Data Science and AI: trends analysis

Sofia Aparicio

Dept. of Computer Science and
Engineering, Instituto Superior Tecnico,
Universidade de Lisboa, Portugal
sofia.aparicio@acm.org

Joao Tiago Aparicio

Instituto Universitario de Lisboa
(ISCTE-IUL)

jtaca@iscte-iul.pt

Carlos J. Costa

Advance/CSG, ISEG (Lisbon School of
Economics & Management),
Universidade de Lisboa
Instituto Universitario de Lisboa
(ISCTE-IUL) ISTAR-IUL
cjcosta@iseg.ulisboa.pt

Abstract — **This study has the primary goal to analyze the growth of data science through the main search trends. This study was conducted by defining in high level the concept of data science as well as its main components. Supported in those elements, we identified the main trends. We used mainly data from google trends to determine the evolution of search by topics, research area, or simple expressions. It allowed us to reckon that artificial intelligence (AI) suffered a lack of interest until 2012. Then it became an increasingly popular field since 2014. This is due to the progression of machine learning and data science. Results show a cumulative search of data science since 2012.**

Keywords - data science; artificial intelligence; programming languages; trends; search.

I. INTRODUCTION

Data science is one of the main topics in computer science. The increasing amount of data available demanded the use of statistics from a new perspective. Data analysis is not only supported by statistic techniques but also by new computing power [1]. In this sense, machine learning also increases its importance, leading to the resurgence of widespread interest in artificial intelligence (AI). This led to several questions: What is the future of data science? What subjects may be related to this one? What is expected for artificial intelligence in the future? Since these questions cannot be answered, we reformulated the following research question: What are the main trends related to Artificial Intelligence and Data Science?

To conduct this study, we analyzed the main search trends of concepts related to data science and artificial intelligence, such as mathematics, computer science and management. Afterwards, it was also relevant to analyze works developed in the areas of machine learning and software development. Specifically, in what concerns software development, the study investigates the context of usage and also the technologies that have been used.

It was possible to observe, on the one hand, the growth of languages related to data science. On the other, the diminishing relevance of languages only used for software development.

II. DEFINING DATA SCIENCE

According to Provost & Fawcett [2], data science is a set of fundamental principles that support and guide the extraction of information and knowledge from data. As according to Dhar [3, p. 64] data science "*focus involving data and, by extension, statistics, or the systematic study of the organisation, properties, and analysis of data and its role in inference, including our confidence in the inference.*" Data science comprises the intersection of several fields of knowledge: data science = {statistics ∩ informatics ∩ computing ∩ communication ∩ sociology ∩ management | data ∩ domain ∩ thinking} [4]. Other researchers emphasise the importance of machine learning and its impact on business [1], [3], [4]. Granville [5, p. 73] presents arguments for the work of being a data scientist "*not statisticians, nor data analysts, nor computer scientists, nor software engineers, nor business analysts. They have some knowledge in each of these areas but also some outside of these areas.*", This is due to the usage of several tools and knowledge that is no pure software development, mathematics nor purely statistics. Data science comprises algorithms implementations and use, as well as other robust techniques to entail predictions to be applied in organisational or societal contexts. Chatfield et al. [6] present an overview of the most common attributes of data scientist as presented in Table 1.

According to Table 1, almost the majority of the authors agree that business domain knowledge is an important attribute [2], [7]–[13] that a data scientist should have. The ability to derive valuable insights, and science computing skills, as well as effective communication skills, are also attributes of the most important in data science [3], [5]–[9], [11], [12], [14], [15]. Other attributes such as statistical modelling knowledge, data visualisation, mathematics, data management, artificial intelligence knowledge, machine learning, analytical traits, and being curious are much referred in literature.

As the starting point, we may define as the interception of Basic fields: Computer Science and IT (CS), Domain/Business Knowledge (BK) and Mathematics and Statistics (MS). In fact, it is possible to identify traditional research ={BK ∩ MS}, Software Development={CS ∩ BK}, Machine Learning = {CS ∩ MS} and obviously (but simplifying), Data Science = {BK ∩ MS ∩ CS} (Figure 1).

Table 1. Data Scientists attributes according to several studies [6]

Attributes	IBM Website (2014)	Cooper (2012)	Davenport (2012)	Davenport and Patil (2012)	Dhar (2013)	Granvill (2014)	Harris et al. (2013)	Laney (2012)	Loukides (2010)	Microsoft Website (2013)	Mohanty et al. (2013)	National Science Board (2005)	Swan and Brown (2008)	Vangelova (2014)	Provost and Fawcett (2013)	Press (2012)	Lev-Ram (2011)	Total
Entrepreneurship/Business domain knowledge			x	x			x	x	x		x		x		x			8
Computer scientist			x	x		x						x		x		x	x	7
Effective communication		x	x	x	x	x	x					x						7
Creating valuable actionable insight			x	x			x		x	x	x	x						7
Curious	x	x		x					x				x		x			6
Statistical and modelling		x			x	x							x	x	x			6
Data visualisation			x	x								x	x		x			5
Mathematics			x	x	x			x						x				5
Data management		x			x	x						x						4
Analytical			x			x							x		x			4
Software engineer			x	x		x						x						4
PhD qualification			x	x				x					x					4
Programmer												x				x		2
Understanding business challenges							x				x							2
Machine learning					x			x										2
Economics				x									x					2
Technologist & quantitative analysts			x															1
Outside of IT knowledge						x												1
Works in a team								x										1
Experience with big data sets								x										1
Optimisation					x													1
Interdisciplinary									x									1
Artificial intelligence					x													1

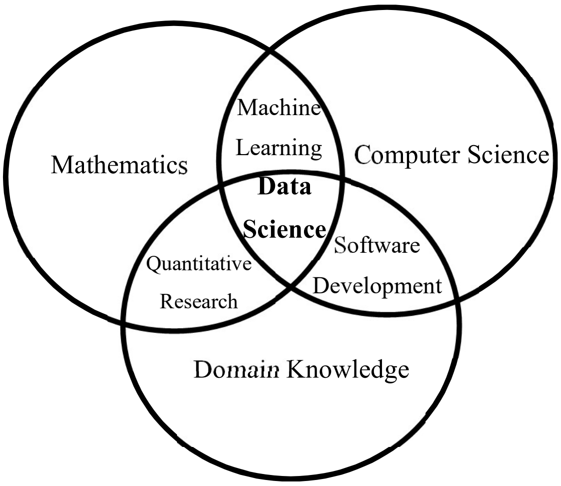

Figure 1. Data Science = {BK∩MS∩CS}

Computer science is a field that includes hardware, programming, artificial intelligence databases, networks. The main application of computer science is software development [16], [17].

Mathematics is a very relevant field of knowledge, supporting commerce, engineering and science. Specifically, statistics is becoming a field of increased usage. From the less computational area like data visualisation and descriptive statistics to the more sophisticated areas of data analysis, the abundance of data allows new possibilities [3], [15], [18].

Machine learning is as a subset of artificial intelligence. It is an approach, where computer systems perform a specific task without using explicit instructions. It relies on patterns and inference. To obtain this result, it consists of the scientific study of algorithms and statistical models. In practice is a connection between computing and statistics[19], [20].

Domain knowledge includes a broad context of usage of computing. Its usage is becoming more and broader. If initially was related especially to scientific computing, computing became popular when entered in business, supporting accountancy and statistics. However, computing has been used in more and more fields [7], [8].

Data science comprises several roles centred on data; those roles entail data analysis and analytics capabilities and competencies [4], [21]. Table 2 presents the main key terms in data science, based on Cao definition and classification [21].

Table 2. Data Science key terms

Data Science	**Analysis**	Processing data by traditional theories, technologies and tools					
	Data Analytics	**Descriptive**	Use statistics to describe data.	**Explicit** (reporting, statistical analysis, alerting, and forecasting)	**Era 1** Explicit analytics	**Objective**: We know what we do not know.	**Low:** complexity degree, intelligence and value **High**: level of visibility
		Predictive	Prediction of the unknown with advanced analytics (theories, technologies, tools and processes)	**Implicit** (predictive + prescriptive) & **Deep** (Aquire in-depth understanding of why and how things happen/will happen.)	**Era 2** Implicit analytics	**Objective**: We do not know what we do not know.	**Low**: level of visibility **Hight**: Complexity degree, intelligence and value
		Prescriptive	Optimisation of data for leading to smart decision-making.				

Cao [21] defines two data analytics Eras: Era 1 is characterised by the use of explicit analytics with descriptive purposes, such as reporting alerting, and forecasting. In the first Era, the primary objective is framework what we know that we do not know (know the unknowns), by processing data with low complexity degree and intelligence, providing at the same time moderate value creation. Era 2 is characterized by the practice of implicit analytics providing predictive and prescriptive data analytics. In Era 2 the primary objective is to extract knowledge for a better understanding of why things happen and how it happens and if they will happen. In this Era the level of visibility is low, and therefore processes and tools provide a higher level of intelligence and supporting smart decisions. AI is powered by the rise of data science, as autonomous and intelligent systems need quality data to train and develop continuously. However, we still have a long way to run in terms of perfection of AI, and people need to understand what systems can and cannot do, and what are the consequences to organisations and society [20]. AI poses several challenges to humankind; one of them is that thinking machines often know more than we do, and recognise that it is a great encounter [20]. Ethical issues also arise with the AI usage in several areas, in human resources decision process, when an algorithm decides the bank credit rate of a house loan based on gender, race, and even when in autonomous driving vehicles when redundant systems fail and the algorithm decides who is going to be killed [22]. Brooks [23] refers seven deadly sins of AI predictions:

1. Overestimating the consequences of technology in the near future and underestimating in the long run;
2. Imagining magic, as some arguments may state that technology in future is magic when it is science;
3. Performance vs competence, usually robots are, today very narrow on what they can do compared to humans;
4. Suitcase words, people tend to think robots think the way humans do;
5. Exponentials, people often think that the autonomous learning process is exponential, but exponentials can collapse when a physical limit is reached;
6. Hollywood scenarios, changes in autonomous and intelligent systems in the real world may not appear as seen in science fiction movies;
7. The speed of deployment, it is important to notice that transformations may not occur as fast as we expect them to be.

In this context, the goal of the present study is to understand the widespread interest of AI and identify the related dimensions to AI. In the next section, our goal is to answer the following question: What are the main trends related to Artificial Intelligence and Data Science? In order to answer this question, we began by defining in high level the concept of data science as well as its main components. Supported in those elements, we identified the main trends. We used mainly data from google trends to determine the evolution of search by topics, research area, or simple expressions.

III. Trends related to mathematics, Computer Science and Domain knowledge

To analyse the trends identifying the fields mathematics, computer science and domain knowledge. There was a difficulty concerning the domain knowledge. So, as long as management is becoming a transversal concept and probably representing what we call domain knowledge, was considered as a possible concept to analyse. The following graph shows the data produced by Google Trends. Numbers represent search interest. Values are relative to the highest point on the chart for worldwide. A value of one hundred is the peak popularity for the term. A value of 25 means that the term is 25% as popular. A score of 0 means there was not enough data for this term. It shows the statistics related to the field study of Mathematics and computer science and the academic discipline of management. As it is present, there is an increase of interest by mathematics.

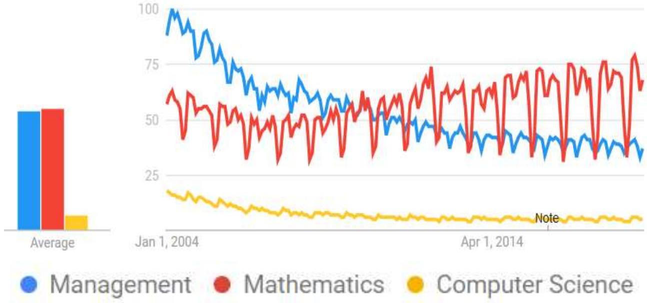

Figure 2. Management, mathematics and computer science

It is interesting to verify that there is a north-south division in what concerns the most search fields and academic disciple. The reason for this division is an interesting question that it is not the subject of the research presented in this paper.

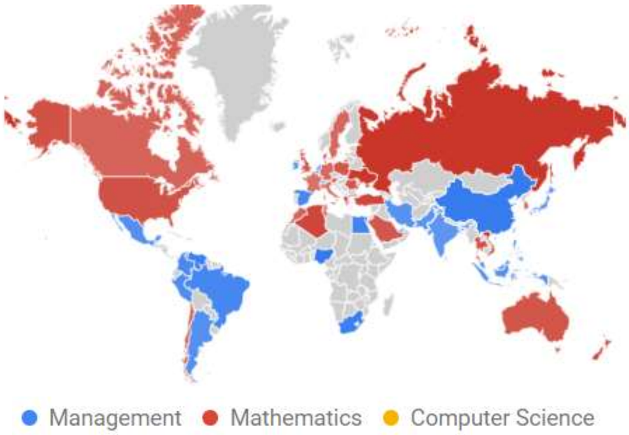

Figure 3. Importance in geographic terms

IV. Evolution of interest by Data Science

In the last 15 years, Artificial Intelligence had decreased and then an increase in popularity. If we analyse the search by teams in the google, we obtain a curve with a U shape. It reduced prevalence until 2012, and its importance increased mainly since 2014. We may try to find the reason, by identifying the search for other topics, like expert systems, data science out machine learning. There is a reduction of interest in topics like the topic "expert system", but an essential increase in topics like data science and machine learning.

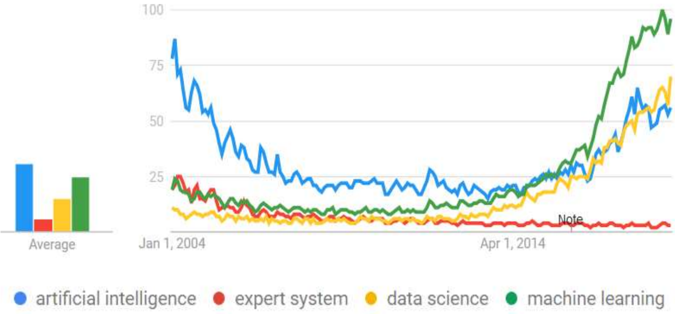

Figure 4. Artificial Intelligence became important again

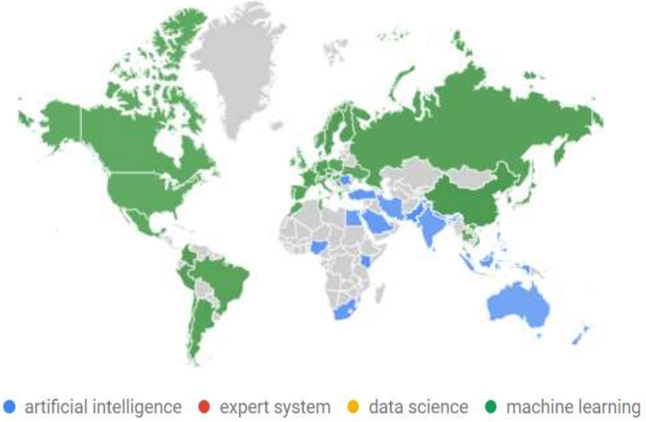
● artificial intelligence ● expert system ● data science ● machine learning

Figure 5. Artificial Intelligence vs machine learning

Some languages are more related to software development than while other languages are specially related to data science. For example, Java and C# are probably the most used languages, respectively in the context of open source community and Microsoft context for software development. An analysis of the computer languages searched in using google, shows that while languages like Java and C# had a decent trend, there is an increase of searches related to Python and R. It means that languages more related to data science have an increase. While languages more related to application development are reducing.

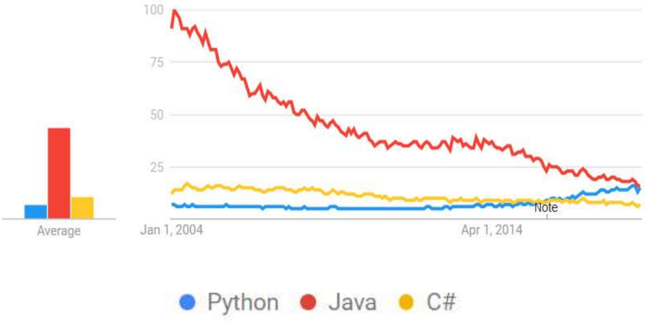

Figure 6. Search trends comparing Python with Java and C#

V. Server-side vs Front End

Analysing the search also allows identifying that there is a reduction of interest in server-side development and the increase of search related to the front-end web development.

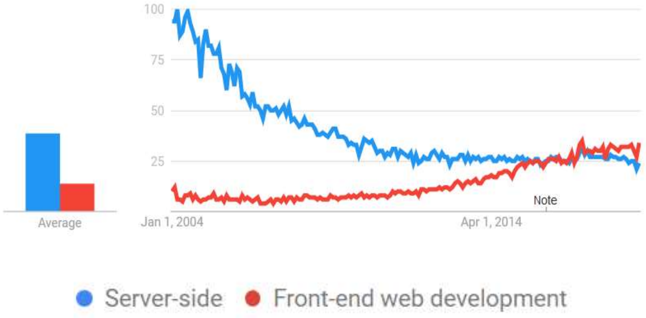

Figure 7. Server-side field vs front end related searches

It is also interesting to identify that the relative importance also changes geographically.

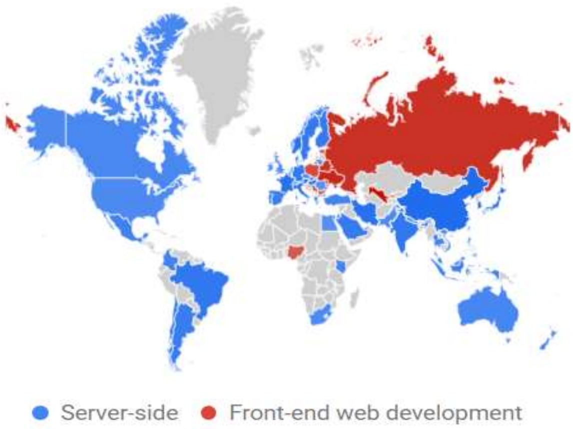

Figure 8. Geographic importance of search on google.

A specific analysis of several technologies allows confirming this idea. JavaScript, stabilised while other languages and technologies are more related to server-side reduced.

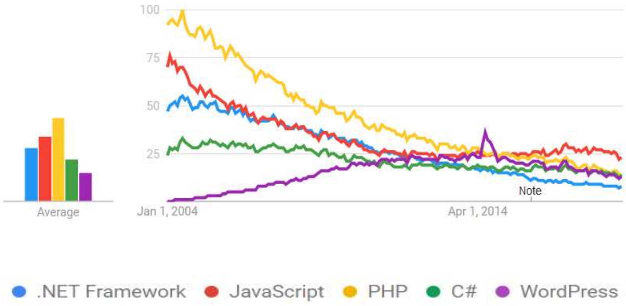

Figure 9. Technologies related to software development

It is also interesting to analyze that there is a different level of interest by all those technologies or languages, according to several countries.

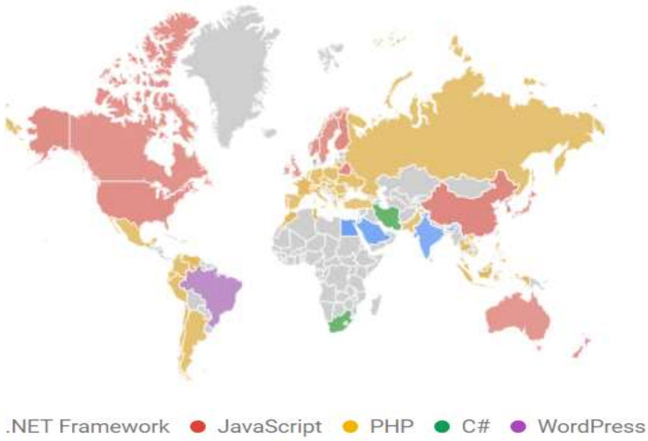

Figure 10. Technologies related to software development, geographic perspective

Analysing only trends search for languages, we verify that Java and C# has a decline in terms of search. R has a stable trend. Javascript shows a U shape., while Python [24] has an increase in terms of search.

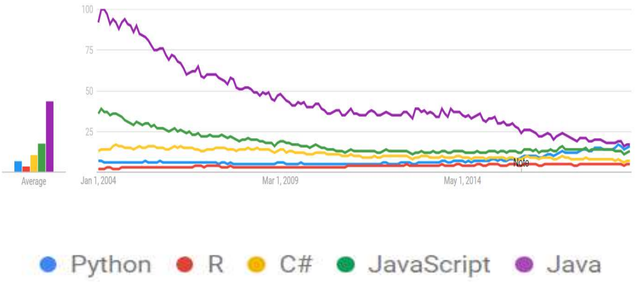

Figure 11. Technologies related to software development.

VI. Web vs Mobile

A few years ago, development effort started in the desktop. Since the second half of the 90s, the importance of the desktop is decreasing. However, now it is also the web development that his decreasing importance compared to mobile development.

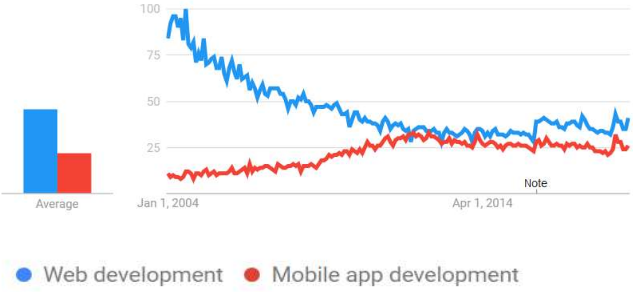

Figure 12. Searches trends comparing web development and mobile app development

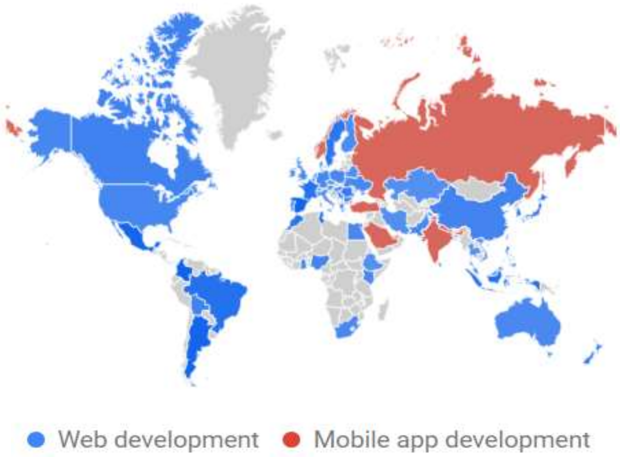

Figure 13. Searches trends comparing web development and mobile app development in geographic perspective

VII. Conclusions

According to the literature review, it was possible to define data science as the intersection of three large fields: computer science and artificial intelligence, mathematics and statistics and domain knowledge. It was possible to identify that this area is growing although its basis areas seem to be decreasing. For example, the interest in computer science is apparently reducing. However, it was also possible to identify that programming languages related to data science are becoming more and more used, has it was already stated in several other papers. Still, while languages and technologies related to software development are being less significant. It is also essential to identify that mobile development is being more critical. It suggests that those devices will become either sensor, obtaining data from users and actuators, communicating with users. Server tends to be mainly repositories and eventually additional processing capacity.

Acknowledgements

We gratefully acknowledge financial support from FCT-Fundação para a Ciencia e Tecnologia (Portugal), national funding through research grant UID/SOC/04521/2019, and UID/MULTI/4466/2019.

References

[1] M. I. Jordan e T. M. Mitchell, «Machine learning: Trends, perspectives, and prospects», *Science*, vol. 349, n. 6245, pp. 255–260, 2015.

[2] F. Provost e T. Fawcett, «Data science and its relationship to big data and data-driven decision making», *Big Data*, vol. 1, n. 1, pp. 51–59, 2013.

[3] V. Dhar, «Data Science and Prediction», *Commun ACM*, vol. 56, n. 12, pp. 64–73, Dez. 2013.

[4] L. Cao, «Data science: Challenges and directions», *Commun. ACM*, vol. 60, n. 8, pp. 59–68, 2017.

[5] V. Granville, *Developing Analytic Talent: Becoming a Data Scientist*. John Wiley & Sons, 2014.

[6] A. Chatfield, V. Shlemoon, W. Redublado, e F. Rahman, «Data scientists as game changers in big data environments», *Fac. Eng. Inf. Sci. - Pap. Part A*, pp. 1–11, Jan. 2014.

[7] M. A. Marques e C. J. Costa, «Social CRM analytics», em 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 2018, pp. 1–6.

[8] T. H. Davenport e D. Patil, «Data scientist», *Harv. Bus. Rev.*, vol. 90, n. 5, pp. 70–76, 2012.

[9] J. G. Harris, N. Shetterley, A. E. Alter, e K. Schnell, «The team solution to the data scientist shortage», *Accent. Inst. High Perform.*, 2013.

[10] D. Laney e L. Kart, «Emerging role of the data scientist and the art of data science», *Gart. Group White Pap.*, 2012.

[11] M. Loukides, *What is data science?* O'Reilly Media, Inc., 2011.

[12] S. Mohanty, M. Jagadeesh, e H. Srivatsa, *Big data imperatives: Enterprise 'Big Data'warehouse, 'BI'implementations and analytics*. Apress, 2013.

[13] A. Swan e S. Brown, «The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs», Set-2008. [Online]. Available: https://eprints.soton.ac.uk/266675/.

[14] D. Simberloff *et al.*, «Long-lived digital data collections: enabling research and education in the 21st century», *Natl. Sci. Found.*, 2005.

[15] L. Vangelova, «Career of the Month», *Sci. Teach.*, vol. 81, n. 7, p. 62, 2014.

[16] H. A. Simon, *The sciences of the artificial*. MIT press, 1996.

[17] A. Newell e H. A. Simon, «Computer Science as Empirical Inquiry: Symbols and Search», Communications, 1976..

[18] M. Ford, *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing Ltd, 2018.

[19] D. Zhang e J. J. Tsai, «Machine learning and software engineering», *Softw. Qual. J.*, vol. 11, n. 2, pp. 87–119, 2003.

[20] H. B. Review, M. E. Porter, T. H. Davenport, P. Daugherty, e H. J. Wilson, *HBR's 10 Must Reads on AI, Analytics, and the New Machine Age (with bonus article «Why Every Company Needs an Augmented Reality Strategy» by Michael E. Porter and James E. Heppelmann)*. Harvard Business Press, 2018.

[21] L. Cao, «Data science: a comprehensive overview», *ACM Comput. Surv. CSUR*, vol. 50, n. 3, p. 43, 2017.

[22] N. Bostrom e E. Yudkowsky, «The ethics of artificial intelligence», em *The Cambridge Handbook of Artificial Intelligence*, K. Frankish e W. M. Ramsey, Eds. Cambridge: Cambridge University Press, 2014, pp. 316–334.

[23] R. Brooks, «Robotics pioneer Rodney Brooks debunks AI hype seven ways», MIT Technology Review. [Online]. Available: https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/..

[24] K. J. Millman e M. Aivazis, «Python for Scientists and Engineers», *Comput. Sci. Eng.*, vol. 13, n. 2, pp. 9–12, Mar. 2011.