytics	USA	USC University of Southern California	Data Sciences and Operations	English	graduate
Master of Science in Analytics	USA	University of the Pacific	School of Engineering and Computer Science	English	graduate
Master's in Business Intelligence and Analytics	USA	American Sentinel University	Healthcare management program	English	graduate
Master of Applied Statistics (M.A.S.)	USA	Colorado State University-Fort Collins	Department of Statistics	English	graduate
Doctor of Computer Science Big Data Analytics	USA	Colorado Technical University	CTU online doctoral	English	graduate
M.S. Data Science	USA	Regis University	College of Computer & Information Sciences	English	graduate
DECISION SCIENCES MS	USA	University of Colorado Denver	Business School	English	graduate
business analytics big data specialization	USA	University of Colorado Denver	Business School	English	graduate
MS in Information Systems – Business Intelligence systems	USA	University of Colorado Denver	Business School	English	graduate
Master of Science in Business Analytics	USA	University of Denver	Daniels College of Business	English	graduate
data mining	USA	Central Connecticut State University	CCSU Faculty	English	graduate
MS IN BUSINESS ANALYTICS	USA	Quinnipiac University	school of business	English	graduate
MS in Business Analytics and Project Management	USA	University of Connecticut	school of business	English	graduate
Master’s in Business Analytics Degree Program	USA	American University	American University Kogod School of Business	English	graduate
A Master’s in Business Analytics	USA	American University	Kogod's online Master of Science in Analytics	English	graduate
Master of science in Data Science	USA	George Washington University	Columbian College of Arts & Sciences	English	graduate
Master of Science in Analytics, Concentration in Data Sciences (MS-DS)	USA	Georgetown University	Grdauate School of Arts and sciences	English	graduate
MS IN BUSINESS ANALYTICS	USA	The George Washington University	GWSB’s MSBA program	English	graduate
BIG DATA ANALYTICS	USA	Florida Polytechnic University	College of Innovation & Technology /Advanced Technology	English	under graduate
Master of Science in Business Intelligence	USA	Full Sail University	Business School	English	graduate
M.S. in Statistical Computing Data Mining Track	USA	University of Central Florida	Department of Statistics	English	graduate
Master of Science in Health Care Informatics	USA	University of Central Florida	UCF's Department of Health Management and Informatics	English	graduate

Master of Science in Information Systems and Operations Management	USA	University of Florida	Department of ISOM	English	graduate
Master of Science in Business Analytics	USA	University of Miami	school of business and administration	English	graduate
Master of Science in analytics join the Big Data Revolution	USA	Georgia State University	Robinson's faculty resources and research	English	graduate
Master's in Analytics	USA	Georgia Tech	Scheller College of Business, the College of Computing, and the College of Engineering	English	graduate
Master of Science in Applied Statistics (MSAS)	USA	Kennesaw State University	Department of Statistics and analytical sciences	English	graduate
Concentration in Business Analytics	USA	University of Georgia	Full-Time MBA Program	English	graduate
MASTER BUSINESS ANALYTICS concentration	USA	Loras College	MBA program	English	graduate
Business Analytics and Information Systems	USA	The University of Iowa	Tippie college of business	English	under graduate
BUSINESS ANALYTICS master's and certificate	USA	The University of Iowa	Tippie college of business	English	graduate
Master of Science in Digital Marketing and Analytics	USA	Aurora University	Dunham School of Business	English	graduate
Master of Science in Business Analytics	USA	Benedictine University	College of Business	English	graduate
master of science Predictive Analytics	USA	DePaul University	college of computing and digital media	English	graduate
Business Intelligence Concentration	USA	DePaul University	college of computing and digital media	English	graduate
M.S. in Data Science	USA	Elmhurst College	School for Professional Studies	English	graduate
Master of Computer Science with a Specialization in Data Analytics	USA	Illinois Institute of Technology	College of Science	English	graduate
Master of Data Science	USA	Illinois Institute of Technology	College of Science	English	graduate
Data Science, M.S.	USA	Lewis University	Computer Science	English	graduate
Business Analytics, M.S.	USA	Lewis University	College of Business	English	graduate
MS in Applied Statistics	USA	Loyola University Chicago	Department of Mathematics and Statistics	English	graduate
Master of Science in Statistics: Analytics Concentration	USA	University of Illinois at Urbana-Champaign	department of statistic	English	graduate
Master of Business Administration, concentration business analytics	USA	University of St Francis	USF online MBA	English	graduate
Business Analytics	USA	Indiana University Bloomington	Full-Time MBA Program	English	graduate
Online MS in Business Analytics	USA	Indiana University Bloomington	KELLEY SCHOOL OF BUSINESS	English	graduate
Data Science M.S.	USA	Indiana University Bloomington	SCHOOL OF INFORMATICS AND COMPUTING	English	graduate
M.S. in Mathematics: Applied Statistics	USA	Indiana University-Purdue University Indianapolis	department of mathematical science	English	graduate
KRANNERT FULL-TIME MBA | BUSINESS ANALYTICS CONCENTRATION	USA	Purdue University-Main Campus	KRANNERT school of management	English	graduate

Master of Science in Data Science	USA	Saint Mary's College	Graduate Programs	English	graduate
BUSINESS ANALYTICS CONCENTRATION	USA	Babson College	MBA program	English	graduate
Bachelor of Science in Data Science	USA	Becker College	Data Science	English	under graduate
Masters in Business Analytics Data Science cluster	USA	Bentley University	school of business	English	graduate
MASTER'S PROGRAM IN COMPUTATIONAL LINGUISTICS	USA	Brandeis University	Graduate School of Arts and Sciences	English	graduate
Master of Science in Strategic Analytics	USA	Brandeis University		English	graduate
MS in Urban Informatics	USA	Northeastern University	data science faculty	English	graduate
PhD Program in data science	USA	Worcester Polytechnic Institute	data science faculty	English	graduate
BS/MS Data Science Degree Program	USA	Worcester Polytechnic Institute	data science faculty	English	under graduate
Data Science Certificate Program	USA	Worcester Polytechnic Institute	data science faculty	English	graduate
Analytics in Knowledge Management	USA	Notre Dame of Maryland University	School of Arts and Sciences	English	graduate
Master of Information Management (MIM)	USA	University of Maryland-College Park	college of information study	English	graduate
Master of Science in Data Analytics	USA	University of Maryland-College Park	Business and Management	English	graduate
Master of Business Administration	USA	Baker College	MBA program	English	graduate
Master of Science in Applied Statistics and Analytics	USA	Central Michigan University	Department of Mathematics	English	graduate
Master of Science in Information Systems (MSIS)	USA	Eastern Michigan University	Computer Information Systems	English	graduate
MS IN BUSINESS ANALYTICS	USA	Michigan State University	Eli broad college of business	English	graduate
INTEGRATED GEOSPATIAL TECHNOLOGY—MS	USA	Michigan Technological University	SCHOOL OF TECHNOLOGY	English	graduate
Master of Science in Information Technology Management	USA	Oakland University	School of Business Administration	English	graduate
Undergraduate Program in Data Science	USA	University of Michigan-Ann Arbor	EECS Department in the College of Engineering and the Department of Statistics in the College of LSA	English	under graduate
MS-Business Analytics	USA	University of Michigan-Dearborn	College of Business	English	graduate
Business Intelligence specialization master of business administartion	USA	Capella University	MBA program	English	graduate
MS in Analytics program	USA	Capella University	school of business & technology	English	graduate
M.S. Health Informatics	USA	The College of Saint Scholastica	graduate program	English	graduate
M.S. in Data Science	USA	University of St. Thomas	School of Engineering	English	graduate
Data Science	USA	Winona State University	Data Science Program	English	under graduate
M.S. IN BUSINESS INTELLIGENCE AND ANALYTICS	USA	Rockhurst University	Helzberg School of Management	English	graduate
Bachelor of Science in Information Science	USA	Elon University	College of Arts & Sciences	English	under graduate

M.S. IN ANALYTICS	USA	North Carolina State University at Raleigh	Institute for advanced analytics	English	graduate
Data Science and Business Analytics (DSBA)	USA	University of North Carolina at Charlotte	Professional Science Master's (PSM) program	English	graduate
Master of Science in Business Analytics Degree	USA	Bellevue University	College of Business	English	graduate
Master of Professional Science in Technology Innovation and Entrepreneurship Degree	USA	Bellevue University	College of Science and Technology	English	graduate
Business Intelligence and Analytics (Master of Science)	USA	Creighton University	Heider college of business	English	graduate
Undergraduate Programs in Mathematics data science concentration	USA	University of Nebraska at Omaha	Department of Mathematics	English	under graduate
Data Science	USA	University of Nebraska at Omaha	Department of Mathematics	English	graduate
Online Master's DegreeMS in Data Analytics	USA	Southern New Hampshire University	graduate program	English	graduate
Information Technology (MS)Database Design	USA	Southern New Hampshire University	graduate program	English	graduate
MBA in Business Intelligence	USA	Southern New Hampshire University	MBA program	English	graduate
BS in Data Analytics	USA	Southern New Hampshire University	Online Bachelor's Degree	English	under graduate
Analytics & Data Sciences	USA	Rutgers University	Professional Science Master's program	English	graduate
Master of Science in Data Science with a concentration in Business Analytics	USA	Saint Peter's University	Data Science graduate program	English	graduate
Master of Science in Information Systems concentration in business intelligence and analytics	USA	Stevens Institute of Technology	school of business	English	graduate
MBA in area Data Analytics	USA	Thomas Edison State College	School of Business and Management	English	graduate
Master of Business Administration Online	USA	Auburn University	Raymond J. Harbert College of Busines	English	graduate
Business Analytics Degree	USA	Auburn University	Raymond J. Harbert College of Busines	English	Undergra duate
Master in Information Systems business analytics concentration	USA	University of Arkansas	Walton College, Graduate school of Business	English	graduate
Big Data	Canada	Simon Fraser University SFU	School of Computer Science	English	graduate
Data Science	online	SMU Southern Methodist University	online	English	graduate
Data Analytics	online	Southern New Hampshire University	online	English	graduate
Business Analytics	Australia	University of Melbourn	Melbourne Business School	English	graduate
Electronic Business Technologies	Canada	University of Ottowa	Telfer School of Management, School of information technology and engineering and the faculty of Law	English	graduate
Data Science and Innovation	Australia	University of Technology, Sydney	Analytics and Data Science	English	graduate
Data Analytics	Canada	Western University Canada	Ivey Business School		
Business Analytics	Canada	York University	Schulich School of Business	English	graduate
Management Analytics	Canada	Queen's University	Smith School of	English	graduate

			Business		
Statistical Machine Learning	Canada	University of Alberta	Department of Computing Science or Department of Mathematical and Statistics	English	graduate

Appendix B – Inventory of academic courses

Course title	Country	University	Unit	Level
Introduction to Data Science	USA	University of Washington	eScience Institute	graduate
Business Intelligence from Big Data	USA	Stanford University	Schoole of Business	graduate
Massive Data Analysis	USA	NYU	Polytechnic School of Engineering	graduate
Precision Practice with Big Data	USA	Stanford University	Medicine	graduate
Analyzing Big Data with Twitter	USA	UC Berkeley	School of Information	all
Web Intelligence and Big Data	USA	Indian Institute of Technology		other
Big Data: Making Complex Things Simpler	USA	MIT	Center for Digital Business	graduate
Introduction to Data Science	USA	Columbia University	Department of Statistics	graduate
Applied Data Science	USA	Columbia University	Department of Statistics	graduate
CAS Big Data Analytics	Switzerland	Hochschule Luzern	Wirtschaft	other
In-Memory Data Management 2015	Germany	Hasso-Plattner-Institut		other
Big Data and Business Analytics	France	HEC Paris	MBA Program	graduate
Marketing Analytics	USA	Bentley University	Graduate School of Business	graduate
INTRODUCTION TO DATA SCIENCE	USA	Worcester Polytechnic Institute	Arts & Sciences	all
BIG DATA MANAGEMENT	USA	Worcester Polytechnic Institute	Arts & Sciences	all
BIG DATA ANALYTICS	USA	Worcester Polytechnic Institute	Arts & Sciences	all
Data Science	USA	Harvard university	Faculty of Arts & Sciences	all
Algorithms for Big Data	USA	Harvard university	Faculty of Arts & Sciences	graduate
Big Data Systems	USA	Harvard university	Faculty of Arts & Sciences	graduate
Process Mining: Data science in Action		Eindhoven University of Technology	oniline coursera	certificate
Mining Massive Datasets		Stanford University	oniline coursera	certificate
Data Mining I, Data Mining II	USA	Arizona State University	W. P. Carey School of Business	graduate
DATA MINING	USA	Bentley University	Graduate School of Business	graduate
Big data analytics	France		ESSEC - CentraleSupélec	graduate
Big Data Tools	France	online	IESEG School of Management	graduate
Data Mining for Big Data	France	University Jean Monnet Saint Etienne & Ecole des Mines de Saint-Etienne	Computer science department	graduate
Systems for distributed Big Data, Statistics, databases, distributed algorithms for large databases, Hadoop, Data Science Kit	France	Data Science		graduate
Data Mining	France	University Paris 13	institut Galilee	graduate
Data Mining/Databases for Big Data/Programming for Data Science	Germany	Albstadt-Sigmaringen University	Faculty of Computer Science	graduate
Data Science	Germany	FH-Brandenburg University of Applied Science	online	certificate
Storage and Mining of Massive Datasets	Germany	University of Luneburg (Leuphana)	Leuphana Graduate School	graduate

Mining Big Datasets	Greece	Athens University of Economics and Business	Department of Management Science and Technology	graduate
Data Mining	Ireland	University College Cork	Department of Computer science	graduate
Big data strategy & Implementation	Netherland	University of Amsterdam	Amsterdam Business School	graduate
Data Mining Tools & Languages	Russia	Novosibirsk State University		graduate
Data Mining – Clustering and Association Analysis	Sweden	Linkoping University	Institute of Technology	graduate
Data Mining & Analytic Technologies	UK	Bournemouth University	Department of Computing & Informatics	graduate
Data Mining	UK	De Montfort University Leicester		graduate
Big Data Applications	UK	Goldsmiths University of London	computing department	graduate
Big Data Management	UK	Heriot Watt University	Department of Computer science	graduate
Data Science Fundamentals/ Data Mining and Analytics/ Programming for Data Scientists	UK	Lancaster University	data science institute	graduate
Big Data	UK	University of Dundee	Computing at the University of Dundee	graduate
Big data	UK	University of Glasgow	School of Computing Science	graduate
Big Data Analytics	UK	University of Reading	School of Systems Engineering	graduate
Concurrent and Data-Intensive Programming	Norway	University in Tromso	Faculty of Science and Technology	all
Data Mining and Business Intelligence	USA	University of Connecticut	school of business	graduate
Data Mining	USA	the george washington university	GW School of Business	graduate
BIG DATA	USA	Georgetwon university	Graduate School of Arts and Sciences	graduate
Big data	UK	City university London	Department of Computer science	graduate
Data-intensive Systems	Denmark	AALBORG UNIVERSITY	Department of Computer Science	graduate
Data Mining for Business Decisions	Denmark	AARHUS UNIVERSITY	School of Business and Social Science	graduate
Data Mining & Business Intelligence	UK	ASTON UNIVERSITY	Aston Business School	graduate
Big Data	Italy	BAICR university	Dipartimento di Scienze Storiche, Filosofico-Sociali, dei Beni Cu	graduate
Data Warehousing and Business Intelligence	Spain	BARCELONA GRADUATE SCHOOL OF ECONOMICS	Graduate School of Economics	graduate
Advanced Methods in Data Mining and Data Warehousing, Text Mining and Web Content Mining, Mining Massive Datasets	Israel	BEN-GURION UNIVERSITY OF THE NEGEV	Department of Information Systems Engineering	graduate
Data Mining & Analytic Technologies	UK	BOURNEMOUTH UNIVERSITY	School of Design, Engineering & Computing,	graduate
Data Mining	Hong Kong	CHINESE UNIVERSITY OF HONG KONG	Department of Statistics at the Chinese University of Hong Kong	graduate
Business Intelligence and Big Data Processing	UK	COVENTRY UNIVERSITY	Faculty of Engineering and Computing	graduate
Data Mining	Sweden	DALARNA UNIVERSITY		graduate
Introduction of data mining	USA	E.J. Ourso College of Business/ Louisiana State University	The Information Systems & Decision Sciences	graduate

			Department	
Data Mining	Ireland	DUBLIN INSTITUTE OF TECHNOLOGY	School of Computing	graduate
Process Mining, Data Mining	Netherl and	EINDHOVEN UNIVERSITY OF TECHNOLOGY	Department of Mathematics and Computer Science	graduate
Data Mining: Applicative Approach	France	EISTI	ENGINEERING SCHOOL - MATHEMATICS - COMPUTER	graduate
Big Data Management & Analytics	Netherl and	ERASMUS UNIVERSITY	Rotterdam School of Management	graduate
Advanced Data Mining Techniques, Databases and Big Data	German y	HTW BERLIN	Treskowallee Campus	graduate
Data Mining Algorithms	Ireland	INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN	Institute of Technology Blanchardstown	graduate
Statistical Models for Data Mining, Introduction to Big Data and Analytics, Basic Algorithms for Data Mining	Spain	INSTITUTO DE EMPRESA	School of Social and Behavioral Sciences	graduate
Data Warehouse Models & Approaches	UK	LEEDS Becket University	Computing & Engineering	graduate
Data Mining - Clustering and Association Analysis	Sweden	LINKOPING UNIVERSITY	ENGINEERING and Computer science	graduate
Big data	Canada	QUEEN'S UNIVERSITY	Smith School of Business	graduate
Data Mining, Data Warehousing,	UK	ROBERT GORDON UNIVERSITY	School of Computing Science and Digital Media	graduate
Data-Intensive Systems	Norway	University of Stavanger	Faculty of Science and Technology	all
Introduction to data Analytics	Turkey	SABANCI UNIVERSITY	Faculty of Engineering and Natural Sciences	graduate
Data mining	Turkey	SABANCI UNIVERSITY	Faculty of Engineering and Natural Sciences	graduate
Big data processing using Hadoop	Turkey	SABANCI UNIVERSITY	Faculty of Engineering and Natural Sciences	graduate
Data Engineering	UK	THE UNIVERSITY OF MANCHESTER	School of Computere Science	graduate
Data Mining/Big Data	German y	UNIVERSITAT KONSTANZ	Department of Computer and Information Sciences	graduate
Relational data mining	Italy	EUROPEAN UNION	University of Eastern Piedmont (UPO)	graduate
Data Mining	Ireland	UNIVERSITY COLLEGE CORK	College of Science, Engineering and Food Science	graduate
Data Mining for Business Analytics	Ireland	UNIVERSITY COLLEGE DUBLIN	UCD Michael Smurfit Gradute Business School	graduate
Information Retrieval & Data Mining	UK	UNIVERSITY COLLEGE LONDON	Department of Computer Science	graduate
Introduction to Data Mining and Machine Learning part 1, Big Data, Introduction to Data Mining and Machine Learning part 2, Hadoop, HDFS, MapReduce, and other Hadoop/SQL technologies,	UK	UNIVERSITY OF DUNDEE	School of Computing	graduate
Big Data	UK	UNIVERSITY OF DUNDEE	School of Computing	graduate
DATA MINING	UK	UNIVERSITY OF EAST ANGLIA	School of Computing Science and Digital Media	graduate
data mining	Finland	UNIVERSITY OF HELSINKI	Department of Computer Science	graduate
Data Mining and Forecasting	UK	UNIVERSITY OF KENT	Kent Business School	graduate

Data Mining and Text Analytics	UK	UNIVERSITY OF LEEDS	School of Computing	graduate
Data Mining and Neural Networks	UK	UNIVERSITY OF LEICESTER	Mathematics department	graduate
Data Mining Overview/ Data Mining/ text mining/ Big data	USA	North Carolina State University	Institute for Advanced Analytics	graduate
Data Mining	UK	UNIVERSITY OF LIVERPOOL	Department of Computer Science	graduate
Big Data Analysis	UK	UNIVERSITY OF LIVERPOOL	Department of Computer Science	graduate
Data Analytics for Business Decision Making	UK	UNIVERSITY OF MANCHESTER	Alliance Manchester business school	graduate
Statistics / Data Mining	Italy	UNIVERSITY OF MILAN-BICOCCA	Scuola di Economia e Statistica	graduate
Machine Learning and Data Mining	New zealand	UNIVERSITY OF OTAGO	Department of Information Science	graduate
DATA MINING	Italy	UNIVERSITY OF PISA	Department of Computer Sciences and Informatics	graduate
Multivariate Statistics for Data Mining	UK	UNIVERSITY OF SOUTHAMPTON	Southampton Business School	graduate
Data Mining	UK	UNIVERSITY OF WARWICK	Department of Computer Science	graduate
Data Mining	USA	Michigan State University/ Eli Broad College of Business	DEPARTMENT OF ACCOUNTING & INFORMATION SYSTEMS	graduate
Data Mining	UK	UNIVERSITY OF WESTMINSTER	Faculty of Science and Technology	graduate
Data Mining	Netherland	UTRECHT UNIVERSITY	Faculty of Science: Information and Computing Sciences	graduate
Data Mining and Decision Support Systems	Austria	VIENNA UNIVERSITY		graduate
Data-Mining Techniques	Netherland	VRIJE UNIVERSITEIT AMSTERDAM	Department of Mathematics	graduate
Data Analytics with High Performance Computing	UK	University of Edinburgh	EPPC	graduate
Big Data	UK	University of Liverpool	Online Programm	graduate
Multivariate Analysis for Big Data	New zealand	Massey University	Massey Business School	graduate
Introduction to Data Science	USA	Columbia University	Data Science Institute	graduate
Big Data Analytics, Advanced Big Data Analytics	USA	Columbia University	Data Science Institute	graduate
Big Data Analytics in Business	USA	Georgia Tech	College of Computing, College of Engineering, and Scheller College of Business	graduate
Data Mining	Netherland	Maastricht University	Department of Data Science and Knowledge Engineering	graduate
Data Mining	Mexico	Autonomous Technological Institute of Mexico	Posgrados ITAM	graduate
Distributed Data Warehouses and Data Mining	Lithuania	Mykolas Romeris University	Institute of Digital Technologies	graduate
Database and Data Mining	Italy	Polytechnic University Of Turin	Alta Scuola Politecnica	graduate
Data Analytics and Data Mining	Ireland	Dublin City University	School of Computing	graduate
Data Mining Algorithms	Ireland	Institute Of Technology Blanchardstown	School of Informatics and Engineering	graduate
Big Data & Analytics	Ireland	Irish Management Institute		graduate
Programming for Big Data	Ireland	National College Of Ireland	School of Computing	certificate
Data Mining	USA	Southern Methodist University	Dedman College of Humanities and Sciences,	graduate

			Lyle School of Engineering and Meadows School of the Arts	
Storing and Retrieving Data	USA	University of California, Berkeley	School of Information	graduate
Data Mining I, Data Mining II	USA	Arizona State University	W.P. Carey School of Business	graduate
Big data technologies	Spain	Instituto de Empresa	school social behavioral & data sciences	graduate
Capture and Data Storage	Spain	Universidad Rey Juan Carlos	Degrees and Continuing Education	graduate
Data Mining, Systems for Big Data	Canada	Simon Fraser University	School of Computing Science	graduate
Applied Statistical Methods for Data Mining	Canada	University Of Alberta	Department of Computing Science	graduate
DATA-INTENSIVE COMPUTING	USA	Illinois Institute of Technology	College of Science	graduate
DATA PREPARATION AND ANALYSIS	USA	Illinois Institute of Technology	College of Science	graduate
DATA MINING	USA	Illinois Institute of Technology	College of Science	graduate
Big data	Brazil	Mackenzie Presbyterian Institute	Information Technology	graduate
Big Data Basics	Australi a	University of South Australia	School of Information Technology and Mathematical Sciences	graduate
Statistical Programming for Data Science	Australi a	University of South Australia	School of Information Technology and Mathematical Sciences	graduate
Data Science for Innovation	Australi a	University of Technology Sydney		graduate
Big data	USA	New York University	department of arts and science	graduate
DATA MINING	USA	Northwestern University	McCormick School of Engineering and Applied Science	graduate
ANALYTICS FOR BIG DATA	USA	Northwestern University	McCormick School of Engineering and Applied Science	graduate
INTRODUCTION TO DATA ANALYTICS	UK	Imperial College London	Imperial College Business school	graduate
Cloud Computing and Big Data	USA	Rutgers University	Graduate School, Professional Science Masters Programs (Master of Business and Science) and School of Communication and Information (Master of Information)	graduate
Data mining	USA	Rutgers University	Graduate School, Professional Science Masters Programs (Master of Business and Science) and School of Communication and Information (Master of Information)	graduate
Data Mining and Business Intelligence	USA	University of Connecticut	School of Business, Department of Operations and Information Management	graduate
Data Mining in R	China	New York University	Shanghai Campus	graduate
Data Mining	China	Chinese University of Hong Kong	department of Statistics	graduate

Big Data Analytics using HADOOP/ Data Mining	India	Great Lakes Institute of Management		certificate
Data Mining and Data Warehousing	India	International School of Information Management		graduate(?)
Data Science	USA	University of Maryland	Robert H. Smith School of Business	graduate
Business Strategies for Big Data	USA	University of San Francisco	College of Arts and Sciences	graduate
Data Mining Methods for Business Applications	USA	University of Tennessee	Department of Business Analytics & Statistics, Haslam College of Business	graduate
Big Data Basics	Australia	University of South Australia	School of Information Technology & Mathematical Sciences	graduate
Data Mining	USA	University of Virginia	Data Science Institute	graduate

Appendix C – Inventory of industrial courses

Organization	Course title	Level
Cloudera	Cloudera Administrator Training for Apache Hadoop	course
Cloudera	Cloudera Developer Training for Apache Spark	course
Cloudera	Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop	course
Cloudera	Cloudera Developer Training for MapReduce	course
Cloudera	Designing and Building Big Data Applications	course
Cloudera	Cloudera Essentials for Apache Hadoop	course
Cloudera	Designing and Building Big Data Applications	course
Cloudera	Certified Professional Data Scientist (CCP:DS)	Certification
Cloudera	Certified Developer for Apache Hadoop (CCDH)	Certification
Cloudera	Certified Administrator for Apache Hadoop (CCAH)	Certification
Cloudera	Certified Specialist in Apache HBase (CCSHB)	Certification
IBM	Introduction to InfoSphere Master Data Management	course
IBM	Introduction to IBM InfoSphere Master Data Management Standard Edition	course
IBM	IBM InfoSphere MDM Standard Edition Architecture and Data Model Design	course
IBM	InfoSphere MDM Reference Data Management	course
IBM	IMS Data Sharing	course
Amazon	Big Data on AWS	course
Amazon	Big Data Technology Foundamentals	course
Amazon	AWS Technical Essentials	course
FhG IAIS	Data-Scientist-Schulungen	
Hortonworks	HDP Developer Java, Apache Pig and Hive, Windows	Courses
Hortonworks	HDP Developer Custom Yarn Applications	Courses
Hortonworks	HDP Developer: Storm and Trident	Courses
Hortonworks	HDP Certified Developer (HDPCD)	Certification
Hortonworks	HDP Certified Developer JAVA (HDPCDJ)	Certification
Hortonworks	HDP Operations:Migrating to the Hortonworks Data Platform	Courses
Hortonworks	HDP Operations: Hadoop Administration	Courses
Hortonworks	HDP Operations:Apache HBase Advance Management	Courses
Hortonworks	HDP Certified Administrator (HDPCA)	Certification
Hortonworks	HDP Analyst: Data Science	Courses
Hortonworks	HDP Analyst: Apache HBase Essentials	Courses
Mapr	DEV3000: Developing Hadoop Applications	Courses
Mapr	MCHBD:MapR Certified HBase Developer	Certification
Mapr	MCHD: MapR Certified Hadoop Developer	Certification
Mapr	MCSD: MapR Certified Spark Developer	Certification
Mapr	MCHA - MapR Certified Hadoop Administrator	Certification
Mapr	DA 4500- Data Analysis with Apache Pig and Apache Hive	Courses
Microsoft	MCSE: Business Intelligence	Certification
Microsoft	MCSE: Data Platform	Certification

SAP	Life-Cycle Data Management (SAP PLM)	Courses
SAP	Certified Application Associate - Modeling and Data Management with SAP BW 7.4	Certification
Pentaho	BA1000 Business Analytics User Console	Course
Pentaho	BA3000 Business Analytics Data Modelling	Course
Pentaho	DI1000 Pentaho Data Integration Fundamentals	Course
Coursera	Web Intelligence and Big Data	course
Coursera	Process Mining: Data science in Action	course
Coursera	Executive Data Science Specialization	Certification
Coursera	Data Science Specialization	Certification
Coursera	Data Science at Scale	Certification
Coursera	Learn Data Science Fundamentals	Certification
Coursera	Big Data Specialization	Certification
Oracle	Big Data Appliance	Certification
Oracle	Meter Data Management	Certification
Oracle	Oracle Data Integrator	Certification
Engineering (School of Ferentino)	Methods of Data Virtualization	Course
Engineering (School of Ferentino)	Data Model	Course
Engineering (School of Ferentino)	Data Warehouse design	Course
Engineering (School of Ferentino)	Advanced Data Warehouse	Course
Google Analytics Academy	Digital Analytics Fundamentals	Course
Google Analytics Academy	Google Analytics Platform Principles	Course
Google Analytics Academy	Ecommerce Analytics: From Data to Decisions	Course
Centre For Development of Advanced Computing(C-DAC)	Using R for Data Visualization and Analytics	Course
Centre For Development of Advanced Computing(C-DAC)	Text Analytics	Course
Centre For Development of Advanced Computing(C-DAC)	Predictive Analytics and Recommender Systems	Course
International School of Engineering	Certificate Program in Big Data Analytics and Optimization	Certifcate
EduPristine	Business Analytics Training	Course
EduPristine	Big Data Hadoop Training	Course
EduPristine	Data Science Course Training	Course
EduPristine	Data Visualization	Course
NIIT	Analytics	Program
NIIT	Data Analytics Essentials	Course
NIIT	Implementing Data Analytics Using R	Course
NIIT	Working with Advanced Business Analytics Techniques	Course
manipal ProLearn	Big Data Analytics using Hadoop	Program
ScaDS	2nd International ScaDS Summer School on Big Data	Summer School

Appendix D – Data Science occupations family as an extension to ESCO classification

Table 13 contains proposed Data Science professions/profiles organised into new proposed hierarchies that can be added to the ESCO classification.

Table 13 Data Science occupations extension to ESCO classification

Top level	Hierarchies existing and new		Occupations group (if any)	Occupations
Managers				
	Production and specialised services managers	Data Science/Big Data Infrastructure Managers		Data Science/Big Data Infrastructure Manager
			Research Infrastructure Managers	RI Manager
				RI Data storage facilities manager
Professionals				
	Science and engineering professionals	Data Science Professionals	Data Science professionals not elsewhere classified	Data Scientist
				Data Science Researcher
				(Big) Data Analyst
				Data Science (Application) Programmer
				Business Analyst
		Database and network professionals	Large scale (cloud) data storage designers and administrators	Large scale (cloud) database designer*)
			Database designers and administrators	Large scale (cloud) database administrator*)
			Database and network professionals not elsewhere classified	Scientific database administrator*)
	Information and communications technology professionals	Data Science technology professionals	Data handling professionals not elsewhere classified	Digital Librarian
				Data Archivist
				Data Steward
				Data curator
Technicians and associate professionals				
	Science and engineering associate professionals	Data Science Technology Professionals	Data Infrastructure engineers and technicians	Big Data facilities Operators
				Large scale (cloud) data storage operators
			Database and network professionals not elsewhere classified	Scientific database operator*)
Clerical support workers				
	General and keyboard clerks			

	Data handling support workers (alternative)	Data and information entry and access	Digital Archivists and Librarians	Digital Librarian
				Data Archivist
				Data Steward
				Data curator

Table 14 provides an example of competences definition for different groups that is improved after initially proposed in D2.1.

Table 14 Competences definition for different Data Science competence groups

Data Analytics (DSDA)	Data Management/ Curation (DSDM)	DS Engineering (DSENG)	Scientific/ Research Methods (DSRM)	DS Domain Knowledge (for example, Business Apps) (DSDK)
Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSDA01 **Use predictive analytics to analyse big data and discover new relations**	DSDM01 Develop and implement data strategy, in particular, in a form of Data Management Plan (DMP)	DSENG01 Use engineering principles to research, design, prototype, data analytics applications, or develop structures, instruments, machines, experiments, processes, systems	DSRM01 Create new understandings and capabilities by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods	DSDK01 Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
DSDA02 **Use appropriate statistical techniques on available data to deliver insights**	DSDM02 Develop and implement data models including metadata	DSENG02 Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	DSRM02 Direct systematic study toward a fuller knowledge or understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSDK02 Use data to improve existing services or develop new services

DSDA03 **Develop specialized analytics to enable agile decision making**	DSDM03 Collect and integrate different data source and provide them for further analysis	DSENG03 Develops specialized data analysis tools to support executive decision making	DSRM03 Undertakes creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	DSDK03 Participate strategically and tactically in financial decisions that impact management and organizations
DSDA04 **Research and analyse complex data sets, combine different sources and types of data to improve analysis.**	DSDM04 Visualise complex and variable data.	DSENG04 Design, build, operate relational non-relational databases	DSRM04 Apply ingenuity to complex problems, develop innovative ideas	DSDK04 Provide scientific, technical, and analytic support services to other organisational roles
DSDA05 **Use different data analytics platforms to process complex data**	DSDM05 Develop and maintain a historical data repository of analysis results	DSENG05 Develop solutions for secure and reliable data access	DSRM05 Ability to translate strategies into action plans and follow through to completion.	DSDK05 Analyse multiple data sources for marketing purposes
		DSENG06 Prototype new data analytics applications	DSRM06 Contribute to and influence the development of organizational objectives	DSDK06 Analyse customer data to identify/optimise customer relations actions

Appendix E – Data Science Body of Knowledge

Table 15 provides a consolidated view of the identified Knowledge Areas in the Data Science Body of Knowledge as it is defined in the Deliverable D2.1. The table contains detailed definitions of the KAG1-DSA, KAG2-DSE, KAG3-DSDM groups that are well supported by existing BoK's and academic materials. General suggestions are provided for KAG4-DSRM, KAG5-DSBP groups that correspond to newly identified competences and knowledge areas and require additional study of existing practices and contribution from experts in corresponding scientific or business domains.

The KAG2-DSE group includes selected KAs from ACM CS-BoK and SWEBOK and extends them with new technologies and engineering technologies and paradigms such as cloud based, agile technologies and DevOps that are promoted as continuous deployment and improvement paradigm and allow organisations to implement agile business and operational models.

The KAG3-DSDM group includes most of the KAs from DM-BoK however extends it with KAs related to RDA recommendations, community data management models (Open Access, Open Data, etc.) and general Data Lifecycle Management that is used as a central concept in many data management related education and training courses.

The presented DS-BoK high level content is not exhaustive at this stage and will undergo further development based on feedback from the Task 3.1 that will use the presented DS-BoK for developing Data Science Model Curriculum (MC-DS). The project will present the current version of DS-BoK to the ELG to obtain feedback and expert opinion. Numerous experts will be invited to review and contribute to the specific KAs definition.

Table 15 Identified DS-BoK Knowledge Areas

KA Groups	Knowledge Areas (KA) from existing BoKs	Additional Knowledge Areas
KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics	BABOK selected KAs • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.	General Data Analytics and Machine Learning KAs • Machine learning and related methods • Predictive analytics and predictive forecasting • Classification methods • Data mining and knowledge discovery • Business intelligence covers data analysis that relies heavily on aggregation and different data sources and focusing on business information; • Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data • Statistical methods, including descriptive statistics, exploratory data analysis (EDA) and confirmatory data analysis (CDA)
KAG2-DSE: Data Science Engineering group including Software and infrastructure	ACM CS-BoK selected KAs: AL - Algorithms and Complexity AR - Architecture and Organization CN - Computational Science GV - Graphics and Visualization	Infrastructure and platforms for Data Science applications group: CCENG - Cloud Computing Engineering (infrastructure and services design, management and operation)

engineering	IM - Information Management PBD - Platform-based Development (new) SE - Software Engineering (extended with SWEBOK KAs) SWEBOK selected KAs • Software requirements • Software design • Software construction • Software testing • Software maintenance • Software configuration management • Software engineering management • Software engineering process • Software engineering models and methods • Software quality	CCAS - Cloud based applications and services development and deployment BDA – Big Data Analytics platforms (including cloud based) BDI - Big Data Infrastructure services and platforms, including data storage infrastructure Data and applications security KAs: SEC - Applications and data security SSM – Security services management, including compliance and certification Agile development technologies • Methods, platforms and tools • DevOps and continuous deployment and improvement paradigm
KAG3-DSDM: Data Management group including data curation, preservation and data infrastructure	DM-BoK selected KAs (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality.	General Data Management KA's • Data Lifecycle Management • Data archives/storage compliance and certification New KAs to support RDA recommendations and community data management models (Open Access, Open Data, etc.)[2] • Data type registries, PIDs • Data infrastructure and Data Factories • TBD – To follow RDA and ERA community developments
KAG4-DSRM: Scientific or Research Methods group	There is no formally defined BoK for research methods	Suggested KAs to develop DSRM related competences: • Research methodology, research cycle (e.g. 4 step model Hypothesis – Research Methods – Artefact – Validation) • Modelling and experiment planning • Data selection and quality evaluation • Use cases analysis: research infrastructures and projects • TBD further extensions
KAG5-DSBP: Business process management group	PMI-BoK selected KAs • Project Integration Management	General Business processes and operations KAs • Business processes and

[2] Example courses provided by RDA community and shared between European Research Infrastructures https://europe.rd-alliance.org/training-programme

	• Project Scope Management • Project Quality • Project Risk Management	operations • Agile Data Driven methodologies, processes and enterprises • Use cases analysis: business and industry • TBD further extensions

Appendix F – Taxonomy of Data Science through job skills analysis

F.1 The Dataset

To extract a taxonomy of skills related with data scientists we used a dataset composed of 1009 job advertisements with “Data scientist” as their job title. The job positions did not have any restrictions on location of employer. The job advertisements where obtained in JSON format and include various information fields such as the employer’s location, the job description, the level of the position, etc. However, only the job description information exists in all the advertisements. In Table 16 Information fields contained in the dataset show the percentage of job ads and their corresponding information fields while Figure 23, Figure 24, Figure 25 and Figure 26 show the analysis of each information field.

Table 16 Information fields contained in the dataset

Information Field	Percentage of ads, %
Employer’s country location	22
Job position’s function	22
Experience level	22
Employer’s size	21

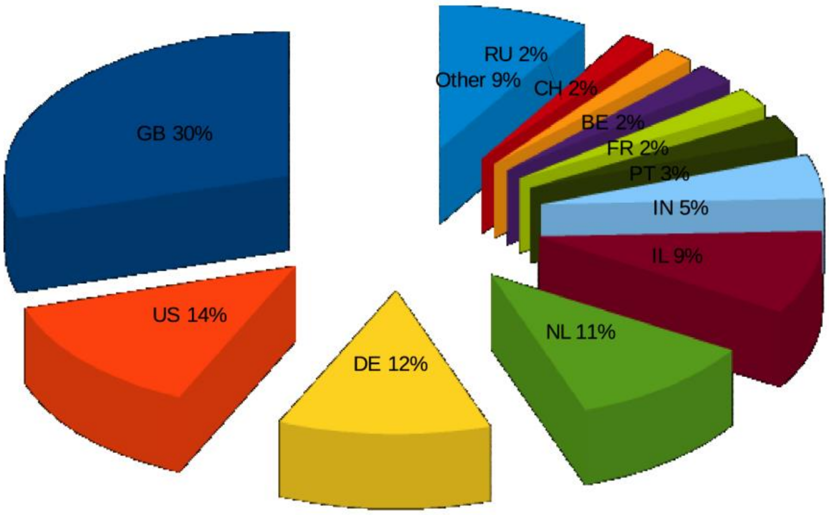


Figure 23 Employer's country locations

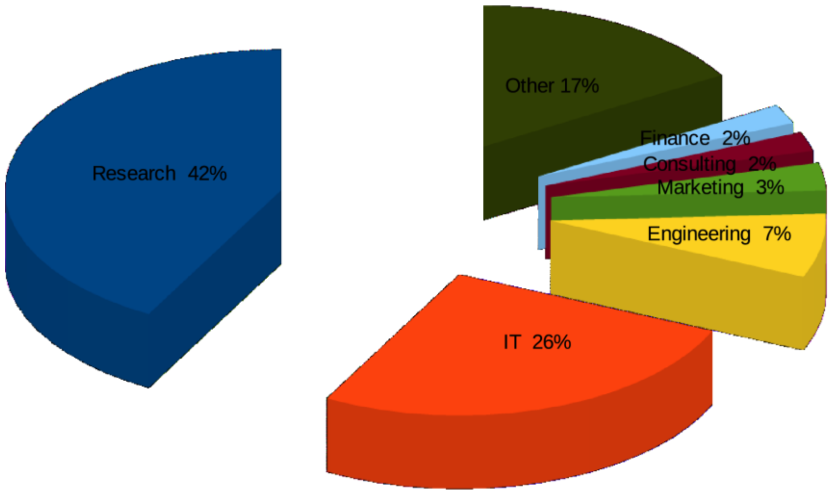


Figure 24 Job position's function

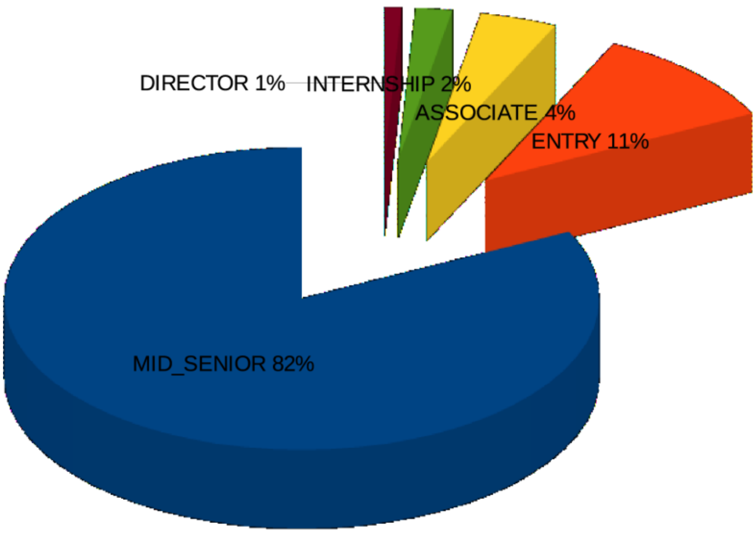


Figure 25 Experience level required

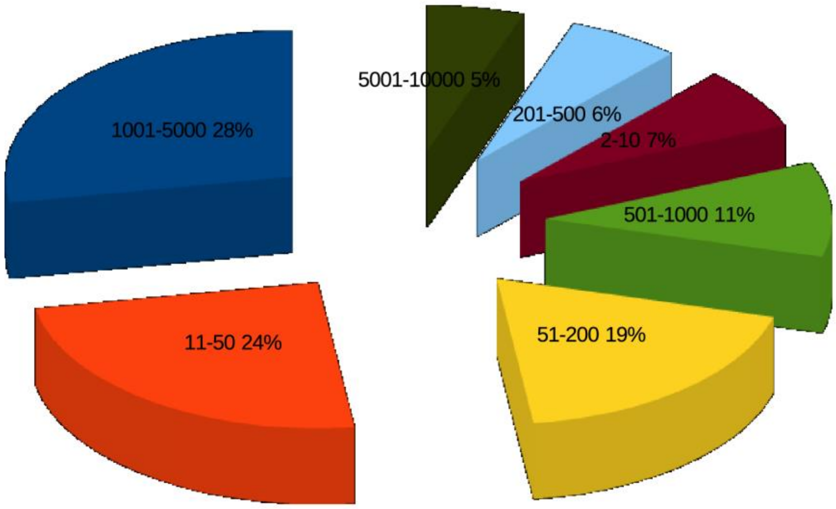


Figure 26 Employer's size in number of employees

From the dataset analysis we can see that most of the job advertisements requesting data scientists originate from Great Britain followed by the US and Germany. Moreover, the majority of job functions is related with research and IT. According to the available information fields the experience level required is medium to senior indicating that considerable years of experience are required. Finally, the employer’s size seems to be between large (1001-5000 employees) and rather small (11-50 employees) organizations.

F.2 Preliminary Results

In order to extract the relevant skills from each advertisement we analysed the job description field contained in all advertisement. The initial term extraction process was based on a simple term frequency count and produced approximately 300,000 terms. After using statistical methods we reduced the terms to approximately 50,000. However, this data set contains a lot of noise and irrelevant terms and had to be manually validated which is a time-consuming process. As a second step we used a hybrid method that uses both linguistic (Part-of-speech tagging) and statistical analysis for ranked term extraction. Table 17 shows the ranked terms extracted by both methods.

Table 17 Top 40 terms extracted with a simple tf method and a hybrid method

Rank	**tf Term Extraction**	**Hybrid Term Extraction**
1	data	data_scientist
2	experience	communication_skills
3	skill	data_sets
4	model	data_analysis
5	scientist	data_science
6	learn	data_sources
7	big_data	data_analytics
8	science	data_analyst
9	customer	ideal_candidate
10	product	computer_science
11	develop	business_problems
12	machine_learning	data_mining
13	process	track_record
14	engineer	team_player
15	understand	data_technologies
16	management	analytics_team
17	service	work_experience
18	build	years_experience
19	research	business_requirements
20	report	programming_language

21	environment	ability
22	analyst	business_intelligence
23	software	business_decisions
24	technique	job_description
25	algorithm	masters_degree
26	sql	programming_languages
27	requirement	team_members
28	program	benefits_package
29	r	programming_skills
30	complex	business_insights
31	solve	business_needs
32	job	data_management
33	decision	business_opportunities
34	responsibility	data_engineers
35	hadoop	experience
36	database	hadoop_ecosystem
37	implement	phd
38	java	presentation_skills
39	senior	skills_experience
40	user	work_environment

Applying in a next phase we grouped together the extracted terms from the hybrid method to form non-hierarchical realizations. Table 18 and Table 19 show a small sample of the extracted groups.

Table 18 Technical and research skills that appear in job advertisements

Data Analytics & ML	**Data Management**	**Research Methods**
artificial neural network	backup	analysis
	core data	aggregate data
algorithm	data architecture	analytical skill
automated systems	data infrastructure	answer
big data analysis	data integrity	applied math
computer science	data management	data acquisition
computer vision	data modelling	data collection
data engineering	data privacy	data point
data mining	data structure	design methods
data visualization	data type	evaluation criteria
expert system	database administration	experiment design
hybrid cloud	database query	exploratory data analysis
kernel trick	database technologies	hypothesis testing
knowledge extraction	database theory	insight
language processing	scrum methodology	problem statement
markov model	search engine technology	research data
probability distribution	sql queries	research design
programming language	streaming data	research methodologies
realtime computing		research scientist
ssis		solution
startup environment		trend analysis
survey methodology		

trend analysis

Table 19 Inter-personal skills, education, specific domains and employment clusters appearing in data scientist job ads

Inter-personal Skills	**Education**	**Domains**	**Employment Relations**
communication skills	business school	ad serving	annual salary
computing platform	bachelor degree	airline	application form
core competencies	doctorate degree	automotive industry	career development
experience	graduate degree	banking	disabilities act
knowledge	master degree	bioinformatics	fringe benefit
management team	ms degree	biotechnology revolution	fringe benefit
problem solver	phd degree	business	job candidate
proficiency	university	credit rating	job evaluation
project management	university degree	drug discovery	job title
team work performance indicator		game design	remuneration package
time management		gaming industry	work environment
work experience		government agencies	
		healthcare-medical	
		insurance company	
		life sciences	
		services industry	
		technology companies	
		travel industry	
		vehicle engineering	

Finally we performed a hierarchical relation discovery using hypernym-hyponym relations included in online dictionaries. Figure 27 shows a sample of hierarchical taxonomy based on a small cluster.

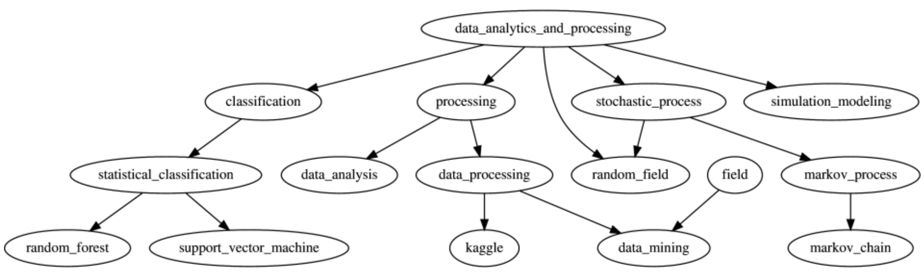


Figure 27 Sample hierarchical taxonomy

Examining the results shown in Table 17, we can see that the majority of terms are related with computer science and math (including statistics) indicating that these skill sets are important in this domain. Moreover, specific programming languages and platforms seem to be included in many advertisements and further investigation could reveal which programming languages and platforms are considered important.

Table 18 reveals that the majority of categories and clusters extracted refer to data manipulation and statistics. Table 19 shows that there is a wide range of domains for which data science is applied and includes both academic research and commercial applications.

Appendix G – Relation between Knowledge Areas in Data Science

When considering "ingredients" of Data Science one can notice that some fields are branches of other fields, sometimes more than one field. The graph in Figure 28 presents dependencies between Knowledge Areas and Knowledge Units defined for DS-BoK. The diagram blocks are extracted from the taxonomy of DS as main KAs. The choice of the KAs is based on ACM taxonomy extended for EDISON and other approaches to DS classification described above. If we consider the Knowledge Units and topics covered by given KAs, we may observe that some of the topics exist in several KAs. Also topics from some KAs are required as "prerequisites" for other KAs. Analysis of these dependencies was a basis for development of the presented DS Taxonomy flow chart.

The main KAs, defined also on the Venn diagram, are located in elliptical frames. Fields of knowledge that are direct branches of the main KAs are presented in rectangles with double frame, and further branches are depicted in rectangles with single frame. Probability and Analytics, that are sub-fields of Statistics, are required as basic knowledge for algorithm design, computation methods, software engineering and networks, that are branches of Computer Engineering, therefore these diagram elements are related to the CE block by colour. The direction of the thick arrows shows which KU is a branch of certain KA. On the other hand, the direction of the thin arrows presents what topics from a given KU is required for another KU or KA.

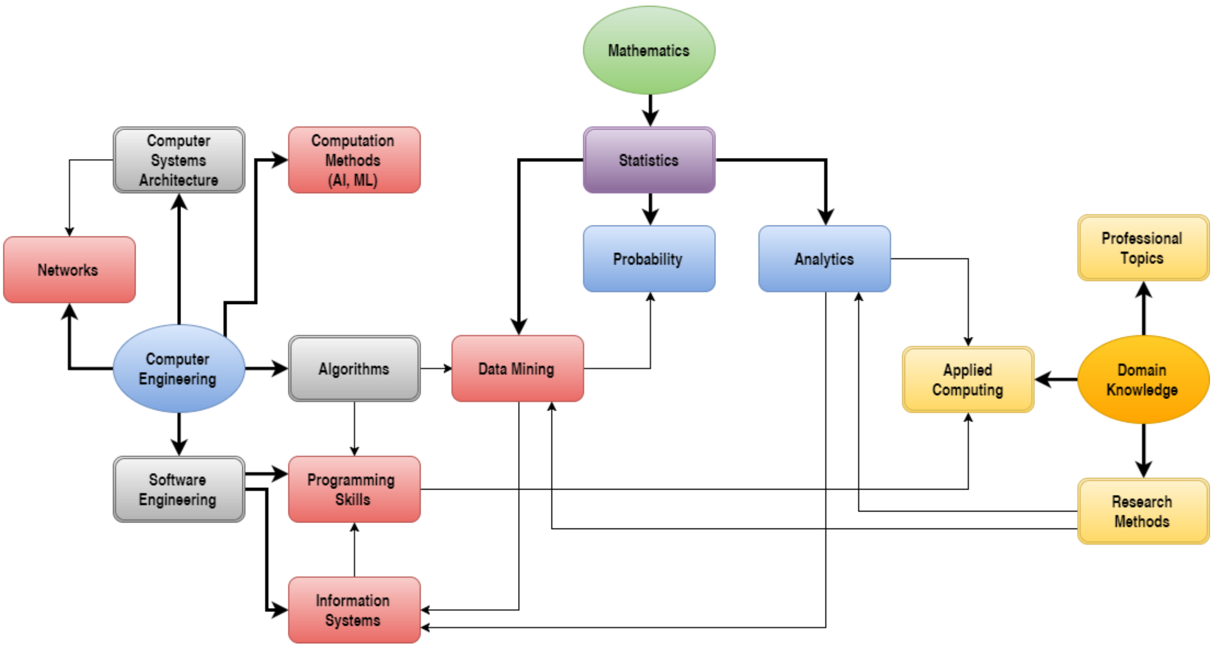


Figure 28 Relation between Knowledge Areas in Data Science

Appendix H – Mapping between learning outcomes and taxonomy

Mapping the Learning Outcomes to Knowledge Areas, as well as Knowledge Areas to Learning Outcomes, is needed to develop an appropriate organization of Data Science teaching programs. Table 12 defines Learning outcomes for five groups related to CF-DS being an extension to e-CF. Depending on the CF-DS requirements for given professions, different Learning Outcomes are desired and various Knowledge Areas should be assigned with appropriate “weights”. The mapping of main Knowledge Areas defined in DS taxonomy to eleven Learning Outcomes defined for CF-DS are presented in Table 20. Such mapping is useful for universities and other educational institutions when developing teaching programs in DS. Depending on the specialization of Data Science, teaching program the topics from certain Knowledge Areas and Knowledge Units should be covered in different courses such that it is most appropriate for the students.

On the other hand mapping the Learning Outcomes for particular CF-DS groups to Knowledge Areas is useful for employers when defining skills required for a given position in the company. This type of mapping is presented in Table 21.

Table 20 Relation between DS taxonomy main Knowledge Areas and Learning Outcomes

Taxonomy KA	Learning Outcomes
Architectures	LO.1, LO.4, LO.11
Parallel architectures	
Distributed architectures	
Software system structures	LO.1
Software architectures	LO.1-LO.6
Software system models	
Ultra-large-scale systems	LO.2, LO.3
Software notations and tools	LO.1
Software creation and management	LO.3, LO.5, LO.6, LO.10
Networks	LO.1
Network protocols	LO.2
Network algorithms	LO.2
Network properties	LO.4
Network structure	LO.4
Network security	LO.11
Probability and statistics	LO.1, LO.2, LO.9
Mathematical software	LO.2, LO.3, LO.7, LO.8
Information theory	
Mathematical analysis	
Distributed computing methodologies	
Data management systems	LO.2, LO.3, LO.5, LO.6, LO.7, LO.8, LO.10, LO.11
Information storage systems	LO.4, LO.5, LO.6, LO.10, LO.11
Information systems applications	LO.5, LO.7, LO.8, LO.9, LO.10
Multimedia information systems	LO.7, LO.8, LO9
Data mining	LO.1, LO.2, LO.3, LO.5, LO.6
Digital libraries and archives	
Information retrieval	LO.10, LO.11
Computing methodologies	LO.1, LO.2, LO.3
Artificial intelligence	
Machine learning	
Modelling and simulation	
Applied computing	LO.1, LO.5, LO.7, LO.9
Social and professional topics	

Research methods	LO.1
Data collection techniques	LO.2
Sampling methods	LO.2
Data analysis and results reporting	LO.2, LO.9

Table 21 Relation between Learning Outcomes and main Knowledge Areas from DS taxonomy

Learning Outcome	Taxonomy KA
LO.1 Choose and execute existing analysis, services and monitoring	Architectures Software system structures Software notations and tools Networks Probability and statistics Data mining Digital libraries and archives Computing methodologies Applied computing Social and professional topics Research methods
LO.2 Apply and develop data analytic methods and applications	Software architectures Software system models Ultra-large-scale systems Network protocols Network algorithms Probability and statistics Mathematical software Information theory Mathematical analysis Distributed computing methodologies Data management systems Data mining Digital libraries and archives Computing methodologies Data collection techniques Sampling methods Data analysis and results reporting
LO.3 Plan, recommend and design data management applications and tools	Software architectures Software system models Ultra-large-scale systems Software creation and management Mathematical software Information theory Mathematical analysis Distributed computing methodologies Data management systems Computing methodologies
LO.4 Assess, design and evaluate Data Science infrastructures	Parallel architectures Distributed architectures Software architectures Software system models Network properties Information storage systems
LO.5 Identify, organize and develop processes for data, information and knowledge management	Software architectures Software system models Software creation and management Data management systems Information storage systems

	Information systems applications Data mining Digital libraries and archives Applied computing Social and professional topics
LO.6 Evaluate, improve, design processes for data, information and knowledge management	Software architectures Software system models Software creation and management Data management systems Information storage systems Data mining Digital libraries and archives
LO.7 Build and organize data models and preservation processes	Mathematical software Information theory Mathematical analysis Distributed computing methodologies Data management systems Information systems applications Multimedia information systems Applied computing Social and professional topics
LO.8 Evaluate, improve and design data models and preservation processes	Mathematical software Information theory Mathematical analysis Distributed computing methodologies Data management systems Information systems applications Multimedia information systems
LO.9 Examine available data, and infer and visualize data insights	Probability and statistics Information systems applications Multimedia information systems Applied computing Social and professional topics Data analysis and results reporting
LO.10 Assess, influence, and prioritize organization improvement and risk management with data	Software creation and management Data management systems Information storage systems Information retrieval
LO.11 Inspect, identify and make use of required security monitoring	Architectures Network security Data management systems Information storage systems Information retrieval