DOI: 10.15649/2346030X.2586
Aibi research, management and engineering journal. Volume 9, Number 3, Pag 49-58, September – December 2021, ISSN 2346-030X

Theoretical foundation for the creation of an academic program in engineering and data science: a bibliometric application.

Frederick Andrés Mendoza-Lozano[1], Jose Wilmar Quintero-Peña[2], Oscar Leonardo Acevedo-Pabón[3], Jose Félix García-Rodriguez[4]
[1,2,3]*Institución Universitaria Politécnico Grancolombiano, Bogota - Colombia*, [4]*Universidad Veracruzana, Veracruz - México*
ORCID: [1]*0000-0001-5087-4476*, [2]*0000-0002-6172-0453*, [3]*0000-0002-1270-4166*, [4]*0000-0002-7319-1472*

Received: June 22, 2021.
Accepted: August 18, 2021.
Published: September 01, 2021.

***Abstract*—** The aim was to define a theoretical approach to data science, which includes object of study and methods, as a previous step for the curricular design of an academic program. The text begins with a review of the literature regarding the evolution of the concept of data and the epistemological foundations of statistics and data analysis through algorithms. The bibliometry of the most relevant scientific production continues, using the thematic characterization approach, using keywords taken from works indexed in SCOPUS. It was found that most of the relevant keywords and themes refer to the methods of data modeling with algorithms and the management of technology for the administration of large databases. The productivity of the analysis of data derived from textual, multimedia, and web information was characterized. The themes related to business applications aimed at knowledge management and business intelligence were also revealed. The concept of data, as an object of study, is extended thanks to the scope of data analysis with algorithms; This method is combined with the classical statistical approach, which provides formal models of better interpretation. It was concluded that the field of application of the new data science is quite broad, particularly when this science is used in interdisciplinary contexts. The above justifies the curricular design of an academic program focused on this subject.

Keywords: Data science, Bibliometry, Data mining, Big data, Classical statistics, Machine learning, Curriculum.

*Corresponding author.
Email: famendoza@poligran.edu.co (Frederick Andrés Mendoza-Lozano).
Peer reviewing is a responsability of the Universidad de Santander.

How to cite this article: F. A. Mendoza-Lozano, J. W. Quintero-Peña, O. L. Acevedo-Pabón y J. F. García-Rodriguez, "Theoretical foundation for the creation of an academic program in engineering and data science: a bibliometric application", *Aibi research, management and engineering journal*, vol. 9, no. 3, pp. 49-58, 2021, doi: 10.15649/2346030X.2586

I. INTRODUCTION

In 2019, the Institución Universitaria Politécnico Grancolombiano (IUPG) decided to create a new undergraduate program in data science. Once it obtains qualified registration from the Ministry of National Education, it will be one of the first undergraduate programs in this field of study, given that a group of higher education institutions offers similar programs at the diploma and postgraduate levels.

Defining a new profession in the country was a challenge, given the need to delimit an object of study that combines curricular contents of traditional programs such as statistics and systems engineering. According to Decree 1330 of 2019, which regulates the quality conditions of higher education programs, in developing the curricular contents, the theoretical foundation on which the object of study is supported, and the epistemological aspects of the field of study must be studied and exposed. Achieving this became the most notable challenge of all the documentation work required to complete the application's presentation for the qualified registration.

A literature review was conducted to structure the theoretical foundation on research perspectives of interdisciplinary scope in data science. Scientific papers published in indexed journals were approached, and bibliometric analysis was implemented. It was possible to delimit a field of study and the methods used to analyze data with these inputs. Thanks to a historical approach, the evolution from classical statistics to data science was illustrated. This work constitutes the cornerstone of the argument that defends the pertinence of creating a new profession with a specific denomination whose profile has a broad field of work.

This article presents the research results that was carried out to provide a theoretical foundation for the data science and engineering program. Here we proceed as follows: we begin with a reflection on the epistemological evolution of data science. Then we propose some bibliometric tools to describe the dynamics of research on quantitative analysis, taking as input the scientific production indexed in SCOPUS. Finally, some conclusions and their implications for the curricular implementation of this new program are presented.

II. THEORETICAL BACKGROUND

a. Statistics to data science

In the 1960s, Tukey [1] proposed a different approach to statistics as a discipline. In essence, he promoted focusing the object of study on data analysis rather than on statistical modeling theory. His position was that the shortcomings of statistical theory and the demand for interdisciplinary development for data analysis would bring extraordinary developments in the short term. For its time, this proposal was controversial and disruptive. However, the evolution of disciplines with a strong quantitative base and, especially, the contributions of computer science to data analysis have shown that his vision was generally correct: "After all, I have come to feel that my central interest is data analysis, in which I include among other things: procedures for analyzing data, techniques for interpreting the results of those procedures, ways of planning data collection. Thus, its analysis is made easier and greater precision is given to the whole machinery and to the statistical (mathematical) results that are applied to data analysis" [1, pp 2].

The transition from statistics to data science is a chapter in the history of the evolution of science that goes from simple and stylized models to understanding reality to theoretical frameworks that account for phenomena of increasing complexity [2]; generally, this type of phenomena is related to biological processes [3] or to social systems that evolve in a non-deterministic way, that is, according to individual decisions that are not predictable [4]. In that sense, the creation of data science is the result of the combination of the classical approach with the algorithmic approach. The former assumes that data are generated according to a stochastic model; the latter does not focus on theorizing about the abstract data-generating mechanism but on extracting relevant information about the relationships between variables and observations, as well as on developing predictive tools. In addition, this new science found its own object of study derived from the evolution of the concept of data.

The roots of the Cartesian scientific method created a reliable mechanism to investigate reality through a procedure that guarantees reliable results, independent of the researcher. In that sense, Descartes [5] put forward the idea of the separability of parts in order to study the mechanisms of reality through the analysis of simple and isolated components. Later on, positivist epistemology magnified science results, claiming that they were the only ones that had validity.

Scientific knowledge is characterized by being rational, systematic, exact, verifiable, and, at the same time, fallible in that it is not dogmatic. The development of science is a constant tension between formal science and factual science [6]. Much scientific work is concerned with structuring models that aspire to represent reality, but their validity or refutation depends on the empirical verification that involves designing a controlled experiment in a laboratory or making measurements directly on phenomena that cannot be reproduced artificially.

As a consequence of experimental scientific work, a body of knowledge associated with measurement errors has emerged, which is present in at least two ways. On the one hand, one must deal with the issue of precision, which is associated with the technologies for experimental design [7]. On the other hand, there is the question of scientific validity based on empirical evidence. That is the extrapolation of the scientific finding to a whole reality when it is based on the results of the observation of an isolated part. Faced with these two objections, scientific work developed a set of theories associated with data collection and analysis. Thus, a growing group of statistical instruments aligned with the compass of scientific thought was consolidated: stylized, simple, and elegant models created under a rational procedure with the rigor of mathematical language.

Statistical tools associated with the measurement and interpretation of numerical data leveraged an extraordinary advance in science. Although the birth of statistics dates back to a recent period in the history of mankind-about two centuries ago, as a response to the need to analyze census data [8]-almost all modern disciplines use and develop core concepts based on statistics.

As indicated above, one of the most relevant problems of scientific work is the possibility of incurring errors. This issue derives from the impossibility of having absolutely precise measurements and is closely related to the simplifications inherent to scientific research. Empirical verifications must always face the effect of the inaccuracies of the measuring instruments, the omitted variables in any quantitative study, and

an analytical, theoretical framework that simplifies as much as possible, at the cost of loss of precision [9]. Moreover, one cannot disregard that most of the phenomena that science deals with are non-deterministic. Even when theories can be tested in the laboratory, measurement data are not always the same since they change according to conditions that cannot be controlled during experimentation and reflect variations over time; to a large extent, science and statistical tools were designed to be able to account for these variations.

Due to their indeterminate nature, many phenomena of reality had to be approached with the rigor that guaranteed that science could advance with high levels of certainty, even if it were impossible to get rid of errors. This need for certainty motivated the development of a complete theory of probabilities, with which empirical verification consolidated a language of its own. Thus, based on the models of formal science, scenarios that contradict the theory are discarded in favor of alternatives that favor it. In short, scientific discourse proposes that something is valid because it is improbable to be otherwise. This involves two stages and two tools: in a comprehensive stage, with the instruments of statistical theory, the probable causes of variations in measurements are analyzed; in a prospective stage, with instruments designed based on the calculation of probabilities, future scenarios are investigated in forecasting exercises.

On the other hand, the scientific method found space in the social sciences, where it has developed extensively up to the present day. This situation is problematic insofar as the transfer of the instruments and procedures of the basic sciences is not always adequately adjusted to the complexity of social phenomena. Nevertheless, having taken all the necessary precautions, specific fields of application have been consolidated and have survived to the present day. Thus, today we have a quantitative theory of human behavior, which is tested utilizing scientific experimentation, and its work is usually included in the area known as psychometrics. Another example is the study of economic phenomena through econometrics.

Like classical or conventional science, the statistical theory has been criticized for its limitations in accounting for reality. The problem lies in the oversimplification of analytical models in contrast to the diversity and complexity of the phenomena being studied. This simplicity gives the researcher greater capacity to explain past phenomena but limits its predictive capacity [10]. In statistical theory, prevention manifests itself when faced with the lack of adjustment of real data to the ideal models in which the purest statisticians work. In this form of research, the model's specificity, its asymptotic properties, and the development of the mathematical architecture prevail. This is not free of discussion since scientific conclusions do not always explain causal relationships, nor are there fallacious interpretations in the statistics used to validate hypotheses [11]-[14].

The tendency to focus on the model assumed to be the "data generator" rather than on the actual data is very pronounced. This is why Breiman [9] describes it as a "culture" among pure statisticians, which extends to interdisciplinary applications in econometrics [15]-[17], sociology [18]-[21], psychology [22], [23] or the biological sciences [24], [25]. Alternatively, the same author postulates that there is another "culture" of an interdisciplinary nature based on data analysis with algorithms. This approach makes it possible to discover, with greater predictive accuracy, the behavior of a variable of interest (dependent) as a function of a set of explanatory variables (independent) through a computational exercise, complex, precise, and susceptible to be optimized through interactions, although much more difficult to interpret, with respect to the instruments of the "culture of data-generating models".

New approaches to data analysis show that, in practice, statistics are insufficient to address all the problems involved in data processing, especially when dealing with a large set of variables with explanatory power [26]. Thus, basic training across all professions can no longer be defined in terms of statistical literacy but in terms of data literacy [27]. Although the development of statistics seems to be framed in the construction of interdisciplinary scientific theory, the proposal to create an undergraduate program argues that there is a data science that delimits an object of study and a profession in which classical methods for data analysis converge with recent computational tools [28].

To understand this, it is necessary to yield to the necessary integration of knowledge that blurs disciplinary boundaries, to understand the social demands of experts in the specific work of data with multiple applications, and to take into account that the concept of "data" as an object of study has changed over time thanks to the development of computational science. The disciplinary influence of modern data analysis techniques is reflected, for example, in economic science, where machine learning tools play a crucial role in solving prediction questions. Although machine learning methods are important, they are insufficient because of econometric challenges such as causality problems, identification of counterfactuals, and economic behavior [29].

In other words, knowledge of the theory plays a fundamental role that, when combined with data modeling tools with algorithms, generates a better understanding of economic problems that may involve the use of large volumes of information. Large-scale data has created an opportunity for economic measurement, even in developing countries. For example, Blumenstockm et al. [30] measure poverty at the individual level using smartphone information. On the other hand, [31] calculate the dynamics of employment and consumption in China in real time. The above are just some of the challenges facing economic measurement. Therefore, it can be stated that data science methods could complement the answers to social questions. In other words, the interdisciplinary application of statistics suggests that it should not be inscribed in a single scientific field; rather, it is a set of theories and instruments of great importance, transversal to many fields of study.

While classical statistics developed a first approach, in which data are assumed to be generated according to a stochastic model to explain numerical data, the possibility of representing other types of information in terms of mathematical expressions significantly broadens the concept of "data". Nowadays, data analysis encompasses processing images, text, audio and geographic information, or a combination of the above, through the massive extraction of information from the web [32]. Thanks to the development of computers, a poem can be both an artistic work valued with the intermediation of emotions and the human spirit or a set of data represented in matrix form for computational analysis. Similar reasoning can be made with respect to images or audio. Moreover, with the technologies for storing large volumes of data, the problem of analysis changed from a static nature, in which data are collected in a specific period for a study delimited in time, to a dynamic one, in which data that is collected in continuous time must be analyzed [33].

The broadening of the concept of data and the interest in studying complex phenomena, which involve both a wide set of variables and a generous number of observations, have resulted in a science with a broad and robust body of knowledge that transcends statistical modeling. This new science has its own object of study based on transdisciplinarity, and its procedure is inscribed in a remarkable uncertainty. This new science is data science. Data science is concerned with understanding and predicting phenomena studied with computational tools; it broadens

the concept of data (relationships between a dependent variable and a large set of explanatory variables) and makes the factual component prevail over the formal one.

Focusing on explaining data behavior tools, they are centered on validating explanatory methods: cross-validation, scoring of proposed models, optimization of model accuracy scoring indicators, and validation of the best modeling strategy in contrasting results after running them all. The procedural steps of this new science can be summarized in three [8]: 1) collection of information, 2) definition of a set of instruments competing to explain the data, and, finally, 3) an evaluation of such candidates to choose the one offering the highest forecast accuracy. This is how the profile of the data scientist expands, as new topics such as data visualization, statistical learning, building stories with data, and data preservation are incorporated into his or her interests [34].

According to [8], the areas of work in Data Science can be classified into six:

1) Collection, preparation, and exploration: experts usually consider that this work occupies 80% of the time spent in the whole analysis process. As discussed above, the concept of data has expanded, and, consequently, the techniques for collection and preparation are diverse and specialized. For example, web scrapping, pubmed scrapping, image processing, and loading and cleaning of text files, among others.

2) Initial representation and transformation: encompasses the work of transforming and preparing the data in a way that reveals the most information. It includes skills in both databases and mathematical representation of non-numeric data.

3) Computation with data: data science is intensive in statistical programming. Currently, the two most widely used languages are Python and R. Any professional trained in these subjects will need to develop advanced knowledge in one of these tools. However, in their professional performance, they will encounter both. Additionally, a large volume of data requires designing and running both compute-saving experiments and cloud computing tools. Before entering the modeling phase in this phase of data science work, the data flow is also defined.

4) Visualization and presentation: through descriptive statistics, clean and organized data bring out a lot of important information. A good presentation of relevant figures can be enough to consolidate a convincing discourse or motivate the direction of a fundamental decision. Of course, the use of statistical information before a non-specialized audience can be subject to manipulation by incurring biases or inaccuracies on causality, exaggerated magnification, omission of important information, and other forms of fallacious argumentation.

5) Modeling: modeling in data science can be inscribed in what [9] called "the two cultures". Either "generative models" are used, which assume that data are generated according to a stochastic process, or predictive models are used based on learning algorithms. Some experiments may use a combination of models of both types to explore which predicts best. In certain circumstances, some accuracy may be sacrificed by the interpretability of generative models [35].

6) Research: data science is under increasing development. It is common to find the international web competitions to convene experts around the prediction problems - previously mentioned - that still exhibit poor levels of accuracy. Likewise, it is of great interest to find algorithms for supervised and unsupervised models which economize computational capacity. In the section on bibliometric analysis, it will be seen that when exploring the results of an indexing database with the word "data", the volume of production around analytics is extraordinarily high. Therefore, both the scientific community and the productive sector are very interested in finding new tools for these topics.

Additionally, the management of large volumes of data poses challenges for regulating the use of public information and the design of official statistics oriented to decision making. Consequently, legal aspects constitute an object of study of interest for the data science curriculum [36]. An alternative way to visualize the areas of performance is to understand the work areas of data engineering and those of data science in a differentiated way. The former is strongly associated with storage, collection, and cleaning, while the latter is related to data research (Figure 1).

Table 1: Division between Data Engineering and Data Science.

Data Engineering	**Data Science**
Processing raw data	Test hypotheses
Running data behind the scenes	Deliver results to business users
Build infrastructure to consolidate and enrich numerous data sets	Apply machine learning algorithms and other analytical approaches
Handle large-scale data processing	Uncover findings in large volumes of data
Monitor and maintain systems	Interpret analysis results
Prepare data for analysis	Develop articulated analysis with visual tools

Source: Prepared by the authors.

III. METHODOLOGY FOR BIBLIOMETRIC ANALYSIS

Bibliometric analysis is commonly used to characterize the bibliographic production of a specific subject. These works quantify the published and indexed production in scientific databases. Depending on the researcher's interest, the work may focus on the impact of publications, the exploration of key authors, the most relevant journals, or the development of sub-themes and their relevance. In the present curriculum proposal, bibliometrics focuses on the last application. Analyzing the deployment of data analysis in research with the greatest international impact will shed light on the structuring of the curriculum and, in particular, on updating the programmatic content of the subjects.

a. Selection of a database, keywords, and observation sales

The two most relevant international indexing references are ISI Web of Knowledge and SCOPUS. Apparently, in Colombia, there is more familiarity with the latter database, and, in addition, the IUPG has institutional access. The results presented below were obtained using the Bibliometrix package, which runs on the R open-access software. The conceptual basis of this instrument is found in the work of Aria and Cuccurullo [37].

The selection of search words is an essential task for developing a bibliometric analysis. This work involves specialized thesaurus analysis and expert validation. Searches were performed to develop the bibliometrics of data analysis by combining several key terms, which referred to the largest areas of parametric and nonparametric statistical theory, with the most relevant designations of supervised and unsupervised data mining models. An important feature of this methodology is that the interpretation of the results transcends the conceptual framework through which the results are produced, as it depends largely on the researcher's expertise.

In the end, the clearest results were obtained using the most generic term possible. The search term was "data". Although this term is too generic and seems a straightforward selection, its results are well interpreted and correspond to the purpose of this work: to know the thematic development of data analysis in all the international scientific production and use it as an input in the curricular structure.

b. Analysis of the conceptual structure

This analysis aims to identify the subtopics of the main theme and group them by similarities, according to the criterion of co-occurrence of keywords [37]. To achieve this, we start from a matrix that crosses all the keywords with the documents so that the co-occurrence of words in the documents becomes relevant.

The keywords of the publications are arranged in a matrix X (keywords vs. documents), where X_{ij} takes the value of 1 (if keyword i is included in document j) or the value of 0 (otherwise). Through a Multiple Correspondence Analysis (MCA), a plane reduced to two dimensions is constructed in which the words are represented more closely, depending on the similarity of their distributions [38]-[40]. MCA allows both an exploratory analysis without assuming restrictions on the data and a simple interpretation of clusters of keywords established according to their position in the two-dimensional factorial plane [41].

c. Centrality and density metrics

The analysis of key topics through the co-occurrence criterion can be visualized as a network. In this way, keywords grouped in clusters, utilizing the k-means algorithm, form groups that acquire density when there is a high co-occurrence of keywords within the same cluster. According to Cobo et al. [38], this metric is interpreted as the level of development within a theme. On the other hand, centrality measures the degree of the interrelation of the keyword of a theme with keywords from other themes.

The equivalence index [38] is defined as

$$e_{ij} = c_{ij}^2 / c_i c_j \qquad (1)$$

Where c_{ij} is the number of documents in which two keywords i and j co-occur. And c_i and c_j represent the number of documents in which each appears.

From an equivalence index, Callon et al. [42] interpret the co-occurrences of keywords as a network. Consequently, they define two classical metrics: 1) centrality, which can be defined as follows:

$$c = 10 * \sum e_{kh} \qquad (2)$$

Where k is a keyword belonging to a topic, h is a keyword belonging to other topics, and 2) the density, which can be defined as follows:

$$d = 100(\sum e_{ij}/w) \qquad (3)$$

Where i and j are keywords belonging to the same topic, and w is the total number of keywords within the topic.

In a Cartesian plane, the quadrants can be represented like this (clockwise, starting with the upper left quadrant):

- The first represents the most developed topics (high density), which are isolated, i.e., highly specialized (low centrality).
- The second represents highly developed and transversal topics: these are the "engines" of the research.
- The third represents low centrality and density topics, i.e., they are either very new or declining in relevance. Even so, it should be kept in mind that this analysis was made with the most relevant papers according to the SCOPUS selection criteria.
- Finally, the fourth quadrant presents the basic (low density) and cross-cutting topics.

d. *Evolution of the themes over time*

The importance of a thematic link can be measured by the elements that the linked topics have in common. According to Cobo et al. [38], the inclusion index is defined as follows: let T^t be the set of topics detected in subperiod t, where $U \in T^t$ represent each topic detected in super-period t. Let $V \in T^{t+1}$ be each topic detected in period $t+1$ It is said that there is a thematic evolution from U to V if there are keywords that occur in both and are associated with the thematic networks. Therefore, V can be considered a topic evolved from U. Keywords $k \in U \cap V$ are considered a thematic nexus or a conceptual nexus, and their level of importance is given by:

$$I = \frac{\#(U \cap V)}{Min(\#U, \#V)} \qquad (4)$$

IV. RESULTS ANALYSIS AND INTERPRETATION

a. Searching with the term "data" in SCOPUS

Because of the breadth of the concept, the search yields results for papers published from 1825 to 2020. In August 2019, this search yielded 10,734,739 publications. The corresponding SCOPUS option was used to select the most relevant papers. The export of results made it possible to consolidate a database with the two thousand most relevant publications throughout history. After applying this filter, Figure 1 shows results since 1970, although the volume of relevant publications grows significantly from 2000 onwards.

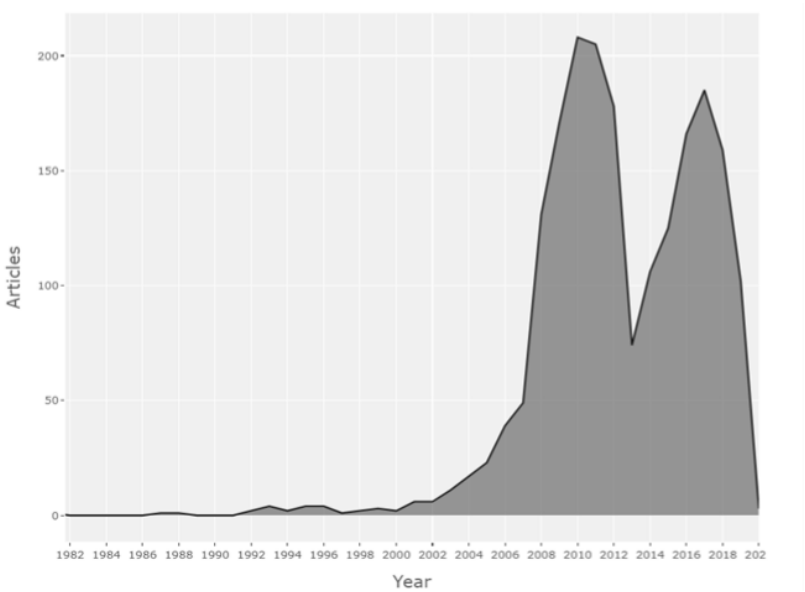

Figure 1: Division between Data Engineering and Data Science.
Source: Prepared by the authors.

The most important author keywords in this set of relevant publications are "big data" and "data mining". Both terms are, in turn, in the list of indexing keywords, which already sets the trend for the most relevant research topics in this data analysis.

Table 2: Most frequent keywords.

Author keywords	Articles	Indexing keywords	Articles
Big data	208	Data mining	611
Data mining	169	Data handling	589
Data quality	124	Big data	465
Data integration	103	Information management	416
Data warehouse	103	Data reduction	326
Data management	87	Data warehouses	272
Data analysis	71	Data processing	252
Cloud computing	64	Data integration	250
Linked data	53	Data quality	235
Big data analytics	37	Data acquisition	232
Data cleaning	37	Data visualization	230
Data processing	37	Data sets	215
Data sharing	37	Digital storage	215
Data fusion	36	Metadata	205
Data model	36	Data structures	188
Open data	36	Data communication systems	168
EData warehousing	34	Visualization	167
Metadata	34	Data privacy	161
Data streams	33	Algorithms	148
Visualization	33	Database systems	143

Source: Prepared by the authors.

SCOPUS uses a sophisticated procedure to assess the relevance of publications. A simple explanation of how it works can be found in the user help portal. In SCOPUS, there are two types of keywords: firstly, those selected by the author(s); secondly, there are indexing keywords selected by content providers. Unlike author keywords, indexing keywords take synonyms, various spellings, and plurals into account.

b. Conceptual structure

The map reveals information from four closely related thematic clusters:

- Cluster 1: groups keywords associated with research in text mining, text analysis, geographic analysis, and web scraping.
- Cluster 2: groups the core topics in data analysis: storage, database structure, query, cleansing, analytics, visualization, cloud computing, artificial intelligence.
- Cluster 3: reveals a closeness, which can also be identified in cluster 2, between data mining, communication systems, information security, and classification algorithms.
- Cluster 4: groups keywords that are of great relevance in administrative sciences. Thus, it reflects the relevance of knowledge management, information systems administration, data-driven decision making.

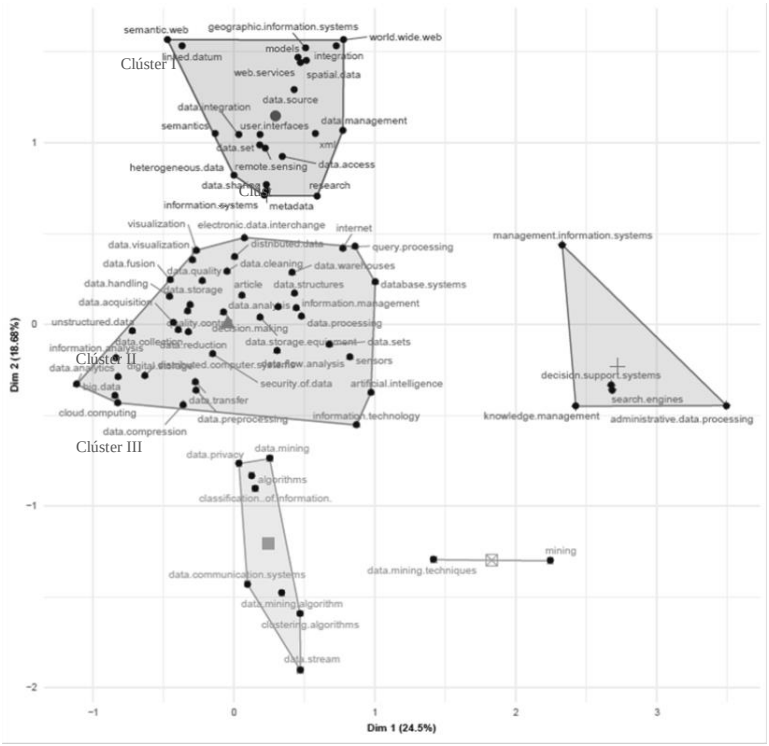

Figure 2: Clusters of keywords.
Source: Prepared by the authors.

c. Thematic map

Figure 3 presents the themes according to their level of centrality and density. The driving themes refer, in the first place, to database processing. This includes processing, acquisition, quality, control, dimension reduction, and its analysis in a broad theme. As topics in disuse, low centrality and density appear to be topics associated with the management of data warehouses and administration. In contrast, very specialized and high-density themes refer to spatial analysis, semantics (database modeling), and metadata. Finally, cross-cutting themes (more central than dense), which could be interpreted as interdisciplinary, refer to data mining, big data, visualization techniques, and communication systems.

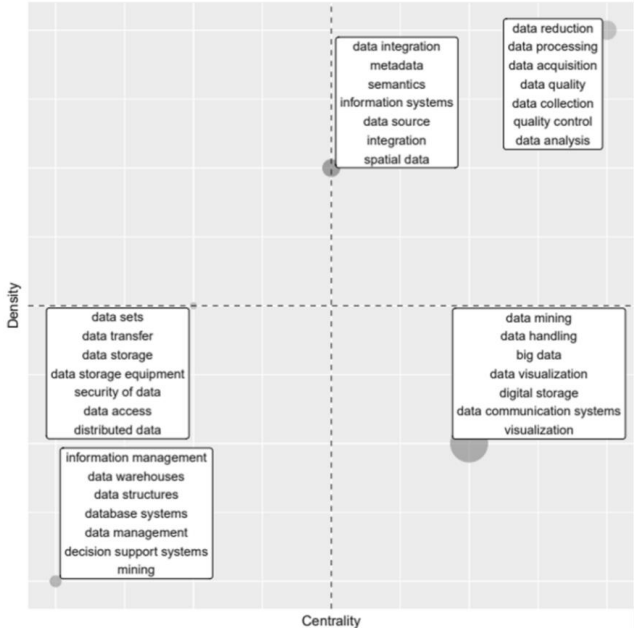

Figure 3: Thematic map. Density vs. Centrality.
Source: Prepared by the authors using SCOPUS data.

d. Evolution of the themes over time

The thematic evolution reflects the relevance of tools for data integration in collaborative reuse projects. It also shows: 1) the consolidation of the use of data mining tools in business intelligence, 2) the migration of the concept of "data management" to a more complex concept known as "data governance", and 3) the thematic migrations it is common that some basic database processing topics migrate to big data.

Table 3: Thematic evolution.

Source theme 1970-2014	Destination theme 2015-2020	Inclusion rate
data integration	data reuse	1
data mining	business intelligence	1
data management	open government data	0.5
data merging	big data	0.5
data mining	semi-structured data	0.25
data model	data Flow	0.25
data processing	big data	0.2
data processing	data fusión	0.2
data cleaning	big data	0.17
data cleaning	data quality	0.17
Reliability	data management	0.17
cloud computing	data lifecycle	0.14
data integration	big data mining	0.14
data management	visualization	0.12
data mining	data models	0.125
data mining	visualization	0.125
data warehousing	data models	0.125

Source: Prepared by authors.

V. CONCLUSIONS

Data science constitutes an expanded field of research with respect to classical statistics in at least two aspects: firstly, its field is no longer limited only to numerical data as an object of study, given that with computational advances, other types of information such as sound and images can be analyzed, when arranged in matrix form; secondly, bibliometrics as a research method in this field reflects that, when searching for the keyword "data", present in a large number of papers indexed in SCOPUS, the topics belong mainly to what [9] calls "the culture of modeling with algorithms". All this appears due to the analysis of the most relevant papers and is evident both in the conceptual organization and in the thematic evolution.

The present work demonstrated that a new object of the study constituted around data has interdisciplinary applications of great impact. Undoubtedly, quantitative research under the expanded concept of "data" finds new problems and methods in other disciplines.

Studying the relevant topics and their evolution over time is the essential input to delivering a curricular design that complies with the requirements of the Colombian regulation contained in Decree 1330 of 2019. The description of the topics and the analysis of the impact levels of each one are the first rigorous reference to establish the teaching fields and the emphasis of the new program. However, this study should be complemented with the analysis of the curricula of a broad and diverse set of local and foreign proposals.

VI. REFERENCES

[1] J. W. Tukey, "The future of data analysis," Ann. Math. Stat., vol. 33, no. 1, pp. 1–67, 1962.
[2] C. Maldonado and N. A. Gómez-Cruz, El mundo de las ciencias de la complejidad. Un estado del arte. Bogotá, Colombia: Universidad del Rosario, 2010.
[3] C. Merow et al., "What do we gain from simplicity versus complexity in species distribution models?," Ecography (Cop.)., vol. 37, no. 12, pp. 1267–1281, 2014, doi: 10.1111/ecog.00845.
[4] K. V Katsikopoulos, "Bounded rationality: the two cultures," J. Econ. Methodol., vol. 21, no. 4, pp. 361–374, 2014, doi: 10.1080/1350178X.2014.965908.
[5] R. Descartes, Discurso del método. Ediciones Colihue SRL, 2004.
[6] M. Bunge, "La ciencia: su método y su filosofía," 1978.
[7] M. Frické, "Big data and its epistemology," J. Assoc. Inf. Sci. Technol., vol. 66, no. 4, pp. 651–661, 2015, doi: 10.1002/asi.23212.
[8] D. Donoho, "50 years of data science," J. Comput. Graph. Stat., vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.
[9] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," Stat. Sci., vol. 16, no. 3, pp. 199–231, 2001, doi: 10.1111/j.1740-9713.2005.00129.x.
[10] K. Mardia and W. Gilks, "Meeting the statistical needs of 21st-century science," Significance, vol. 2, no. 4, pp. 162–165, 2005, doi: 10.1111/j.1740-9713.2005.00129.x.
[11] W. M. Briggs, "Everything wrong with p-values under one roof," Studies in Computational Intelligence, vol. 809. Springer Verlag, 340 E. 64th Apt 9A, New York, United States, pp. 22–44, 2019, doi: 10.1007/978-3-030-04200-4_2.
[12] T. Derrick, "The criticism of inferential statistics," Educ. Res., vol. 19, no. 1, pp. 35–40, 1976.
[13] J. R. Jamison, "The use of inferential statistics in health and disease: a warning," South African Med. J., vol. 57, no. 19, pp. 783–785, 1980.
[14] B. L. Hopkins, B. L. Cole, and T. L. Mason, "A critique of the usefulness of inferential statistics in applied behavior analysis," Behav. Anal., vol. 21, no. 1, pp. 125–137, 1998.
[15] A. Charpentier, E. Flachaire, and A. Ly, "Econometrics and machine learning," Econ. Stat., vol. 2018, no. 505–506, pp. 147–169, 2018, doi: 10.24187/ecostat.2018.505d.1970.
[16] D. Qin, "Let's take the bias out of econometrics," J. Econ. Methodol., vol. 26, no. 2, pp. 81–98, 2019, doi: 10.1080/1350178X.2018.1547415.
[17] S. Athey and G. W. Imbens, "Machine Learning Methods That Economists Should Know about," Annu. Rev. Econom., vol. 11, pp. 685–725, 2019, doi: 10.1146/annurev-economics-080217-053433.
[18] M. Molina and F. Garip, "Machine Learning for Sociology," Annual Review of Sociology, vol. 45. Annual Reviews Inc., Department of Sociology, Cornell University, Ithaca, NY 14853, United States, pp. 27–45, 2019, doi: 10.1146/annurev-soc-073117-041106.
[19] S. Mützel, "Facing big data: Making sociology relevant," Big Data Soc., vol. 2, no. 2, p. 2053951715599179, 2015.
[20] D. A. McFarland, K. Lewis, and A. Goldberg, "Sociology in the era of big data: The ascent of forensic social science," Am. Sociol., vol. 47, no. 1, pp. 12–35, 2016.
[21] K. Healy and J. Moody, "Data visualization in sociology," Annu. Rev. Sociol., vol. 40, pp. 105–128, 2014.
[22] P. Barrett, "What if there were no psychometrics? Constructs, complexity, and measurement," J. Pers. Assess., vol. 85, no. 2, pp. 134–140, 2005, doi: 10.1207/s15327752jpa8502_05.
[23] N. Bolger, "Data analysis in social psychology," Handb. Soc. Psychol., vol. 1, pp. 233–265, 1998.
[24] D. Bzdok and J. P. A. Ioannidis, "Exploration, Inference, and Prediction in Neuroscience and Biomedicine," Trends Neurosci., vol. 42, no. 4, pp. 251–262, 2019, doi: 10.1016/j.tins.2019.02.001.
[25] A. L. Boulesteix and M. Schmid, "Machine learning versus statistical modeling," Biometrical J., vol. 56, no. 4, pp. 588–593, 2014, doi: 10.1002/bimj.201300226.
[26] J. Wang and Q. Tao, "Machine learning: The state of the art," IEEE Intell. Syst., vol. 23, no. 6, pp. 49–55, 2008.
[27] R. Gould, "Data literacy is statistical literacy," Stat. Educ. Res. J., vol. 16, no. 1, pp. 22–25, 2017.
[28] P. Bühlmann, "Comments on: Data science, big data and statistics," Test, vol. 28, no. 2, pp. 330–333, 2019, doi: 10.1007/s11749-019-00646-6.
[29] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," J. Econ. Perspect., vol. 31, no. 2, pp. 87–106, 2017, doi: 10.1257/jep.31.2.87.
[30] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," Science (80-.)., vol. 350, no. 6264, pp. 1073–1076, 2015, doi: 10.1140/epjds/s13688-017-0125-5.
[31] L. Dong, S. Chen, Y. Cheng, Z. Wu, C. Li, and H. Wu, "Measuring economic activities of China with mobile big data," arXiv Prepr. arXiv1607.04451, 2016, doi: 10.1140/epjds/s13688-017-0125-5.
[32] B. Yu, "Embracing statistical challenges in the information technology age," Technometrics, vol. 49, no. 3, pp. 237–248, 2007, doi: 10.1198/004017007000000254.
[33] S. Tonidandel, E. B. King, and J. M. Cortina, "Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science," Organ. Res. Methods, vol. 21, no. 3, pp. 525–547, 2018, doi: 10.1177/1094428116677299.
[34] B. Beaton, A. Acker, L. Di Monte, S. Setlur, T. Sutherland, and S. E. Tracy, "Debating data science: A roundtable," Radic. Hist. Rev., vol. 2017, no. 127, pp. 133–148, 2017, doi: 10.1215/01636545-3690918.
[35] D. V Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electron., vol. 8, no. 8, 2019, doi: 10.3390/electronics8080832.
[36] P. J. H. Daas, M. J. Puts, B. Buelens, and P. A. M. van den Hurk, "Big data as a source for official statistics," J. Off. Stat., vol. 31, no. 2, pp. 249–262, 2015, doi: 10.1515/JOS-2015-0016.
[37] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," J. Informetr., vol. 11, no. 4, pp. 959–975, 2017, doi: 10.1016/j.joi.2010.10.002.
[38] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field," J. Informetr., vol. 5, no. 1, pp. 146–166, 2011, doi: 10.1016/j.joi.2010.10.002.
[39] V. Batagelj and M. Cerinšek, "On bibliographic networks," Scientometrics, vol. 96, no. 3, pp. 845–864, 2013, doi: 10.1007/s11192-012-0940-1.

[40] K. Börner, C. Chen, and K. W. Boyack, “Visualizing knowledge domains,” Annu. Rev. Inf. Sci. Technol., vol. 37, no. 1, pp. 179–255, 2003, doi: 10.1002/aris.1440370106.
[41] C. Cuccurullo, M. Aria, and F. Sarto, “Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains,” Scientometrics, vol. 108, no. 2, pp. 595–611, 2016.
[42] M. Callon, J. P. Courtial, and F. Laville, “Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry,” Scientometrics, vol. 22, no. 1, pp. 155–205, 1991.