Effect of Data Science Teaching for Non-STEM Students
A Systematic Literature Review

Luiz Barboza

CESAR School
Recife, Brazil 50030–390
Email: lcbj@cesar.school

Erico Souza Teixeira

CESAR School
Recife, Brazil 50030–390
Email: est@cesar.school

***Abstract*—The evolution of computing capacity allowed specialists in certain areas to benefit from this advance, although with little knowledge about data analysis technologies. In this way, our role as software scientists, more than increasing computational power, is to facilitate the access of people from other areas to these technologies and, with this combined effort, bring more relevant results to society. With this objective in mind, a systematic literature review was carried out to understand if (RQ1), how (RQ3) and why (RQ2) data science is being taught to students of non-STEM (Science, Technology, Engineering and Mathematics). The bases used in this research were ACM and IEEE, dismissing the articles that met the exclusion criteria. These criteria were: a) articles focused on the use of technology to improve the learning process in general; b) articles targeting different groups than the one prioritized here, non-STEM; c) educational improvements obtained with different proposals other than the introduction of data science.**

Keywords*–*Data Science; Non-STEM; Teaching.

I. INTRODUCTION

The popularity of data science courses has increased over the last five years (2015 to 2020), as we can see on the graph generated by Google Trends shown in Figure 1.

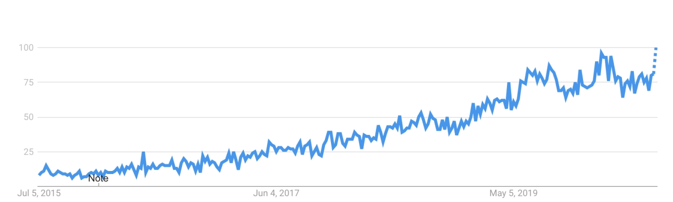

Figure 1. Worldwide search term, Data Science Course (source: Google Trends)

This is true for industry and academia, particularly in STEM courses, where this discipline has a solid base and even a reference curriculum [1] as the main guide. On the other hand, this type of knowledge is still not widespread in non-STEM areas. In fact, data science applied in different domain areas, is one of three data science pillars, as seen in Figure 2.

Bearing this in mind, a systematic review of the literature is presented based on the current state of the art of *if*, *how* and *why* data science is being offered in non-STEM courses. In the next sections, the method for research, selection, extraction and synthesization will be detailed to answer the three research questions.

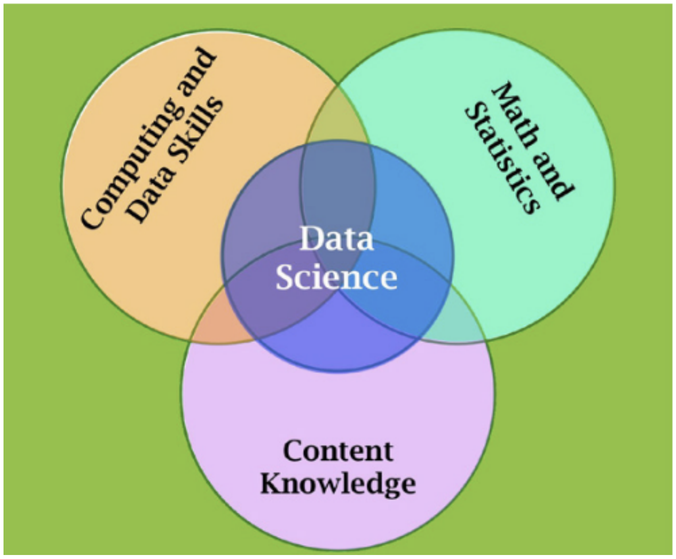

Figure 2. Data Science comprised of Computation, Domain Knowledge and Math/Statistics [2]

II. BACKGROUND

Since their emergence in the 1950s [3], machine learning algorithms have had limited applications, as they depended on computational power to process large volumes of data [4]. From the beginning of this century, the increase in computing power and the demand for understanding and relating the large mass of information available, machine learning solutions have become more sophisticated [5] and more popular in their use [6]. This popularization allowed machine learning to reach an important milestone, the possibility of access for people without specific training in science or data technology. Today, it is possible for people from different areas, such as Economics, Administration, Health, Philosophy, Architecture, among others, to be able to extract information from their data without the need for prior in-depth knowledge of data science. Even tough, this paper focuses on non-STEM students in general, we can consider economists as an example of it as the background context presented as follows. Like physicists, economists acquire non-experimental data generated by processes they want to understand. The mathematician John von Neumann defined a game [7] as: 1) a list of players; 2) a list of actions available to each player; 3) a list of how the accumulated winnings for each player depend on the actions of all players; and, 4) a time protocol that tells who chooses what and when. This definition corresponds to what economists call the economic system, a social understanding of who chooses what and when.

In addition, economists would like to conduct experiments to study how a hypothetical change in the rules of the game or in a pattern of behavior observed by some ”players”, for example, government regulators or a central bank, can affect the patterns of behavior of other players. Thus, the “structural model builders” in economics seek to infer, from historical patterns of behavior, a set of invariant parameters for hypothetical situations in which a government or regulatory body follows a new set of rules. ”Structural models” look for invariant parameters to help regulators and market designers understand and predict data patterns in historically unprecedented situations.

Like physicists, economists use models and data to learn. These models are then used to explain new data. Then, new models are built as evolution of their predecessors. This allows us to learn from the depressions and financial crises of the past. Nowadays, with big data, faster computers and better algorithms, patterns can be seen where only noise was previously heard.

The work presented here proposes to evaluate how the study of data analysis, even if not in-depth, can better train students from different non-technical areas of study, such as Economics and Administration.

III. REVIEW METHOD

A. Research Questions

The research questions analysed here were: **RQ1)** How is knowledge in data science being taught to non-technical target audiences, particularly economics and business students? **RQ2)** What are the learning improvements that these students are experiencing with the use of data science in different areas of their studies? **RQ3)** What was the method used in teaching data science?

B. Search Protocol

Using IEEE and ACM as the main source of research without year of publication threshold, the work here will look for documents related to data science knowledge that are being introduced to either secondary or higher level education targeting non-technical audiences, such as students of economics or business administration, defined by the following research string:

(”data science”) AND (teaching OR education) AND (economics OR administration OR humanities OR non-technical)

The protocol applied here was comprised of four steps: 1) Apply the search string: apply the string according to the objectives; 2) Filter based on the criterion: Filter the articles by the inclusion and exclusion criteria by analyzing their abstracts; 3) Validate answers to the research questions: read the selected texts, checking if they answer the research questions. If so, extract them as a reference for the final article resulting from a systematic review; 4) Synthesize: apply the thematic synthesis method in order to summarize the research findings.

TABLE I. SELECTION PROTOCOL RESULTS

	IEEE	ACM
Initial set of papers	300	130
Passed inclusion criteria	11	18
Final list of papers	3	6

After the inclusion/exclusion criteria review, the article set was filtered if it answered one of the research questions. The results are summarized in Table I.

C. Selection

The criteria for the inclusion of the article are:

- The article should be written in the English language.

- The article should have its scope focusing on data science studies of non-technical target audiences.

The criteria for the exclusion of the article are:

- Articles focusing on the use of technology to improve the learning process.

- Articles targeting different groups other than the one prioritized here, namely, non-STEM.

- Educational improvements achieved through different proposals other than data science introduction.

- Data Science application without the explicit goal of educational purposes.

D. Extraction

At this step of the process, specific extracts of the analysed papers were identified as being a valid answer to any of the three research questions. As an example, [8] could be cited here, specially as the author starts beautifully with this sentence: *”Because no data exists in a vacuum, each Data Analytics major must choose an applied domain in which to specialize. The goal of this specialization is to understand the types of questions that data are used to answer in that discipline, and how data are collected and interpreted in this context. There are currently seven available domains: Anthropology and Sociology; Biology; Economics; Philosophy; Physics; Political Science; and Psychology”*

The domains mentioned by [8] adhered to the data science studies in different proportions, as depicted in Figure 3.

E. Synthesis

In order to answer the first research question, the following classification was applied: a) school level to which it was applied; b) location/scope; c) concepts taught; and d) target audience. The coding applied to analyze the answers of the second research question was: e) the achieved results; and f) how they were measured. Finally, the last research question had its own coding, g) the method used in teaching data science.

IV. RESULTS

A. RQ1) The IF

In order to answer this research question, the following aspects were analysed: education level, location/scope, concepts taught and target audience.

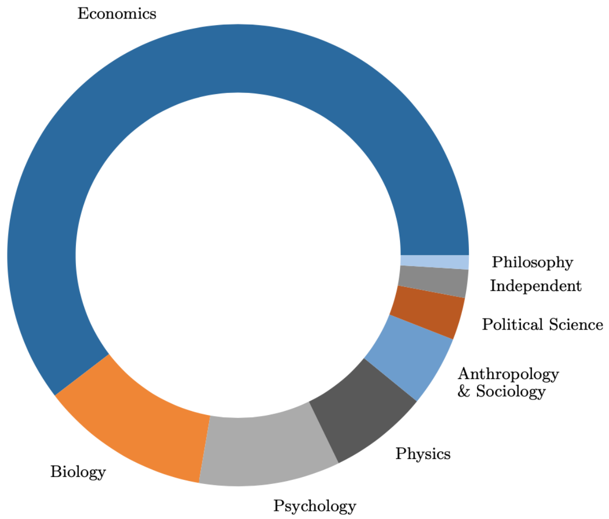

Figure 3. Proportion of courses applying data science on a particular study [8]

1) School Level: Junior High School (from 5th up to 9th graders): As seen in [9], Data Science is being taught to school kids, from 10 up to 15 years old. *"We organized a half-day long data science tutorial for kids in grades 5 through 9 (10-15 years old). Our aim was to expose them to the full cycle of a typical supervised learning approach - data collection, data entry, data visualization, feature engineering, model building, model testing and data permissions"*. The main goal of this experience is to expose young kids to data analysis reasoning and to an intuitive overview of the data science process. *Senior High School (from 10th up to 12th graders)*: A deeper approach can be observed in [10], in which programming (python), data analysis and problem solving are experienced by high school students. Serving as a bridge between programming intuition and logic to actual imperative coding, as stated by the author *"this course is a crucial component of the K-12 computational thinking pathways we are developing at our school district, which take students from block-based programming and computational thinking (elementary school) to text-based programming and applications of computer science (high school). Our mandatory 8th grade course serves as a bridge between these two components"*. *College/University level (undergraduate and graduate)*: At the college/university level is where we can see most of data science teaching. The most relevant aspect for the scope analysed here is *if* it is being taught to non-STEM students. This particular item will be reviewed under the target audience topic of this study. Anyhow, this kind of practice can be observed at undergraduate level by numerous authors [2][8][11][12][13][14]. To depict an example of it we can cite [12] *"data analysis and visualization techniques could be applied in an English literature class in order to help students better understand contextual information, analyze characters' social networks, and visualize literary techniques"*. At the graduate level, we can mention the experience reported by [15]: *"We are developing educational materials for data science to provide broad and practical training in data analytics non-CS students. This includes students majoring in science and engineering who want to acquire skills to analyze data, such as biology, chemistry, and geosciences. This also includes students in the humanities that would like to pursue data-driven research, such as journalism students interested in social media analysis"*.

2) Location/Scope: All the studies analysed report localized experiences, in the sense that none of the studies reported a broader experiment throughout a larger region rather than the local institution in which the study itself was used as base for the research. Most of the studies considered were based in the following states of the USA [14]: California [9][12][15], Washington DC [10] and Ohio [8]. The remaining of the studies were performed in Europe, in the following countries: Germany [2], Switzerland [13] and Finland [11].

3) Concepts Taught: The level of depth in which the concepts are being taught can be categorized in the following groups: Data Analysis, Programming, Big Data, Data Science and Machine Learning. *Data Analysis*: It can be defined [16] *"as a set of mathematical/statistical procedures, generally used as computer programs, embracing elementary but particularly multidimensional statistical techniques that require an iterative application in order to statistically process the data and extract information from the data set. This method involves the use of mathematical/statistical rules generally applicable and not subject dependent as procedures for the assessment of data and the acquisition of new information"*. With this definition in mind, some studies focused on analysing historical data and extracting knowledge from it, such as [2][9][11][12]. *Programming*: Constructing programs is recognized as a complex task, as mentioned by [17] over 30 years ago: *"All software construction involves essential tasks, the fashioning of the complex conceptual structures that compose the abstract software entity, and accidental tasks, the representation of these abstract entities in programming languages and the mapping of these onto machine languages within space and speed constraints"*. Nevertheless, it can present its intuition and rationale, in order to encourage early logical thinking, as [10] has been doing for high school students. *Big Data Engineering*: Parallel processing of large amounts of data has been disrupted by the iconic paper published by Google researchers about its now open source technology, MapReduce [18]: *"MapReduce is a programming model and an associated implementation for processing and generating large data sets"*. This is a key concept when talking about Data Engineering, which is also a discipline being taught as a foundation concept of Data Science Programs, as is being done by [15]. *Data Science*: According to [19], *"Data science updates the concept of data mining in the light of the availability of big data, that differ from data by their automatic generation through social networks, sensors and other data generating tools. In this sense, data science can be defined extending Giudici (2003), as an integrated process that consists of a series of activities that go from the definition of the objectives of the analysis, to the selection and processing of the data to be analysed, to the statistical modelling and summary such data and, finally, to the interpretation and evaluation of the obtained statistical measures'"*.In that sense, it comprises a more complete process, in which it processes large amounts of data in order to infer new knowledge for the business context. According to this view, some studies [14] offer a more complete program combining all previous concepts. *Machine Learning*: It is considered a subset of Data Science specialized in identifying patterns in data as described by [20] *"knowledge discovery process as the chain of accessing data from various sources, integrating and maintaining data in data warehouses, extracting patterns by machine learning methods"*. Some programs cover that important topic, including supervised, unsupervised methods and reinforced learning, as [8][13].

4) Target Audience: The last aspect, and probably the most important one, analysed in order to answer the research question RQ1, is to which target audience the data science content is being taught to, if to non-STEM or STEM only. On this topic we can observe different areas, from liberal arts, business and life sciences, that are being complemented with this kind of content. In [15], Journalist students use big data to understand social impact in a collaborative environment. [12] reports the use of data analysis to make social civic issues more tangible. Different areas such as: Anthropology and Sociology; Biology; Economics; Philosophy; Physics; Political Science; and Psychology were pointed in [8]. Besides humanities areas mentioned before, [14] presents evidences of data science being applied to to life science related courses, such as: Medical Statistics; Marine biology; Biostatistics; Genomics; Psychology; and Neuroscience. And lastly, [13] acknowledges a wide variation of courses being supported by data science studies: Physics; Biochemistry and physics; Environmental sciences; Earth sciences; Statistics; Mathematics; Biomedical engineering; Bioinformatics; Materials Architecture Management; and Social-Political Sciences.

B. RQ2) The WHY

The aspects analysed in order to answer this research question were the results achieved and how they were measured.

1) Results Achieved: Most of the success criteria adopted by the analyzed studies was the feedback of the students about the level of learning on the presented data science content [2][9][11][12][14][15]. In particular, we could mention [13] as an example *"The course has received so far two official evaluations by the students conducted on behalf of ETH Zurich. The general satisfaction 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"*. In two other cases [8][10], since data science programs were being offered for the first time, what was mainly measured were the number of consecutive offerings and the popularity that of the courses, for example in [8]: *"Over the first four semesters of the program, we have offered 13 sections of Introduction to Data Analytics, enrolling approximately 240 students in total. At the end of the program's second academic year (2017–2018), there were already about 100 declared DA majors among the first year, sophomore, and junior classes. Overall, 37% of our majors are women, and this percentage has been rising with each class year. At the end of the 2018–2019 academic year, our first year with graduating seniors, we anticipate that we will enroll approximately 130 total declared majors, and that we will graduate 27"*. Besides the success level, based on student satisfaction or courses popularity, some studies collected lessons learned and improvements to be incorporated to the programs, as cited by [13]: *"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"*.

2) Measurement Techniques: The achieved results mentioned in the previous section were measured in different ways. In some cases as a qualitative survey of students feedback, as in [11]: *"According to students feedback, the courses one and two went well. Both, the ADA as a subject, and the course structure were thanked. Most of all, the students appreciated the absence of a final exam"*. In some other cases, a more quantitative approach was made, even without the concern of being statistically validated, as in [13] and [9]. In comparison with cases that had this level of validation, as in [10]: *"We analyzed each construct using a repeated-measure ANOVA with a type 2 sum of squares, using time-of-survey (pre- or post-survey) as the within-subjects factor and gender, prior familiarity with Python, and the trimester they took the course in as between-subject factors. Post-hoc testing was done using a t-test (paired when the independent variable was time-of-survey), with the Bonferroni correction to address family-wise error rate"*. Lastly, for the course [8] that had popularity as its main success criterion, they simply performed an accrual offering after offering of the program.

C. RQ3) The HOW

1) Proprietary Methodologies: Most experiences apply proprietary methodologies [2][8][9][10][11][12][13][14][15] that are in some extent a variation of ACM Data Science Curricula [1], which originally was designed for technical undergraduate educational formation. As an example of a proprietary methodology we can mention [11]: *"To achieve a good learning atmosphere leading to effective learning, we use pedagogic methods, such as, collaborative learning, pair programming, and learning by doing. During the day, we are aloud to find something we haven't even planned. This approach draws us near to the ideology where data scientist is thought as 'part analyst, part artist'"*.

2) ACM Data Science Curricula: The ACM Data Science Curricula [1], comprises the following knowledge areas:

- Computing Fundamentals, including: Programming, Data Structures, Algorithms, and Software Engineering
- Data Acquirement and Governance
- Data Management, Storage, and Retrieval
- Data Privacy, Security, and Integrity
- Machine Learning
- Data Mining
- Big Data, including: Complexity, Distributed Systems, Parallel Computing, and High Performance Computing
- Analysis and Presentation, including: Human-Computer Interaction and Visualization
- Professionalism

V. LIMITATIONS AND THREATS TO VALIDITY

As a process to apply the techniques of a formal SLR, it was an interesting experience. Even if it were performed by applying a strict methodology, it relies only on technical research databases, ACM and IEEE. Some domain specific databases were used as reference, however no relevant studies were found. In that sense, this could be a threat to the validity of this study.

VI. CONCLUSION

In conclusion, teaching data science to different areas of knowledge other than the technical ones (non-STEM) is already collecting its fruits, and still has room for further growth. It is interesting to observe how it is being applied to different levels of students, from primary school and high school up to undergraduate and post-graduate courses. Another interesting point is that it is being offered to different target audiences, from economics, to medicine, social studies and so on. In terms of benefits, it is possible to see that the level of learning and interest on the subject are aspects that have being monitored by the providers of such courses. Not only that, but also the lessons learned in terms of how the teaching methodology could improve in order to present this kind of content to non-STEM students. Finally, the technique used to measure those results varies from practitioner to practitioner, ranging from no measurement at all up to statistically validated quantitative research.

VII. ACKNOWLEDGMENTS

This article was produced as the final activity of the first class, systematic literature review, of the doctorate program at CESAR School. Hence, we would like to thank the course professors, Cesar Franca and Paula Carvalho, as well as the course coordinator, Felipe Ferraz.

REFERENCES

[1] A. Clear, A. S. Parrish, J. Impagliazzo, and M. Zhang, “Computing Curricula 2020: Introduction and Community Engagement,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 653–654, event-place: Minneapolis, MN, USA. [Online]. Available: https://doi.org/10.1145/3287324.3287517

[2] J. Engel, “Statistical literacy for active citizenship: A call for data science education,” Statistics Education Research Journal, vol. 16, no. 1, 2017, pp. 44–49.

[3] B. G. Buchanan, “A (very) brief history of artificial intelligence,” Ai Magazine, vol. 26, no. 4, 2005, pp. 53–53.

[4] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” 2004.

[5] I. Mierswa, “May 9, 2017,” library Catalog: ingomierswa.com. [Online]. Available: https://ingomierswa.com/2017/05/09/

[6] A. Ng, “What Artificial Intelligence Can and Can’t Do Right Now,” Harvard Business Review, Nov. 2016, section: Analytics. [Online]. Available: https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

[7] J. Von Neumann and O. Morgenstern, Theory of games and economic behavior (commemorative edition). Princeton university press, 2007.

[8] J. Havill, “Embracing the Liberal Arts in an Interdisciplinary Data Analytics Program,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. Minneapolis, MN, USA: Association for Computing Machinery, Feb. 2019, pp. 9–14. [Online]. Available: https://doi.org/10.1145/3287324.3287436

[9] S. Srikant and V. Aggarwal, “Introducing Data Science to School Kids,” in Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, ser. SIGCSE ’17. Seattle, Washington, USA: Association for Computing Machinery, Mar. 2017, pp. 561–566. [Online]. Available: https://doi.org/10.1145/3017680.3017717

[10] “Pythons and Martians and Finches, Oh My! Lessons Learned from a Mandatory 8th Grade Python Class | Proceedings of the 51st ACM Technical Symposium on Computer Science Education.” [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3328778.3366906

[11] M. Marttila-Kontio, M. Kontio, and V. Hotti, “Advanced data analytics education for students and companies,” in Proceedings of the 2014 conference on Innovation & technology in computer science education, ser. ITiCSE ’14. Uppsala, Sweden: Association for Computing Machinery, Jun. 2014, pp. 249–254. [Online]. Available: https://doi.org/10.1145/2591708.2591746

[12] S. J. Van Wart, “Computer Science Meets Social Studies: Embedding CS in the Study of Locally Grounded Civic Issues,” in Proceedings of the eleventh annual International Conference on International Computing Education Research, ser. ICER ’15. Omaha, Nebraska, USA: Association for Computing Machinery, Aug. 2015, pp. 281–282. [Online]. Available: https://doi.org/10.1145/2787622.2787751

[13] E. Pournaras, “Cross-disciplinary higher education of data science - beyond the computer science student,” Data Sci., vol. 1, 2017, pp. 101–117.

[14] S. Kross and P. J. Guo, “Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges,” in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, May 2019, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300493

[15] Y. Gil, “Teaching Parallelism without Programming: A Data Science Curriculum for Non-CS Students,” in 2014 Workshop on Education for High Performance Computing, Nov. 2014, pp. 42–48.

[16] V. Vitali, “Formal methods for the analysis of archaeological data: Data analysis vs expert systems,” Computer Applications and Quantitative Methods in Archaeology, 1990, pp. 207–209.

[17] F. P. Brooks, “No silver bullet essence and accidents of software engineering,” Computer, vol. 20, no. 4, Apr. 1987, p. 10–19. [Online]. Available: https://doi.org/10.1109/MC.1987.1663532

[18] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, Jan. 2008, pp. 107–113. [Online]. Available: https://dl.acm.org/doi/10.1145/1327452.1327492

[19] P. Giudici, “Financial data science,” Statistics and Probability Letters, vol. 136, may 2018, pp. 160–164.

[20] G. Kauermann and T. Seidl, “Data Science: a proposal for a curriculum,” International Journal of Data Science and Analytics, vol. 6, no. 3, Nov. 2018, pp. 195–199. [Online]. Available: https://doi.org/10.1007/s41060-018-0113-2