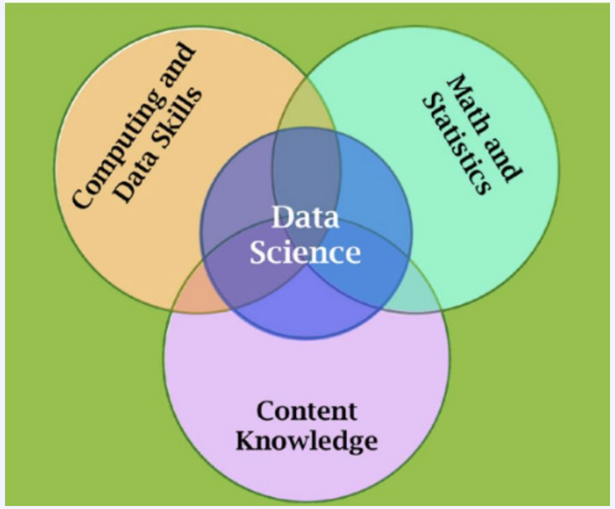
Conference: The Ninth International Conference on Data Analytics - 2020

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



Center of Advanced Studies and Systems of Recife

Authors: Luiz Barboza - lcbj@cesar.school
Erico S Teixeira - est@cesar.school

About the Authors

Luiz Barboza

PhD student in Software Engineering at Recife Center for Advanced Studies and Systems (CESAR), Recife, Brazil. MSc in Software Engineering from the Technological Research Institute (IPT), São Paulo, Brazil.

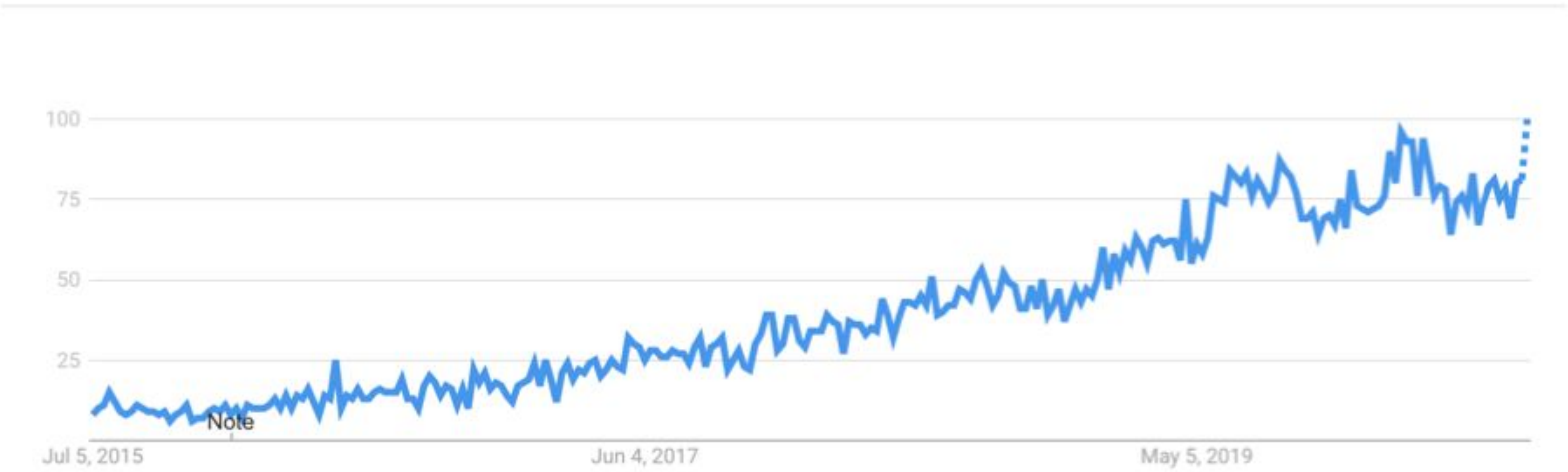
Erico S Teixeira

PhD in Computational Chemistry at Federal University of Pernambuco (UFPE), Recife, Brazil. Phd Sandwich at University of Florida, Gainesville, USA. Msc. in Computer Science at Federal University of Pernambuco (UFPE), Recife, Brazil.

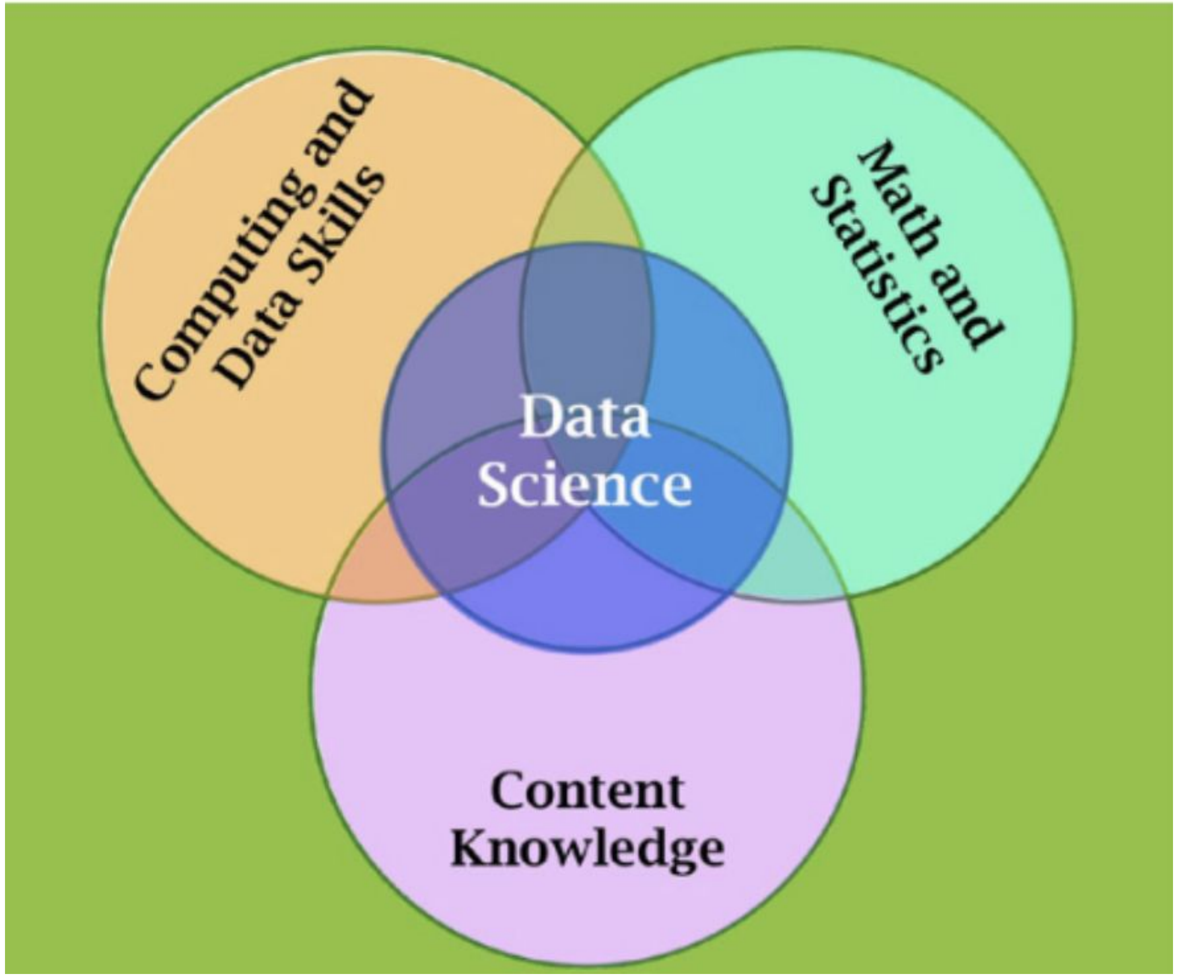
Purpose of the Paper

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.

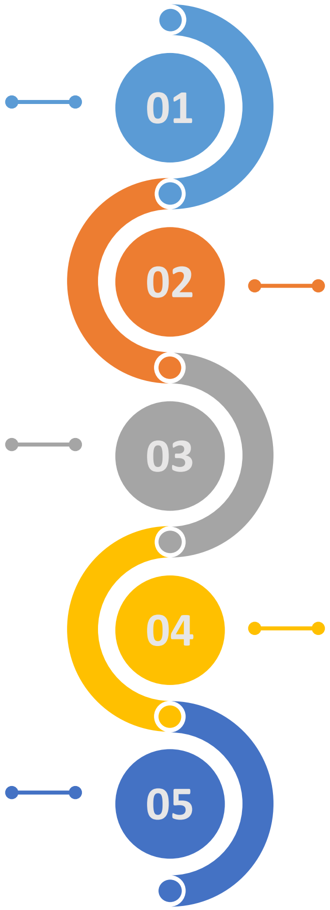
Data Science interest in the last 5 years



Data Science Tripod



Literature Review Methodology



01 Development of the Protocol

02 Identification of the criteria for inclusion and exclusion

03 Search for relevant studies

04 Data extraction

05 Synthesis

Research Questions

- **RQ1 (IF)**

  How is knowledge in data science being taught to non-technical target audiences?

- **RQ2 (WHY)**

  What are the learning improvements that these students are experiencing with the use of data science in different areas of their studies?

- **RQ3 (HOW)**

  What was the method used in teaching data science?

C.E.S.A.R school

Inclusion/Exclusion Criteria

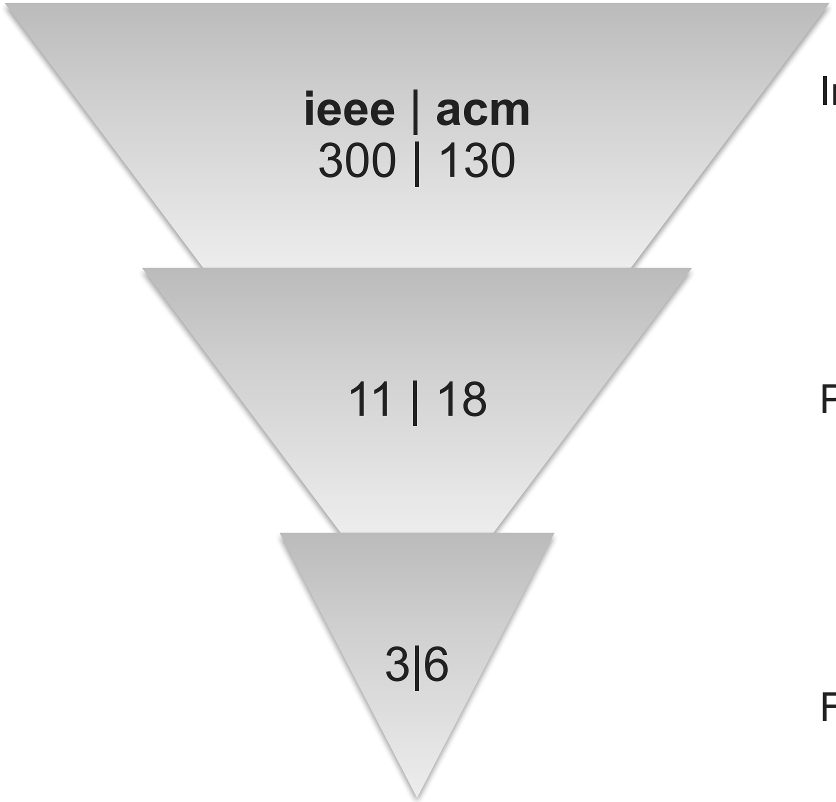
The criteria for the inclusion of the article are:

- Written in the English language.
- Focusing on data science studies of non-technical target audiences.

The criteria for the exclusion of the article are:

- Focusing on the use of technology to improve the learning process.
- Targeting different groups other than non-STEM.
- Educational improvements achieved other than data science introduction.
- Data Science application without the explicit goal of educational purposes.

Researched Databases



Initial set of papers

Passed inclusion criteria

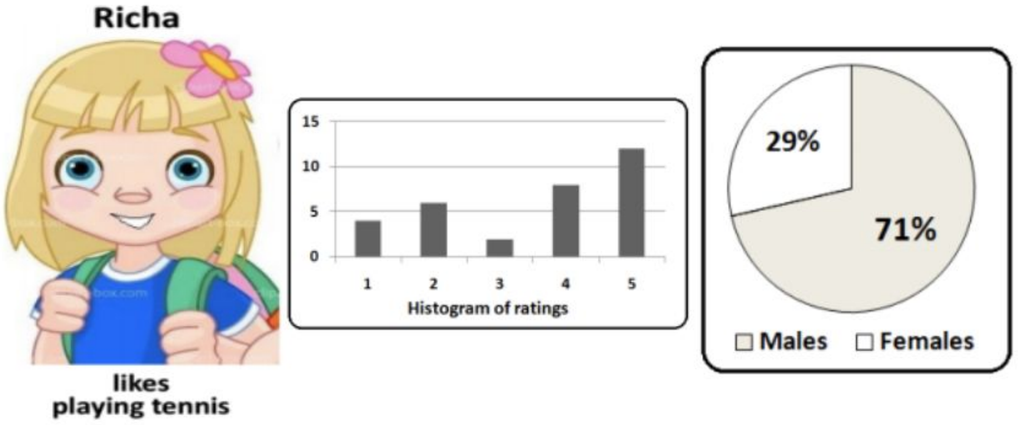
Final list of papers

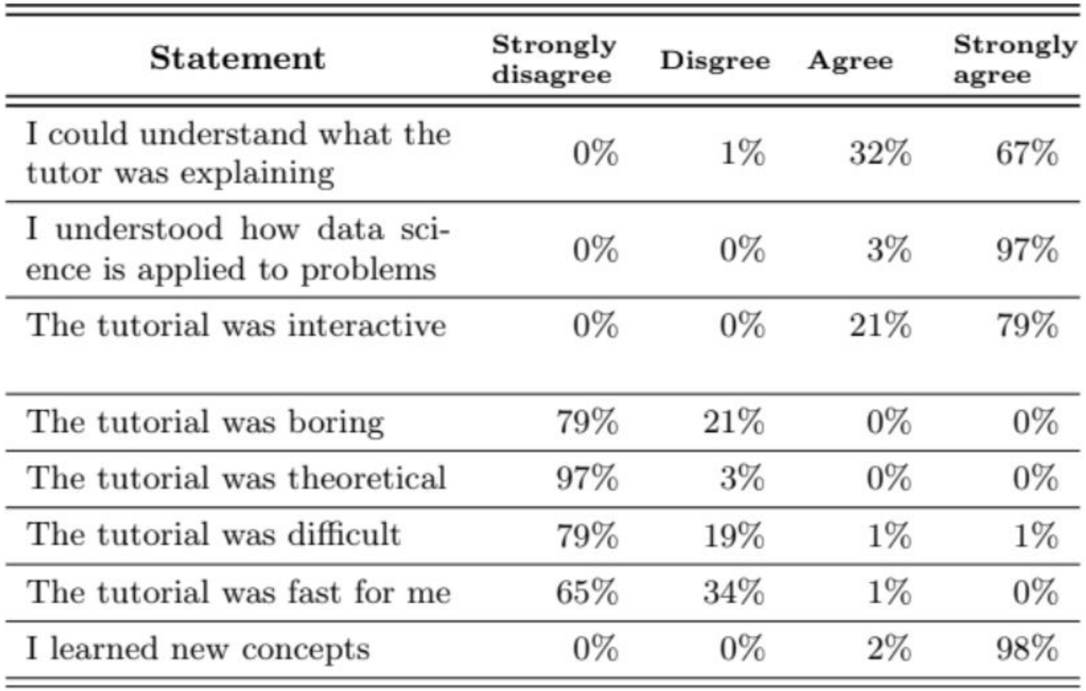
Results: RQ1) The IF School Level

Junior High School

As seen in [10], 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”*

[10] S. Srikant and V. Aggarwal. Introducing Data Science to School Kids.





Statement	Strongly disagree	Disgree	Agree	Strongly agree
I could understand what the tutor was explaining	0%	1%	32%	67%
I understood how data science is applied to problems	0%	0%	3%	97%
The tutorial was interactive	0%	0%	21%	79%
The tutorial was boring	79%	21%	0%	0%
The tutorial was theoretical	97%	3%	0%	0%
The tutorial was difficult	79%	19%	1%	1%
The tutorial was fast for me	65%	34%	1%	0%
I learned new concepts	0%	0%	2%	98%

Results: RQ1) The IF School Level

[11] A. Nanavati, A. Owens, M. Stehlik. Pythons and Martians and Finches, Oh My! Lessons Learned from a Mandatory 8th Grade Python Class

Senior High School

A deeper approach can be observed in [11], 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

Results: RQ1) The IF School Level

College/University

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 [12], [13], [9], [8], [2], [14].

- *[8] E. Pournaras. Cross-disciplinary higher education of data science - beyond the computer science student.*
- *[9] J. Havill. Embracing the Liberal Arts in an Interdisciplinary Data Analytics Program*
- *[12] M. Marttila-Kontio, M. Kontio, and V. Hotti. Advanced data analytics education for students and companies.*
- *[13] S. J. Van Wart. Computer Science Meets Social Studies: Embedding CS in the Study of Locally Grounded Civic Issues.*
- *[14] S. Kross and P. J. Guo. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges.*
- *[2] J. Engel. Statistical Literacy for Active Citizenship: A Call for Data Science Education.*

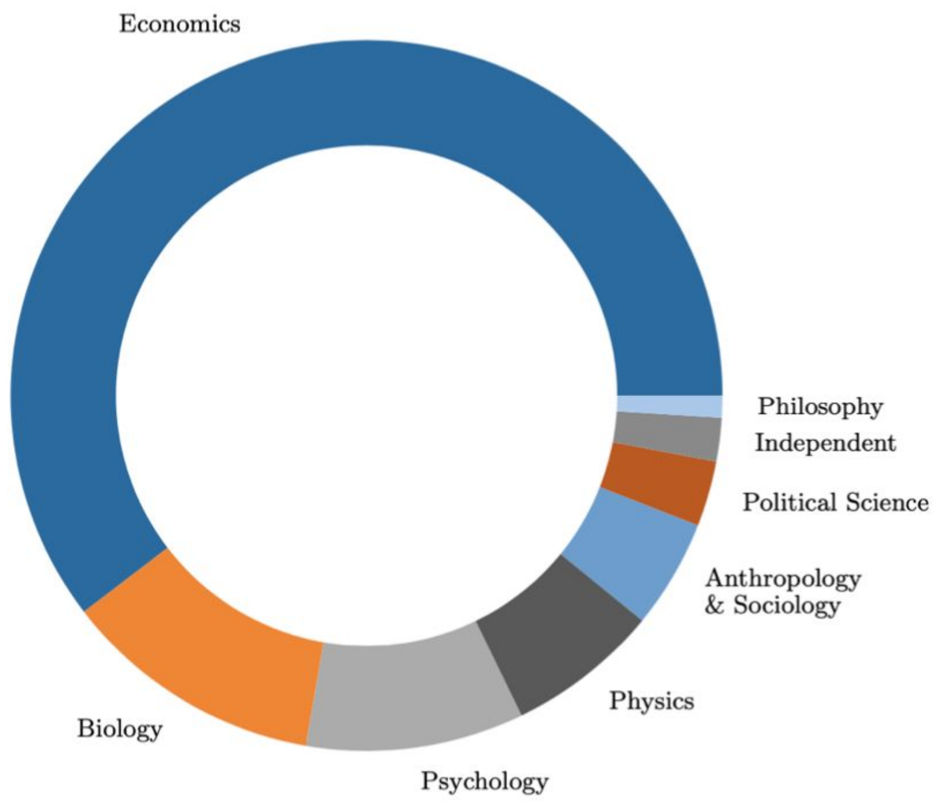
C.E.S.A.R school

Results: RQ1) The IF School Level

[9] J. Havill. Embracing the Liberal Arts in an Interdisciplinary Data Analytics Program

College/University

[9]“*Because no data exists in a vacuum, each DA 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*



Results: RQ1) The IF Location

As Most of the studies considered were based in the following states of the **USA** [14]: California [10], [15], [13], Washington DC [11] and Ohio [9].

The remaining of the studies, were performed in **Europe**, in the following countries: Germany [2], Switzerland [8] and Finland [12].”

[12] M. Marttila-Kontio, M. Kontio, and V. Hotti. Advanced data analytics education for students and companies.

Adjective: **tietävä**
in the know (informed, aware)

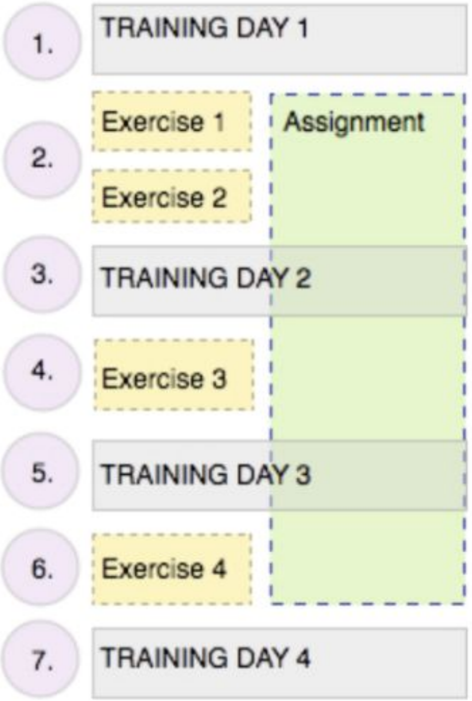
The Tietävä project is funded by European Social Fund via the Centre for Economic Development, Transport and the Environment.

Results: RQ1) The IF Concepts Taught

[12] M. Marttila-Kontio, M. Kontio, and V. Hotti. Advanced data analytics education for students and companies.

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 [10], [12], [13], [2]



Results: RQ1) The IF Concepts Taught

[11] A. Nanavati, A. Owens, M. Stehlik. Pythons and Martians and Finches, Oh My! Lessons Learned from a Mandatory 8th Grade Python Class

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 [11] has been doing for high school students.

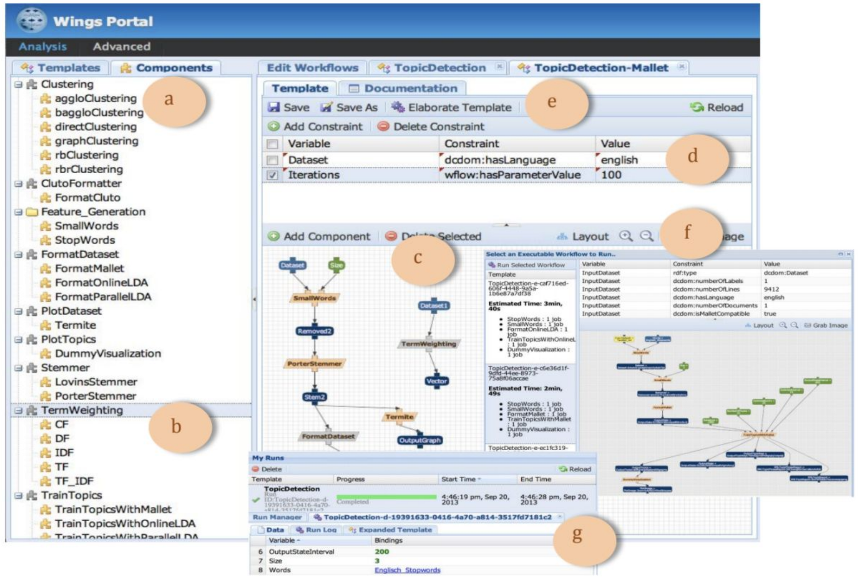
Curriculum Design - Finch Robot

Results: RQ1) The IF Concepts Taught

[15] Y. Gil. Teaching Parallelism without Programming: A Data Science Curriculum for Non-CS 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].

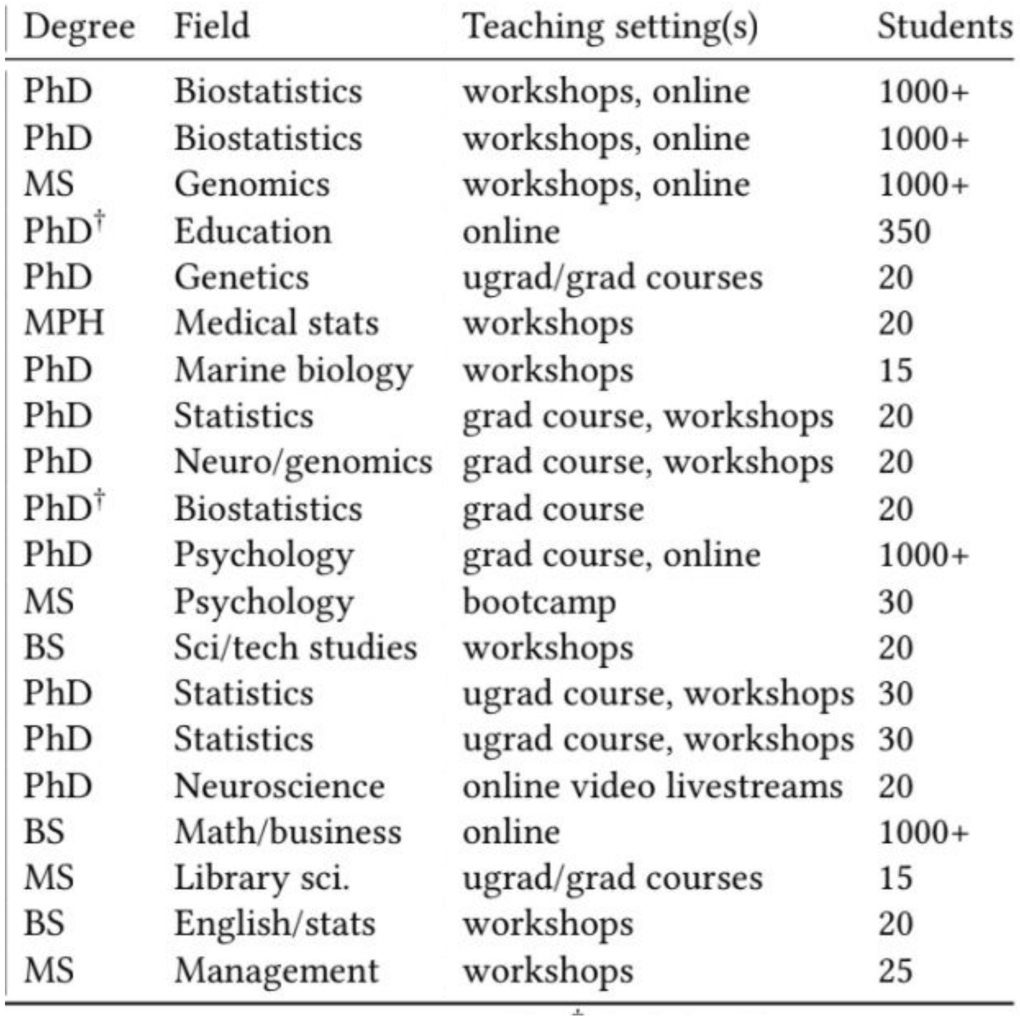


Results: RQ1) The IF Concepts Taught

[14] S. Kross and P. J. Guo. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges.

Data Science

[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 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.



Degree	Field	Teaching setting(s)	Students
PhD	Biostatistics	workshops, online	1000+
PhD	Biostatistics	workshops, online	1000+
MS	Genomics	workshops, online	1000+
PhD†	Education	online	350
PhD	Genetics	ugrad/grad courses	20
MPH	Medical stats	workshops	20
PhD	Marine biology	workshops	15
PhD	Statistics	grad course, workshops	20
PhD	Neuro/genomics	grad course, workshops	20
PhD†	Biostatistics	grad course	20
PhD	Psychology	grad course, online	1000+
MS	Psychology	bootcamp	30
BS	Sci/tech studies	workshops	20
PhD	Statistics	ugrad course, workshops	30
PhD	Statistics	ugrad course, workshops	30
PhD	Neuroscience	online video livestreams	20
BS	Math/business	online	1000+
MS	Library sci.	ugrad/grad courses	15
BS	English/stats	workshops	20
MS	Management	workshops	25

Results: RQ1) The IF Concepts Taught

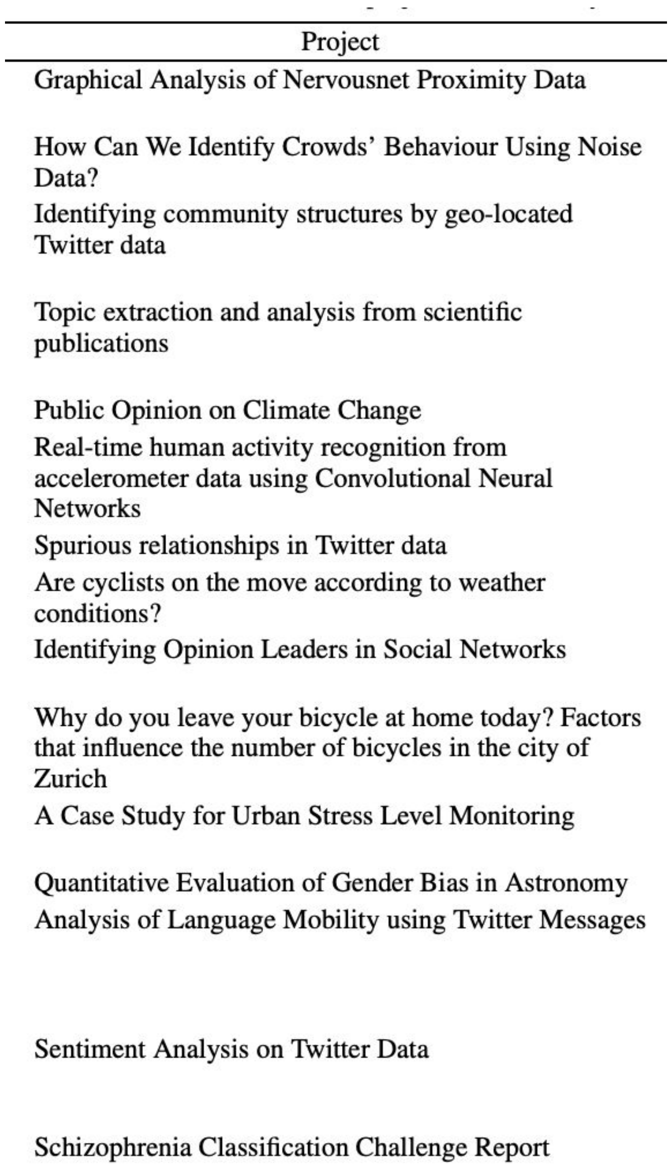
[8] E. Pournaras. Cross-disciplinary higher education of data science - beyond the computer science student.

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 [9], [8].

.

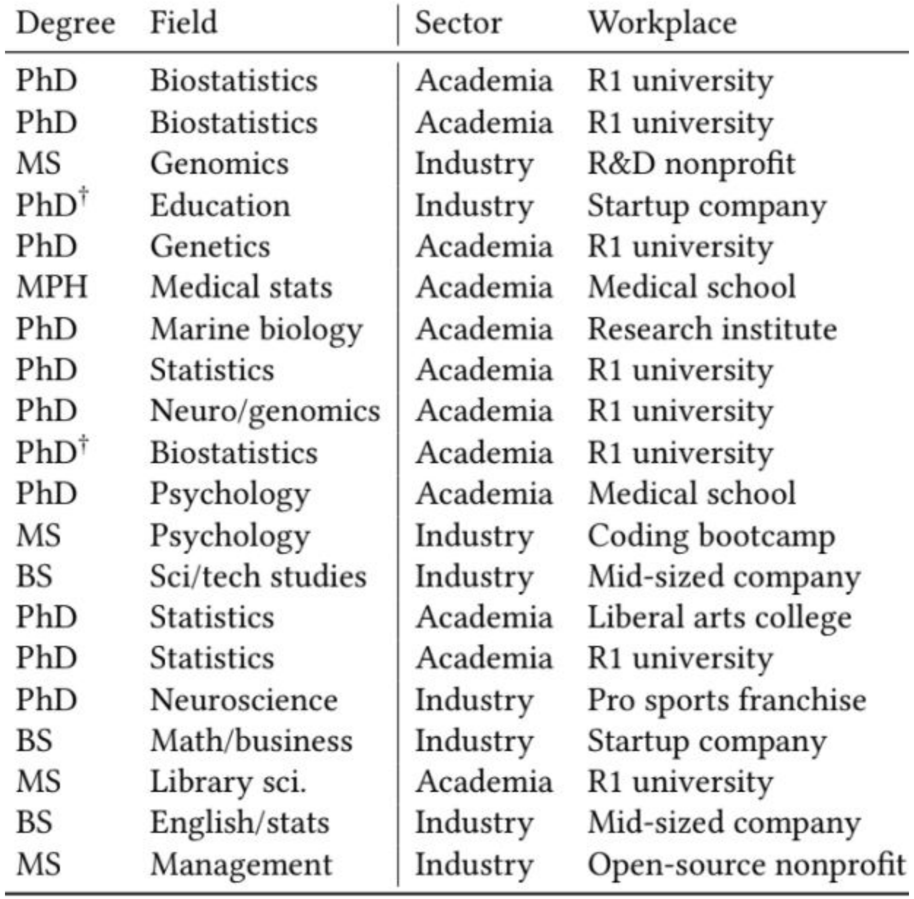


Project
Graphical Analysis of Nervousnet Proximity Data
How Can We Identify Crowds' Behaviour Using Noise Data?
Identifying community structures by geo-located Twitter data
Topic extraction and analysis from scientific publications
Public Opinion on Climate Change
Real-time human activity recognition from accelerometer data using Convolutional Neural Networks
Spurious relationships in Twitter data
Are cyclists on the move according to weather conditions?
Identifying Opinion Leaders in Social Networks
Why do you leave your bicycle at home today? Factors that influence the number of bicycles in the city of Zurich
A Case Study for Urban Stress Level Monitoring
Quantitative Evaluation of Gender Bias in Astronomy
Analysis of Language Mobility using Twitter Messages
Sentiment Analysis on Twitter Data
Schizophrenia Classification Challenge Report

Results: RQ1) The IF Target Audience

[14] S. Kross and P. J. Guo. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges.

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.



Degree	Field	Sector	Workplace
PhD	Biostatistics	Academia	R1 university
PhD	Biostatistics	Academia	R1 university
MS	Genomics	Industry	R&D nonprofit
PhD†	Education	Industry	Startup company
PhD	Genetics	Academia	R1 university
MPH	Medical stats	Academia	Medical school
PhD	Marine biology	Academia	Research institute
PhD	Statistics	Academia	R1 university
PhD	Neuro/genomics	Academia	R1 university
PhD†	Biostatistics	Academia	R1 university
PhD	Psychology	Academia	Medical school
MS	Psychology	Industry	Coding bootcamp
BS	Sci/tech studies	Industry	Mid-sized company
PhD	Statistics	Academia	Liberal arts college
PhD	Statistics	Academia	R1 university
PhD	Neuroscience	Industry	Pro sports franchise
BS	Math/business	Industry	Startup company
MS	Library sci.	Academia	R1 university
BS	English/stats	Industry	Mid-sized company
MS	Management	Industry	Open-source nonprofit

Results: RQ2) The WHY
Results Achieved

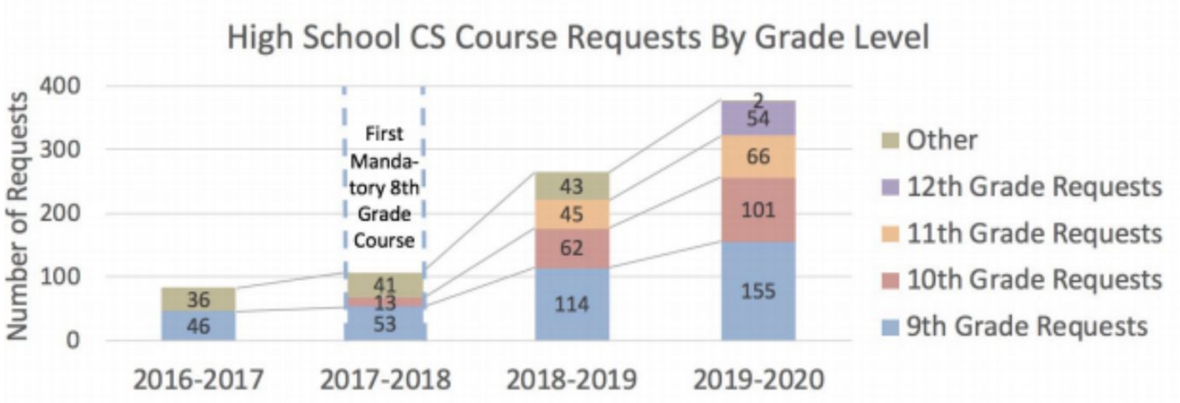
Most of the success criteria adopted by the analyzed studies was the **feedback of the students** about the level of learning on the presented data science content [10], [15], [12], [13], [2], [14].

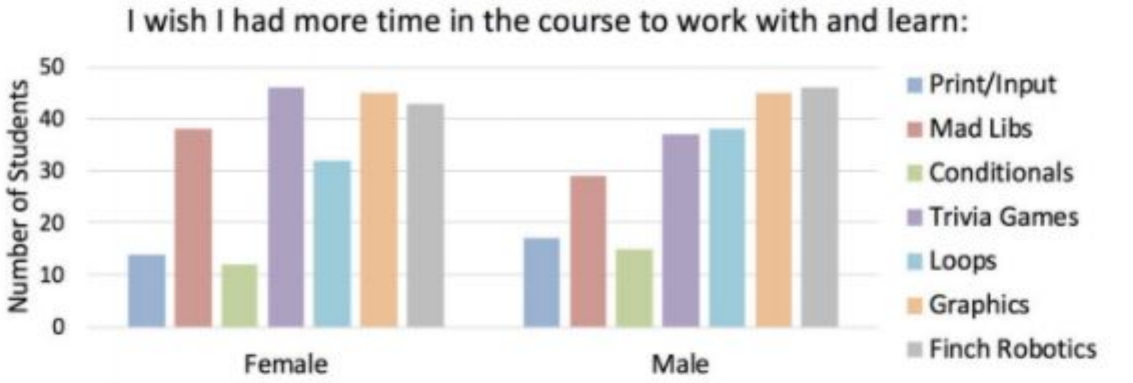
In particular, we could mention [8] as an example *"The course has received so far two official evaluations by the students conducted on behalf of ETH Zurich. The general satisfaction has been 4.4/5.0 and the lecturers' evaluation 4.5/5.0 on the following aspects: understandable and clear explanation of the subject, learning goals, lecture significance, motivation to active participation, and material made available"*

Results: RQ2) The WHY Measurement Techniques

In some cases as a **qualitative survey of students feedback**, in others, a more **quantitative** approach was made, even without the concern of being statistically validated, as in [8] and [10]. In comparison with cases that had this level of validation, as in [11]: ”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.”.

[11] A. Nanavati, A. Owens, M. Stehlik. Pythons and Martians and Finches, Oh My! Lessons Learned from a Mandatory 8th Grade Python Class





Results: RQ3) The HOW

Proprietary Methodologies

Most experiences proprietary methodologies [10], [15], [12], [13], [9], [8], [2], [14], [11] 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 [12]: *"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'"*.

Results: RQ3) The HOW

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

Conclusions

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 postgraduate 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.