Data scientists in the scientific literature: LDA topic modelling on the semantic scholar database

Len Krawczyk
ORCID: 0009-0001-5450-6475
College of Inter-area Individual Studies in the Humanities and Social Science
University of Warsaw

Łukasz Iwasiński
ORCID: 0000-0003-2126-7735
Faculty of Journalism, Information and Book Studies
University of Warsaw

Mateusz Szymański
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

Abstract

Purpose/Thesis: This paper explores the representation of data scientists in scientific literature. It aims to answer the following questions: How has the number of publications on data scientists evolved over time? How are papers regarding data scientists distributed over different fields of study? In what context are data scientists represented in the scientific literature?
Approach/Methods: The authors used Latent Dirichlet Allocation (LDA) topic modelling to the resources available within the Semantic Scholar API.
Results and conclusions: There has been an increase in the number of publications on data scientists since 2008. A robust connection between data scientists and information technology, as well as biomedical research, was found. Little literature discusses data scientists in a sociocultural context.
Originality/Value: To our knowledge, no studies have been devoted to the representation of data scientists in scientific literature. The research may contribute to the conceptualisation of this notion.

Keywords:

Data Science. Latent Dirichlet Allocation. Semantic Scholar. Text Mining. Topic Modeling.

Text received on 14th of October 2024.

ZIN 2024, 62(2), 47–74
DOI: 10.36702/zin2024.02.09

1. Introduction

The rapid increase in the amount of data produced globally requires new forms of data management to derive value from it. Extracting knowledge from extensive databases demands skilled professionals capable of creating statistical models to uncover structured and unstructured data patterns. These professionals are commonly referred to as data scientists. However, due to the relative novelty of this phenomenon, the term 'data scientist' does not yet have a fixed definition (Hazzan et al., 2023). Given the rapid evolution of data science as a profession, definitions and roles continue to shift, reflecting its dynamic nature and widespread influence across diverse domains. Usually, data science is defined as a multidisciplinary field (Cleveland, 2001). Data scientists are typically proficient in applying statistical, analytical, and machine-learning techniques to draw insights from data (Donoho, 2017; Ho et al., 2019), often intending to create value in a commercial context (Reyes & Felipe, 2018).

In scientific literature, data scientists are primarily treated as a professional group (Espinoza & Gellegos, 2019). Efforts to define data scientists often involve analysing their skills and qualifications by examining quantitative data from various sources, such as job offers (Ho et al., 2019) and heterogeneous sources (Ismail & Zainal Abidin, 2016; Coelho Da Silveira et al., 2020). There is a scarcity of qualitative research on data scientists, although a few studies do exist (Pereira, Cunha, & Fernandes, 2020; Żulicki, 2022; Lowrie, 2017). Despite their undeniable impact on everyday life (Śledziewska & Włoch, 2020) and the broader scientific community (Hazzan & Mike, 2023), there is limited research on data scientists themselves outside the commercial context.

Big data is an essential factor not only in today's global economy but also in knowledge production (Krumholz, 2014; Priestley & McGrath, 2019). Data scientists wield powerful tools with uncertain implications (Boyd & Crawford, 2012) that have the potential to reshape the world. Therefore, we believe it is crucial to explore this topic further to understand better who shapes modern knowledge and how science reflects dynamic global changes. This paper aims to examine the representation of data scientists in scientific literature. It also strives to explore associations between data science and other fields of study, which may contribute to the conceptualisation of this term. To achieve this, we have employed a data scientist's toolkit, including text mining techniques and Latent Dirichlet Allocation (LDA) topic modelling, to analyse a vast repository of scholarly data accessible through the Semantic Scholar API. This approach leverages both computational power and theoretical insight, providing a robust framework for capturing and analysing the complex web of themes and relationships embedded within the literature on data science.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for large collections of discrete data, especially text corpora. In the original paper written

by Blei et al. (2003) on Latent Dirichlet Allocation, the authors trained the model on several datasets for different purposes, including text corpora. The main text data sets used for training and evaluating LDA included, among others, scientific abstracts from the C. elegans community, which contained 5,225 documents. The use of scientific abstracts as part of the evaluation and demonstration of LDA's capabilities paper does illustrate its suitability for analysing scientific abstracts by uncovering latent themes or topics within a large collection of text documents. The advantage of using LDA is the fact that it is a powerful technique for unsupervised analysis, making it one of the most extensively used text-mining tools in research on scholarly data (Thakur & Kumar, 2022) and currently a recognised scientometric tool in library and information sciences (Lamba & Madhusudhan, 2018a; Lamba & Madhusudhan, 2018b; Miyata et al., 2020; Sugimoto et al., 2011). This has been reflected in numerous studies using this method for thematic clustering of scientific articles in multidisciplinary literature research (Anupriya & Karpagavalli, 2015; Griffiths & M. Steyvers, 2004) as well within domain-specific context such as information communication technologies (Lim & Maglio, 2018; Liu et al., 2016; Cortez et al., 2018; Chen, Wang & Lu, 2016), biomedical sciences (Ebrahimi, Dehghani & Makkizadeh, 2023; García et al., 2020; Yoon & Suh, 2019; Zou, 2018), management (Cho et al., 2017; Joo et al., 2018; Moro et al., 2015) environmental sciences (Chang et al., 2021; Dayeen et al., 2020; Jeon et al., 2018; Syed et al., 2018). It was used for the classification of scientific papers, as well as for finding patterns of rhetorical moves (Louvigne et al., 2013).

Using LDA to analyse data on the representation of data scientists in the scientific literature is a fitting approach, especially considering the multidisciplinary nature of data science, as it spans fields such as statistics, computer science, machine learning, business intelligence, and domain-specific applications like healthcare, finance, and social sciences. LDA is particularly well-suited to uncover hidden topics across large corpora of text, making it practical for identifying the diverse themes and sub-disciplines present in the literature that may not be immediately apparent through manual analysis.

Given that we utilise the Semantic Scholar API to access vast amounts of scholarly data, LDA's ability to handle large data sets makes it a suitable choice. With its broad scope and frequent updates, data science literature can be overwhelming to classify and interpret manually, but LDA allows for scalable and automated topic identification.

By applying LDA, we can discover latent themes that may connect data science to other fields of study. Doing so allows us to understand better how data science interacts with, influences, and is influenced by other fields. LDA for this type of analysis is appropriate because it leverages the model's strengths in identifying latent topics within large, multidisciplinary datasets. It captures data science's complex and interconnected nature, providing valuable insights into its influence,

development, and conceptualisation in scholarly discourse. To our best knowledge, no such research has been published.

2. Research objectives and methodology

Our research questions are as follows:

- Q1. How has the number of publications on data scientists evolved over time?
- Q2. How are papers regarding data scientists distributed over different fields of study?
- Q3. In what context are data scientists represented in the scientific literature?

To answer these questions, we used text mining. It is a collection of techniques designed to recognise patterns within unstructured and semi-structured textual data. It aims to uncover previously undiscovered knowledge (Fan et al., 2006). Exploring patterns in the scientific literature often involves topic modelling. The fundamental concept behind topic modelling revolves around developing a probabilistic generative model for a collection of textual documents. In topic modelling, documents are conceived as blends of topics, where each topic represents a probability distribution across words (Thakur & Kumar, 2022). Our methodology involves performing topic modelling on scientific abstracts to identify topics that can be discerned within scientific literature on data scientists. Automated methods, of course, come with inherent limitations. An evident drawback is the lack of control over the quality of the data being analysed. The potential for incorporating unsuitable data into the analysed dataset is ever-present due to the nature of automated data extraction methods, particularly when dealing with extensively unstructured resources. To circumvent the complexities associated with data and feature extraction from online sources, we utilised the Semantic Scholar database, which can be accessed through the Semantic Scholar Academic Graph API (S2AG). The documents were also automatically gathered, but the architecture of Semantic Scholar facilitates further processing by design (Kinney et al., 2023).

Semantic Scholar is based on an advanced data processing system that consistently acquires documents and metadata from various sources. Semantic Scholar collaborates with over 50 publishers, data providers, and aggregators, integrating content from more than 500 academic journals, university presses, and scholarly societies worldwide. Notable partners include the Association for Computational Linguistics, ACM, arXiv, BioOne, bioRxiv, BMJ Journals, University of Chicago Press, CiteSeerX, Clinical Trials Transformation Initiative, DBLP, De Gruyter, Frontiers, HAL, HighWire, IEEE, Karger, medRxiv, Microsoft, Papers With Code, Project MUSE, PubMed, SAGE Publishing, Science, Scientific.Net, SciTePress,

Springer Nature, SPIE, SSRN, Taylor & Francis Group, MIT Press, The Royal Society Publishing, Wiley, and Wolters Kluwer. These partnerships enhance the discoverability of scholarly content and provide valuable insights into how researchers engage with academic materials (Semantic Scholar, n.d.).

This system extracts text and metadata, standardises and clarifies details such as authors, institutions, and venues, categorises the subject area of each paper, produces a textual overview of its significant findings, and carries out additional functions. The Semantic Scholar database encompasses over 200 million articles, approximately 80 million authors, and around 550,000 publication venues (Kinney, 2023). This breadth of content renders the database extensive and provides comprehensive coverage of scientific resources.

We requested access to the Semantic Scholar API key. Although we were granted access to make up to 100 requests per second, downloading a dataset of the scale we were targeting – potentially up to 200 million entries – posed significant logistical and temporal challenges. Specifically, at this rate, it would take approximately 23 days of continuous, uninterrupted data requests to retrieve the entire corpus. This limitation highlights several practical issues, including the risk of network interruptions or API service limitations, which could lead to incomplete data collection or require retries, further extending the retrieval timeline. Moreover, handling such a large volume of data presents challenges regarding data storage capacity, processing power, and data management during analysis. Given these constraints, we focused on defining a more targeted dataset using specific keywords and limiting the number of entries retrieved.

With the vast volume of available literature, a focused keyword approach allowed us to create a manageable and thematically relevant corpus while preserving analytical depth. Our initial approach involved employing the keyword "data scientist" as a search query, as manual checks indicated that the volume of results for "data scientist" was the same as for "data scientists." We recognise that keyword dependence may inadvertently exclude some related studies. However, given that data science remains a relatively novel and niche topic, we decided to download the 10,000 most relevant entries for the keyword "data scientist" per year. We believe that the specificity of "data scientist" minimises ambiguity, enabling a more focused analysis aligned with the study's objectives. Thus, while keyword dependence may introduce bias, it also reveals valuable insights into the disciplinary contexts, research focus, and evolving engagement with data science across various fields. We gathered publications spanning from January 2005 to August 2023. The obtained database contained 188,066 entries for further analysis.

The following inclusion criteria for entries to the corpus were established:

(1) An entry must have a non-empty abstract.
(2) An entry must contain the phrase „data scientist" in the title or the abstract.
(3) An entry must be associated with a publication venue in some way – the

field venue or publication_venue has to be non-empty.
(4) An entry must be recognised as written in English.

Feature extraction of the data was provided by Semantic Scholar. We decided to select the following features for the analysis and filtering: paperId (an identifier of a paper), title, abstract, fieldsOfStudy, publicationTypes, publicationVenue (an identifier of a journal), venue (journal name) and year of publication. After filtering, duplicates were removed from the corpus.

For the analysis and processing, we used Python language with specific libraries. For text pre-processing and analysis, the following libraries were used:

(1) re for text cleaning,
(2) nltk for tokenisation and stop-words cleaning,
(3) spacy for lemmatisation,
(4) wordcloud for data visualisation,
(5) langdetect for language detection.

For LDA analysis, we used:

(1) re for text cleaning,
(2) nltk for tokenisation and stop-words cleaning,
(3) spacy for lemmatisation,
(4) wordcloud for data visualisation,
(5) langdetect for language detection.

The pipeline for analysis was taken in the following steps:

(1) Database acquisition from Semantic Scholar API (188 066 most relevant entries to keyword "data scientist").
(2) Filtering by inclusion criteria (1–3) mentioned above, performed on lowercased abstracts and titles in order to gather all relevant data (but the abstracts were saved with capitalisation for further analysis).
(3) Language identification and filtering out non-English publications.
(4) Lowercasing abstracts to avoid distinguishing words with the same meaning.
(5) Word filtering:
    a. Removing "-" in the middle of words to preserve words so they would not be treated as separate tokens.
    b. Removing one-character words.
    c. Removing numbers and special characters.
    d. Removing stop-words (most used words in English) to exclude words that occur most frequently and create unnecessary noise in the data.
(6) Lemmatization – aggregating various grammatical forms of a word to treat it as a single entity, denoted by the word's lemma or its base form as found in a dictionary.
(7) Removing extra stop-words (data, science, etc.) to eliminate highly common words often associated with data scientists, which could introduce unwanted noise.

(8) Counting total word occurrences to determine other potential stop-words and create a word cloud.
(9) Tokenization (splitting text into separate words).
(10) n-gram counting (n = 2,3) to uncover the most common bigrams and trigrams.
(11) Joining meaningful n-grams as a single token to preserve tokens with separate meanings, eg., machine learning – machinelearning.
(12) Token filtering to avoid noise in the data:
   a. Removing tokens that occurred less than five times in the corpus.
   b. Token has to occur in at least 3 unique documents.
   c. The token has to be longer than 3 characters or be included in a list of meaningful tokens (such as ml, nlp, or ai).
(13) Deleting duplicate entries from the database to avoid using the identical article metadata in the analysis more than once.
(14) Performing LDA with sklearn.
(15) Visualizing the results with pyLDAvis.
(16) Iterative experimenting with the number of clusters to collaboratively find the number most suitable for interpretation.
(17) Qualitative interpretation of the topics.

3. Results

Through the process of filtering the initial database by removing entries with empty abstracts or empty publication venues, we gathered a collection of 76,817 scholarly metadata files. We then applied an additional condition, requiring each entry to contain the phrase "data scientist" in either the abstract or the title, resulting in a refined database of 2,239 entries. As Semantic Scholar categorises its entries into fields of study, the distribution of documents within the corpus we created for this research is depicted in Table 1.

Table 1. Document count by field of study in relation to the volume of the whole corpus.

field of study	# of docs with the phrase "data scientist"	total # of docs in the corpus	% of docs with the phrase "data scientist"
Computer Science	1654	43034	3.84
Medicine	339	25355	1.34
Not Assigned	299	7031	4.25
Mathematics	91	3578	2.54
Engineering	86	2415	3.56
Sociology	47	1466	3.21

field of study	# of docs with the phrase "data scientist"	total # of docs in the corpus	% of docs with the phrase "data scientist"
Business	31	1295	2.39
Biology	29	5637	0.51
Psychology	29	2685	1.08
Political Science	25	1411	1.77
Physics	16	1229	1.30
Geography	14	1571	0.89
Geology	3	550	0.55
Economics	3	544	0.55
Art	2	87	2.30
Materials Science	1	375	0.27
History	1	264	0.38
Philosophy	0	78	0.00

Note: A paper may have one, none or multiple fields of study assigned. Source: self-authored.

We encountered the overrepresentation of Computer Science publications in our corpus. It was expected as it reflects the historical and disciplinary roots of data science, primarily anchored in computational and technical domains. Conversely, fewer publications in such fields of studies as Social Studies, Geology and Philosophy may stem from different terminologies or less frequent engagement with the explicit term "data scientist".

A significant proportion of publications in our corpus belong to the domain of Computer Science (1654), comprising 3.84% of the total publications in this field. The field of Medicine follows in terms of the number of publications, though with a significantly lower document count (339). This trend is unsurprising, given the widespread integration of technical advancements and AI solutions in medical research and diagnostics (Lai et al., 2021). However, it is important to note that the database may exhibit a bias due to the potential underrepresentation of papers from other fields of study. This bias could be attributed to Semantic Scholar's original focus as a database for computer science, geoscience, and neuroscience, which only expanded to include biomedical literature starting in 2017 (Fricke, 2018). Despite this, the proportion of entries containing the term "data scientist" in relation to the entire corpus remains relatively low (1.34%) compared to fields like Mathematics (91 publications/2.54%), which, despite having fewer total publications, demonstrates better relative representation. Engineering has 86 publications, representing 1.08%. Other fields exhibit very low numbers (<50) of abstracts or titles containing the phrase "data scientist". This distribution of documents across various fields reveals a scarcity of research on data scientists within socio-economic and socio-cultural contexts. It highlights a significant research gap in this area.

Numbers of publications per year are displayed in Table 2. The first publication involving data scientists in the abstract or title was published in 2008. The number of publications started to rise gradually in 2012, which seems related to the beginning of "the era of big data" which started in this period (Floridi, 2014). Interest in the subject was the highest in 2020. In 2022, the number of publications noticeably decreased, but in August 2023, there were 237 publications, and it is reasonable to expect that it has risen by the end of the year.

Table 2. Number of publications per year.

Year	**#**
2008	1
2009	3
2010	2
2011	4
2012	11
2013	39
2014	52
2015	85
2016	133
2017	188
2018	243
2019	308
2020	356
2021	349
2022	255
2023	237

Source: self-authored.

By performing the steps mentioned in the "Research objectives and methodology" section, we acquired outcomes displaying word cloud visualising the 50 most frequent words (Figure 1), 25 of which are presented in Figure 2. As anticipated, the most prevalent term prior to lemmatisation and word filtration is "data", followed by "scientist". To enhance the relevance of the analysis, we decided to exclude these words by generating additional customised stop words.

Another step in the analysis was lemmatising the vocabulary and performing *n*-gram counting to uncover prevalent co-occurring phrases. This process offered valuable insights by helping to create supplementary tokens that encapsulate meaningful expressions. The most common bigrams often refer to specific phenomena, such as artificial intelligence or machine learning, and therefore, they should not

be split into separate tokens. We created a list of such phrases as specific tokens. The results of *n*-gram counting are shown in Figures 3 to 5.

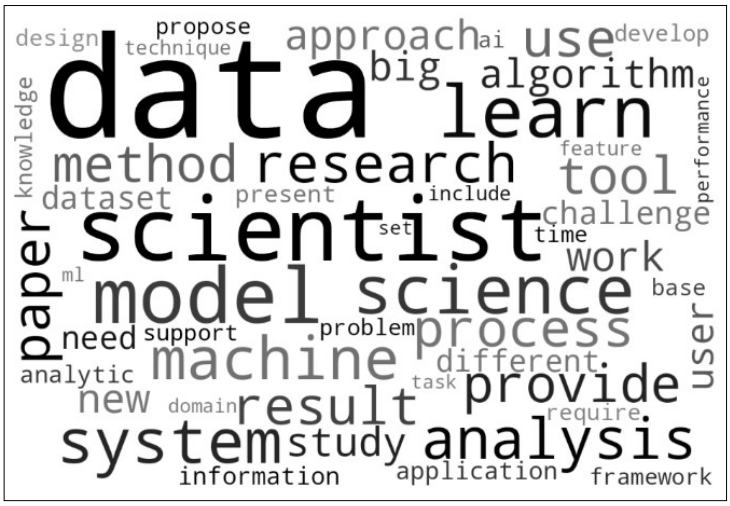

Figure 1. Word cloud of 50 most common words.

Source: self-authored.

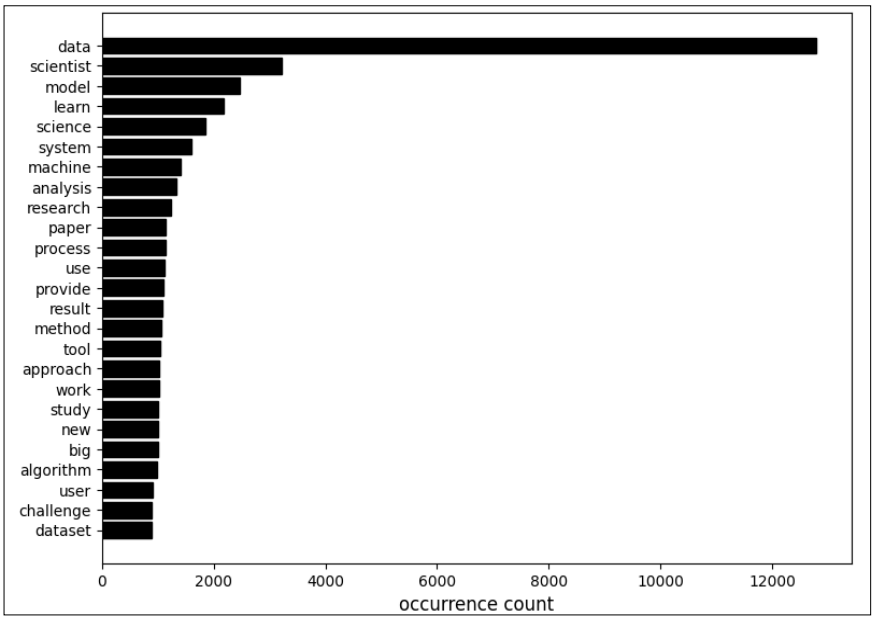

Figure 2. 25 most common words.

Source: self-authored.

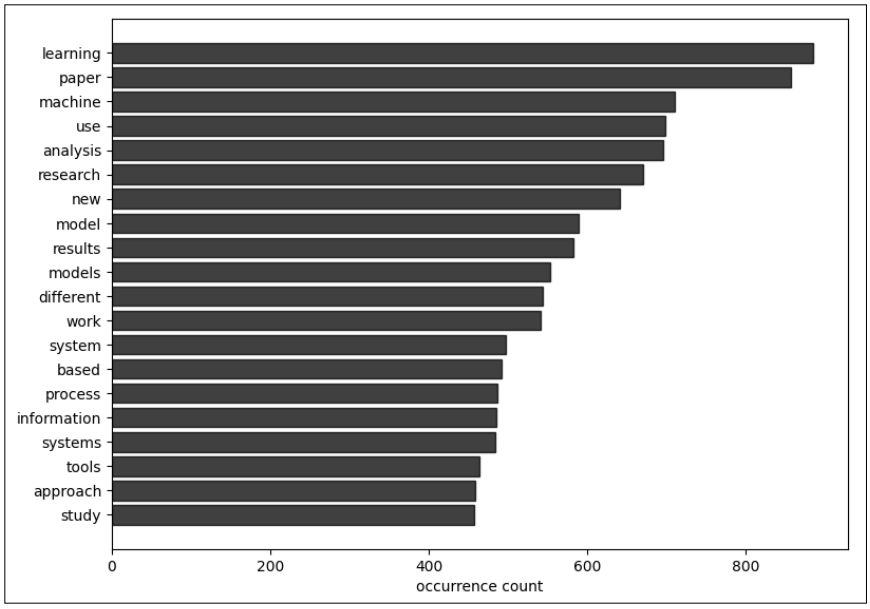

Figure 3. The 20 most common unigrams.

Source: self-authored.

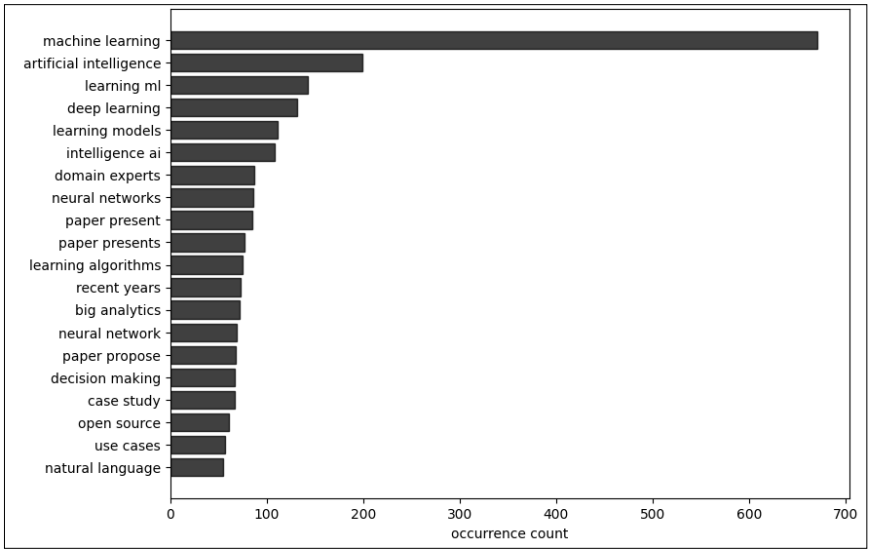

Figure 4. The 20 most common bigrams.

Note: Some bigrams include acronyms such as "AI" or "ML," which may not appear naturally in the text but are instead artefacts of automatic preprocessing.

Source: self-authored.

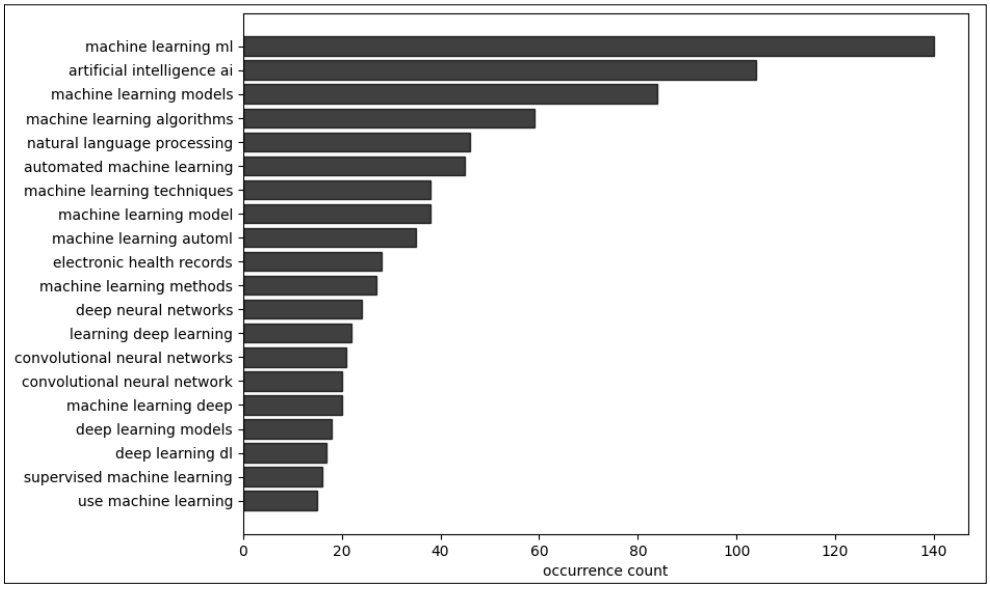

Figure 5. The 20 most common trigrams.

Note: Some trigrams include acronyms such as "AI" or "ML," which may not appear naturally in the text but are instead artefacts of automatic preprocessing.

Source: self-authored.

To make LDA work effectively, careful token filtering is necessary. We decided that a token should appear at least 5 times in the corpus to be worth considering. Also, for a token to be relevant, it has to appear in at least 3 documents, has to be longer than 3 characters or be part of the list of specific, meaningful short expressions, such as "ml" (*machine learning*), "ai" (*artificial intelligence*) or "nlp" (*natural language processing*). This approach aimed to filter out unnecessary acronyms while keeping the meaningful ones.

Following token filtering, we proceeded to perform Latent Dirichlet Allocation topic clustering. LDA is a generative probabilistic model that defines a topic as a distribution of words. Within this framework, each document in the corpus is a mixture of topics, and each topic is a mixture of words from the entire corpus vocabulary. More precisely, for each topic, a non-negative probability is assigned to each word from the vocabulary, and each document is a convex combination of topics (Blei et al., 2003).

The number of topics (clusters) is chosen arbitrarily. When selecting the optimal number of clusters or (topics) researchers have a range of quantitative and qualitative methods at their disposal, depending on the character of the problem. David Blei, author of the LDA original paper, states in another article, "The standard for selecting a solution is not so much accuracy as a *utility: Does the model*

simplify the data in a way that is interpretable, passes tests of internal and external validity, and is useful for further analysis?" (Blei & Lafferty, 2009). This highlights that practical interpretability and usefulness should often take precedence over rigid accuracy metrics. Therefore, determining the optimal number of clusters for this type of study relies on the researcher's qualitative assessment rather than a prescribed heuristic (Wiedemann, 2016). However, interpretations must be approached cautiously, relying on subject-area specialists on the team (DiMaggio et al., 2013). Because we are a multidisciplinary team comprising three individuals with diverse backgrounds, including two researchers with experience in human and social sciences and a machine learning student with professional data science expertise, we adopted a process of collaborative, iterative experimentation to determine the number of clusters. Through this process, we arrived at a selection of 20 clusters, a decision that emerged as the most harmonious fit with the dataset's content, demonstrating a coherent and meaningful structure. This choice reflects both qualitative and data-driven considerations, ensuring a robust and insightful interpretation of the data.

The coverage *C(t)* of a topic *t* is defined as follows:

$$C(t) = \frac{\sum_d (d) \cdot p(t \mid d)}{\sum_{t'} \sum_d (d) \cdot p(t' \mid d)}$$

where $|d|$is a document length (in tokens) and $p(t|d)$ is a measure of assignment of a document *d* to a topic *t.* This measures how large a portion of documents in a corpus is captured by the topic. The LDA model allows for the adjustment of the term's relevance, which can help synthesise the idea behind a topic. Siever & Shirley (2014) defined the relevance as follows:

Let ϕ_{kw} denote the probability of term $w \in \{1, ..., V\}$ for topic $k \in \{1, ..., K\}$, where *V* denotes the number of terms in the vocabulary, and let p_w denote the marginal probability of term *w* in the corpus. The relevance of term *w* to topic *k* given a weight parameter λ(where $0 \leq \lambda \leq 1$) is defined as:

$$r(\mathrm{w}, \mathrm{k}, \lambda) = \lambda \log (\phi_{kw}) + (1-\lambda) \log \left(\frac{\phi_{kw}}{p_w} \right)$$

where λ determines the weight given to the probability of term *w* under topic *k* relative to its lift, which is the ratio of a term's probability within a topic to its marginal probability across the corpus. Setting $\lambda = 1$ results in ranking terms in decreasing order of their topic-specific probability, and setting $\lambda = 0$ ranks terms by their lift (Siever & Shirley, 2014).

We decided to include results for $\lambda = 1$ and $\lambda = 0.5$. The value of represents a balance between words with a high probability of occurrence in the topic (which

may also appear frequently in other topics) and words that are more distinctive to the chosen topic. This approach can be particularly advantageous when the words with the highest probability are overly general, making it challenging to uncover the underlying theme of a topic.

As a result of our experimentation with a number of clusters, we uncovered 20 clusters (topics), which we labelled and assigned to 5 different categories. Some topics fit more than one category. One topic (Graphical Data and Security) was excluded from the analysis because of non-coherent words and weak coverage (2.4%). Results of LDA topics modelling on 20 clusters and words $\lambda = 1$ and $\lambda = 0.5$ are presented in Table 3.

Table 3. LDA topics, relevant words and coverage in the corpus.

topic summary	words with $\lambda = 1$	words with $\lambda = 0.5$	coverage
Big Data Analytics in Healthcare This topic is focused on big data research in healthcare, the application of AI models for knowledge discovery in medical contexts and its challenges.	big, research, clinical, health, analytics, analysis, methods, tools, ai, challenges, studies, use, information, knowledge, healthcare, researchers, including study, business, social	clinical, big, health, studies, research, analytics, healthcare, challenges, medicine, social media, care, including, researchers, methods, review, risk, tools, scientific, ai, artificial intelligence	11.6%
Systems, Databases and Scalability This topic is focused on large-scale systems, databases, and addresses issues related to scalability.	systems, query, system, analysis, processing, large, analytics, users, queries, time, graph, distributed, database, different, performance, python, algorithms, applications, user, big	query, queries, graph, distributed, processing, database, python, execution, systems, large, spark, languages, users, system, exploration, scalable, parallel, interactive, analytics, apache	11.1%
ML: Classification and Prediction This topic is focused on machine learning models, feature selection and engineering, prediction and classification.	models, machine learning, model, feature, prediction, algorithms, results, features, accuracy, dataset, learning, methods, system, process, different, predictive, datasets, ml, performance, classification	models, feature, prediction, accuracy, classifiers, machine learning, features, predictive, model, dataset, selection, algorithms, feature engineering, classification, classifier, regression, results, learning, random, schema	7.9%

topic summary	words with λ = 1	words with λ =0.5	coverage
ML: Automation and Pipelines This topic is focused on machine learning automation, pipeline development, and documentation.	ml, machine learning, automl, model, system, pipelines, process, pipeline, time, models, code, approach, systems, learning, automated, performance, documentation, solutions, tools, support	ml, automl, machine learning, pipelines, pipeline, documentation, automation, automated machine, hyperparameter, ml models, sales, automated, model, code, tuning, drilling, metalearning, cleaning, system, automate	6.1%
Privacy Preserving The topic is focused on privacy concerns, data analysis, and the application of technology to manage sensitive information.	esearch, privacy, social, big, analysis, information, paper, work, network, management, based, applications, storage, networks, access, technologies, methods, use, datasets, questions	privacy, social, network, big, research, storage, networks, tensor, information, differential, privacy preserving, journalists, sensitive, paper, work, secure, analysis, qualitative, management, topological	5.7%
Deep Learning and Image Classification The topic is focused on deep learning, including image classification, (convolutional) neural networks, and performance evaluation.	deeplearning, model, machine learning, analysis, images, classification, datasets, models, performance, accuracy, neural network, techniques, methods, based, image, paper, approaches, tasks, dataset, neural networks	deep learning, images, neural network, image, convolutional, classification, neural networks, segmentation, deep, datasets, accuracy, trained, speech, imaging, encoding, model, bias, machine learning, performance, chat gpt	5.6%
Algorithms and Statistical Methods The topic is focused on algorithms, mathematical tools and probabilistic analysis, including machine learning methods.	algorithm, model, learning, machine learning, results, paper, based, matrix, analysis, work, dataset, different, process, approach, study, time, experiments, techniques, methods, method	matrix, matrices, algorithm, markov, reduction, experiments, stochastic, kernel, breast, linear, probability, projection, india, scenario, educational, estimation, learning, summaries, transition, dimension	5.1%
COVID Pandemic The topic is focused on the COVID-19 pandemic, including health analysis and disease detection in the context of AI.	covid, pandemic, health, detection, different, study, people, mining, use, model, ai, work, analysis, public, approach, based, results, important, information, process	covid, pandemic, detection, professions, coronavirus, screening, healthy, health, pregnant, seizure, people, vaccine, said, spread, mining, infectious, population, chest, interventions, covidnet	4.8%

topic summary	words with λ = 1	words with λ =0.5	coverage
Education and Skill Development The topic is focused on education, skill development, research projects, and programs to enhance learning.	students, research, education, skills, learning, university, paper, course, project, training, information, programs, help, statistics, student, model, provide, need, article, fairness	students, education, course, university, student, skills, courses, curriculum, research, graduate, universities, project, programs, programme, teaching, fairness, college, institutions, learning, statistics	4.8%
Profession, Job Requirements and Roles The topic is focused on job requirements, technology utilisation, and the roles of engineers in creating technological solutions in companies.	ai, job, engineers, software, systems, design, technologies, development, companies, artificial intelligence, use, analysis, work, technology, need, requirements, big, research, roles, process	ai, job, engineers, software, companies, technologies, artificial intelligence, roles, design, designers, software engineering, fair, jobs, requirements, trust, systems, transport, technology, development, company	4.5%
Notebooks and Programming Methods The topic is focused on code notebooks, programming tools, and explainability in computational analysis.	notebooks, methods, research, notebook, programming, different, model, python, explainability, computational, jupyter, software, design, models, framework, development, code, explanations, systems, tools	notebooks, notebook, explainability, jupyter, programming, explanations, toolkit, python, serverless, serving, book, explanation, computational, adaptive, readers, methods, coding, metrics, software, pruning	4.5%
Quality Assessment, Effectiveness and Transparency The topic is focused on quality assessment in various applications, using algorithms and statistical methods in the context of challenges and transparency.	quality, different, systems, models, machine learning, use, algorithms, methods, model, challenges, approach, chapter, paper, time, statistical, real, transparency, solutions, problems, management	quality, chapter, trading, traffic, transparency, book, volatility, periodic, feminism, taxonomy, army, production, forecasts, concerns, real, discussed, road, regularisation, coherent, reader	4.4%

topic summary	words with λ = 1	words with λ =0.5	coverage
Tools for Business Analytics The topic is focused on tools and technologies' applications in business. It covers big data analytics, business needs, and collaboration within organisations.	tools, business, big, analytics, different, process, paper, analysis, research, need, organisations, work, challenges, learning, information, collaboration, article, social, use, knowledge	blockchain, business, tools, organisations, collaboration, unstructured, big, content, analytics, theories, organisational, competencies, centre, big analytics, shared, alternatives, face, article, spreadsheets, behaviour	4.4%
Healthcare Informatics and Patient Care The topic is focused on healthcare informatics, patient care, medical data, and digital solutions in the field.	healthcare, clinical, research, health, medical, care, informatics, patients, systems, information, system, big, patient, medicine, knowledge, development, computer, digital, group, team	healthcare, informatics, clinical, medical, care, patients, health, insurance, medicine, vehicle, group, national, radiation, patient, research, biomedical, nursing, translational, nurses, collaborations	3.7%
Health Information Management and Privacy The topic is focused on health information management, privacy concerns in healthcare systems.	health, information, results, management, privacy, development, model, framework, skills, research, digital, use, analysis, AI, design, patient, training, performance, platform, study	centres, phishing, privacy, residents, firm, health, digital, composition, patient, plant, management, skills, ehealth, radiology, aiml, physicians, twin, personalised, leadership, safety	3.7%
Stock Prices Forecasting The topic is focused on business analytics, forecasting stock prices, data visualisation, including predictive modelling.	model, training, models, based, business, machinelearning, paper, framework, digital, visualisation, process, stock, analytics, problems, approach, big, research, solutions, platforms, prediction	model, training, stock, prices, bidaml, lstm, oil, tweets, models, arima, forecasting, rmse, auto, firms, business, sports, geosparkviz, cnns, digital, timeseries	2.9%
Biomedicine: Cancer and Molecular Data Analysis The topic is focused on computational analysis of cancer, genes and other molecular data using computational frameworks.	code, framework, package, available, features, https, analysis, cancer, pipeline, cell, best, opensource, time, dataset, complex, networks, computational, graph, results, brain	package, cell, code, https, embeddings, cancer, molecular, brain, tpot, expression, cells, antipatterns, variants, genes, gene, framework, intensity, pvldb, pipeline, embedding	2.9%

topic summary	words with λ = 1	words with λ =0.5	coverage
Graphical Data and Security (Excluded) This topic was excluded from the analysis because of non-coherent words.	model, users, analysis, system, security, information, packages, function, common, paper, values, results, table, object, functions, user, ggplot, approach, graphical, code	ggplot, packages, graphical, security, plotting, table, cyber, object, columns, iris, function, scripts, graphics, plot, textual, cray, plugin, flood, values, plots	2.4%
Ethics The topic is focused on ethics, environmental impact, reporting, and the societal role of technology.	ethics, ethical, climate, field, statistics, impact, issues, society, different, reporting, big, work, biodiversity, computing, social, community, interdisciplinary, need, develop, cloud	ethics, climate, ethical, biodiversity, reporting, official, society, athletes, carbon, epistemic, impacts, accreditation, session, civiliser, literacy, exercise, whatif, mlai, heat, arise	2.2%
IoT, Devices and Malware Analysis The topic is focused on IoT, benchmark devices and malware analysis.	IoT, malware, programming, etal, based, use, visual, devices, statistical, interactive, metadata, users, time, scheme, need, provide, approach, dataset, user, prototype	malware, iot, scheme, labelling, etal, compliance, array, chart, visual, devices, recovery, prototype, spam, plotly, expressions, programming, benchmarks, citation, skill, metadata	2.2%

Source: self-authored.

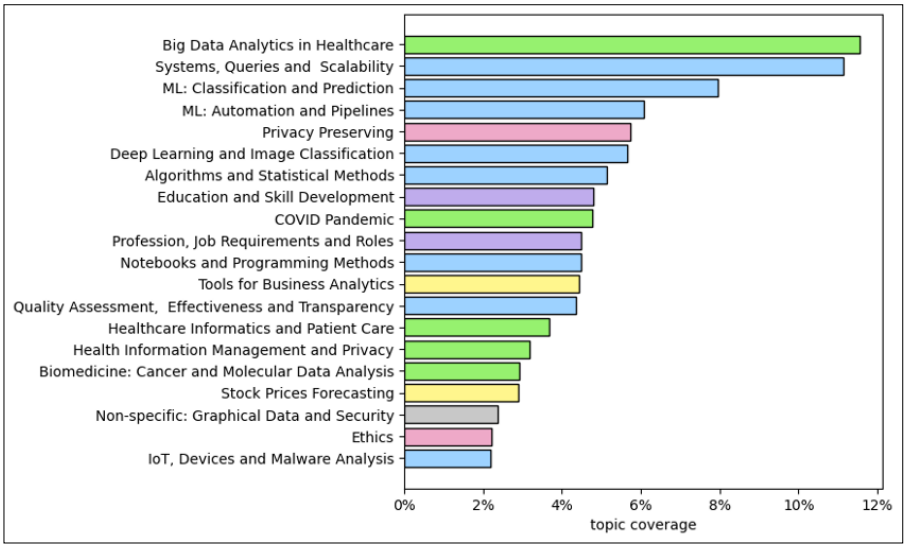

Figure 6. The coverage of topics.

Source: self-authored.

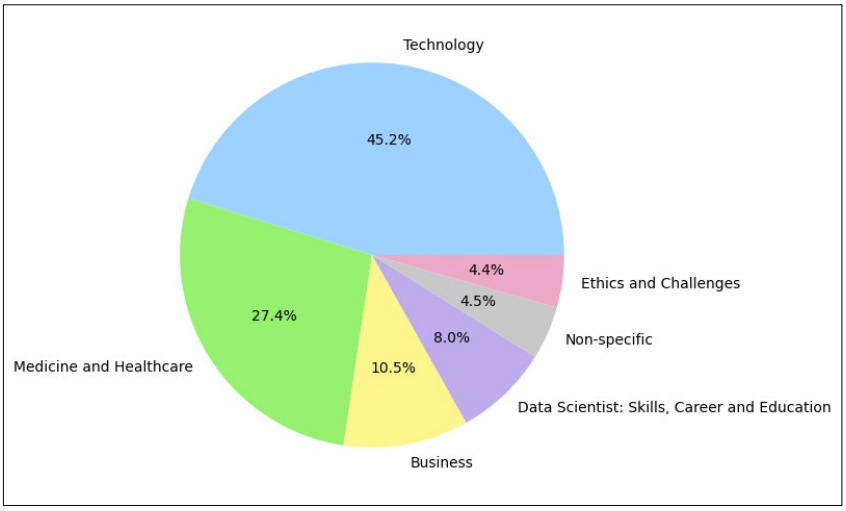

Figure 7. The coverage of topics grouped into categories.

Source: self-authored.

The topic categories are as follows:

(1) Information Technology (8/20 topics):
- Systems, Databases and Scalability,
- Machine Learning: Classification and Prediction,
- Machine Learning: Automation and Pipelines,
- Deep Learning and Image Classification,
- Algorithms and Statistical Methods,
- Notebooks and Programming Methods,
- Quality Assessment, Effectiveness and Transparency,
- IoT, Devices and Malware Analysis.

The Information Technology category covers 8 of the 20 topics and delves into various crucial technological aspects shaping data science. Systems, Databases, and Scalability relate to architecture issues essential to the work of data scientists. Topics such as Machine Learning: Classification and Prediction and Machine Learning: Automation and Pipelines highlight issues related to artificial intelligence. Topics like Deep Learning, Image Classification and Algorithms, and Statistical Methods explore the advanced techniques used in data analysis. Notebooks and Programming Methods emphasise the role of programming tools, while Quality Assessment, Effectiveness, and Transparency touch on issues related to data analysis and algorithms challenges. Lastly, IoT, Devices, and Malware Analysis showcases how data science extends into emerging fields, revealing its versatile applications.

(2) Medicine and Healthcare (Topics 5/20):
- Big Data Analytics in Healthcare,

- COVID Pandemic,
- Healthcare Informatics and Patient Care,
- Health Information Management and Privacy,
- Biomedicine: Cancer and Molecular Data Analysis.

The Medicine and Healthcare category, covering 5 out of the 20 topics, delves into critical aspects at the intersection of data science and healthcare. Big Data Analytics in Healthcare and Healthcare Informatics and Patient Care underscore how data-driven insights enhance healthcare delivery. Health Information Management and Privacy refer to the sensitive realm of protecting patients' data. The analysis uncovered that data scientists had an impact on the COVID-19 pandemic and vaccine development. The category also includes Biomedicine: Cancer and Molecular Data Analysis, spotlighting data scientists' contribution to understanding complex diseases and their diagnostics. This category showcases the profound influence data scientists have on improving healthcare outcomes. It also emphasises that data scientists make significant contributions to medicine-related research. The pivotal role of data science in biomedical research, facilitated by artificial intelligence tools, significantly enhances knowledge production and scientific advancement in this field (Lai et al., 2021). Nonetheless, this progress is accompanied by many ethical concerns and challenges, particularly privacy and security (Krumholz, 2014).

(3) Ethics and Challenges (Topics 4/20):
- Ethics,
- Privacy Preserving,
- Health Information Management and Privacy,
- Quality Assessment, Effectiveness and Transparency.

The Ethics and Challenges category, comprising 4 of the 20 topics, delves into essential dimensions of ethical considerations within data science. Topics like Ethics relate to ethical dilemmas that arise when handling data. Privacy Preserving underscores the importance of safeguarding individual privacy while extracting insights from data. Quality Assessment, Effectiveness, and Transparency also highlight the ongoing pursuit of maintaining data science practices' quality, effectiveness, and transparency. This category highlights data science practitioners' ethical and practical challenges in pursuing responsible and impactful data-driven decision-making supported by interpretable algorithms. This is coherent with numerous ethical issues addressed toward big data in various contexts, such as data management (Nair, 2020), health research (Rothstein, 2015) and privacy and security preservation (Joshi, 2020).

(4) Business (Topics 3/20):
- Tools for Business Analytics,
- Stock Prices Forecasting,
- Profession, Job Requirements and Roles.

The Business category, encompassing 3 out of the 20 topics, delves into key aspects of data science within a business context. Topics such as Tools for Business Analytics stress the significance of data-driven tools in shaping business decisions. Stock Prices Forecasting explores the application of data science in predicting financial market trends. Additionally, Profession, Job Requirements, and Roles shed light on the evolving landscape of data science roles within the business sphere and their role in teams and collaboration with domain experts. This category showcases how data science is leveraged to inform business strategies, which aligns well with studies assessing the profound impact of data science on business (Mishra, 2021).

(5) Data Scientist: Skills, Career and Education (Topics 2/20):
   – Education and Skill Development,
   – Profession, Job Requirements and Roles.

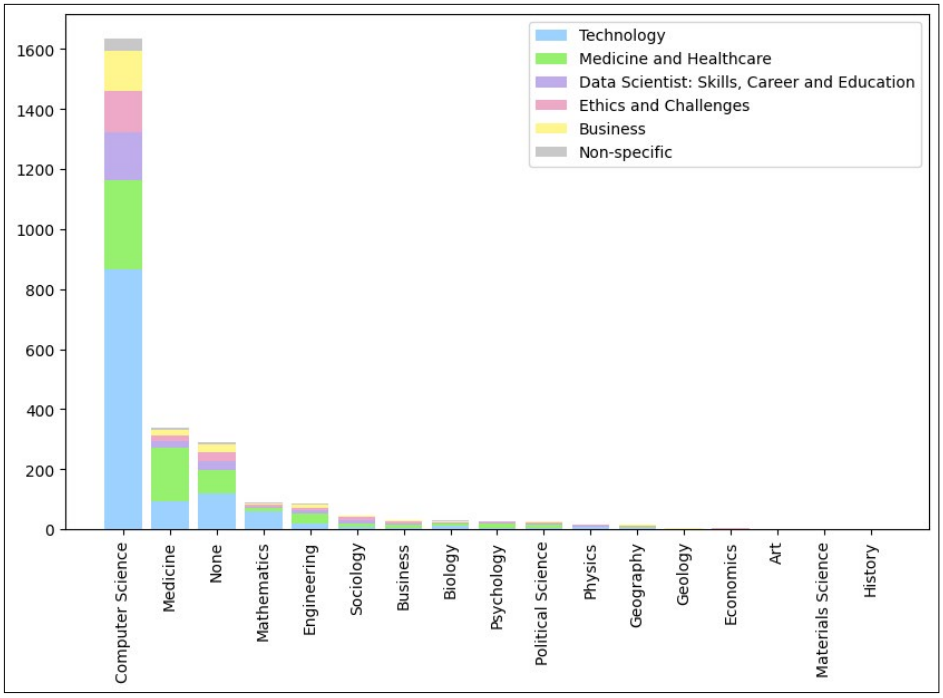

Figure 8: Topic categories in different fields of studies. Since each topic is a mixture of topics, the y-axis represents a weighted combination of the number of occurrences and topic shares.

Source: self-authored.

The Data Scientist: Skills, Career, and Education category encompasses 2 of 20 topics. Topics such as Education and Skill Development put on the spot the pathways and skills required for a data science career. Profession, Job Requirements,

and Roles delve into the dynamic roles and evolving requirements within the data science profession. As was said before, professionals in this field significantly impact business, and employment in this field is rising (Mishra, 2021). This category offers insights directly into the social world of data scientists, regarding them as distinct subjects of study.

After grouping topics into categories, we investigated their overlap with the documents' fields of study as a form of external validation for the number of clusters. The results were satisfactory, as the distribution of topics reasonably matched the assigned fields of study. Topics related to Technology covered over half of the publications categorised under Computer Science, indicating strong alignment. Similarly, publications within the field of Medicine predominantly covered topics related to Health and Medicine. Furthermore, Technology-related topics showed substantial relative coverage in the field of Mathematics, whereas Engineering appeared to be more heavily influenced by topics related to Medicine and Health.

4. Study limitations

While extensive and comprehensive, using a corpus built from the Semantic Scholar database presents several limitations that must be considered when interpreting results. While Semantic Scholar extensively covers many disciplines, some fields or subfields may be less thoroughly represented. Moreover, recommendations and suggestions offered by Semantic Scholar based on artificial intelligence may reflect algorithmic biases, potentially favouring certain types of content.

One key limitation in constructing our database is keyword dependence, as the corpus relies on the search term "data scientist". This approach may inadvertently exclude relevant studies that use synonymous or related terms, thus introducing bias. This leads to challenges with generalizability, as findings may not extend well to broader or related areas of data science, particularly those using different terminology or less common phrases. Finally, evolving terminology poses a challenge, as the meaning and context of terms like "data scientist" have likely changed from 2005 to 2023, potentially affecting the interpretation of trends and roles captured in the corpus. These limitations underscore the need for caution and contextual awareness when analysing such data.

The method used for analysis also comes with limitations. The "bag of words" approach used in Latent Dirichlet Allocation has limitations due to its simplification of text data. This approach does not consider the order or arrangement of words in a document, which can impact the interpretation of topics and overlook nuances in meaning as it treats each word as independent of its neighbouring words, disregarding the inherent sequential or contextual information present in natural language. This can lead to a loss of meaning, as words' meanings often depend on

their surrounding context. In this approach, each document is represented solely by the frequency of words, ignoring other valuable features like sentence structure, document length, or other text features.

There are several problems with analysing abstracts with LDA. Abstracts typically have limited text length, often consisting of only a few sentences or paragraphs. Due to this brevity, statistical methods like LDA can be susceptible to noise, resulting in accidental words or proper names being incorrectly attributed as highly relevant to a topic by the model. These terms may be coincidental or have low overall significance for the meaning of the entire document. Another consideration is that in cases where the corpus of abstracts is limited, rare or emerging topics might not have enough occurrences to generate coherent topics in LDA, resulting in their underrepresentation. Also, abstracts tend to follow specific language patterns, making them relatively homogeneous. This homogeneity can lead to LDA identifying topics aligned with generic scientific discourse rather than capturing more specific content.

To enhance the quality and relevance of the results, conducting a coverage comparison with alternative databases would be beneficial.

5. Summary

The analysis has shown a consistent upward trajectory in the number of publications centred on data scientists since 2008. The peak was observed in 2020, with the total number of publications being $n = 356$. With a minor regression observed in 2022 ($n = 255$), data scientists are still an area of interest in scientific literature, reaching 237 publications in August 2023.

A plethora of publications regarding data scientists reside within the domain of Computer Science ($n = 1654$). The second field is Medicine (339). A substantial portion of the corpus entries (299) lacks ascribed fields of study, constituting the third most prevalent category in the subject. Mathematics, Engineering, and other fields show a modest presence, while other disciplines exhibit minimal representation.

The study unveils an extensive landscape of literature that delves into the Information Technology category. Also, a distinct link between data scientists' works and the realm of biomedical research was found. This connection can be observed through various subfields, such as cancer genomics, patient data management, and the data-driven response to the COVID-19 pandemic. Moreover, a robust connection between data scientists and the business sector is evident. The documents within the corpus address an array of themes, ranging from data science applications in business intelligence to the roles of data scientists in teams in work environments. Ethical dilemmas and challenges arising from the proliferation of big data are prominently featured within the literature concerning data scientists. These

discussions encompass concerns regarding data privacy preservation, spanning diverse contexts, including medical domains. Furthermore, code quality and algorithm transparency deliberations contribute to this ethical discourse. However, only two thematic categories make data scientists a central subject of study. They revolve around their professional roles and job requirements. There is also a thread regarding the courses and training for data scientists.

Generally, little literature discusses data scientists in a sociocultural context, with only a small number of publications within the field of Sociology (n = 47) and a lack of distinct topics on the subject. We consider this a striking gap in the literature because we believe it is important to study data scientists as social actors, given how much they shape knowledge and decision-making in various areas, such as medicine and business, as seen in this study. Another issue making data scientists an interesting subject in social sciences is the ethical implications of data collection, analysis, and use, which are critical in today's digital age. Sociology can explore how data scientists navigate ethical dilemmas related to privacy, consent, and bias, contributing to discussions on responsible data practices and regulations. Understanding their role can provide insights into how data-driven decisions shape societal dynamics and structures. They also create models that predict and explain human behaviour based on data patterns. Studying their methodologies can shed light on the underlying assumptions and biases that influence these models, thereby enhancing our understanding of how human behaviour is quantified and analysed.

References

Anupriya, P., & Karpagavalli, S. (2015). LDA-based topic modelling of journal abstracts. In *Proceedings of the 2015 International Conference on Advanced Computing and Communication Systems* (pp. 1–5). IEEE.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi: 10.1080/1369118X.2012.678878.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–94). Taylor & Francis.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Cho, K.-W., Bae, S.-K., & Woo, Y.-W. (2017). Analysis of topic trends and topic modelling of KSHSM Journal Papers using text mining. *The Korean Journal of Health Service Management*, *11*(4), 213–224. doi: 10.12811/kshsm.2017.11.4.213.

Chen, J., Wang, T. T., & Lu, Q. (2016). THC-DAT: A document analysis tool based on topic hierarchy and context information. *Library Hi-Tech*, *34*, 64–86.

Coelho Da Silveira, C., Marcolin, C. B., Da Silva, M., & Domingos, J. C. (2020). What is a data scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias*, *8*(1), 25–39. doi: 10.5585/iptec.v8i1.17263.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. doi: 10.1111/j.1751-5823.2001.tb00477.x.

Cortez, P., Moro, S., Rita, P., King, D., & Hall, J. (2018). Insights from a text mining survey on Expert Systems research from 2000 to 2016. *Expert Systems*, *35*(3), e12280. doi: 10.1111/exsy.12280.

Cho, K.-W., Bae, S.-K., & Woo, Y.-W. (2017). Analysis of topic trends and topic modelling of KSHSM Journal Papers using text mining. *The Korean Journal of Health Service Management*, *11*(4), 213–224. doi: 10.12811/kshsm.2017.11.4.213.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. doi: 10.1080/10618600.2017.1384734.

Dayeen, F. R., Sharma, A. S., & Derrible, S. (2020). A text mining analysis of the climate change literature in industrial ecology. *Journal of Industrial Ecology*, *24*(2), 276–284. doi: 10.1111/jiec.12998.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. doi: 10.1016/j.poetic.2013.08.004.

Ebrahimi, F., Dehghani, M., & Makkizadeh, F. (2023). Analysis of Persian bioinformatics research with topic modelling. *BioMed Research International*, *2023*(1), 3728131. doi: 10.1155/2023/3728131.

Espinoza Mina, M. A., & Gallegos Barzola, D. D. P. (2019). Data scientist: A systematic review of the literature. In M. Botto-Tobar, G. Pizarro, M. Zúñiga-Prieto, M. D'Armas, & M. Zúñiga Sánchez (eds.), *Technology Trends* (Vol. 895, pp. 476–487). Springer International Publishing.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76–82.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Fricke, S. (2018). Semantic scholar. *Journal of the Medical Library Association: JMLA*, *106*(1).

García, D., Massucci, F. A., Mosca, A., Rafols, I., Rodrıguez, A., & Vassena, R. (2020). Mapping research in assisted reproduction worldwide. *Reproductive BioMedicine Online*, *40*(1), 71–81. doi: 10.1016/j.rbmo.2019.10.013.

Hazzan, O., & Koby, M. (2023). Data science as a research method. In O. Hazzan & M. Koby (eds.), *Guide to Teaching Data Science* (pp. 121–135). Springer International Publishing.

Ho, A., Nguyen, A., Pafford, J. L., & Slater, R. (2019). A data science approach to defining a data scientist. *SMU Data Science Review*, *2*(3).

Ismail, N. A., & Zainal Abidin, W. (2016). Data scientist skills. *IOSR Journal of Mobile Computing & Application*, *3*(4), 52–61. doi: 10.9790/0050-03045261.

Joo, S., Choi, I., & Choi, N. (2018). Topic analysis of the research domain in knowledge organization: A latent Dirichlet allocation approach. *Knowledge Organization*, *45*(2), 170–183. doi: 10.5771/0943-7444-2018-2-170.

Jeon, H. J., Kim, D. Y., Han, K. J., Han, D. W., Son, S. W., & Lee, C. M. (2018). An analysis of indoor environment research trends in Korea using topic modelling: Case study on

abstracts from the journal of the Korean Society for Indoor Environment. *Journal of Odor and Indoor Environment*, *17*(4), 322–329. doi: 10.15250/joie.2018.17.4.322.

Joshi, M. V. (2020). Security/privacy issues and challenges in big data. *International Research Journal of Engineering and Technology*, *07*(06).

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... & Weld, D. S. (2023). The Semantic Scholar open data platform. *arXiv*. https://arxiv.org/pdf/2301.10140.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, *33*(7), 1163–1170.

Lai, Y., Kankanhalli, A., & Ong, D. (2021). Human-AI collaboration in healthcare: A review and research agenda. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (2021).

Lamba, M., & Madhusudhan, M. (2018a). Metadata tagging of library and information science theses: Shodhganga (2013-2017). In *Beyond the Boundaries of Rims and Oceans: Globalizing Knowledge with ETDs*.

Lamba, M., & Madhusudhan, M. (2018b). Application of topic mining and prediction modelling tools for library and information science journals. In M. R. Murali Prasad, A. Munigal, R. Nalik, M. Madhusudhan, & G. Surender Rao (eds.), *Library Practices in Digital Era* (pp. 395–401). BS Publications.

Lim, C., & Maglio, P. P. (2018). Data-driven understanding of smart service systems through text mining. *Service Science*, *10*(2), 154–180. doi: 10.1287/serv.2018.0208.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modelling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608. doi: 10.1186/s40064-016-3252-8.

Luna-Reyes, L. F. (2018). The search for the data scientist: Creating value from data. *ACM SIGCAS Computers and Society*, *47*(4), 12–16. doi: 10.1145/3243141.3243145.

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, *4*(1), 2053951717700925. doi: 10.1177/2053951717700925.

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. In *Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 503–512). IEEE.

Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modelling and visualization. *Scientometrics*, *125*(1), 665–687. doi: 10.1007/s11192-020-03657-5.

Nair, S. R. (2020). A review on ethical concerns in big data management. *International Journal of Big Data Management*, *1*(1), 8–25.

Pereira, P., Cunha, J., & Fernandes, J. P. (2020). On understanding data scientists. In *Proceedings of the 2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1–5). IEEE.

Priestley, J. L., & McGrath, R. J. (2019). The evolution of data science: A new mode of knowledge production. *International Journal of Knowledge Management*, *15*(2), 97–109. doi: 10.4018/IJKM.2019040106.

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *Journal of Law, Medicine & Ethics*, *43*(2), 425–429.

Semantic Scholar. (n.d.). *Publisher partners* [online]. Retrieved from: https://www.semanticscholar.org/about/publishers [11.11.2024].
Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70).
Śledziewska, K., & Włoch, R. (2020). *Gospodarka cyfrowa. Jak nowe technologie zmieniają świat*. Warsaw University Press. doi: 10.31338/uw.9788323541943.
Syed, S., Borit, M., & Spruit, M. (2018). Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016. *Fish and Fisheries*, *19*(4), 643–661. doi: 10.1111/faf.12280.
Thakur, K., & Kumar, V. (2022). Application of text mining techniques on scholarly research articles: Methods and tools. *New Review of Academic Librarianship*, *28*(3), 279–302.
Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences*. Springer Vs.
Yoon, J. E., & Suh, C. J. (2019). Research trend analysis by using text-mining techniques on the convergence studies of AI and healthcare technologies. *Journal of Information Technology Services*, *18*(2), 123–141.
Żulicki, R. (2022). *Data science: Najseksowniejszy zawód XXI wieku w Polsce. Big data, sztuczna inteligencja i PowerPoint*. Wydawnictwo Uniwersytetu Łódzkiego.
Zou, C. (2018). Analyzing research trends on drug safety using topic modelling. *Expert Opinion on Drug Safety*, *17*(6), 629–636. doi: 10.1080/14740338.2018.1458838.

Data scientists w literaturze naukowej: modelowanie tematyczne LDA w bazie danych Semantic Scholar

Abstrakt

Cel/Teza: Niniejszy artykuł analizuje reprezentację *data scientists* (specjalistów ds. analizy danych) w literaturze naukowej. Celem jest odpowiedź na następujące pytania: Jak zmieniała się liczba publikacji na temat *data scientists* na przestrzeni lat? Jak publikacje dotyczące *data scientists* są rozproszone w różnych dziedzinach nauki? W jakim kontekście *data scientists* są przedstawiani w literaturze naukowej?
Koncepcja/Metody badań: Zastosowano modelowanie tematów metodą utajonej alokacji Dirichleta (LDA) do zasobów dostępnych w ramach API Semantic Scholar.
Wyniki i wnioski: Od 2008 roku obserwuje się wzrost liczby publikacji na temat *data scientists*. Odkryto silny związek pomiędzy *data scientists* a technologią informacyjną oraz badaniami biomedycznymi. Niewiele publikacji porusza temat *data scientists* w kontekście społeczno-kulturowym.
Oryginalność/Wartość poznawcza: Zgodnie z naszą wiedzą, dotychczas nie prowadzono badań poświęconych reprezentacji data scientists w literaturze naukowej. Przeprowadzone badanie może przyczynić się do konceptualizacji tego pojęcia.

Słowa kluczowe

Data science. Eksploracja tekstu. Modelowanie tematyczne. Semantic Scholar. Utajona alokacja Dirichleta.

LEN KRAWCZYK, osoba absolwencka kierunku filozofia nowych mediów na Uniwersytecie Śląskim, obecnie studiuje socjologię cyfrową na Uniwersytecie Warszawskim. Wiceprzewodnicząca Koła Naukowego Socjologii Cyfrowej. Stypendysta ministra za wybitne osiągnięcia naukowe. Publikowała w czasopiśmie Psychiatria Danubina.

ŁUKASZ IWASIŃSKI, absolwent kierunku Business and Technology w International Faculty of Engineering na Politechnice Łódzkiej (tytuł magistra inżyniera – 2006 r.) oraz socjologii na Uniwersytecie Łódzkim (tytuł magistra – 2007 r., stopień doktora – 2013 r.). Pracuje jako adiunkt w Katedrze Informatologii Wydziału Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Wykładał m.in. na Uniwersytecie Łódzkim (na kierunkach socjologia oraz dziennikarstwo i komunikacja społeczna), Politechnice Łódzkiej (na kierunku organizacja i zarządzanie), w Społecznej Akademii Nauk (w ramach programu MBA). Autor książki „Socjologiczne dyskursy o konsumpcji" (2016) oraz kilkudziesięciu rozdziałów w monografiach zbiorowych i artykułów z dziedziny socjologii, kulturoznawstwa i informatologii.

MATEUSZ SZYMAŃSKI, absolwent studiów magisterskich z matematyki na Uniwersytecie Śląskim, obecnie student kierunku machine learning na Uniwersytecie Warszawskim. Pracuje jako data scientist i programista Python. Publikował w czasopiśmie Psychiatria Danubina.

Contact details:
mm.krawczy10@student.uw.edu.pl
l.iwasinski@uw.edu.pl
Uniwersytet Warszawski
Wydział Dziennikarstwa Informacji i Bibliologii
ul. Nowy Świat 69, 00-927 Warszawa