*March 2017*

# Cross-discipline Higher Education of Data Science – Beyond the Computer Science Student

Evangelos Pournaras [a,1],

[a] *Professorship of Computational Social Science*
*ETH Zurich, Zurich, Switzerland*

**Abstract**

The majority of economic sectors are transformed by the abundance of data. Smart grids, smart cities, smart health, Industry 4.0 impose to domain experts requirements for data science skills in order to respond to their duties and the challenges of the digital society. Business training or replacing domain experts with computer scientists can be costly, limiting for the diversity in business sectors and can lead to sacrifice of invaluable domain knowledge. This paper illustrates experience and lessons learnt from the design and teaching of a postgraduate cross-disciplinary data science course at a top-class university. The course design is approached from the perspectives of the constructivism and transformative learning theory. Students are introduced to a guideline for a group research project they need to deliver, which is used as a pedagogical artifact for students to unfold their data science skills as well as reflect within their team their domain and prior knowledge. The course content is designed to be self-contained for students of different discipline. Without assuming certain prior programming skills, students from different disciplines are qualified to practice data science with open-source tools at all stages: data manipulation, interactive graphical analysis, plotting, machine learning and big data analytics. Quantitative and qualitative evaluation with interviews outlines invaluable lessons learnt.

**Keywords.** education, data science, cross-discipline, big data, research methodology, learning, constructivism theory, transformative theory

## 1. Introduction

The pervasiveness of Internet of Things and ubiquitous computing brings unprecedented transformations in several sectors of economy. Nowadays, design, operational, management and regulatory decisions in smart cities, smart grids, smart health services and Industry 4.0 rely on streams of massive data. This radically alters the skills set of domain experts required to automate, analyze and optimize such complex systems [6]. Data science becomes of a paramount importance. Experts' skills on statistics are not adequate as data may be unstructured, very large in size, may require real-time processing and advanced machine learning techniques that go beyond descriptive statistics [6]. On the

[1] Evangelos Pournaras, Professorship of Computational Social Science, Clausiusstrasse 50, 8092, Zurich, Switzerland; E-mail: epournaras@ethz.ch.

one hand, training domain experts to new skills of data science is costly and not all enterprises have the resources for this purpose. On the other hand, replacing the domain experts with computer scientists, who are formally trained on data science may result in imbalances in the job market and lack of diversity, cohesion and domain knowledge. Note that several groundbreaking discoveries in the areas of complex networks and biology, for instance mapping human genome have been made with data science methodologies applied by domain experts rather than computer scientists. There is ongoing research on introducing formal models for such discoveries [18].

Given the evident lack of plurality and interest for data scientists in the job market [34,20], academic institutes need to respond to their role and educate a broad range of scientists in data science with novel didactic and pedagogical approaches tailored beyond the computer science student. The experience aggregated by the design of such a novel educational course is the focus of this paper.

The design and teaching of a postgraduate data science course in a cross-disciplinary context can benefit from the constructivism and transformative learning theory. By using research methodogies applied in data science research projects used as pedagogical artifacts, students can benefit from the concepts of the two learning theories: learners' prior knowledge and experience as well as habits of mind and point of view.

This paper illustrates the experience and lessons learnt from the design and teaching of a cross-discipline data science course at ETH Zurich. It shows how the course overcomes challenges such as creating a self-contained content or choosing software tools for teaching data manipulation and graphical analysis. Moreover, the diversity of (i) the students, (ii) the projects selected and (iii) the project teams is discussed along with its role to students' success. The course evaluation and students' feedback are illustrated with quantitative and qualitative information aggregated from official university evaluations as well as personal interviews conducted for the purpose of this paper. Several lessons learnt are derived related to the content size, the difficulty level, the role of diversity, the choice of software tools, the role of the research projects as a pedagogical artifact as well as the data requirements that students and lecturers need to take care of.

This paper is outlined as follows: Section 2 discusses the perspective of learnign theory on data science. Section 3 introduces the course "Data Science in Techno-socio-economic Systems" designed and taught at ETH Zurich. Section 4 illustrated the self-contained course content. Section 5 illustrates the guide of data science research projects and shares experiences on students' work. Section 6 illustrates a quantitative and qualitative evaluation of the course. Section 7 summarizes the lessons learnt and the societal implications of cross-disciplinary data science education. Finally, Section 8 concludes this paper.

## 2. Perspectives of Learning Theory on Data Science

To the best of the author's knowledge, there is very limited relevant work on how learning theory applies to data science education, and especially in the cross-discipline education of data science. This section discusses the perspective of two relevant theories in this context, *costructivism* and *transformative* learning theory.

A constructivism pedagogical approach to data science benefits from the prior knowledge and experiences on which learners of different background have been exposed [4]. Beyond the technical computer science content of data science, the background knowledge that learners bring is invaluable to contextualize and develop data science knowledge and practices, articulate a domain-specific reasoning via posing valid research questions and hypotheses from learners' field of expertise [8].

Data science education with learning methods from the perspective of Papert's constructionism [26] promote the use of data science tools and techniques as self-learning artifacts to facilitate the construction of new knowledge [1]: learning the actual data science methods and generate new knowledge from their applicability on a domain. In this sense, data science can be seen as an intellectual environment that students actively use as an evocative object [33] to solve a domain problem, while this use entails practicing data science and therefore the development of new skills and knowledge. This view is in line with the learning approach of 'to-think-with' and 'to-learn-with' technology (data science in this context) [30].

The transformative learning theory is relevant to the education of data science in a cross-disciplinary context. The theory explains how learners revise and interpret meaning [31] and articulates learning as the cognitive process of effecting change in a frame of reference composed of two dimensions: *habits of mind* and *points of view*. Emotions are involved [16] and ideas may not be easily accepted if the pedagogical approach does not encounter the diversity of the learners, i.e. values, associations and concepts they have formed [22]. This is especially relevant for cross-disciplinary data science education.

For instance, consider the cognitive process of data speculation by learners of different disciplines, for instance, a (visual) exploratory analysis of residential energy demand data. Self-reflection on this process is evident given the habits of mind and view point of learners. An electrical engineer may speculate about system robustness, for instance power peaks causing blackouts. In contrast, an economist may interpret data in economic terms, meaning a power peak may imply low energy prices. And a social scientists may link power peaks to human behavior, for instance human mobility (returning back from work), residential activities and an overall certain lifestyle. Moreover, a computer scientist may find these data as privacy-intrusive given a prior knowledge on inference techniques capable of, for instance, detecting with high accuracy TV programs chosen by exlusively using data of the TV power consumption [12]. Such ethical concerns may influence the individuals' choices of the data analytics methods applied over sensitive personal data. The learning process as well as the educational content should encounter for this diversity and has the capacity to be integrative of different learners' experiences.

## 3. A Cross-disciplinary Data Science Course

This section outlines a relevant cross-disciplinary data science course created at ETH Zurich. The 3-credit course entitled *Data Science in Techno-socio-economic Systems* is designed for MSc students and it is part of the department[2] "Humanities, Social and Political Sciences" (GESS). The course was designed in 2014 and has been running for the three years of 2015-2017 during spring semesters. The lecturers, including the author, are two computer scientists with specialization in distributed systems and big data as

[2] Available at `https://www.gess.ethz.ch/en/` (last accessed: March 2017)

well as a physicist. All lecturers have experience in multi-disciplinary research and work for the group[3] "Computation Social Science" (COSS).

ETH Zurich offers a wide range of courses within computer science curricula in the peripheral area of data science, for instance, data mining, big data, machine learning and others. Most of these courses are offered by renowned international experts and they concern state-of-the-art methods and techniques of data science, mainly from an academic viewpoint, i.e. analytical expressions and complexity analysis of machine learning algorithms. For this reason, most of these courses are exclusively designed for students with a strong mathematical or computer science background. A similar trend is observed in the vast majority of science and technical universities. The course "Data Science in Techno-socio-economic Systems" is designed to establish a broader scope of data science education that is highly cross-disciplinary, practical, yet, research-oriented. Without being an introductory course to the aforementioned more advanced and computer science-oriented courses, "Data Science in Techno-socio-economic Systems" minimizes the content overlap, while providing evident learning opportunities to students attended more advanced courses to unfold advanced knowledge and skills in a new educational context.

Table 1 illustrates the educational background of students participating in the course each year. The following observations can be made: (i) Students from physics and computer science cover together almost half of the students throughout the years. (ii) The course gains significant popularity from students following the direction "Management, technology and economics". (iii) Diversity in the directions increases over the years.

**Table 1.** Students' educational direction participating in the course "Data Science in Techno-socio-economic systems". The top-3 educational directions are indicated by the grey boxes.

| Direction | 2015 (%, $n = 13$) | 2016 (%, $n = 37$) | 2017 (%, $n = 43$) |
|---|---|---|---|
| Physics | 30.8 | 18.9 | 25.6 |
| Biochemistry and physics | 7.7 | 0.0 | 0.0 |
| Environmental sciences and engineering | 7.7 | 5.4 | 2.3 |
| Energy science and technology | 0.0 | 0.0 | 7.0 |
| Earth sciences | 0.0 | 0.0 | 2.3 |
| Statistics | 0.0 | 2.7 | 7.0 |
| Mathematics | 0.0 | 0.0 | 4.7 |
| Biomedical engineering | 0.0 | 2.7 | 0.0 |
| Computational biology and bioinformatics | 7.7 | 2.7 | 0.0 |
| Computer science | 15.4 | 27.02 | 18.6 |
| Electrical engineering | 15.4 | 10.81 | 4.7 |
| Robotics systems and control | 0.0 | 0.0 | 2.3 |
| Mechanical engineering | 7.7 | 10.81 | 2.3 |
| Civil engineering | 0.0 | 2.7 | 2.3 |
| Materials | 7.7 | 0.0 | 0.0 |
| Architecture | 0.0 | 2.7 | 0.0 |
| Management, technology and economics | 0.0 | 13.5 | 14.0 |
| Humanities, social and political sciences | 0.0 | 0.0 | 4.7 |
| Other | 0.0 | 0.0 | 2.3 |

[3] Available at `http://www.coss.ethz.ch` (last accessed: March 2017)

Table 2 illustrates the students' semester status, who have participated. There are three main categories of students attending the course: (i) 6th semester BSc students (last year BSc students), (ii) 2nd semester MSc students (first year MSc students) and (iii) PhD students. Therefore, the course succeeds to attract a broad range of students groups including mature BSc students, MSc students who early plan to get involved with data science education and PhD students who acquire data science skills in the their PhD project or intend to learn research methodologies applied in data science.

**Table 2.** Students' semester status participating in the course "Data Science in Techno-socio-economic systems". The top-3 semesters are indicated by the grey boxes.

| Semester | 2015 (%, $n = 13$) | 2016 (%, $n = 37$) | 2017 (%, $n = 43$) |
|---|---|---|---|
| 1st - BSc | 0.0 | 0.0 | 0.0 |
| 2nd - BSc | 0.0 | 0.0 | 0.0 |
| 3rd - BSc | 0.0 | 2.70 | 0.0 |
| 4th - BSc | 0.0 | 5.41 | 0.0 |
| 5th - BSc | 0.0 | 0.0 | 0.0 |
| 6th - BSc | 15.39 | 2.70 | 20.93 |
| 7th - BSc | 0.0 | 0.0 | 0.0 |
| 8th - BSc | 7.69 | 4.41 | 0.0 |
| 1st - MSc | 0.0 | 2.70 | 0.0 |
| 2nd - MSc | 61.54 | 51.35 | 46.51 |
| 3rd - MSc | 0.0 | 5.41 | 2.33 |
| 4th - MSc | 0.0 | 2.70 | 13.95 |
| Doctoral student | 7.69 | 21.62 | 13.95 |
| Other | 7.69 | 0.0 | 2.33 |

Sections 4-6 illustrate the content and research projects of the course as well as the course evaluation and students' feedback.

## 4. Self-contained Data Science Education

The goal of the course is to teach a large spectrum of postgraduate students data science and guide them to develop skills with which they can independently practice data science starting from data collection to oral/written presentation of results. The educational content relies on epistemic concepts of science and research in order to cultivate a problem-oriented mindset as well as confidence by learning rigorous methodologies and conventions: formulating research questions/hypotheses, presenting results in a quantitative manner, combining the behavioral and design science research strategies.

Moreover, the education relies exclusively on open and free software. It has been shown that closed and proprietary software compromises empirical research and replication of results in data science [7].

Content-wise, the course covers the following topics:

- **Fundamentals**: data collection, data modeling, data manipulation, databases, plotting and visualization.
- **Data mining and machine learning**: classification, decision-trees, clustering, prediction, neural networks and others.

- **Big data analytics**: MapReduce, parallel computing, data streaming and social media.
- **Applications**: smart grids, smart cities, traffic systems, twitter analysis, mobility, localization, activity recognition, privacy-preserving social sensing and mining.

The course makes no assumption of pre-requisite knowledge, e.g. certain programming skills, and therefore, it is designed to build up a minimal knowledge at all stages of data science. Data manipulation proves to be one of the most challenging educational material to design here. It is the "makes your hands dirty" job, a critical requirement for the direct involvement of students in data science as it involves processing tabular data, removing missing values or outliers, aggregating, filtering, formatting and transforming data. Although most programming languages, e.g. python, scala, java, etc., provide advanced data manipulation methods, teaching a general-purpose programming language in a cross-disciplinary audience of students entails several drawbacks and limitations: (i) steep learning curve, especially for non-computer scientists, (ii) potential distraction from the main goals of the course, (iii) hard to make a choice for a programming language that would meet the expectations and desires of an heterogeneous audience [35], (iv) may raise motivational issues for students who already perform data manipulation with a programming language of their convenience.

The solution here has been AWK[4], a standardized interpreted unix programming language that is very easy to learn, it is a mature language with plenty of online learning material and serves the educational purpose of the course. AWK handles all I/O operations as well as resource management, i.e. memory, and therefore allows students to entirely focus on calculations over data. AWK has been a new experience for the vast majority of students, even for computer scientists with systems background who learn about the use of AWK in a new context: data science. Although there is very limited [17] formal educational material of teaching AWK for data science, several online sources[5] praise the features of the language for this purpose.

Concerning databases, the focus is on running SQL queries. The free visual environment of MAMP[6] is used that comes with easy installation and configuration of Apache, PHP and MySQL. Moreover, the educational material on plotting includes Gnuplot[7] that is free and open-source, supports graphical analysis, can interoperate with AWK and can provide high-quality graphics. Gephi[8] is shown as a tool for studying and visualizing complex networks, although an extensive coverage of this topic is out of the scope of the course.

The part on data mining and machine learning provides an overview to supervised and unsupervised machine learning algorithms and how to choose one for a certain problem. Teaching these algorithms from a mathematical perspective is covered in depth in other courses that require a narrower focus. Instead, this course aims at qualifying students to make informed decisions about the algorithms they use and their implications. For instance, it is shown that clustering of time series subsequences is meaningless de-

[4] Available at `https://www.gnu.org/software/gawk/manual/` (last accessed: March 2017)

[5] Examples: `http://www.gregreda.com/2013/07/15/unix-commands-for-data-science/`, `http://datascienceatthecommandline.com`, `http://john-hawkins.blogspot.ch/2013/09/using-awk-for-data-science.html` (last accessed: March 2017)

[6] Available at `https://www.mamp.info/en/` (last accessed: March 2017)

[7] Available at `http://gnuplot.sourceforge.net` (last accessed: March 2017)

[8] Available at `https://gephi.org` (last accessed: March 2017)

spite a long lasting research that adopts this method [19]. Similarly, extracting a user profile from historical discrete data by using the centroids of clusters may result in a profile that does not correspond to reality, as the mean may not appear in the historic data. The cluster medoids may be more relevant in such a case [28]. Data over-fitting and under-fitting are discussed as well.

The part on big data analytics covers batch vs. real-time data processing using Hadoop[9] and Storm[10]. Students are familiarized with the Hadoop architecture, job running, and terminology in addition to gaining a deep understanding of the MapReduce model. Several examples in pseudo-code are given, however, scripts and batch execution routines are illustrated as black boxes that students could use in their projects with lightweight modifications and the help of the tutors. Finally, emphasis is given on the challenges of big data such as distinguishing causation from correlations, especially spurious ones that appear when big data allow a massive number of variable combinations.

The course is highly application-oriented. One of the goals of the course is to develop domain knowledge and expertise and therefore every lecture involves one or more application examples in which the data science methodologies are underlined and discussed. Several of these examples come from the multi-disciplinary research of the tutors [28,27,11], adjusted in an education context. They are accompanied by the datasets and the software code for the repeatability and expansion of the results by the students. Moreover, earlier successful students' projects are presented to develop a psychological construct of self-efficacy in the course [2]. This proves to be particularly essential for students' confidence and expectations at the beginning, when they define the goals of their projects.

## 5. Data Science Research Projects

Students are evaluated based on a group research project of 2-3 people that they have to carry out throughout the semester. The grade of the project covers entirely the grade of the course and it is formed based on the following criteria: (i) *scientific clarity*-25%, (ii) *technical clarity*-25%, (iii) *writing and content presentation*-25%, (iv) *oral presentation*-25%, (v) *data generation and collection*-10% bonus. The latter acknowledges the key role that data construction plays on scientific practices of theory development [8].

Unlike a more conventional project report, the research project for this course has several formal requirements that reflect on the knowledge quality of the course material required to successfully practice data science as well as effectively present results written and orally. In this sense, the research project can be seen as a pedagocical artifact reflecting on the constructivism learning theory [4]. The project requirements are defined by an educational guideline that is based on fundamental epistemic concepts and conventions of research philosophy and strategy applied in the context of data science. The outline of the guide is the following:

1. *Define the challenge.*
2. *Define the outcome and its significance.*
3. *Reason about a data science approach.*

[9] Available at `http://hadoop.apache.org` (last accessed: March 2017)
[10] Available at `http://storm.apache.org` (last accessed: March 2017)

4. *Select the data sources.*
5. *Define evaluation metrics and measurements.*
6. *Build the data analytics pipeline.*
7. *Perform validation and evaluation.*
8. *Draw conclusions and future work.*

In Step 1, defining the challenge involves the formulation of a research question or hypothesis accompanied by related work. In other words, the students are encouraged to develop a problem-solving mindset from the very beginning of their project work. However, the guideline is not strict at this point as it is known that new research questions and hypotheses may become apparent during the exploratory data analysis. The outcome and the significance of a project in Step 2 concerns the broader positioning of students' work in society. Is the outcome an artifact, such as an algorithm or an engineered system? Or just a better understanding of an observed phenomenon? Can the results be used to design a new policy? And for whom is the outcome relevant? An end user, a policy-maker or a system operator? A certain problem can be studied with other approaches beyond data science, for instance, agent-based simulation or analytical approaches. Justifying the intractability of a mathematical problem or reasoning about the data science approach is part of Step 3 of the guide. At Step 4 of the guide stands the selection of data sources. Datasets need to meet project requirements, for instance, quality, size, format, granularity level etc. For instance, when a smart grid data science project focuses on residential energy consumption, data about the aggregate load of a power grid are not relevant. In this case, smart meter residential energy data are required or advanced methods for energy disaggregation [21]. The guideline encourages students to be quantitative in their illustrations and for this reason they need to define specific meaningful metrics and measurements in Step 5. For instance, the relationships in a social network can be measured with topological and graph spectral properties. When these networks are temporal, the respective temporal metrics should be applied [25]. The data analytics pipeline in Step 5 is the design of the data manipulation, processing and analytics performed. Step 6 suggests a high-quality illustration of quantitative results. Step 7 completes the guide with explicitly stating the conclusions and future work.

The guide is presented to the students at the very first lecture followed by lectures with project examples that adhere to the guideline. Students are asked to deliver an early one-page proposal at the 3rd week of the semester outlining the challenge they tackle, why they employ a data science approach, what the related work is in the problem area and what data they intend to use. At the end of the semester, students deliver their report and present their project to the course instructors as well as all other students in the class.

Table 3 illustrates students' projects during the first and the second year of the course. The following two key observations can be made: (i) The diversity of the projects is very high ranging from analysis of twitter data, mobile sensing, Internet of Things, analysis of scientific publications, analysis of traffic and environmental data, medical applications and other. (ii) The diversity of the students' background in the teams formed is not high. The group formation trend is that students either form teams with colleagues they know from their own studies program or they are more comfortable when they work together with people from the same background, even if they choose a project that is not in the domain area of their expertise, for instance Project 4 and 15.

However there are some exceptions worth mentioning and discussing. For instance, Project 11 has been a challenging project, though very successful and resulted in a pres-

**Table 3.** Students' projects and diversity in their educational direction.

| Number | Project | Student 1 | Student 2 | Student 3 |
|---|---|---|---|---|
| 1 | Graphical Analysis of Nervousnet Proximity Data | Computer Science | | - |
| 2 | How Can We Identify Crowds Behaviour Using Noise Data? | Electrical engineering | Physics | - |
| 3 | Identifying community structures by geo-located Twitter data | Environmental science and engineering | Materials | - |
| 4 | Topic extraction and analysis from scientific publications | Biochemistry and physics | Computational biology and bioinformatics | Electrical engineering |
| 5 | Public Opinion on Climate Change | Physics | Physics | - |
| 6 | Real-time human activity recognition from accelerometer data using Convolutional Neural Networks | Computer science | - | - |
| 7 | Spurious relationships in Twitter data | Physics | Physics | Physics |
| 8 | Are cyclists on the move according to weather conditions? | Computer science | Computer science | Mechanical engineer |
| 9 | Identifying Opinion Leaders in Social Networks | Computer science | Computer science | Computer science |
| 10 | Why do you leave your bicycle at home today? Factors that influence the number of bicycles in the city of Zurich | Environmental Sciences and engineering | Environmental Sciences and engineering | - |
| 11 | A Case Study for Urban Stress Level Monitoring | Mechanical engineering | Architecture | Mechanical engineering |
| 12 | Quantitative Evaluation of Gender Bias in Astronomy | Physics | Physics | Physics |
| 13 | Analysis of Language Mobility using Twitter Messages | Management, technology and economics | Management, technology and economics | Management, technology and economics |
| 14 | Sentiment Analysis on Twitter Data | Computer science | Computer science | Computational biology and bioinformatics |
| 15 | Schizophrenia Classification Challenge Report | Computer science | Electrical engineering | Electrical engineering |

tigious scientific publication in conference proceedings [13]. This project conducted by an architect and two mechanical engineers and involves an advanced data collection process in the context of smart cities using mobile phones, wearables and several environmental sensors carried by participants in the study. The goal of the project is to measure several urban qualities in a city path, for instance, greenery, stress, noise pollution and other. Project 13 is another successful project ran be students of the same, but highly inter-disciplinary study program and resulted in a scientific publication as well [23]. The goal of the project is to measure the spatio-temporal language mobility evolution and detect real-world events as well as tourism patterns via Twitter and the analysis of 10TB of tweets[11].

[11] Available at `https://archive.org/details/twitterstream.` (last accessed: March 2017)

There are also projects though that faced some serious challenges. For example, Project 10 studied correlations between the use of bicycles and the weather in Zurich. It proved not straightforward for the students to go beyond descriptive statistics and a regression analysis of the data without additional supervision effort. In this particularly case, the low diversity of the group played a critical role. At the end the students managed to compare classification results on weather phenomena between k-means clustering and the Gaussian Data finite mixture model fitted by the EM algorithm [10]. Project 2 also proved to be especially challenging for the students due to the low data quality by privacy-preservation constraints introduced during the data collection process [24]. In this case, the initial hypothesis was whether the activity of the Chaos Communication Congress[12] could be detected via noise sensors, for instance, parallel sessions, breaks, human interactions and other. During the 2014 edition of the congress, the Nervousnet[13] team deployed a smart phone platform for an anonymous privacy-preserving data collection that involves smart phone sensors as well as a GPS-free privacy-preserving localization mechanism using bluetooth beacons [24]. As the platform relies on a volunteering and participatory data collection process, the collected dataset is highly sparse, yet contains data from a wide range of sensors. By exclusively narrowing the scope of the project down to the noise sensor, students could not detect events with a high accuracy and statistical confidence despite the heavy interpolation applied. However, during the project, they started experimenting with the other sensor data available and turned their project on mining sensor data into a sensor fusion project that improved the detection accuracy significantly. A lesson learnt here is that practicing data science under the constraint of privacy-preservation requires an explicit addressing in the data science course design with more advanced techniques and the education of alternative approaches, for instance, privacy-preserving data analytics using summarization and differential privacy [27].

## 6. Course Evaluation and Students' feedback

The course has received so far two official evaluation by the students conducted on behalf of ETH Zurich. The general satisfactions has been 4.4/5.0 and the lecturers' evaluation 4.5/5.0 on the following aspects: understandable and clear explanation of the subject, learning goals, lecture significance, motivation to active participation, and material made available.

For the purpose of this paper, the author conducted interviews with 5 students attended the course during the first two years to acquire further information about its effectiveness within a cross-disciplinary educational scope and the research-oriented methodology on teaching data science. The educational background of the interviewees is outlined in Table 4.

The interview is guided by the following agenda questions:

1. *How effective was the course for you?*
2. *Was this course too easy or too hard for you?*
3. *If you are a computer scientist, what was beneficial and limiting factors after attending a data science course designed beyond the computer scientist?*

[12] Available at `https://www.ccc.de/en/` (last accessed: March 2017)

[13] Available at `http://www.nervousnet.ethz.ch` (last accessed: March 2017)

**Table 4.** Background of interviewees.

| Interviewee | Direction |
|---|---|
| 1 | Architecture |
| 2 | Mechanical engineering |
| 3 | Computer science |
| 4 | Mechanical engineering |
| 5 | Management, technology and economics |

4. *If you are not a computer scientist, what was beneficial and limiting factor for you after attending a data science course taught by computer scientists and including other computer science students?*
5. *How successful was the course to provide you the minimum set of skills to practice data science?*
6. *How effective was in this course to learn doing research by practicing data science?*
7. *How effective was in this course to learn data science by doing some research?*

Questions 6 and 7 reflect on how students perceive constructionism learning methods [26], data science as an evocative object [33] and the learning approach of 'to-think-with' and 'to-learn-with' technology (data science in this context) [30]. Questions 3 and 4 reflect on how students experience the transformative learning approach [31]. They are formulated from the perspective of the computer scientist vs. non-computer scientist to encounter the diversity [22] and the dimensions within the frame of reference of the cognitive learning process [31]: habits of mind and points of view.

During the interviews, the effectiveness of the course was communicated by Interviewee 2 as "*a very nice change from my normal study life*" and "*learnt a lot of interdisciplinary skills*". Interviewee 5 gave emphasis on group work by stating that "*got to work closely with people, learn how they work and think, brainstorming and share ideas, make friends*". Interviewee 4 realized that "*Later on in my studies I could see a lot of areas where a data science approach could be useful, so I am happy to tell that I acquired some basic skills to practice data science.*", while Interviewee 1 experienced the course as "*a good overview for techniques in data science*". However, Interviewee 3 noticed that "*I would like to have known more about the exact technical details of the example applications that were presented.*" and "*as a computer science student I really like the implementation details.*" This indicates that it is very challenging to capture the right level of detail for a broad range of students with different expectations.

Interviewee 5 mentioned that the course had the "*appropriate*" difficulty level and Interviewee 3 found the course "*relatively easy*" as the interviewee felt "*familiar with most technical details discussed during the course*". In contrast, Interviewee 2 said that the "*course is one of the most complex and time intensive courses I did in GESS*" but also mentioned that "*the level was perfect to learn new skills*". In similar line is Interviewee 4 by stating that "*the course was pretty hard for me, but the effort was worth it*". Interviewee 1 identified that the course "*was challenging but not too hard, especially since we had to work in groups for our final projects*".

Question 3 answered by Interviewee 3 revealed that "*a lot of time the course was focused on data science technical stuff I was already familiar with*". However the student also stated that "*A benefit was seeing all those applications on the real world, which we don't usually focus on in computer science*". Question 4 was answered by the rest of the interviewees. Interviewee 2 mentioned that "*I profited from knowledge and experience of other participants and lecturers and learned new terms of data science and statistics*" but also stated that "*I did not have a lot of background knowledge about the hard skills and libraries. This wasn't limiting for the course but for the project we did, because we needed a lot of time to find the right sources and libraries*". Interviewee 5 listed as benefits "*(i) coaching from the COSS team, (ii) learn about social data, opportunities, limitations, (iii) presentations by the other teams: gives ideas, insights and (iv) learn some new tools, e.g. mawk*". It was mentioned as limiting factor the "*Too much material covered during the lectures*". Interviewee 4 found beneficial the "*The way of thinking of a data scientist*" while Interviewee 1 "*would have preferred to have hands on examples/tutorials of various techniques in a given programming language*".

The answers on Question 5 share similar remarks on the beneficial and limiting factors of Questions 3 and 4. Interviewee 5 mentioned "*better to focus on the top-3 tools and techniques*" with Interviewee 1 agreeing, though adding that "*it helped point students in the right direction for which techniques might be useful for which questions*". Interviewee 2 reacted very positively by stating "*Very successful, I learned the skills to approach a big data problem and subdiving it into smaller problems. I also got to know nice tools to do that in real problems such as Weka and scikit (a python library).*" and Interviewee 4 "*I could see a lot of areas where a data science approach could be useful, so I am happy to tell that I acquired some basic skills to practice data science.*"

Answers to Questions 6 and 7 have a high heterogeneity, evidently showing how students perceive the link between data science and research methodologies. For Interviewee 1, "*was nice to learn techniques from other disciplines*" and "*nice to have a project to test your solutions*", though "*could have helped to have intermediate homework/assignments as well*". Interviewee 2 finds doing research by practicing data science as "*quite effective*", however, leaning data science by doing some research proves to be "*quite time consuming*". Interviewee 4 confidently states that "*if there wasn't our project, I would not understand cluster-based approaches as I do so now*" and supports that doing projects is "*the best way to learn*". Interviewee 5 believes that this course should have "*part of the lecture devoted to how to conduct research in general*", though practicing data science cannot obviously cover the whole broad spectrum of research methodology. In contrast, Interviewee 5 perceived the learning of data science by doing some research as "*One of my best ETH experiences so far*".

## 7. Lessons Learnt and Societal Implications

The design and teaching of the course "Data Science in Techno-socio-economic Systems" in a cross-disciplinary audience of students results in several lessons learnt that can be summarized as follows:

- **Content size and level**: Although the students express in overall a high satisfaction about the size of the material and the difficulty level, a few remarks indicate that there is space for improvement. The feedback suggests that working/lab sessions during the class may motivate further the computer scientists to improve their knowledge as well as the non-computer scientists to practice their skills during the course. Moreover, if the educational curriculum allows an increase in the course credits, the students could work on deliverables during the semester and undertake formal exams at the end of semester as a more systematic way to track and capitalize progress throughout the semester.

- **Diversity**: It is shown that the proposed course design motivates a high level of diversity in the educational background of the participating students as well as in the projects students choose. However, the diversity of the teams can improve by accommodating the formation of teams, stretching more the role of diversity in the evident success of the projects or by incentivizing with a bonus grade cross-disciplinary teams.
- **Software tools**: It is essential for tutors to cultivate in students the knowledge, skills and critical thinking required to independently make informed choices about the use of the most appropriate software tools for a certain data science task. Rigorous evaluation and replication of results benefits from free and open-source tools [7]. A data scientist trained and relying on a broad spectrum of open-source software tools is more versatile in the job market than a data scientist trained for commercial software solutions. For a lecturer teaching data science in a cross-disciplinary audience, the choice of software tools is not straightforward and to certain extent it is a trade-off, for instance, the choice of AWK for data manipulation and Gnuplot for plotting and graphical analysis.
- **Research as a pedagogical artifact**: The research project proves to be a highly rewarding experience for students to learn data science and evidently has a high pedagogical value within a cross-disciplinary educational context. Students acknowledge the challenge to apply rigorous scientific methodologies on which they may not have earlier formal training. However, the research projects undoubtedly have a motivational value, provide freedom to students to unfold their interests and they are an actual opportunity to apply data science skills in real-world problems. Research projects run the risk of being too ambitious in the scope of a semester course, may rely on false assumptions or lack focus and therefore require a significant level of supervision effort, especially at the beginning during which tutors need to accommodate students' confidence and convey a spirit of self-efficacy.
- **Data requirements**: Data itself impose both explicit and implicit constraints on what a data scientist can learn from the data. Factors such as quality, dimensionality, granularity as well as functional/non-functional requirements during data collection, for instance, informational self-determination and privacy-preservation alter the opportunity space on what someone can learn from data. Scraping the surface of the available data and putting under scrutiny different graphical views, aggregation levels and data transformations shapes the solution to a data science problem or even reshapes a new solution to a different problem that was not evident or intended before this process.

The facilitation of cross-disciplinary data science education in university curricula qualifies a new generation of versatile professionals with the capability to communicate and work together with a broader range of experts. Moreover, making accessible data science to a wider range of domain experts can reduce business training costs. Similarly, academic education of data science with open-source and free software tools can reduce business costs on expensive commercial software suites.

Data science education using research methodologies cultivates to citizens a higher awareness about what data mean, a cognitive reasoning based on empirical evidence, critical thinking and constructive doubt. In other words, it cultivates these mental capacities to withstand the challenges of our nowadays digital societies [15] concerning the interpretation and wise use of information from (social) media [9], populism leading to ineffective voting [5,3], privacy and autonomy violations from big data profiling technologies or profit-oriented recommender systems [14,29], manipulative actions and means of propaganda in social networks and beyond [32].

## 8. Conclusion

This paper concludes that cross-disciplinary data science education is highly challenging and requires a very different approach in the design of study courses than data science education exclusively for computer scientists. However, this paper shows that cross-disciplinary data science education is feasible and highly rewarding for students. The perspective of costructivism and transformative learning theory proves effective for the design of a course with these requirements. Learning data science in this cross-disciplinary context has a value by itself as the students' diversity and the blend of skills in collaborative research projects create multifaceted learning opportunities that cannot unfold otherwise. This is empirically shown via the design, development and teaching of a new cross-disciplinary data science course at a top-class university and the experiences aggregated throughout the lifetime of the course.

This paper contributes lessons learnt such as how to make choices in regards to the content size and difficulty level, the diversity of students, students' projects and project teams, the choice of software tools for different data science tasks, the use of research projects as a pedagogical artifact and how data requirements influence what a student can learn from data. Cross-disciplinary data science education qualifies more versatile data scientists in the job market, can reduce business costs for training and ultimately cultivate a more democratic and participating citizen prepared to respond to the upcoming challenges of the digital society [15].

## Acknowledgment

Special thanks go to Dr. Izabela Moise for her contribution and our collaboration to design and run the MSc course "Data Science in Techno-socio-economic Systems". The author would also like to thank Prof. Dr. Dirk Helbing, the Computational Social Science (COSS) group and the department of Humanities, Social and Political Sciences (GESS) at ETH Zurich for hosting and supporting this course. Last but not least, I am grateful to the interviewees, as well as all students attending this course for making cross-disciplinary education of data science such an enriching and constructive teaching experience.

## References

[1] Edith Ackermann. Piagets constructivism, paperts constructionism: Whats the difference. *Future of learning group publication*, 5(3):438, 2001.

[2] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[3] Jamie Bartlett. Populism, social media and democratic strain. *Democracy in Britain: Essays in honour of James Cornford*, pages 91–96, 2014.

[4] Robert E Bleicher and Joan Lindgren. Success in science learning and preservice science teaching self-efficacy. *Journal of science teacher education*, 16(3):205–225, 2005.

[5] Gavin Brown. Review of education in mathematics, data science and quantitative disciplines: Report to the group of eight universities. *Group of Eight (NJ1)*, 2009.

[6] Thomas H Davenport and DJ Patil. Data scientist: The sexiest job of the 21st century-a new breed of professional holds the key to capitalizing on big data opportunities. but these specialists aren't easy to findand the competition for them is fierce. *Harvard Business Review*, page 70, 2012.

[7] Renato P Dos Santos. Big data as a mediator in science teaching: A proposal. 2014.

[8] Richard Duschl. Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education*, 32(1):268–291, 2008.

[9] Umberto Eco. *Faith in fakes*. Random House, 2014.

[10] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[11] Sébastien Gambs, Marc-Olivier Killijian, Izabela Moise, and Miguel Nuñez del Prado Cortez. Mapreducing gepeto or towards conducting a privacy analysis on millions of mobility traces. In *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International*, pages 1937–1946. IEEE, 2013.

[12] Ulrich Greveler, Peter Glösekötterz, Benjamin Justusy, and Dennis Loehr. Multimedia content identification through smart meter power usage profiles. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[13] Danielle Griego, Varin Buff, Eric Hayoz, Izabela Moise, and Evangelos Pournaras. Sensing and mining urban qualities in smart cities. In *Proceedings of the 31st IEEE International Conference on Advanced Information Networking and Applications-(AINA 2017)*. IEEE, 2017.

[14] Natali Helberger, Kari Karppinen, and Lucia DAcunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, pages 1–17, 2016.

[15] Dirk Helbing and Evangelos Pournaras. Society: Build digital democracy. *Nature*, 527:33–34, 2015.

[16] Knud Illeris. Transformative learning in the perspective of a comprehensive learning theory. *Journal of Transformative education*, 2(2):79–89, 2004.

[17] Jeroen Janssens. *Data Science at the Command Line*. O'Reilly Media, September 2014.

[18] Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery. *arXiv preprint arXiv:1612.08544*, 2016.

[19] Eamonn Keogh, Jessica Lin, and Wagner Truppel. Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 115–122. IEEE, 2003.

[20] John Yohahn Kim and Choong Kwon Lee. An empirical analysis of requirements for data scientists using online job postings. *International Journal of Software Engineering and Its Applications*, 10(4):161–172, 2016.

[21] J Zico Kolter and Matthew J Johnson. Redd: A public data set for energy disaggregation research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, volume 25, pages 59–62, 2011.

[22] Jack Mezirow. Transformative learning: Theory to practice. *New directions for adult and continuing education*, 1997(74):5–12, 1997.

[23] Izabela Moise, Edward Gaere, Ruben Merz, Stefan Koch, and Evangelos Pournaras. Tracking language mobility in the twitter landscape. In *Proceedings of the 4th International Workshop on Data Science and Big Data Analytics (DSBDA 2016)*. IEEE, 2017.

[24] Federico Musciotto, Saverio Delpriori, Paolo Castagno, and Evangelos Pournaras. Mining social interactions in privacy-preserving temporal networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 1103–1110. IEEE, 2016.

[25] Vincenzo Nicosia, John Tang, Cecilia Mascolo, Mirco Musolesi, Giovanni Russo, and Vito Latora. Graph metrics for temporal networks. In *Temporal networks*, pages 15–40. Springer, 2013.

[26] Seymour Papert. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc., 1980.

[27] Evangelos Pournaras, Jovan Nikolic, Pablo Velásquez, Marcello Trovati, Nik Bessis, and Dirk Helbing. Self-regulatory information sharing in participatory social sensing. *EPJ Data Science*, 5(1):14, 2016.

[28] Evangelos Pournaras, Matteo Vasirani, Robert E Kooij, and Karl Aberer. Decentralized planning of energy demand for the management of robustness and discomfort. *IEEE Transactions on Industrial Informatics*, 10(4):2280–2289, 2014.

[29] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer, 2015.

[30] M Rosa. *Constructing Identities through online Role Playing Game: relationships with the teaching and learning of mathematics in a distance learning course*. PhD thesis, UNESP - São Paulo State University, 2008.

[31] Edward W Taylor. Transformative learning theory. *New directions for adult and continuing education*, 2008(119):5–15, 2008.

[32] Daniel Trottier and Christian Fuchs. *Social media, politics and the state: protests, revolutions, riots, crime and policing in the age of Facebook, Twitter and YouTube*, volume 16. Routledge, 2014.

[33] Sherry Turkle. *Evocative objects: Things we think with*. MIT press, 2011.

[34] Wil MP Van der Aalst. Data scientist: The engineer of the future. In *Enterprise Interoperability VI*, pages 13–26. Springer, 2014.

[35] Barbara Wixom, Thilini Ariyachandra, David Douglas, Michael Goul, Babita Gupta, Lakshmi Iyer, Uday Kulkarni, John G Mooney, Gloria Phillips-Wren, and Ozgur Turetken. The current state of business intelligence in academia: The arrival of big data. *Communications of the Association for Information Systems*, 34(1):1, 2014.