2016 IEEE 8th International Conference on Cloud Computing Technology and Science

Quantitative and Qualitative Analysis of Current Data Science Programs from Perspective of Data Science Competence Groups and Framework

Tomasz Wiktorski, Anoosheh Shirazi
Department of Electrical Engineering and Computer Science
University of Stavanger
4036 Stavanger, Norway
Email: tomasz.wiktorski@uis.no

Yuri Demchenko, Adam Belloum
Faculty of Science
University of Amsterdam
1090 GH, Amsterdam, The Netherlands

***Abstract*—Data Science is becoming a field connecting multi-year development in areas such as Big Data and Data Analytics, and also applied domains like Bioengineering. Data Science education programs are rapidly being created on all levels. Usually it happens through reuse or renaming and can result in curricula that lack proper balance of competences, which balance is necessary for future data scientists. Our quantitative analysis of over 300 programs worldwide shows that at least one of the three core data science competence groups is under-represented in the majority of programs. Moreover, general business courses are often suggested to students to cover the domain competence group, which in most cases results in superficial treatment of this competence group. Our further qualitative analysis demonstrates that learning outcomes for most of the courses are usually not defined or defined improperly.**

1. Introduction

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 catalog 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 hard to understand the current state of Data Science education. Degree of coverage of competence groups is a main example of missing information among existing lists. 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, which we suggest to the community. Use of these words has geographical connotations and we settle on these definitions only for clarity in this paper.

1.1. EDISON project

Establishing Data Scientist as a profession is a long reaching proposition; EDISON is conceived to pave the way for such long term challenge. In particular, EDISON works to achieve the following objectives: (1) define the Data Science Competence Framework, Body of Knowledge and Model Curricula; (2) propose a framework and an ICT environment for re-skilling and certifying Data Scientists; (3) develop a sustainable education model for Data Science and Data Intensive technologies.

1.2. Related work

There exist several lists of data science programs online. Probably the most comprehensive is "Colleges with Data Science Degrees" [1]. It provides a greater breadth than we aim at, thanks to being curated over several years, but does not aim at an in-depth analysis of listed offerings. In our list, we have excluded many generic programs in computer science or information science with only minor elements of data analysis or domain knowledge, which are present on other lists. While such programs might with time develop in the Data Science direction, they do not provide a meaningful basis for analysis of existing Data Science programs.

To our best knowledge this is the first such comprehensive attempt of quantitative and qualitative analysis of Data Science programs worldwide.

1.3. Organization

After the introduction, in Section 2 we describe the EDISON Inventory, in particular its organisation, methods of population, and how we provide the EDISON Inventory as a service to Data Sciencd community. The Data Science Competence Framework, which is the basis for analysis in this paper, is described in Section 3. In Section 4 we describe the results of quantitative analysis of various aspects of courses including coverage of competence groups and in Section 5 we shortly present the results of the qualitative analysis of selected programs. We summarise the main points in Section 6.

9781-5090-1445-3/16$31.00 © 2016 IEEE (DTW'16)

2330-2186/16 $31.00 © 2016 IEEE
DOI 10.1109/CloudCom.2016.106

2. Inventory

In this section we describe the EDISON Inventory, focusing on organisation and methods for population. Data collected in the Inventory is a basis for further analysis.

2.1. Organization of inventory

EDISON inventory resides primarily online to allow for frequent updates. EDISON inventory contains information about:

1) academic programs
2) academic courses
3) industrial courses
4) books

Data for over 300 programs and over 100 academic and industrial courses were collected by EDISON partners and further contributions were provided by the community. The inventory of programs is the subject of analysis in this paper.

The following data elements were collected for each program, with infrequent exceptions where some elements were not available:

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 publicly on the project's website and anyone interested is allowed to submit request for updates.This is further described in section 2.4. In addition, we have collected the following data to facilitate the necessary analysis:

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

It is impossible to ensure that such inventory is ever fully complete, but to our knowledge it is the most comprehensive effort of that sort existing today. These data are a basis for quantitative analysis in the case of degree of coverage of competence groups and qualitative analysis in the case of intended learning outcomes.

2.2. Methods for population

Population of the EDISON Inventory is a continuous process, in which we aim to engage the Data Science community. 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 initial list population was primarily based on a search, based on a set of terms including, but not limited to: data science, machine learning, data analysis/analytics, business intelligence, and business analysis/analytics. While breadth of the coverage was important, simultaneously we focused on the depth of each entry. In particular, we focused on the 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).

Further, the Inventory was extended through a network of partners with knowledge about specifics of various countries. Due to language differences, such offerings might be underrepresented in a general English-based search.

2.3. Inventory as a community service

The initial goal of the inventory was to serve as a basis for analysis presented in this paper, which further contributes to the development of i.a. model curricula. Nevertheless, such inventory can be a resource on its own right. We make it available to the community on the EDISON website [2]. It is possible to browse and filter the Inventory and also submit corrections and new entries.

3. Basis for analysis

The basis for quantitative analysis of entries in the EDISON Inventory is the Data Science Competence Framework (CF-DS) presented in [3]. The framework is described in greater detail in Section 4.4 of *Deliverable D2.1 - Data Scientist Competences and Skills Framework (CF-DS) and BoK definition (first version)* [4]

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 (relating mostly to computer and software engineering), and Data Science domain expertise (which can vary depending on the particular focus of the program). Each course in the program 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 competence meta-groups: 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 scientists. Small differences in coverage are natural. We propose 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 analysed 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.

4. Quantitative analysis of degree-giving programs

4.1. Origin of programs

Figure 1 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 have not yet been 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.

As explained in Section 2 the Inventory is a result of a combination of search results together with input from EDISON and EDISON Liason Group (ELG), which is a group consisting of independent experts that represent the major stakeholders in Data Science that work as a consulting body for the project and will create a basis for future independent expert group for universities and for European Commission. Results from search give particular weight to programs conducted in English, which are naturally most common in the UK. At the same time, many partners from e.g. the Netherlands and Italy, result in good coverage of these countries.

4.2. 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 2 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 specialised department.

In Figure 3 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.

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

4.3. Coverage of domain knowledge

Each program in the inventory was analysed 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 4 and Figure 5 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. An 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.

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 overemphasise generic management and business skills. There is little conceptual connection between courses offered to cover domain knowledge and those covering other competence groups.

Figure 1. Origin of European Programs

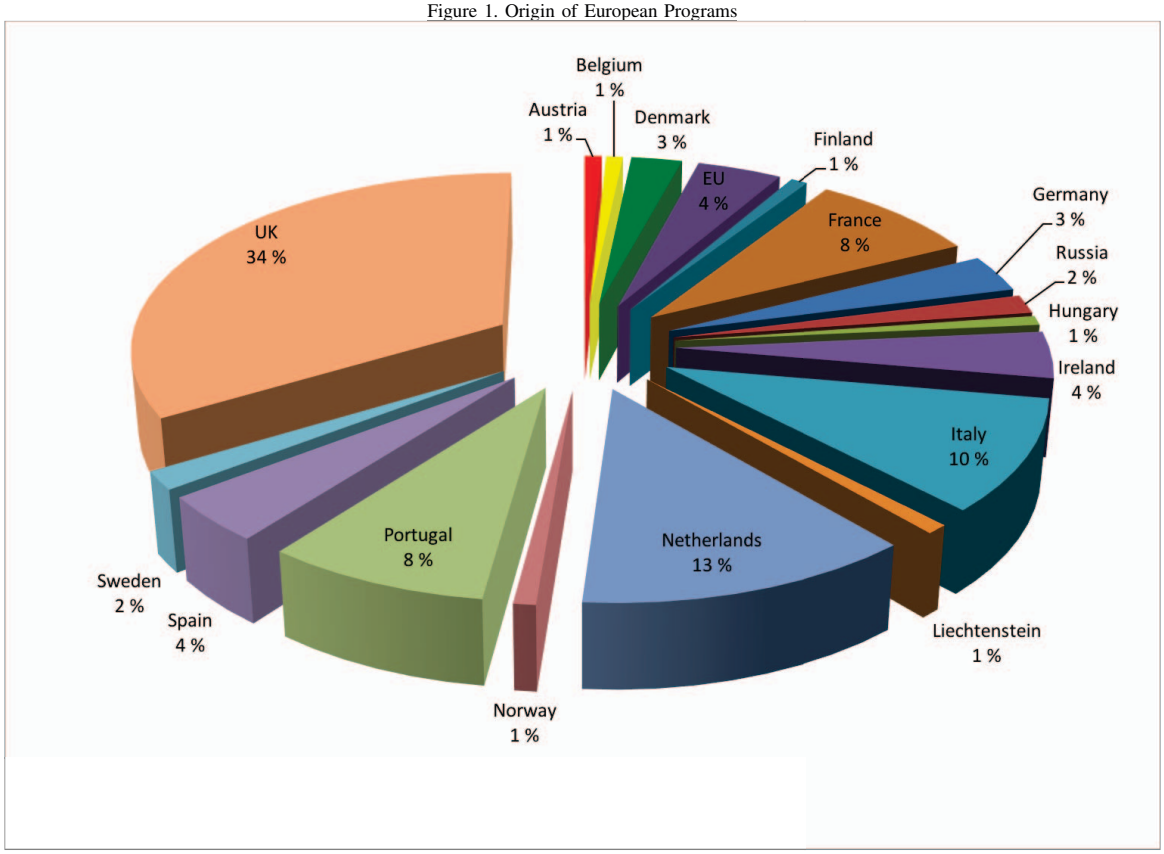


Figure 2. Source of European Programs

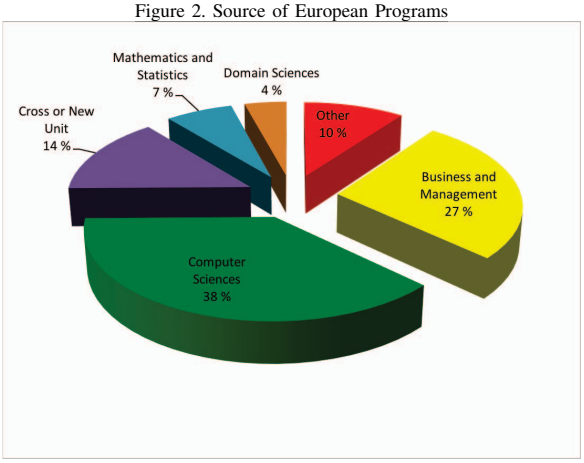
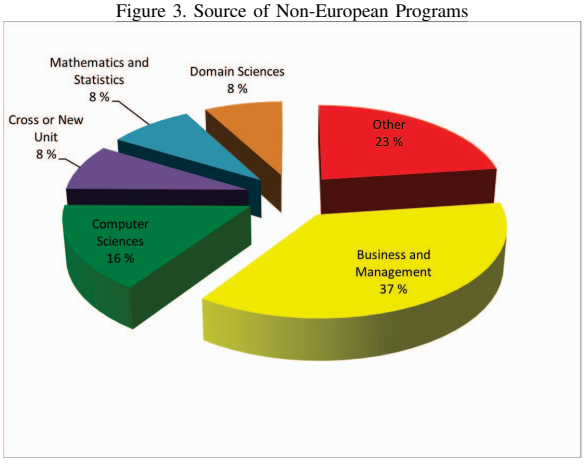


Figure 3. Source of Non-European Programs



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 specialised courses in economics, financial analysis or similar.

Many programs appear to equate data scientists with

Figure 4. Balance of European Programs

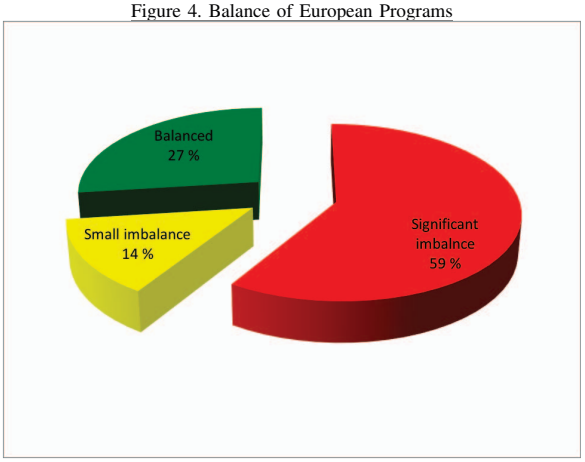
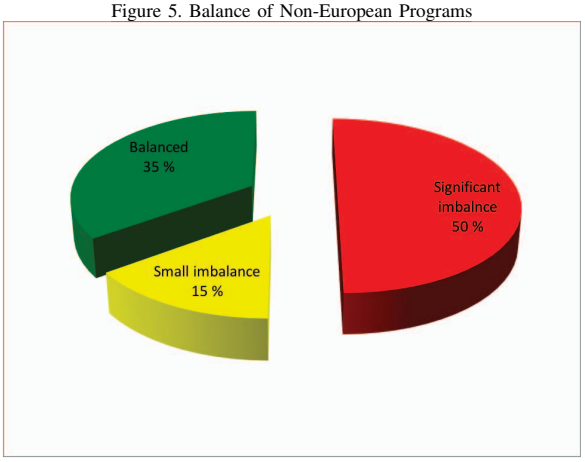


Figure 5. Balance of Non-European Programs



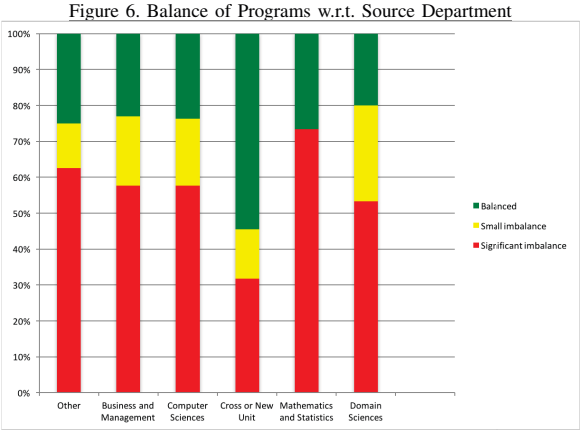
business analysts. While business analysis might be considered a special case of Data Science, the opposite is certainly not correct.

Finally, in Figure 6 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 over-interpreted in the early stages of Data Science curricula development.

5. 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

Figure 6. Balance of Programs w.r.t. Source Department



goals, competences and objectives. Only 8% of European programs have such definitions and the corresponding number for non-European programs is 16%, mostly due to US influence. It is far fewer than expected, considering that all academic programs should formalise learning outcomes. Due to limited data only general conclusions can be extracted.

When we evaluated the quality of learning outcomes w.r.t. Blooms 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.

6. Conclusions

We created the EDISON Inventory of Data Science education resources. The main focus of the Inventory was academic programs because analysis of existing programs is an important component for designing Model Curricula. The Inventory was published as a service to Data Science community. It is also open for correction and inclusion of new entries.

Subsequently, we analysed programs in the EDISON inventory. We noticed significant differences between Europe and outside Europe (mostly United States) in departments 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.

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 standardisation.

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 competences 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, which is 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 Blooms 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. An example of such apporach for a single course is presented e.g. by Wlodarczyk and Hacker [5].

Acknowledgments

We acknowledge the financial support from European Commission Horizon 2020 under Grant Agreement 675419 (EDISON).

References

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

[2] EDISON. Building the data science profession, http://edison-project.eu, Last visited on 28.02.2016.

[3] Manieri, A., et al. (2015, November). Data Science Professional Uncovered: How the EDISON Project will Contribute to a Widely Accepted Profile for Data Scientists. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 588-593). IEEE.

[4] Demchenko, et al. (2016, February) Deliverable D2.1 - Data Scientist Competences and Skills Framework (CF-DS) and BoK definition (first version), http://edison-project.eu/data-science-competence-framework-cf-ds

[5] Wlodarczyk, T. W., & Hacker, T. J. (2014, December). Problem-based learning approach to a course in data intensive systems. In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on (pp. 942-948). IEEE.