analytics

MDPI

Article

Uncovering Patterns and Trends in Big Data-Driven Research Through Text Mining of NSF Award Synopses

Arielle King [1] and Sayed A. Mostafa [2,*]

[1] Applied Science & Technology, North Carolina A&T State University, Greensboro, NC 27411, USA; anjones2@aggies.ncat.edu
[2] Department of Mathematics & Statistics, North Carolina A&T State University, Greensboro, NC 27411, USA
* Correspondence: sabdelmegeed@ncat.edu

Abstract: The rapid expansion of big data has transformed research practices across disciplines, yet disparities exist in its adoption among U.S. institutions of higher education. This study examines trends in NSF-funded big data-driven research across research domains, institutional classifications, and directorates. Using a quantitative approach and natural language processing (NLP) techniques, we analyzed NSF awards from 2006 to 2022, focusing on seven NSF research areas: Biological Sciences, Computer and Information Science and Engineering, Engineering, Geosciences, Mathematical and Physical Sciences, Social, Behavioral and Economic Sciences, and STEM Education (formally known as Education and Human Resources). Findings indicate a significant increase in big data-related awards over time, with CISE (Computer and Information Science and Engineering) leading in funding. Machine learning and artificial intelligence are dominant themes across all institutions' classifications. Results show that R1 and non-minority-serving institutions receive the majority of big data-driven research funding, though HBCUs have seen recent growth due to national diversity initiatives. Topic modeling reveals key subdomains such as cybersecurity and bioinformatics benefiting from big data, while areas like Biological Sciences and Social Sciences engage less with these methods. These findings suggest the need for broader support and funding to foster equitable adoption of big data methods across institutions and disciplines.

Keywords: big data; grant funding; award synopsis; higher education; topic modeling; word co-occurrence networks

check for updates

Academic Editor: Domenico Ursino

Received: 18 November 2024
Revised: 23 December 2024
Accepted: 2 January 2025
Published: 6 January 2025

Citation: King, A.; Mostafa, S.A. Uncovering Patterns and Trends in Big Data-Driven Research Through Text Mining of NSF Award Synopses. *Analytics* **2025**, *4*, 1. https://doi.org/10.3390/analytics4010001

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Big data has been described as "the next big thing in innovation" [1] and "the fourth paradigm of science" [2]. Across fields, practitioners and scholars recognize the vast potential of big data and continue to explore its advancements and applications [3–9]. Leading firms, such as IBM, Xerox, and Google, have leveraged big data analytics to innovate more effectively, creating new products and services that have solidified their competitive edge [10]. In today's digital landscape, big data has transformed how we collect, process, and utilize information, reshaping entire sectors.

The relevance of big data in modern society is evident in its contributions to scientific discovery, economic growth, social influence through personalized recommendations and social media algorithms, and rapid technological advancements. As a crucial element of national competitiveness, big data offers immense strategic and operational value, generating both academic interest and corporate investment.

Previous studies relay the impact of big data research in specific disciplinary contexts. Generally, papers discussing higher education and big data examine how big data analytics can enhance educational outcomes such as teaching and learning or institutional operations using international databases [9,11,12]. Most influential to our study is the work of Mohammadi and Kirmini [13] who used text-mining techniques to explore the scope of big data across disciplines. Our study is the first to analyze patterns and trends of big data-driven research across academic fields within U.S. institutions of higher education, through the lens of funded awards from the U.S. National Science Foundation (NSF). It also offers insights into varying data-driven research subdomains within academic research areas. By analyzing trends and themes in NSF-funded big data-driven research through text mining, descriptive analysis, and topic modeling, we aim to highlight how different types of higher education institutions (HEIs) contribute to federally funded big data-driven research to promote broader participation and bolster national competitiveness on a global scale.

The rest of this article is organized as follows: The Section 2 provides an overview of big data, the NSF's role in funding big data-driven research, and related literature and lists the study's guiding research questions. The Section 3 details the rationale, data, and analysis procedures. Finally, the article concludes with the Results and Discussion, highlighting the study's key findings, implications, and limitations.

2. Background

2.1. Big Data

The exponential growth in data volume, variety, and velocity has introduced the defining phenomenon of the big data era [14]. The urgency to transform vast streams of unstructured data into valuable insights has never been greater, driven largely by the proliferation of data generated from internet searches, cloud computing, social media posts, sensor readings, and online shopping. Big data encompasses data that are too large and fast-moving for conventional database systems to handle [15]. Yang [16] adds that managing the complexities of big data requires advanced technologies capable of processing massive volumes at unprecedented speeds. Recently, veracity has joined volume, velocity, variety, value, and variability as part of the "V" definition of big data, summarizing its core attributes [5]. Although scholars have not agreed on a single definition, most agree that big data involves "large amounts of digital data generated by technological devices that necessitate specific algorithms or computational processes to answer relevant research questions" [17].

2.2. Big Data's Impact

Initially, big data and business analytics were primarily business and marketing tools used to maximize profit and enhance consumer experiences. In 2011, McKinsey & Company projected a 140,000 to 190,000 individuals with deep analytical skills shortage by 2018 to support the expanding demand for big data [15]. Recognizing this need, the U.S. government launched initiatives in 2012 and 2014 to capitalize on big data's potential, including the White House Big Data Report and the Department of Education's Office of Educational Technology [11]. As data became a critical competency across STEM fields, the demand for skilled data scientists surged [18]. In response, numerous U.S. institutions of higher education established programs in Business Intelligence, Data Analytics, and Data Science, leading to rapid growth in formal Data Science programs, from zero in 2011 to 40 across 30 institutions by 2014 [19]. These developments marked the emergence of data science as a field dedicated to addressing industry and research demands.

2.3. Big Data-Driven Research Across Disciplines

Over the past decade, disciplines outside of business have increasingly adopted big data methods to drive innovation and discovery. Organizations that employ advanced analytics can harness big data to generate new insights, products, and services [20]. Across various fields, researchers have demonstrated the potential of big data to advance knowledge. In bioinformatics, deep learning has driven breakthroughs in sequence analysis and structure prediction [21]; in engineering, big data applications support both passive and active processes [22]; in education and psychology, big data has helped visualize critical learning periods, particularly in response to the COVID-19 pandemic [9]. For educational outcomes, digital assessment environments have shown promise in enhancing learning by leveraging educational data [23]. Similarly, in economics, theory-guided machine learning has improved our understanding of heterogeneous treatment effects [24,25], while in geoscience, data-driven methods enable the construction of knowledge graphs [26].

Additionally, big data's applications have been explored through quantitative and qualitative research across fields such as business [27], the public sector [7], management [6], marketing [3], biomedical sciences [28], and education [4]. However, these studies are often limited to specific disciplinary perspectives. In 2017, ref. [29] found that big data's interdisciplinary nature is rapidly maturing with computer science as a fundamental discipline. A 2022 study addressed this gap by examining big data research across disciplines through journal publications in the Web of Science from 2012 to 2017 [13].

2.4. NSF and Big Data

The NSF is the only federal agency dedicated to advancing basic research and education across the full spectrum of STEM disciplines. As part of its mission, NSF has implemented agency-wide programs to support big data's potential for accelerating scientific discovery and innovation. In 2012, the NSF launched the Critical Techniques, Technologies, and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) initiative. This evolved into the Harnessing the Data Revolution (HDR) program in 2018, one of the NSF's 10 Big Ideas to foster data-driven discovery to address fundamental scientific questions. Through HDR and other initiatives, the NSF has consistently supported data science, funding 23% of all federally sponsored basic research conducted by U.S. colleges and universities [30].

2.5. Research Using NSF Data

NSF awards provide a valuable administrative data source for studying various aspects of U.S. research funding. Such data offer insights into issues of social inequality, human behavior, and the impact of social policies [31,32]. However, limited publications have focused specifically on the impact of federal funding on research output and disciplinary development. A 2018 study applied topic modeling to NSF abstracts to identify trends in ocean science [33]. Klami & Honkela [34] explored the relationship between research content and NSF divisions using self-organizing maps to improve project classification. Other studies have used NSF award abstracts to explore funding trends in nanotechnology [35], research politicization, and education [36,37].

2.6. Significance and Contribution

The core concepts of empowerment, discovery, impact, and excellence are emphasized in the NSF's strategic plan to advance the United States' leadership in science and engineering. It is widely known that research activities at U.S. universities and colleges are crucial for driving economic development, fostering innovation, enhancing educational outcomes, and promoting partnerships between academia and industry [38–41]. This combination

underscores the broad societal impact of university research. Previous studies on big data research in higher education often focus on singular topics, such as learning analytics or improving institutional operations [12]. Given the U.S. demand for qualified data analysts and increased investment and interest in big data's potential to drive global, economic, and social impacts, examining big data-driven research trends within U.S. HEIs is crucial. However, to the best of our knowledge, no prior research has analyzed big data-driven research trends within higher education through a funding agency's lens. This study addresses that gap in knowledge by exploring big data-driven research trends within U.S. HEIs via NSF awards, guided by the following research questions:

RQ1: What is the trend or pattern of funded big data-driven research within NSF-defined research domains from 2006 to 2022?

RQ2: Whether and how the patterns in big data-driven research vary by the institution's research classification (R1, R2, or other) and population served (Predominately White Institution (PWI), Historically Black College and University (HBCU) or Hispanic Serving Institution (HSI)).

RQ3: What subdomains are most prominent in NSF-funded big data-driven research?

By addressing these research questions, this study will shed light on the evolving landscape of big data-driven research within U.S. HEIs. Specifically, by examining trends, institutional variations, and prominent subdomains of NSF-funded big data-driven research, the findings can inform policymakers, educators, and funding agencies about where resources and efforts are being concentrated and where gaps or opportunities might exist. Given the increasing reliance on data analytics to address pressing global, economic, and societal challenges, understanding these trends is critical for shaping future research priorities and fostering inclusive participation across diverse institutional types.

3. Data and Methods

This study uses natural language processing (NLP) techniques to perform content analysis on award synopses funded by the National Science Foundation (NSF). Specifically, the study examines trends and patterns in NSF-funded big data research from 2002 to 2022 within seven NSF-designated research areas (directorates): Biological Sciences (BIO), Computer and Information Science and Engineering (CISE), Engineering (ENG), Geosciences (GEO), Mathematical and Physical Sciences (MPS), Social, Behavioral, and Economic Sciences (SBE), and STEM Education (EDU).

3.1. Research Design and Rationale

A quantitative approach was chosen for this study, employing automated text mining tools in R software to visualize and address the research questions. Klenke [42] defines content analysis as a research methodology that systematically analyzes large volumes of text, often by extracting frequent keywords. As applied here, content analysis uses text-mining algorithms to identify concepts and themes within NSF award abstracts. One primary method, topic modeling, is an unsupervised machine learning technique that identifies groups of words likely to appear together within documents, thereby forming topics. For example, a 2018 study applied topic modeling to NSF abstracts to uncover research topics in ocean sciences [33]. Topic modeling automates clustering, producing topics that closely approximate manually coded themes [43].

3.2. Data

Award data from fiscal years 2006 to 2022 were downloaded from the NSF website. Award data were included if it was a standard grant with an abstract, title, and the "Performance Institution" or "Institution" columns contained institutions listed in the Carnegie

Classifications of Higher Education. The dataset comprised 88,548 awards across seven NSF directorates: BIO, CISE, ENG, GEO, MPS, SBE, and EDU. See Table 1 for the distribution of these awards by directorate.

Table 1. Distribution of NSF awards by directorate between 2006 and 2022.

Directorate	n	Percent
ENG	18,349	20.72
MPS	18,140	20.49
CISE	17,603	19.88
GEO	12,201	13.78
SBE	9,466	10.69
BIO	8,480	9.58
EDU	4,309	4.87
Total	88,548	100

NSF award data were merged with the Carnegie Classification of Institutions of Higher Education dataset [44]. Using Carnegie's institutional research activity and population served groups, institutions are classified by their level of research activity (R1 for "very high research activity", R2 for "high research activity", or neither) and by the population they serve (Historically Black Colleges and Universities (HBCUs), Hispanic-serving institutions (HSIs), or neither). HBCUs, HSIs, and other groups not used in this study are historically or enrollment-designated Minority-Serving Institutions (MSIs) through federal appropriations. This study considers these categories because of their integral role in the American Education System.

3.3. Analysis Procedures

Text mining packages within R software (version 4.3.1) were used for data preprocessing and analysis [45]. The keywords "big data", "analytics", "machine learning", "predictive modeling", "artificial intelligence", and "data science" were used to identify big data-driven research. These keywords were selected based on scientometrics mapping and text-analytics results of big data research trend studies in [13,46] and best capture the holistic meaning of big data [8,13,47,48]. If the synopsis contained at least one of these keywords, it was labeled a big data-driven award or, for short, a "big data award"; otherwise, it was labeled as a non-big data-driven award.

Descriptive statistics and co-occurrence word networks (visualizations of word co-occurrence) were used to address RQ1 and RQ2. A co-occurrence word network is a visual representation of words that frequently appear adjacent in a given context. The co-occurring word networks can reveal concepts, topics, and recurring themes over time in big data awards.

To address RQ3, topic modeling was employed to identify research subdomains and their proportions within each NSF directorate portfolio of awards. Employing LDA compliments the study's main objective by discovering latent topics that may not align with predefined divisions of the NSF. It will also help explore how specific subdomains within research fields are affected by big data or becoming more data-driven. Furthermore, LDA's capabilities are strengthened when applied to datasets over time, capturing evolving research trends [49]. Topic modeling is an unsupervised machine learning technique commonly used in natural language processing (NLP) for text clustering. Topic modeling rests on two main principles:

1. Each topic is a cluster of words that shares some semantic domain.
2. Each document is a mixture of the topics, i.e., the document contains words from various topics in different proportions. For this study, documents are the synopses of NSF awards.

We employed Latent Dirichlet Allocation (LDA) [50] for the NSF awards corpus topic modeling. We give a brief description of the method as follows: Consider a corpus $\mathcal{C}$ consisting of D documents, where each document d contains N_d words, with $d = 1, \ldots, D$. LDA models this corpus over K topics through the following generative process:

1. For each document d, draw a topic proportion vector θ_d from a Dirichlet distribution parameterized by α.
2. For each topic $k = 1, \ldots, K$, draw a topic distribution β_k, representing the distribution over vocabulary terms, from a Dirichlet distribution with parameter η.
3. For each word w_n in document d (where $n = 1, \ldots, N_d$),
   i. Select a topic z_n for the word from a multinomial distribution governed by θ_d.
   ii. Draw the word w_n itself from a multinomial distribution determined by the topic distribution β_{z_n}.

This process captures the underlying topic structure of a collection by associating each document with a mixture of topics and each topic with a distribution over words.

We used the tidy principles of text mining in R to apply LDA to the NSF synopses data [51]. The LDA model calculates probabilities of a term generated from a topic (β_{z_n}) and computes probabilities that a topic belongs to a document (θ_d). Research subdomains are created using each topic's top β values. Each document is assigned a topic based on the highest θ value. The authors labeled the topics at their discretion using the top terms generated by the LDA model. Tables A1–A7 in Appendix A present LDA results for subdomains within each NSF directorate. The tables include topic labels, top terms (with the largest probabilities) for the topic, and topic weights (topic count divided by sum of topic counts within the directorate). Topic modeling is a computationally intensive technique, which required limiting the data for RQ3 to awards from 2012 to 2022 (n = 64,777), a period marking a substantial rise in big data publications [9,13,52]. The selection of this period aligns well with key developments in big data research, thereby improving the accuracy of the topic modeling.

4. Results

4.1. Trends and Patterns in Funded Big Data-Driven Research (RQ1)

4.1.1. General Trends in Funded Big Data-Driven Research

The results for RQ1 focus on trends in NSF-funded big data awards over time. Figure 1 shows a consistent increase in the proportion of big data awards across all directorates, with the Computer and Information Science and Engineering (CISE) directorate leading significantly. CISE's share of NSF-funded big data awards grew from 5% in 2006 to 45% in 2022, likely influenced by early advancements in cloud computing within information systems [53]. From 2012 onward, several directorates show notable increases in big data funding. Between 2006 and 2022, the percentage growth in big data awards for each directorate was as follows: CISE (39.92%), ENG (19.24%), BIO (7.92%), EDU (13.29%), MPS (13.29%), SBE (8.1%), and GEO (6.98%). Figure 2 further confirms a positive trend in big data awards, with an 18.98% overall growth from 2006 to 2022, reflecting a shift toward data-driven research across disciplines.

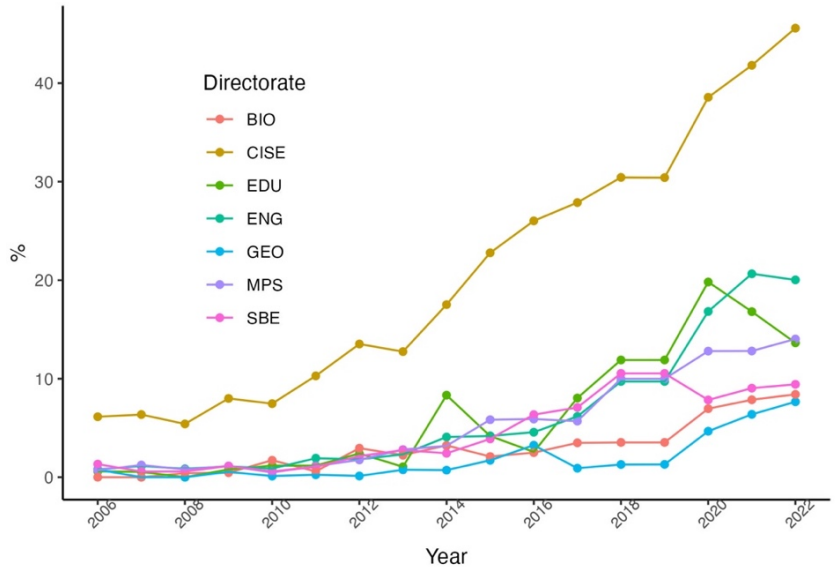


Figure 1. The trend in big data-driven awards made by the various NSF directorates. Note: the percentage is calculated as (no. of big data-driven awards/total no. of awards) × 100.

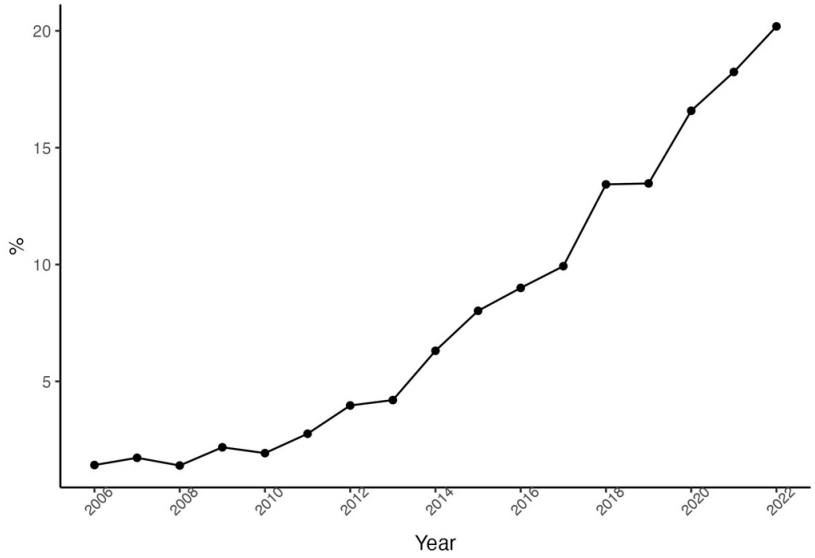


Figure 2. The trend in big data-driven awards across all NSF directorates. Note: the percentage is calculated as (no. of big data-driven awards/total no. of awards) × 100.

4.1.2. Big Data Keyword Trends over Time

Figure 3 shows the trend of keywords used to identify big data awards. From 2011 onwards, keyword use in award abstracts steadily increases, echoing the findings of [48]. The most frequently used keywords across all directorates include "machine learning" and "artificial intelligence". Specifically, machine learning increased by 10.85%, artificial intelligence by 6.49%, data science by 4.55%, analytics by 2.76%, big data by 1.73% (first appearing in 2008), and predictive modeling by 0.2%. Figure 4 displays keyword trends by directorate from 2006 to 2022, with CISE leading in all keywords. Interestingly, big data follows a bell curve, indicating a recent decline, likely due to its broad application across domains [17]. Predictive modeling remains sparse, with growth under 1.05% in any directorate.

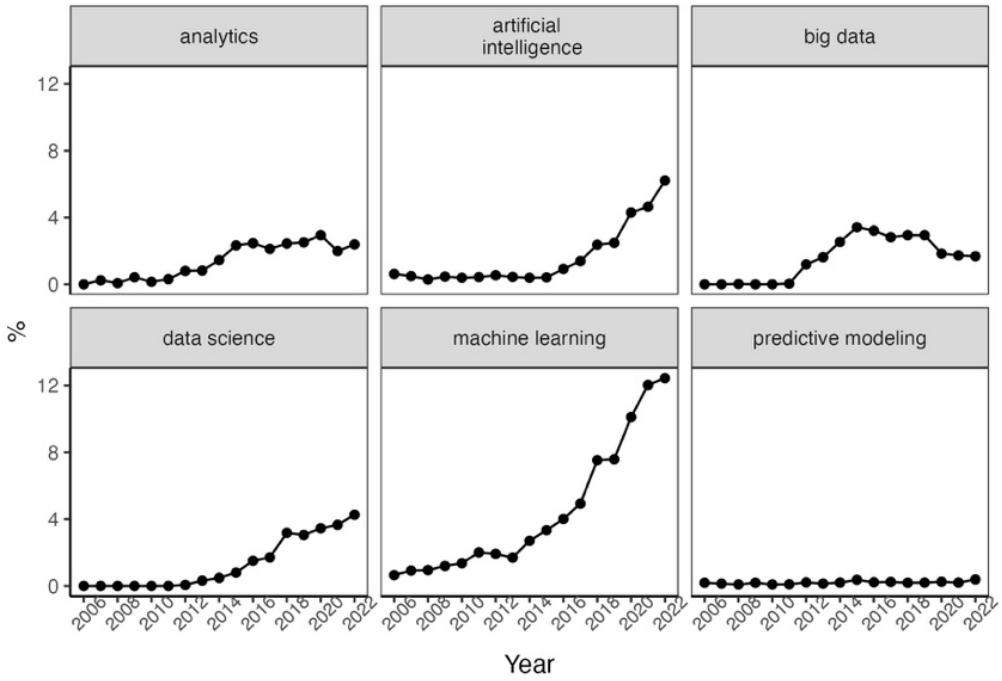


Figure 3. The trend in big data-driven awards across all NSF directorates by keywords: Note: the percentage is calculated as (no. of awards containing the keyword/total no. of awards) × 100.

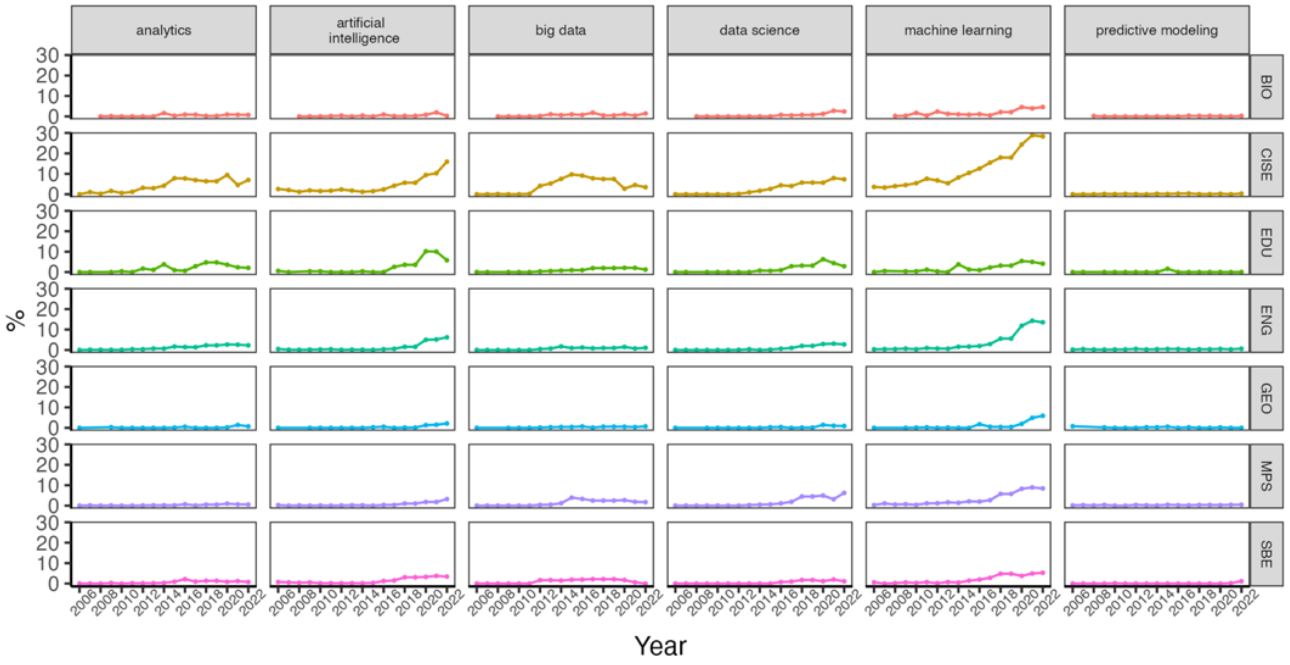


Figure 4. The trend in big data-driven awards by keyword within each NSF directorate. Note: the percentage is calculated as (no. of awards containing the keyword/total no. of awards) × 100.

4.1.3. Big Data Themes over Time

Figures 5–11 show co-occurring word networks of big data awards for each directorate, comparing two time periods: 2012–2017 and 2018–2022. Each of the co-occurring word networks has nodes (words), edges (the connection or relationship between the words), and weights (the degree representing the number of connections that a given node has and the strength of the connection represented as frequency, n). Across directorates, the frequency of data-driven methods increased after 2017. For example, in EDU, word clusters around "data" occurred over 50 times before 2018, while after 2018, these nodes appeared

over 100 times. Additionally, several isolated word pairs emerged post-2017, indicating a diversification of big data topics and tools. In ENG (Figure 5) and CISE (Figure 8), environmentally conscious themes of "energy efficiency" reveal efforts made to reduce the carbon footprint of the digital data era. Within SBE, Figure 6, we note that big data usage is increasing in social efforts such as food security, the justice system, and mental health. This shift reflects the continuing expansion of big data applications across fields through federally funded projects, contributing to the ongoing big data revolution.

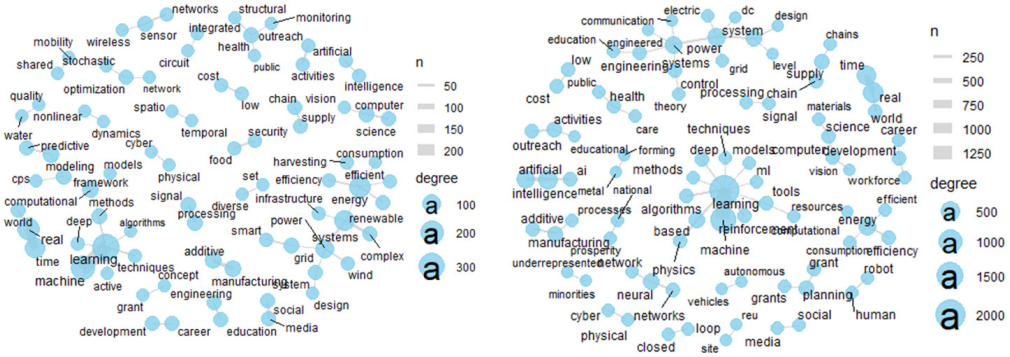


Figure 5. Most frequent co-occurring words in the ENG directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

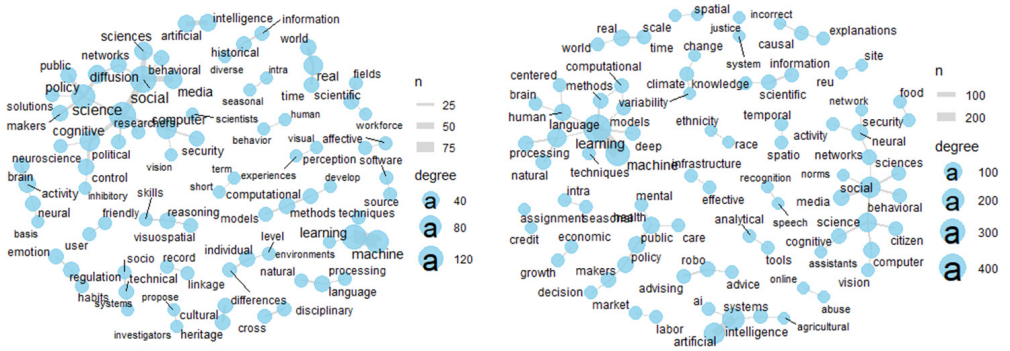


Figure 6. Most frequent co-occurring words in the SBE directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

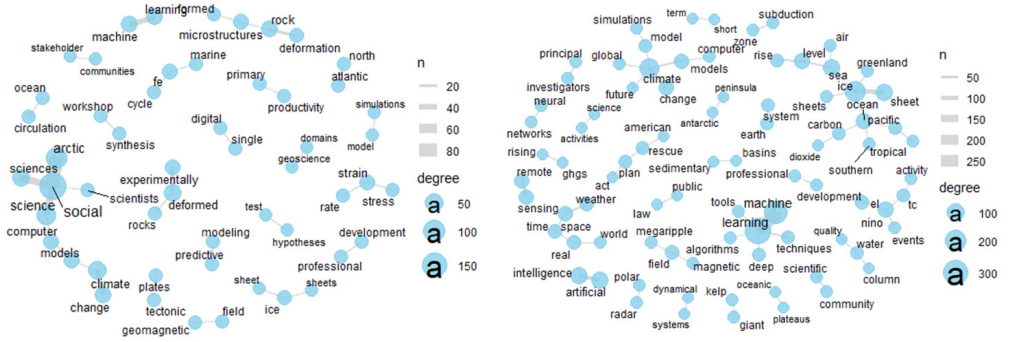


Figure 7. Most frequent co-occurring words in the GEO directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

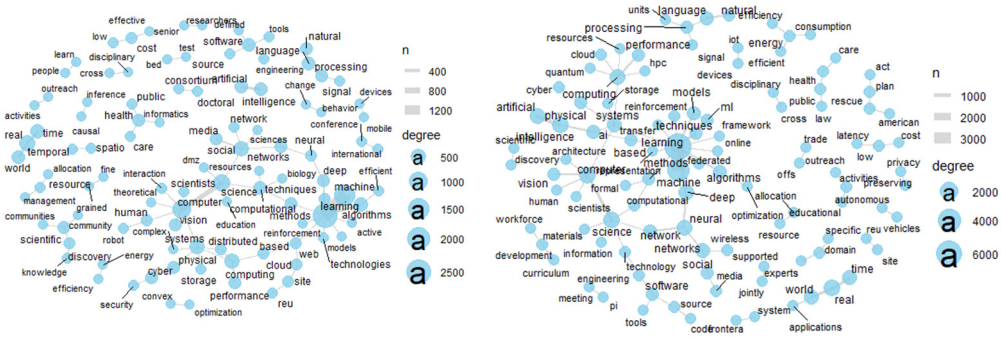


Figure 8. Most frequent co-occurring words in the CISE directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

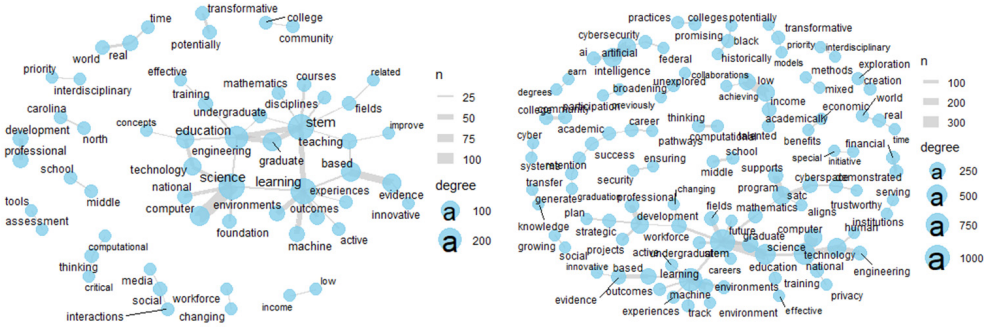


Figure 9. Most frequent co-occurring words in the EDU directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

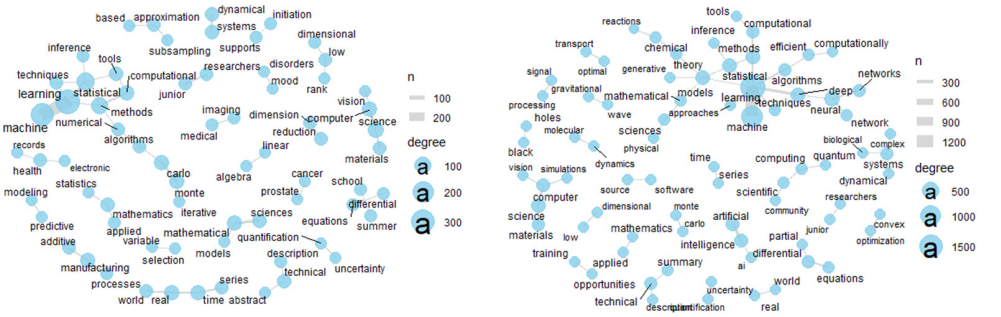


Figure 10. Most frequent co-occurring words in the MPS directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

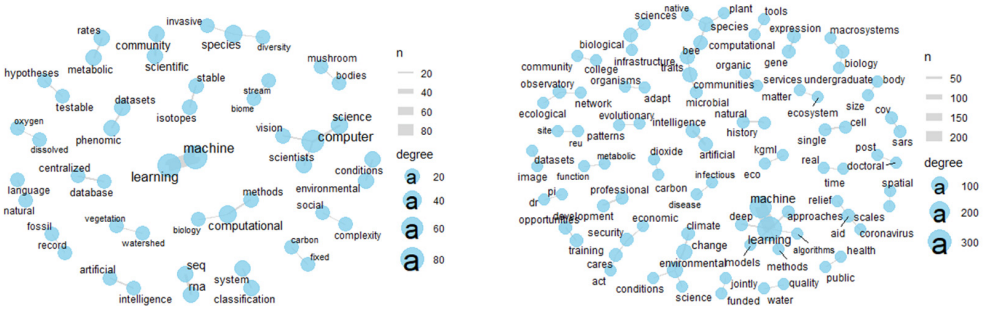


Figure 11. Most frequent co-occurring words in the BIO directorate's big data-driven awards from 2012–2017 (**left**) and 2018–2022 (**right**).

4.2. Big Data-Driven Research Trends by Institutional Characteristics (RQ2)

4.2.1. Research Classification

NSF big data awards were examined using the research classification of U.S. institutions. Figure 12 shows that R1 institutions (doctoral universities with very high research

activity) account for the majority of big data awards, comprising 80% of total awards and 90% of big data awards. Figure 13 shows that R1 institutions consistently lead in big data award percentages, except in 2014, when R1 and R2 institutions were nearly equal, while "neither" institutions lagged. After a drop in 2015, R2 institutions show a steady increase in big data funding, although the gap between R1 and R2 has narrowed in recent years, while it has widened further for "neither" institutions. Figure 14 provides a breakdown by the directorate, showing that CISE has the highest rate of big data awards across all classifications, though it exhibits an oscillating pattern for "neither" institutions not seen in R1 or R2. These results indicate disparities in NSF-funded big data research across institutions with different research classifications.



Figure 12. Distribution of awards by the award status and the institution's research classification.

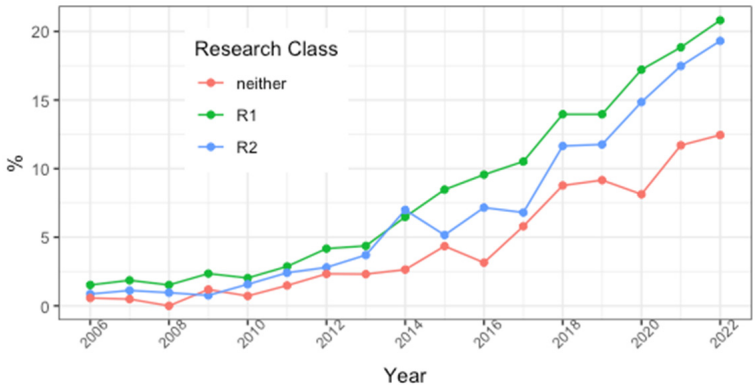


Figure 13. The trend in big data-driven awards by the recipient institution's research classification.

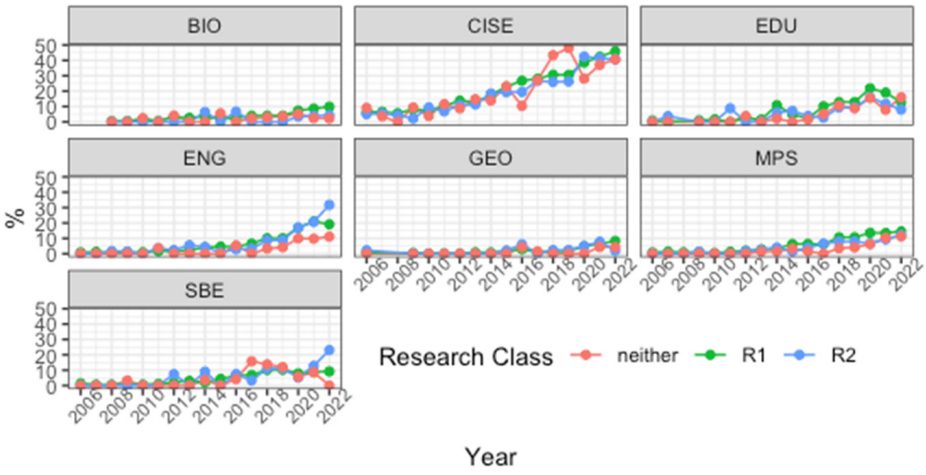


Figure 14. The trend in big data-driven awards by the recipient institution's research classification within each NSF directorate.

4.2.2. Population Served

Big data awards were also analyzed based on the populations served by institutions, classified as HBCU, HSI, or neither (neither HBCU nor HSI). It is readily seen from Figure 15 that HBCU institutions demonstrate a unique trend, with noticeable peaks surpassing HSI and "neither" institutions in 2011 and maintaining a lead from 2014 to 2019. By 2022, HBCUs led in big data award proportions with a 25% increase, aligning with increased federal investment under initiatives like the American Rescue Plan and debt relief programs targeting HBCUs [54]. Figure 16 further stratifies big data awards by the directorate, with CISE again leading across all groups, especially among HBCUs (see Figure 15). HSI and "neither" institutions exhibit lower big data award growth in fields like BIO and GEO.

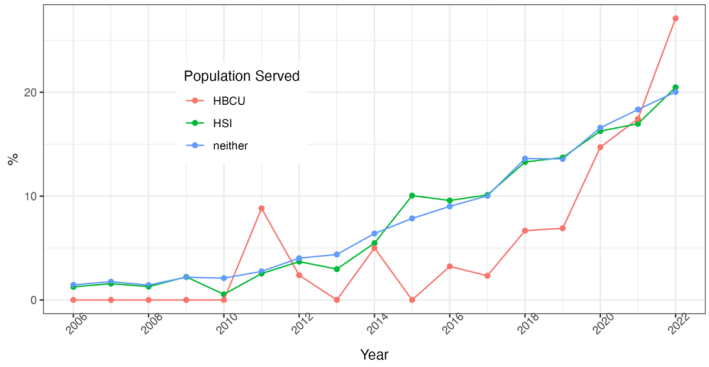


Figure 15. The trend in big data-driven awards by the recipient institution's served population.

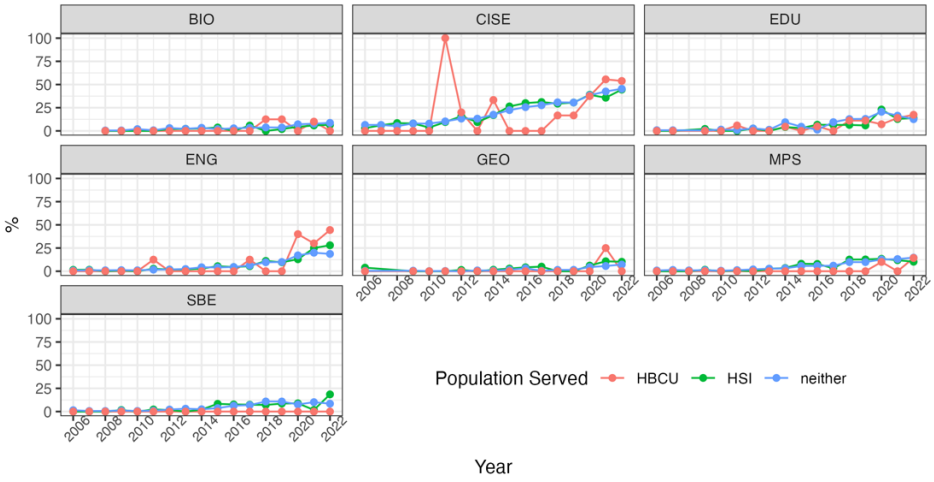


Figure 16. The trend in big data-driven awards by the recipient institution's served population within each NSF directorate.

4.3. Big Data-Driven Research Within Subdomains of NSF-Defined Research Areas (RQ3)

Subdomains within each NSF directorate were identified using Latent Dirichlet Allocation (LDA) topic modeling. Tables A1–A7 in Appendix A display subdomain topic labels, top terms for each subdomain, and subdomain weights, summarizing trends across NSF directorates. We examined the proportion of big data awards per subdomain before and after 2017 to assess shifts in big data engagement over time (Figures 17–23).

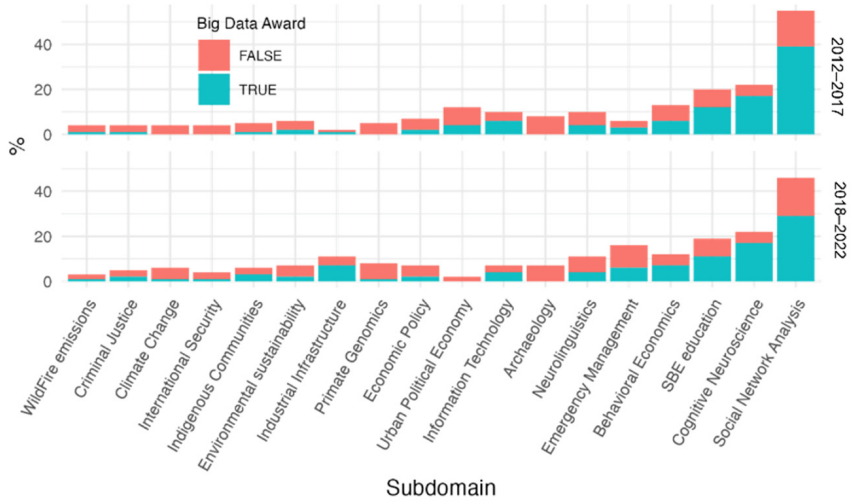


Figure 17. Distribution of big data-driven awards within the SBE directorate's research subdomain.

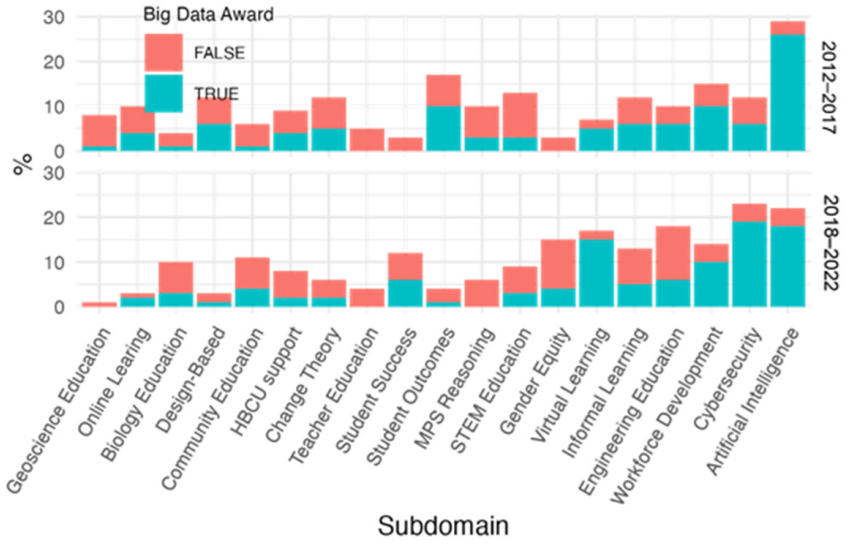


Figure 18. Distribution of big data-driven awards within the EDU directorate's research subdomain.

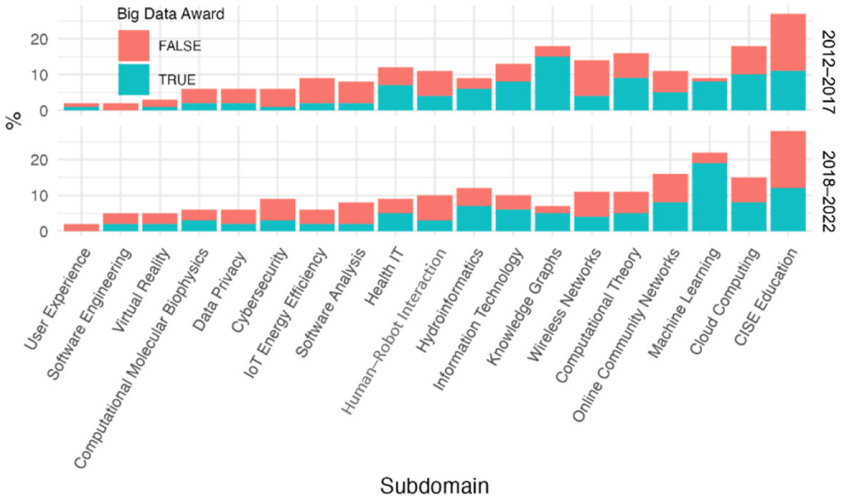


Figure 19. Distribution of big data-driven awards within the CISE directorate's research subdomain.

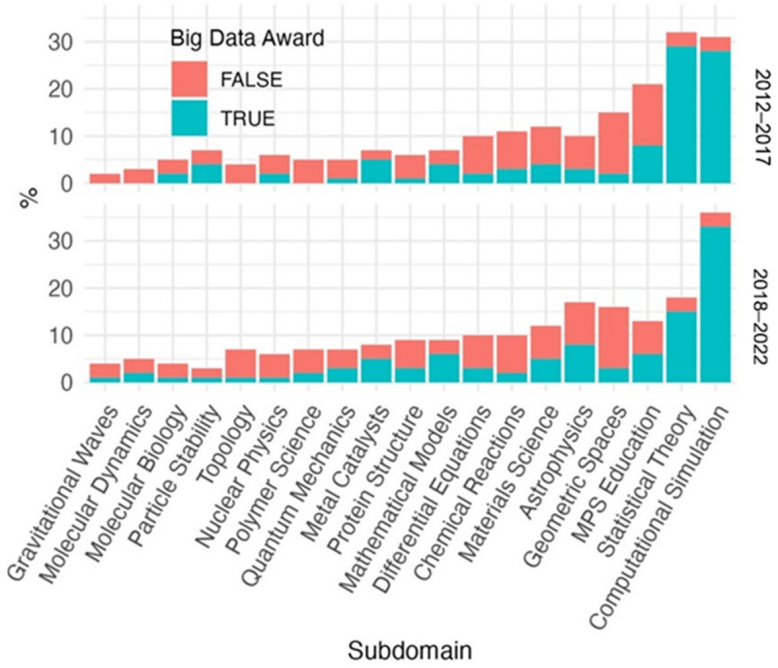


Figure 20. Distribution of big data-driven awards within the MPS directorate's research subdomain.

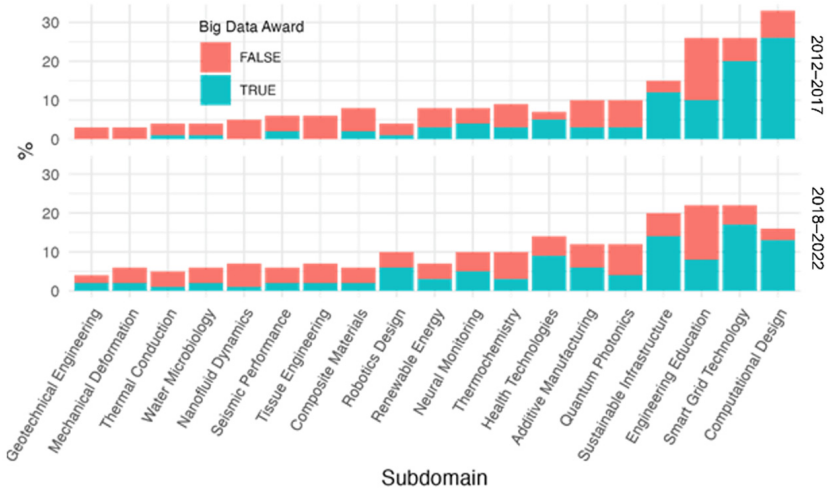


Figure 21. Distribution of big data-driven awards within the ENG directorate's research subdomain.

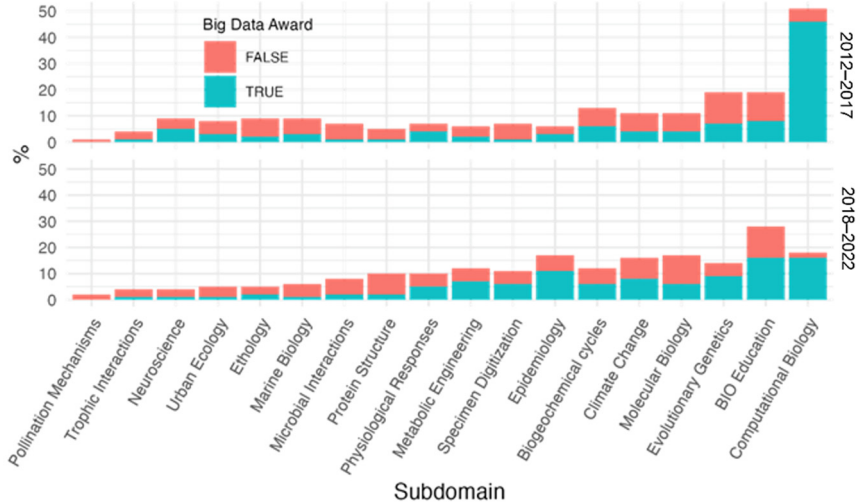


Figure 22. Distribution of big data-driven awards within the BIO directorate's research subdomain.

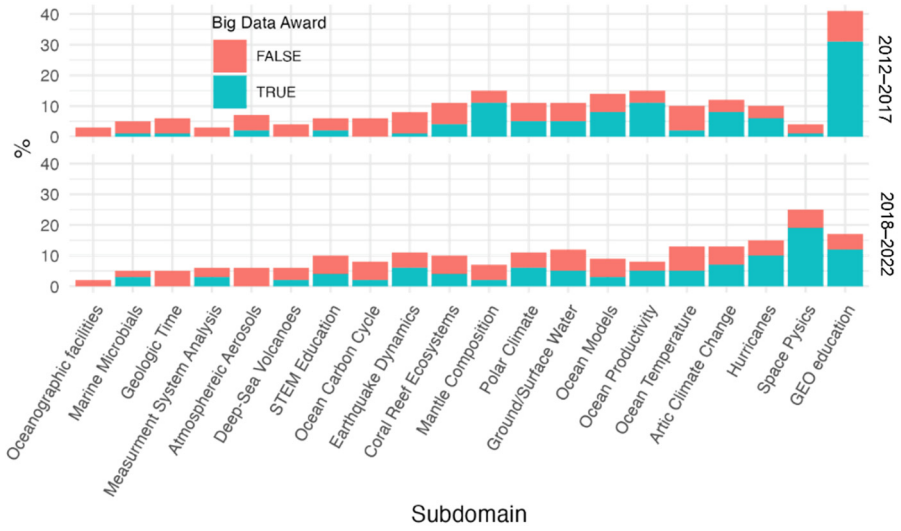


Figure 23. Distribution of big data-driven awards within the GEO directorate's research subdomain.

The results indicate that certain subdomains have increased their share of big data awards post-2017, while others have remained constant or declined. Notable subdomains with higher proportions of big data awards after 2017 include the following:

- SBE: Emergency management, cognitive neuroscience, industrial infrastructure, and climate change.
- EDU: Diversity/inclusion, cybersecurity, virtual learning, community education, and student success.
- CISE: Machine learning, cybersecurity, hydroinformatics, and online community networks.
- MPS: Computational simulations, astrophysics, mathematical modeling, and quantum mechanics.
- ENG: Sustainable infrastructure, robotics design, geotechnical engineering, and health technologies.
- BIO: Bioeducation, climate change, specimen digitization, and epidemiology.
- GEO: Space physics, hurricanes, deep-sea volcanoes, STEM Education, and earthquake dynamics.

The observed growth in big data-driven research awards across NSF directorates in subdomains post-2017 highlights significant shifts in priorities and innovations. Advancements in computational power and data availability are driving the rise in machine learning, hydro informatics (CISE), and health technologies (ENG). Increased focus on diversity, inclusion, and student success (EDU) and epidemiology (BIO) suggests responsiveness to societal challenges such as educational equity and global health crises. Climate change (SBE and BIO) and sustainable infrastructure (ENG) highlight the application of big data to pressing environmental and sustainability issues.

Some subdomains, however, have shown declines in big data award percentages, such as cognitive learning in EDU, knowledge graphs in CISE, computational biology in BIO, and computational design in ENG. This decline may reflect that certain sub-topics have reached a level of maturity where big data methods are no longer seen as novel. It can also result from an increase in interdisciplinary research, as noted in previous studies where large research areas like BIO and ENG displayed high interdisciplinarity [34]. For instance, Lima and Rheuban [33] reported that 65% of NSF ocean science awards became more interdisciplinary over time. Additionally, the diverse, emerging topics using big data techniques within all directorates suggest an increase in innovation and the adoption of big data across subdomains.

5. Discussion

This study employed an exploratory quantitative approach to analyze patterns, trends, and disparities in big data-driven research at U.S. higher education institutions (HEIs) using NSF award data. The findings reveal that the Computer and Information Science and Engineering (CISE) directorate leads in leveraging big data methodologies, with machine learning and artificial intelligence emerging as dominant themes across institutional classifications and research domains. The prominence of these trends can be attributed to the exponential growth in data generation through digital activities, computational power, and societal demand for automation and actionable insights. Consequently, the proliferation of machine learning and artificial intelligence has heightened awareness of the ethical concerns associated with big data. Notably, "preserving privacy" has frequently appeared as a co-occurring term in big data awards in CISE since 2017 (Figure 8). Moreover, data privacy (within CISE) and cybersecurity (within CISE and EDU) have emerged as critical subdomains to tackle the privacy risks inherent in big data. The continued prevalence of machine learning and artificial intelligence in both academic research and practical applications aligns with findings from prior studies that observed similar patterns in computer science fields [55,56]. These results underscore the central role of these technologies in shaping the landscape of modern research and innovation [57,58].

For the research subdomain, common topics emerged in data-driven fields, such as discipline-based education, computational methods, and statistical analysis. Additionally, an increasing number of isolated topics, such as social media, energy, learning, and security, highlight the diversification of big data applications. R1 institutions and non-HBCU/non-HSI institutions maintain a substantial lead in big data awards. The recent exponential increase in funding for HBCUs suggests that national investments to enhance diversity, equity, and inclusion in education and the workforce are positively impacted.

While certain fields and institution types are thriving in big data-driven research, the results indicate that directorates like Biological Sciences (BIO), Geosciences (GEO), Mathematical and Physical Sciences (MPS), and Social, Behavioral, and Economic Sciences (SBE) awarded relatively less NSF funding for big data-driven projects. The results reveal significant disparities in NSF-funded big data-driven research across institutional classifications, disciplines, and subdomains. First, we present possible contextual factors that drive disparities in institution types, followed by research areas and their subdomains.

The Carnegie classification system is intended to provide a neutral taxonomy; however, it has inadvertently contributed to disparities in research trends and award allocation among institutions. By emphasizing metrics on research expenditures and doctoral degrees conferred, the system tends to favor well-resourced institutions, marginalizing smaller or less-funded colleges and universities. A second factor is the perceived prestige of R1 institutions. This creates a cycle where privileged institutions advance while others struggle to gain recognition. A third factor is mission alignment. Institutions focusing on teaching or community engagement may find their contributions undervalued in a system prioritizing research output. This misaligns an institution's mission and classification, affecting funding and policy decisions. While recent initiatives have improved access for MSI institutions, as results show in our results, targeted efforts to simplify funding processes, promote equity, and recognize diverse institutional missions are critical to fostering more inclusive research ecosystems.

To address the disparities among research areas, we discuss some anecdotal contexts. Biological Sciences (BIO) and Geosciences (GEO) often deal with highly complex, non-standardized datasets, making integrating big data methods challenging. In SBE disciplines, data sensitivity (e.g., privacy concerns in human behavior studies) limits the scope of data-sharing and aggregation needed for robust big data applications. Subdomains in BIO

and GEO have historically been underfunded for computational advancements, limiting the development of data infrastructure necessary for adopting big data techniques. SBE fields may be skeptical of the scalability or rigor of big data applications, which can deter funding and broader adoption. Big data methods are inherently interdisciplinary, requiring collaboration across statistics, computer science, and domain-specific expertise. Less represented subdomains may lack established networks to foster such interdisciplinary research. GEO and BIO often prioritize traditional field-based research over computational approaches, slowing the adoption of advanced data analytics. SBE subdomains may lack well-defined big data use cases, limiting their ability to demonstrate clear value in funding proposals. Non-R1 institutions, which often house researchers in underrepresented subdomains, lack the computational infrastructure and grant-writing expertise to compete effectively for big data-related funding.

We provide recommendations for infrastructure development, interdisciplinary training, and enhancing research networks to increase support for underrepresented subdomains and disciplines. The first recommendation is establishing centralized computational resources and data-sharing platforms tailored to the subdomain's needs. For instance, in GEO, developing integrated geospatial data repositories with computational tools for climate modeling and ecosystem analysis. In BIO, bioinformatics infrastructure could address complex genomic and biodiversity datasets. In SBE, implementing secure platforms for aggregating and analyzing sensitive social science data is perilous, particularly in addressing privacy concerns.

Second, establishing research networks is critical to fostering collaboration and inclusivity. Consortiums can bridge gaps between non-R1 and MSIs by leveraging shared expertise and resources. Grant writing experts can provide templates, training sessions, and dedicated support for researchers to make grant applications more competitive.

Lastly, emphasizing the societal impacts of data-driven research is vital. Projects that address climate change, emergency management, or public health can align with funding agencies' broader goals, demonstrating their critical roles in addressing societal changes. These recommendations can help dismantle the disparities in big data-driven research in underrepresented fields, encourage collaboration across disciplines, promote fairness, and make NSF-funded research more impactful overall.

This study underscores the need to increase awareness and support for underrepresented research areas and institutions, encouraging them to adopt data-intensive research methods and consider strategies for securing federal funding to support such initiatives.

Furthermore, this work aims to inspire emerging research fields to explore how big data techniques might enhance their studies. This preliminary analysis suggests a need for future research on how non-R1 institutions can leverage big data methods in education, social media, artificial intelligence, and machine learning to promote greater equity in academic research opportunities.

To deepen our understanding of big data's role across disciplines, future studies should explore how large, complex datasets are being utilized in methodologies and practice in different contexts. Using publicly available funding data to analyze awards outside one's research area can foster creativity, collaboration, and innovation to solve contemporary problems.

6. Limitations

The primary objective of this study is to employ state-of-the-art analytics techniques, including NLP, LDA, and data visualizations, to uncover the trends, patterns, and disparities in data-driven funded research. This study is limited to the abstracts of NSF standard grants led by U.S. institutions of higher learning between 2012 and 2022. Abstracts may not

contain sufficient detail to capture the full scope of the research or its implications, leading to incomplete insights. Furthermore, NSF award decisions are influenced by broader political, economic, and societal trends that are not directly observable in abstracts. In addition to the ambiguity of big data-driven research across disciplines' intent and context, the keywords used to identify data-driven research may not capture all relevant awards due to alternate terminology or implicit applications of the keywords. For example, predictive modeling is a general method and does not specify individual big data techniques such as linear regression, neural networks, or decision trees. Oftentimes, data-driven studies are non-exhaustive and concentrate on a subset of big data's themes of information, technology, methods, and impact, albeit societal, economic, and organizational changes [59].

Author Contributions: Conceptualization, A.K. and S.A.M.; Data curation, A.K. and S.A.M.; Formal analysis, A.K. and S.A.M.; Methodology, A.K. and S.A.M.; Supervision, S.A.M.; Writing—original draft, A.K.; Writing—review and editing, A.K. and S.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work of Arielle King was funded by the North Carolina A&T State University Chancellor's Distinguished Fellowship, a Title III HBGI grant 10 from the U.S. Department of Education.

Data Availability Statement: The data used in this study are publicly available from the National Science Foundation award search site: https://www.nsf.gov/awardsearch/download.jsp. Accessed on 1 October 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This appendix provides the results of topic modeling for identifying subdomains under NSF-defined research areas (see Sections 3.3 and 4.3 in the main text). Specifically, Tables A1–A7 below summarize each directorate's topic labels, terms, and weights (topic count/sum of topic counts within the directorate).

Table A1. Topics modeling results: MPS topic labels.

Topic	Top Terms	Weight
Astrophysics	Galaxies, stars, observations, formations, gas, students, data, team	8.23%
Chemical Reactions	Chemistry, reactions, metal, students, catalysts, synthesis	7.42%
Computational Simulation	Methods, computational, algorithm, learning, develop, model, application, optimization, efficient	5.34%
Differential Equations	Equations, nonlinear, numerical, differential, mathematical, solutions, fluid, analysis, methods, models	7.25%
Geometric Spaces	Theory, geometry, study, geometric, spaces, pi, mathematics, dimensional, manifolds, topology	12.5%
Gravitational Waves	Gravity, wave, black, neutron, physics, holes, relativity	2.08%
Materials Science	Materials, properties, technical, electronic, device, solid, temperature	7.08%
Mathematical Models	Network, mathematical, model, time, stochastic, system, spectrum, control, develop	3.02%
Metal Catalysts	Students, science, material, engineering, development, award, support, engineering, program, education, development, impacts	2.85%
Molecular biology	Cell, biology, material, DNA, student, systems, design	3.10%

Table A1. *Cont.*

Topic	Top Terms	Weight
Molecular Dynamics	Molecule, dynamics, chemical, computational, model, chemistry, experimental, species, methods	2.64%
MPS Education	Conference, workshop, theory, student, support, summer, REU, physics	9.71%
Nuclear Physics	Quantum, physics, plasma, matter, nuclear, measurements, electron, laser, experiments, energy	4.16%
Particle Stability	Particles, surface, water, environmental, liquid, coatings, model, ice	2.25%
Topology	Light, optical, energy, students, nanoparticles, chemical, surface, spectroscopy, properties	4.55%
Statistical Theory	Data, statistical, methods, models, analysis, dimensional, inference, develop,	4.35%
Quantum Mechanics	Systems, theory, quantum, physics, random, study, model, dynamical, particle	3.84%
Protein Structure	Chemistry, protein, molecule, structure, students, program.	5.02%
Polymer Science	Materials, polymer, properties, organic, molecular, application, chemistry, structure, student,	4.61%

Table A2. Topics modeling results: ENG topic labels.

Topic	Top Terms	Weight
Information Technology	Systems, networks, data, wireless, communication, power, distributed, design	5.02%
Structural Design	Building, structures, data, earthquake, soil, engineering, seismic, design	4.51%
Particle Dynamics	Flow, fluid, particle, dynamics, heat, transport, experiments	5.89%
Biosensors	Sensing, detection, optical, device, resolution, micro	5.42%
Thermodynamics	Chemical, gas, reaction, energy, carbon, fuel, production, catalyst, biomass	4.85%
Infrastructure	Data, infrastructure, social, risk, urban, public, impacts, decision, communities, disaster	4.57%
Robotics	Control, human, robots, motion, system, soft, design	3.04%
Supply Chain	Supply production, chain, care, service, models, develop, cost, food	2.68%
Nanoparticles	Nanoparticles, molecule, protein, surface, interactions, DNA, properties, assembly	5.13%
Engineering Education	Workshop, conference, students, engineering, international, support, meeting, science, education, learning, professional, development, REU, STEM	15.30%
Industry Technology	University, center, students, industry, engineering, technology, proposed, program	9.71%
Optimization	System, model, control, methods, optimization, computational, time	7.04%
Manufacturing	Design, manufacturing, support, evaluation, process	2.25%
Materials Science	Materials, properties, mechanical, polymer, process, applications, fundamental, manufacturing	9.77%
Tissue Mechanics	Cell, tissue, mechanical, cancer, disease, stem development	6.49%
Quantum Devices	Quantum, devices, optical, materials, light, semiconductor, photonic, applications	6.72%

Table A3. Topics modeling results: CISE topic labels.

Topic	Top Terms	Weight
CISE Education	Students, science, learning, computer, school, education, computing, teachers, support, undergraduate	6.02%
Computational Molecular Biophysics	Computational, simulation, materials, molecular, Software, methods, physics, biological, science	3.00%
Cloud Computing	Compute, data, performance, systems, memory, Applications, cloud, hardware, parallel	7.84%
Wireless Networks	Network, wireless, communication, spectrum, internet, traffic, data, design	7.35%
Online community Networks	Social, design, support, people, information, public, community, understanding, technology, online	7.07%
Computational Theory	Algorithms, theory, computation, optimization, applications, efficient, science, computer, method	6.46%
Human Robot Interaction	Human, control, robot, system, physical, autonomous, real, time	6.11%
Information Technology	Data, network, science, university, infrastructure, community, support, campus, resources, computing	5.31%
Software Analysis	Software, systems, code, techniques, programming, tools, analysis, verification, program	5.37%
User Experience	Doctoral, consortium, information, students, community, participants, feedback, science, gravitational, human	1%
Data Privacy	Privacy, users, mobile, data, device, web, information, access, techniques	3.54%
Cybersecurity	Security, attacks, quantum, systems, cyber, information, cybersecurity, techniques	4.19%
Health IT	Human, system, patient, medical, data, clinical, time	1.9%
Hydroinformatics	Data, science, scientific, community, cyberinfrastructure, water, software, tools, support	5%
Software Engineering	Software, community, tools, engineering, design, infrastructure, support, development, source, evaluation	2%
Virtual Reality	Virtual, visual, speech, computer, video, vision, 3d, reality, systems, recognition	2%
Machine Learning	Learning, machine, models, data, algorithms, deep, ai, model, methods, evaluation	5%
Knowledge Graphs	Data, analysis, information, techniques, methods, mining, knowledge, graph, algorithms, search	5%

Table A4. Topics modeling results: BIO topic labels.

Topic	Top Terms	Weight
Bio Education	Scientists, workshop, meeting, conference, biology, support, career, students, students, training, REU	11.8%
Biogeochemical Cycles	Carbon, soil, water, ecosystem, nitrogen, forest, streams, organic, climate	6.42%
Climate Change	Species, change, climate, environmental, tree, responses, effects, drought, ecological, temperature	7.34%
Computational Biology	Data, methods, tools, develop, models, analysis, computational, community, software	4.7%

Table A4. *Cont.*

Topic	Top Terms	Weight
Epidemiology	Disease, host, virus, immune, infection, pathogen, transmission, COVID, diseases	4.22%
Ethology	Behavioral, social, species, animals, study, understanding Reproductive, individuals	5.23%
Evolution Genetics	Species, evolution, genetic, diversity, traits, genomic, study, populations	8.92%
Marine Biology	Students, marine, university, instrument, system, undergraduate, science, training, support	5.4%
Metabolic Engineering	Synthetic, systems, engineering, biology, metabolic, chemical, design, develop, molecular	4.4%
Microbial Interactions	Plant, microbial, species, diversity, communities, fungi, interactions, host	5.9%
Molecular Biology	Gene, expression, RNA, genome, DNA, cell, function	8.57%
Neuroscience	Neurons, system, students, sensory, activity, mechanisms, memory, animal, behavior	3.69%
Physiological Responses	Plant, stress, crop, growth, response, signaling, students, molecular	4.05%
Pollination Mechanisms	Plant, flowering, data, pollinators, species, time	1.29%
Protein Structure	Cell, proteins, cellular, molecular, students, understanding, signaling, iron	5.6%
Specimen Digitization	Collections, specimens, biodiversity, species, collection, museum, digitization	5.15%
Trophic Interactions	Food, species, prey, interactions, models, experiments, communities, students, ecosystems, understanding	3.12%
Urban Ecology	Ecological, urban, species, human, natural, coastal, land, data, ecosystem, change	4.24%

Table A5. Topics modeling results: SBE topic labels.

Topic	Top Terms	Weight
Social Network Analysis	Data, methods, analysis, social, information, develop, tools, statistical, science, network	17.6%
SBE Education	Science, students, stem, program, workshop, education, training, university	8.6%
Archaeology	Social, archaeological, political, study, local, communities, data, ancient, society	7.12%
Neurolinguistics	Speech, linguistics, English, words, children, understanding, learning, processing	6.1%
Behavioral Economics	Behavior, decision, people, models, social, theory, information, individuals	5.9%
Cognitive Neuroscience	Cognitive, learning, Neural, Human, memory, visual, understanding, information, activity, development	5.8%
Emergency Management	Social, public, covid, risk, political, pandemic data, survey, support, information	5.8%
Primate Genomics	Human, genetic, data, primate, species, biological, study, understanding	5.6%
Urban Political Economy	Urban, political, social, local, cities, public, economic, development	5.5%
Economic Policy	Economic, policy, financial, data, effects, income, market, labor	4.7%
Environmental Sustainability	Environmental, food, land, Energy, Social, Systems, change, climate, communities, development	4.4%

Table A5. *Cont.*

Topic	Top Terms	Weight
Climate Change	Human, environmental, archaeology, change, sites, times Climate, past, data	4.3%
Information Technology	Firms, innovation, market, information, technology, trade, policy, industry, data	3.8%
Indigenous Communities	Indigenous, native, linguistic, American, conference, community, documentation, knowledge	3.5%
Criminal Justice	Legal, law, justice, criminal, court, police, enforcement	3.2%
International Security	Conflict, violence, international, countries, care, security, political, military, medical, war	3.1%
Wildfire Emissions	Spatial, forest, doctoral, climate, human, fire, land, dissertation	2.6%
Industrial Infrastructure	Infrastructure, water, workers, technology, systems, support, impacts, human, public, social	2.4%

Table A6. Topics modeling results: GEO topic labels.

Topic	Top Terms	Weight
GEO EDUCATION	Students, science, workshop, program, geoscience, community, scientists, university, scientific, education	7.8%
Ocean Temperature	Ocean, deep, data, seafloor, hydrothermal, ridge, sea, samples, cruise, program	7.6%
Coral Reef Ecosystems	Marine, species, coral, ecosystem, communities, reef, understanding, environmental	6.2%
Earthquake Dynamics	Earthquake, seismic, slip, zone, data, deformation, subduction, plate, understanding	6.1%
Ground/Surface Water	Water, sediment, river, coastal, erosion, transport, groundwater, flow, rivers, processes	6.0%
Ocean Carbon Cycle	Iron, ocean, FE, water, trace, isotope, chemical, oxygen, elements, isotopes	5.9%
Ocean models	Climate, ocean, variability, model, pacific, circulation, Atlantic, north, tropical	5.8%
Polar Climate	Arctic, communities, change, social, Alaska, human, climate, environmental, local, community	5.7%
Atmospheric Aerosols	Atmospheric, aerosol, organic, cloud, chemistry, air, compounds	5.1%
Artic Climate Change	Ice, sea, arctic, sheet, level, climate, ocean, Greenland, change, model	4.9%
Geologic Times	Climate, records, past, change, time, cores, data, proxy, lake	4.8%
Hurricanes	Climate, soil, precipitation, arctic, fire, weather, permafrost, water, vegetation, land	4.5%
Space Physics	Support, solar, space, instrumentation, university, instrument, funded, system, acquisition	4.3%
Mantle Composition	Mantle, crust, earth, subduction, rocks, plate, seismic, tectonic, deformation, processes	4.2%
Deep-Sea Volcanoes	Volcanic, eruption, processes, rocks, volcanoes, volcano, study, geochemical, understanding	3.8%
Ocean Productivity	Carbon, ocean, co2, organic, nitrogen, production, microbial, biogeochemical, cycle, water	3.6%
Marine Microbials	Antarctic, polar, field, antarctica, study, southern, public, region, students, time	2.9%

Table A6. *Cont.*

Topic	Top Terms	Weight
Measurement System Analysis	Data, system, time, community, development, based, develop, analysis, tools, methods	2.7%
Oceanographic facilities	Carbon, ocean, co2, organic, nitrogen, production, microbial, biogeochemical, cycle, water	2.5%

Table A7. Topics modeling results: EDU topic labels.

Topic	Top Terms	Weight
Cybersecurity	Cybersecurity, security, students, cyber, education, modules, learning, hands, systems, privacy	6.24%
Virtual Learning	Learning, virtual, data, spatial, manufacturing, human, understanding, support, 3d, develop	2.91%
Workforce Development	Training, students, program, education, graduate, science, industry, skills, development, workforce	4.78%
STEM Education	Learning, stem, children, science, study, studies, cognitive, development, knowledge, program	3.51%
STEM Education	Stem, students, graduate, program, school, underrepresented, programs, careers, education, university	4.46%
Student Success	Students, stem, student, low, retention, income, science, support, success, academic	3.86%
Teacher Education	Teachers, teacher, stem, school, teaching, science, mathematics, university, program, Noyce	4.46%
MPS Reasoning	Students, physics, mathematics, reasoning, development, learning, instructional, student, science, instruction	5.81%
Change Theory	Education, stem, change, network, teaching, undergraduate, national, study, nsf, support	5.28%
Geoscience Education	Learning, students, workshops, student, geoscience, development, teaching, materials, based, professional	4.54%
Student Outcomes	Students, stem, data, study, student, learning, outcomes, college, career, examine	5.42%
Biology Education	Students, undergraduate, biology, student, based, science, learning, institutions, stem, community	4.18%
HBCU Support	Undergraduate, award, students, support, institution, black, university, provide, historically, experiences	5.67%
Engineering Education	Students, student, learning, stem, engineering, courses, education, chemistry, undergraduate, skills	7.3%
Online Learning	Students, development, student, data, learning, design, materials, education, content, online	4%
Artificial Intelligence	Students, learning, data, ai, student, computer, online, system, support, programming	4.96%
Informal Learning	Learning, science, workshop, computer, computing, computational, design, education, community, informal	6.24%
Gender Equity	Stem, women, support, education, program, equity, participation, experiences, engineering, impacts	6.70%
Design-Based	Engineering, students, learning, design, based, technology, student, knowledge, courses, mathematics	4.22%
Community Education	Stem, community, program, students, education, college, institutions, support, colleges, university	5.46%

References

1. Gobble, M.M. Big data: The next big thing in innovation. *Res. Technol. Manag.* **2013**, *56*, 64–67. [CrossRef]
2. Strawn, G.O. Scientific Research: How Many Paradigms? *Educ. Rev.* **2012**, *47*, 26.
3. Amado, A.; Cortez, P.; Rita, P.; Moro, S. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *Eur. Res. Manag. Bus. Econ.* **2018**, *24*, 1–7. [CrossRef]
4. Baig, M.I.; Shuib, L.; Yadegaridehkordi, E. Big data in education: A state of the art, limitations, and future research directions. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 1–23. [CrossRef]
5. Bello-Orgaz, G.; Jung, J.J.; Camacho, D. Social big data: Recent achievements and new challenges. *Inf. Fusion* **2016**, *28*, 45–59. [CrossRef]
6. Choi, T.; Wallace, S.W.; Wang, Y. Big data analytics in operations management. *Prod. Oper. Manag.* **2018**, *27*, 1868–1883. [CrossRef]
7. Fredriksson, C.; Mubarak, F.; Tuohimaa, M.; Zhan, M. Big data in the public sector: A systematic literature review. *Scand. J. Public Adm.* **2017**, *21*, 39–62. [CrossRef]
8. Kalantari, A.; Kamsin, A.; Kamaruddin, H.S.; Ale Ebrahim, N.; Gani, A.; Ebrahimi, A.; Shamshirband, S. A bibliometric approach to tracking big data research trends. *J. Big Data* **2017**, *4*, 1–18. [CrossRef]
9. Li, J.; Jiang, Y. The Research Trend of Big Data in Education and the Impact of Teacher Psychology on Educational Development During COVID-19: A Systematic Review and Future Perspective. *Front. Psychol.* **2021**, *12*, 753388. [CrossRef]
10. Ciampi, F.; Demi, S.; Magrini, A.; Marzi, G.; Papa, A. Exploring the impact of big data analytics capabilities on business model innovation: The mediating role of entrepreneurial orientation. *J. Bus. Res.* **2021**, *123*, 1–13. [CrossRef]
11. Eynon, R. The rise of Big Data: What does it mean for education, technology, and media research? *Learn. Media Technol.* **2013**, *38*, 237–240. [CrossRef]
12. Tulasi, B. Significance of Big Data and Analytics in Higher Education. *Int. J. Comput. Appl.* **2013**, *68*, 21–23. [CrossRef]
13. Mohammadi, E.; Karami, A. Exploring research trends in big data across disciplines: A text mining analysis. *J. Inf. Sci.* **2022**, *48*, 44–56. [CrossRef]
14. .Abourezq, M.; Idrissi, A. Database-as-a-Service for Big Data: An Overview. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 157–177. [CrossRef]
15. Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A.H. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; Mckinsey Global Institute: Washington, DC, USA, 2011.
16. Yang, L. Big Data Analytics: What Is the Big Deal? 30 December 2013. Available online: https://english.ckgsb.edu.cn/knowledge/article/big-data-analytics-whats-the-big-deal/ (accessed on 1 October 2024).
17. Favaretto, M.; De Clercq, E.; Schneble, C.O.; Elger, B.S. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE* **2020**, *15*, e0228987. [CrossRef]
18. Jang, H. Identifying 21st Century STEM Competencies Using Workplace Data. *J. Sci. Educ. Technol.* **2016**, *25*, 284–301. [CrossRef]
19. Tang, R.; Sae-Lim, W. Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Educ. Inf.* **2016**, *32*, 269–290. [CrossRef]
20. Davenport, T.H.; Harris, J.G.; Morison, R. *Analytics at Work: Smarter Decisions, Better Results*; Harvard Business Press: Boston, MA, USA, 2010.
21. Li, Y.; Huang, C.; Ding, L.; Li, Z.; Pan, Y.; Gao, X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **2019**, *166*, 4–21. [CrossRef]
22. Shang, C.; You, F. Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era. *Engineering* **2019**, *5*, 1010–1016. [CrossRef]
23. Gobert, J.D.; Sao Pedro, M.A. Digital assessment environments for scientific inquiry practices. In *the Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*; Rupp, A.A., Leighton, J.P., Eds.; Wiley: West Sussex, UK, 2017; pp. 508–534.
24. Athey, S.; Imbens, G.W. Machine learning methods for estimating heterogeneous causal effects. *Stat* **2015**, *1050*, 1–26.
25. Belloni, A.; Chernozhukov, V.; Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **2014**, *81*, 608–650. Available online: http://www.econis.eu/PPNSET?PPN=819207500 (accessed on 1 October 2024). [CrossRef]
26. Zhou, C.; Wang, H.; Wang, C.; Hou, Z.; Zheng, Z.; Shen, S.; Cheng, Q.; Feng, Z.; Wang, X.; Lv, H. Geoscience knowledge graph in the big data era. *Sci. China Earth Sci.* **2021**, *64*, 1105–1114. [CrossRef]
27. Frizzo-Barker, J.; Chow-White, P.A.; Mozafari, M.; Ha, D. An empirical study of the rise of big data in business scholarship. *Int. J. Inf. Manag.* **2016**, *36*, 403–413. [CrossRef]
28. van Altena, A.J.; Moerland, P.D.; Zwinderman, A.H.; Olabarriaga, S.D. Understanding big data themes from scientific biomedical literature through topic modeling. *J. Big Data* **2016**, *3*, 23. [CrossRef]

29. Hu, J.; Zhang, Y. Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. *Scientometrics* **2017**, *112*, 91–109. [CrossRef]
30. National Science Board. *FY 2022 Performance and Financial Highlights*; NSF 23-003; National Science Foundation: Alexandria, VA, USA, 2022. Available online: https://nsf-gov-resources.nsf.gov/2023-03/FY22%20PerfFinHighlights_web-Final-3-9-23.pdf (accessed on 1 October 2024).
31. Card, D.; Chetty, R.; Feldstein, M.S.; Saez, E. Expanding access to administrative data for research in the United States. In *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*; SSRN-Elsevier: Rochester, NY, USA, 2010. [CrossRef]
32. Einav, L.; Levin, J. The data revolution and economic analysis. *Innov. Policy Econ.* **2014**, *14*, 1–24. [CrossRef]
33. Lima, I.D.; Rheuban, J.E. Topics and trends in NSF ocean sciences awards. *Oceanography* **2018**, *31*, 164–170. [CrossRef]
34. Klami, M.; Honkela, T. Self-Organized Ordering of Terms and Documents in NSF Awards Data. In Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007), Bielefeld, Germany, 3–6 September 2007. [CrossRef]
35. Huang, C.; Notten, A.; Rasters, N. Nanoscience and technology publications and patents: A review of social science studies and search strategies. *J. Technol. Transf.* **2011**, *36*, 145–172. [CrossRef]
36. Rasmussen, L. Increasing Politicization and Homogeneity in Scientific Funding: An Analysis of NSF Grants, 1990–2020. Center for the Study of Partisanship and Ideology (CSPI). Report No. 4. 2021. Available online: https://www.cspicenter.com/p/increasing-politicization-and-homogeneity-in-scientific-funding-an-analysis-of-nsf-grants-1990-2020 (accessed on 5 November 2024).
37. Sherwood, R.D.; Hanson, D.L. A review and analysis of the NSF portfolio in regard to research on science teacher education. *Electron. J. Res. Sci. Math. Educ.* **2008**, *12*, 1–19. Available online: https://ejrsme.icrsme.com/article/view/7764 (accessed on 1 October 2024).
38. González, C. Undergraduate Research, Graduate Mentoring, and the University's Mission. *Science* **2001**, *293*, 1624–1626. [CrossRef]
39. Link, A.N.; Scott, J.T.U.S. Science Parks: The Diffusion of an Innovation and Its Effects on the Academic Missions of Universities. *Int. J. Ind. Organ.* **2003**, *21*, 1323–1356. [CrossRef]
40. Smilor, R.W.; O'Donnell, N.P.; Stein, G.M.; Welborn, R.S. The Research University and the Development of High-Technology Centers in the United States. *Econ. Dev. Q.* **2007**, *21*, 203–222. [CrossRef]
41. Zhu, T.; Zhang, X.; Liu, X. Can University Scientific Research Activities Promote High-Quality Economic Development? Empirical Evidence from Provincial Panel Data. *Rev. Econ. Assess.* **2022**, *1*, 34–50. [CrossRef]
42. Klenke, K. *Qualitative Research in the Study of Leadership*; Emerald Group Publishing Limited: Bradford, UK, 2016.
43. Volkova, N.P.; Rizun, N.O.; Nehrey, M.V. Data science: Opportunities to transform education. *CTE Workshop Proc.* **2019**, *6*, 48–73. [CrossRef]
44. The Carnegie Classification of Institutions of Higher Education. October 2023. Available online: https://carnegieclassifications.acenet.edu/ (accessed on 1 October 2024).
45. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022; Available online: https://www.R-project.org/ (accessed on 1 October 2024).
46. Singh, V.K.; Banshal, S.K.; Singhal, K.; Uddin, A. Scientometric mapping of research on 'Big Data'. *Scientometrics* **2015**, *105*, 727–741. [CrossRef]
47. Park, H.W.; Leydesdorff, L. Decomposing social and semantic networks in emerging "big data" research. *J. Informetr.* **2013**, *7*, 756–765. [CrossRef]
48. Wamba, S.F.; Akter, S.; Edwards, A.; Chopin, G.; Gnanzou, D. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *Int. J. Prod. Econ.* **2015**, *165*, 234–246. [CrossRef]
49. Alattar, F.; Shaalan, K. Emerging Research Topic Detection Using Filtered-LDA. *AI* **2021**, *2*, 578–599. [CrossRef]
50. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
51. Silge, J.; Robinson, D. *Text Mining with R: A Tidy Approach*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
52. Ahadi, A.; Singh, A.; Bower, M.; Garrett, M. Text mining in education—A bibliometrics-based systematic review. *Educ. Sci.* **2022**, *12*, 210. [CrossRef]
53. Buyya, R.; Yeo, C.S.; Venugopal, S.; Broberg, J.; Brandic, I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Gener. Comput. Syst.* **2009**, *25*, 599–616. [CrossRef]
54. US Department of Education. FACT SHEET: Biden-Harris Administration Highlights a Record of Championing Historically Black Colleges and Universities (HBCUs). 2023. Available online: https://www.ed.gov/news/press-releases/fact-sheet-biden-harris-administration-highlights-record-championing-historically-black-colleges-and-universities-hbcus (accessed on 1 October 2024).
55. O'Driscoll, A.; Daugelaite, J.; Sleator, R.D. 'Big data', Hadoop and cloud computing in genomics. *J. Biomed. Inform.* **2013**, *46*, 774–781. [CrossRef] [PubMed]
56. Rodríguez-Mazahua, L.; Rodríguez-Enríquez, C.; Sánchez-Cervantes, J.L.; Cervantes, J.; García-Alcaraz, J.L.; Alor-Hernández, G. A general perspective of Big Data: Applications, tools, challenges, and trends. *J. Supercomput.* **2016**, *72*, 3073–3113. [CrossRef]

57. López Belmonte, J.; Segura-Robles, A.; Moreno-Guerrero, A.; Parra-González, M.E. Machine Learning and Big Data in the Impact Literature. A Bibliometric Review with Scientific Mapping in Web of Science. *Symmetry* **2020**, *12*, 495. [CrossRef]
58. Khanfar, A.A.; Kiani Mavi, R.; Iranmanesh, M.; Gengatharen, D. Determinants of artificial intelligence adoption: Research themes and future directions. *Inf. Technol. Manag.* **2024**, 1–21. [CrossRef]
59. De Mauro, A.; Greco, M.; Grimaldi, M. What is Big Data? A consensual definition and a review of key research topics. *AIP Conf. Proc.* **2015**, *1644*, 97–104. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.