IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 8, 2022, accepted March 7, 2022, date of publication March 10, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158675

Reproducibility in Computing Research: An Empirical Study

WULLIANALLUR RAGHUPATHI[1], (Member, IEEE), VIJU RAGHUPATHI[2], AND JIE REN[1]
[1]Gabelli School of Business, Fordham University, New York, NY 10023, USA
[2]Koppelman School of Business, Brooklyn College, City University of New York, Brooklyn, NY 11210, USA

Corresponding author: Viju Raghupathi (vraghupathi@brooklyn.cuny.edu)

ABSTRACT In computing, research findings are often anecdotally faulted for not being reproducible. Numerous empirical studies have analyzed the reproducibility of a variety of research. Our objective, in this study, is to quantify the current state of reproducibility of research in computing based on prior research, using three reproducibility factors—Method, Data and Experiment—to measure three different degrees of reproducibility. Twenty-five variables traditionally utilized to document reproducibility are identified and grouped into three factors, namely Method, Data and Experiment. These variables describe the extent to which these factors are documented for each paper. Approximately 100 randomly selected research papers from the International Conference on Information Systems series, for the year 2019, are surveyed. Our findings suggest that none of the papers documented all the variables. In fact, the results show that relatively few variables for each factor are documented. Some of the variables vary across different categories of papers, and most papers fail in at least one of the factors. Reproducibility scores decrease with increased documentation requirements. Reproducibility may improve over time, as researchers prioritize reproducibility and utilize methods that ensure reproducibility. Research documentation in computing is remarkably limited, resulting in a dearth of reproducible factors. Future research may study the shifts and trends in reproducibility over time. Meanwhile, researchers and publishers must increase their focus on the reproducibility aspects of their papers. This study contributes to our understanding of the status quo of reproducibility in computing research.

INDEX TERMS Reproducibility, computing, method, data, experiment.

I. INTRODUCTION

While reproducibility is historically accepted as a measure of trustworthy science, in recent years there has been a renewed and urgent focus on this area of research [1]–[3]. Certainly, reproducibility should automatically be a critical consideration of every research paper [4]. Not only does reproducibility allow researchers to build on published results but it also facilitates the review process [5], [6]. Reproducible research is becoming an imperative, ensuring transparency and building trust. In addition, reproducibility supports the sharing of methodologies, optimizing collaboration and the rapid dissemination of research [7]. Recently, however, researchers in various disciplines have raised concerns about the reproducibility of published results [8]–[10]. A 2016 survey in Nature found that many of these scientists across a wide range of disciplines had a personal experience of failing to reproduce a result, and that most scientists believed that science was currently facing a 'significant' reproducibility crisis [11], [12]. Key outlets such as the WSJ [13], the Economist [14] and the Atlantic [15]–[17] have all published extended pieces on reproducibility. Thus, reproducibility is not only a challenge in computing; rather, it is pervasive challenge across most disciplines. The fields of psychology [18], biology [19], [20], biomedicine [9], neuroscience [21], drug development [22], chemistry [23], climate science [24], economics [25] and education [26] among others, have reported reproducibility problems [20]. A recent study estimated the cost of funding irreproducible research at approximately $28 billion a year in the U.S. alone [27], [28]. A well-known effort to replicate findings from prominent social and cognitive psychology studies showed fewer significant findings and smaller effect sizes than the original studies [18]. And while reproducibility is considered a fundamental aspect

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

of reliable research, studies show that a substantial number of published research results cannot be reproduced [11], [18], [29]–[35]. This circumstance is particularly true for the papers presented at major conferences and published in top journals. In many cases, even the primary researchers are unable to reproduce their own findings [22], [36]–[38]. In principle, it should be possible to specify a methodology with sufficient detail that anyone can reproduce it exactly, and yet, practically speaking, there are fundamental, technical and social barriers to doing so [38].

The reproducibility problem is more pronounced in computing research perhaps because the computing discipline is multidisciplinary, and the artifacts, both tangible and intangible, are developed and validated in the context of socio-technical approaches to research [39]–[44]. Computing research spans sub-disciplines that include business computing, compilers, embedded and real-time systems, networking, operating systems, user-centered applications and mobile and web applications, among others [45]. In the sub-discipline of software development, advances in research and applications are aided by algorithms, programming languages, tools and models of quality assurance and testing and so forth. But design and coding are subjective processes. Reproducibility is difficult to effectuate when there is no proper documentation of design specifications, pseudocode, prototype, etc. of how the artifacts are developed [45]. Version control is an added challenge. Compounding the problem are big datasets. Meanwhile, the computational methods necessary to process and analyze those datasets has prompted new ways of considering reproducibility [3]. In addition to a substantial lack of reproducibility in computing research, identifying reproducibility problems is itself challenging [2], [3]. This is due to the lack of both frameworks and methods, and the tools necessary to identify reproducibility problems. Further, while there is a lot of buzz about reproducibility, there are very few studies that have actually assessed reproducibility [46]–[48] and there are a scant number of frameworks or models to evaluate reproducibility [49]–[52]. This empirical study attempts to fill many of these gaps. This is a descriptive analytic study that sheds light on the current state of reproducibility in computing by examining papers from a recent conference on information technology. Adapting an existing model in the study of reproducibility in artificial intelligence research and applications [31]–[33], we develop and offer a framework and check list for undertaking reproducibility studies in computing in general. Further, this framework is operationalized with an applied, hands-on check list to evaluate the studies. The usefulness is its potential to be applied to a research paper or a report prior to submission and peer review, or publication. In other words, it provides ex ante support as opposed to other models (e.g., code testing) that are ex post [49], [50], [53]. The model described here can be applied to all aspects of a research paper or publication, namely, data, experiment (or analysis) and method. Lastly, the framework outlined and applied here is not restricted to one or another sub-discipline of computing; rather it can be applied across the board [39], [54]. The framework and check list described here will be useful to both researchers and practitioners in the reproducibility assessments of their work. We build on prior work to argue that numerous factors—including those falling under documentational, experimental and methodological categories—prevent a high degree of reproducibility of computing research [55], [56]. Empirical work that studies reproducibility in the different sub-disciplines of computing has been sporadic and ad hoc at best. This study aims to fill that gap by investigating and shedding light on the nature and dimensions of the reproducibility of current research in the computing discipline. We examine papers published as part of the proceedings of a prestigious business computing and information systems conference, applying an adapted reproducibility evaluation framework and methodology from [31] and [32].

The rest of the paper is organized as follows: First, we provide a comprehensive review of reproducibility. We then provide an overview of the reproducibility framework used in this study and follow this section with a description of research methods. We then provide the results and analysis of our study. Finally, we highlight the scope and limitations of our study and offer conclusions.

II. LITERATURE REVIEW

Publications are at the epicenter of academic life, observe [44]. Computing is in a unique position among scientific disciplines because researchers in the discipline typically eschew the publication process and disseminate their cutting-edge research at conferences. Unlike peer-reviewed publications with multiple review layers, conferences utilize an entry process with a single review stage. Thus, conferences have had a profound impact on the way research is conducted by computing researchers and have provided those researchers with a distinct advantage. To be competitive in the academic world, researchers must play the publishing game, which emphasizes numerical metrics of success [44]. The pressure to publish innovative ideas is biased towards bringing preliminary findings to the public arena as quickly as possible and circumventing the thoughtful, if relatively lengthy, peer evaluation and review process that has been the cornerstone of good research. Compounding this situation is the trend that novelty is replacing research grounded in theory [44]. The inevitable outcome is the degradation of research quality. Simultaneously, computer-based models and tools are being used in scientific research at an exponential rate, but reproducibility methods have not kept pace, leading to skepticism about the results generated by computational methods [39], [57]. As a result, a currently popular discourse is the promotion of awareness and policies designed to intervene, such as the contemporary Association for Computing Machinery (ACM) policy on the scrutiny of outputs and systems and badging [58]. The ACM construes research to be reproducible [29] when its findings can be generated by another team utilizing a different dataset. Journals, too, have begun to demand better documentation and, to the extent possible,

more openness (e.g., making data available publicly). IEEE has also set up The Ad Hoc Committee on Open Science and Reproducibility. The goal of the 2020 Ad Hoc Committee on Open Science and Reproducibility ''is to analyze models, practices and experiences in supporting open science and reproducibility within the IEEE Computer Society (CS) and at peer societies and publishers'' [59]. Against this backdrop, many studies have emerged that look at aspects of reproducibility across the different sub-disciplines of computing. For example, in a recent IEEE study approximately 60% of IEEE conferences, magazines and journals have no policies and procedures in place to ensure research reproducibility [60]. In another example, [61] report that fewer than approximately 15% of MobiHoc papers (2000-2005) that utilized simulations (114 out of 151 papers) for MANET analysis were repeatable. [62] verified 134 papers published in the IEEE Transactions on Image Processing and found that only 33% of the papers published the datasets, while only 9% of the papers made available the code needed for reproducibility. Recently, [5] looked at about 600 papers from ACM conferences and journals and identified repeatability weaknesses in approximately 32% of the papers. Their study also found that a few researchers were unwilling to share their code and data. In instances where they *were* shared, too little information was provided to repeat the experiment. [63] evaluated the computational reproducibility of 204 papers and their ability, as independent researchers, to acquire the resources necessary to reproduce a paper's findings. The authors were able to retrieve the tangible products from 44% of the sample but only able to reproduce the results for 26% [63]. [64] analyzed data from Scopus, which showed that the reproducibility problem was prevalent in several other fields as well. [65] suggested that it was difficult to confirm most results in current conferences. Recent studies [5], [66], [67] have also shown that the peer-review process by itself is incapable of ensuring reproducibility, an obvious point given the process is not designed to check for reproducibility. Additionally, according to [68], the ''publish or perish'' mentality is a significant problem: ''Innovative findings produce the rewards of publication, employment and tenure; replicated findings produce a shrug.'' [67] and [69] suggest that in the future reproducible submissions should always be the default and that doing reproducible research will become imperative [6]. To that end, scientists, institutions and funding agencies have been pushing for the development of methodologies and tools that preserve software artifacts. Still, the consensus is that long-term reproducibility remains, in computing research, elusive [70]. This is a problem given that the scientific method depends on reproducibility to back up the development of scientific knowledge. When scientists cannot conduct the same experiment and obtain the same findings as the initial researchers, the event implies the hypothesis is false [71]. Therefore, the failure to reproduce findings affects the very integrity of science [9], [67], [72], [73]. To wit, there are very few empirical studies of reproducibility in computing [74], [75], and the few studies that were done focused on granular methods to test reproducibility such as access to data, code compilation, software quality testing, etc. [39], [76]. Furthermore, not only are there very few studies on reproducibility but there are also even fewer methods to study reproducibility [77]–[79]. Thus, there is a paucity of studies as well as methods making it beneficial to undertake additional studies and develop broader methods. This is a motivation for our study. To reiterate, this study makes a modest attempt to shed light on both aspects. In addition, it is only recently that attempts are being made to develop automated tools to assist with reproducibility [77], [79]–[81]. However, nearly all of these are at the development stage [79], [81]. In summary, very few studies exist, as highlighted above, and they have typically focused on ex post reproducibility. This study is different in that it conducts a reproducibility evaluation before a paper is published. To ensure reliability in computing research, steps must be taken to increase the reproducibility of the research [31]–[33], [82]–[85]. In the meantime, the current - and unfortunate - state of reproducibility in computing research must be documented.

Our goal with this study is to assess the current state of reproducibility in empirical computing research. Our chief *proposition* is that the documentation in computing research is insufficient to reproduce the published findings; that is, current documentation practices at top business, computing, and academic conferences cede much of the published findings to non-reproducibility. We surveyed research papers from the most prestigious information systems conference, namely ICIS, to test the proposition. Our research contributions are multi-fold: (i). we assess the contemporary status of reproducibility in computing research and provide a panoramic overview by conducting an empirical analysis; (ii). we develop a framework and operationalize it with a check list to verify reproducibility in computing papers, and (iii). we investigate the implications of reproducibility for computing research and offer prescriptive recommendations.

A. OVERVIEW OF REPRODUCIBILITY

There is consensus among researchers that empirical results ought to be reproducible but the definition and meaning of reproducibility is not clearly understood [18], [31]–[34].

For this study, we define reproducibility in empirical computing research as:

''*the ability of an independent research team to produce the same results using the same research method based on the documentation made by the original research team* (adapted from [32])."

The reproducibility evaluation framework developed by [31], [32] and utilized to analyze reproducibility in artificial intelligence research is the basis for this study. This part of the narrative is largely paraphrased from their seminal work. The key point to emphasize is that a separate group of researchers ought to be able to generate the same findings as the initial researchers primarily using the original documentation. The documentation, therefore, is key to ensuring that

the independent team can conduct the exact same research and obtain the same results as the original team [31], [32]. Typical computing research documentation is comprised of three parts: the documentation of the research method that the original research team developed and aims to validate; the data (if any) that is used in the research; and a description of an experiment in text and code form. When the findings of the initial research and those of the reproduced results are similar, one can conclude it is possible to reproduce the initial research.

B. REPRODUCIBILITY DOCUMENTATION

Documentation is the key starting point to reproducibility. To reproduce the results of the research, the documentation must include relevant information and must be specified to a granular level. Researchers must clearly identify what is relevant and how fine-grained the documentation must be to make sure that results can be reproduced using only this information [32]. Following this framework, we also grouped the documentation into three categories: *Method, Data* and *Experiment*. The documentation for the research method includes the description of the computing research method as well as its research question [31], [32]. Additionally, data, along with the documentation describing the data and how it can be used, are necessary for reproducibility. Therefore, data engineering and preprocessing are important. The goal is to make available the cleaned exact dataset. Version control is also necessary. Finally, to compare results, the actual output of the research is required [31], [32]. If the research involves conducting an experiment, proper documentation detailing the exact steps involved, including the analysis and results, must be made available [31], [32]. The hardware and software used must be properly specified. While methods and data are required in most research studies, experiments in computing research, more likely dealing with tangible artifacts, are typically more ad hoc. Overall, the extent of documentation in terms of method, data and experiment sits on a continuum of degrees of reproducibility. The 'gold standard' is the ability to share documentation for all three categories in an open and transparent way (e.g., putting everything in a cloud environment) [86], [87]; but the cost of such an infrastructure could be high. Plus, maintenance and updates require ongoing attention.

Following the lead of [31], [32] the documentation factors —methods, data and experiment - enable the definition of the three degrees to which the original results can be reproduced. The degrees are quantified into a numerical score as described in the two Gundersen and Kjensmo papers. *R1: Experiment reproducible* implies the inclusion of all three factors, and by following the document, independent researchers can reproduce the results; *R2: Data reproducible* includes method and data and implies the research is potentially a data-driven empirical study. Alternative researchers ought to be able to arrive at similar findings using this documentation; and *R3: Method reproducible* implies that the method alone is documented, and an independent set of researchers may reproduce the results using this documentation. Figure 1 depicts how the three degrees relate to one another and which degree of reproducibility requires what type of documentation.

	Method	Data	Experiment
R1 (Experiment reproducibility)	X	X	X
R2 (Data reproducibility)	X	X	
R3 (Method reproducibility)	X		

FIGURE 1. The three degrees of reproducibility (Source: [15], [16]).

Drawing from the literature and basing our research squarely on the adaptation of the model developed and tested in [31], [32], our goal, as stated, is to quantify the state of reproducibility of empirical computing research. We mean to show that the documentation of computing research is not of a high-enough quality to reproduce the reported results, and that the current documentation practices at a top business computing and information systems conference do not support the outcome that reported research results will be reproducible.

III. RESEARCH METHODS

Following [32], [55], an observational study in the form of a manual survey of research papers was conducted to generate quantitative data about the state of the documentation quality of business computing research. Each paper was read several times to extract the values for the variables in each factor. The research papers were reviewed, and a set of 25 variables were manually identified. To compare results among papers and groups of papers, we used three reproducibility metrics - R1D, R2D and R3D - to score the documentation quality. As stated, the research method in this study is adapted, with several modifications, from [31], [32]. (For more details regarding the methodology, please refer to those papers.) Using a data-driven approach, visualization & descriptive analytics [88], [89], well-established methods of analysis, were applied to this dataset of papers to gain insight into reproducibility [88], [90]. The emerging field of visual analytics allows us to graphically represent the data and thereby visualize the results to gain insight [90]–[92]. By integrating a proper design with visual techniques, charts and statistics can be generated [93], [94]. Visual analytics help aggregate, process and represent large amounts of data in easy-to-understand charts [90], [92], [94]. The overall objective is to tell the stories through visualization [90], [93]. Compared to other types of analytics, descriptive analytics tends to be more data driven; its focus is on describing the data 'as is' with no preconceived assumptions [91]. Descriptive analytics via visualization eases the understanding of historical and current trends to make meaningful decisions [89], [93], [94].

A. SURVEY

To evaluate the hypothesis, we surveyed a total of 125 papers from the 2019 Association for Information Systems (AIS) proceedings of the International Conference on Information Systems (ICIS 2019) (https://aisel.aisnet.org/icis2019/). The ICIS's own description (https://aisnet.org/page/ICISPage) supports our choosing this set of papers:

"The International Conference on Information Systems (ICIS) is the most prestigious gathering of information systems academics and research-oriented practitioners in the world. Every year its 270 or so papers and panel presentations are selected from over 800 submissions."

Studying a sample of documents from this conference, wherein papers are chosen after a rigorous review process, was deemed appropriate. Because the number of papers under each topic in ICIS 2019 varies, we randomly selected 5 to 11 papers in each topic to maintain a balance of topics and avoid selection bias. As a result, a total number of 19 topics and 125 papers were reviewed. Of these 125, 100 papers comprised empirical research, and 25 were conceptual. A panel of researchers manually classified the papers into empirical and conceptual research types. After dropping the conceptual papers, researchers proceeded to analyze the reproducibility performance of the 100 empirical papers. Table 1 shows the number of published papers (the population size) and the number of surveyed papers (sample size).

TABLE 1. Population size and sample size of papers.

	ICIS 2019 (Population Size)	Selected for reviewed (Sample Size)
# of topics	26	19 (aggregated into 6 major topics as shown in Table 2).
# of papers	431	100 (empirical papers).

The ICIS 2019 identified 26 total topics. During data collection, five of the topics were dropped because there were fewer than five papers on each topic. The remaining 19 topics were aggregated into six major topics, as shown in Table 2.

We also analyzed the papers by paper length (full vs. short) and topic (six topics). Figure 2 shows the breakdown of the papers by topic (six topics) and paper length (full vs. short). 'full' indicates that the article is complete, while 'short' indicates that it is just part of the full article. Short papers typically have a length of about 10 pages; full papers run about 18 pages. There is a 50:50 balance of full and short papers in the 100-paper sample reviewed. Of the 100 empirical papers surveyed, most fall under the topics of analytics, data science and smart systems (27%); business models, digital transformation and innovation (26%); and other topics (21%). The distribution among the other three topics—cybersecurity, privacy and ethics of IS (11%), sustainable and societal impact of IS (8%) as well as human computer interfaces (7%)—is relatively small. Note that regrouping the topics

TABLE 2. Aggregation of topics.

	TOPIC	Sub-topics
1-SUS	Sustainable and Societal Impact of IS	Sustainable and Societal Impact of IS
2-CPE	Cyber-security, Privacy and Ethics of IS	Cyber-security, Privacy and Ethics of IS
3-BDT	Business Models, Digital Transformation and Innovation	Business Models and Digital Transformation Crowds, Social Media & Digital Collaborations Digital Learning Environment & Future IS Curriculum DLT, Blockchain & FinTech Innovation and Entrepreneurship
4-ADS	Analytics, Data Science and Smart Systems	Analytics & Data Science Smart Service System & Service Science Mobile IoT & Ubiquitous Computing Digital government & smart cities
5-HCI	Human Computer Interface	Human Computer / Robot Interactions and Interfaces Human Behavior & IS
6-GEN	Other Topics	General Topics IS in Healthcare IS Development & Implementation Economics & IS Design Science Research Future of Work

caused an imbalance in the number of papers surveyed. While each of the three dominating topics includes more than three sub-topics defined by ICIS, the other three topics include only one or two sub-topics.

B. FACTORS AND VARIABLES

Adapting the process in [31], [32], we treated the three types of documentation, namely Method, Data and Experiment, as the factors specified by 25 different variables. Sixteen of the variables from prior studies were deemed fit for the study of reproducibility in Information Systems research. An additional 12 IS-domain relevant variables were added, for a total of 25 variables. Table 3 shows the factors, variables and their description.

Unless otherwise specified, each variable in Table 3 was encoded as a 1 or 0, where 1 represents an explicit mention of the variable in the paper, and 0 represents no explicit mention.

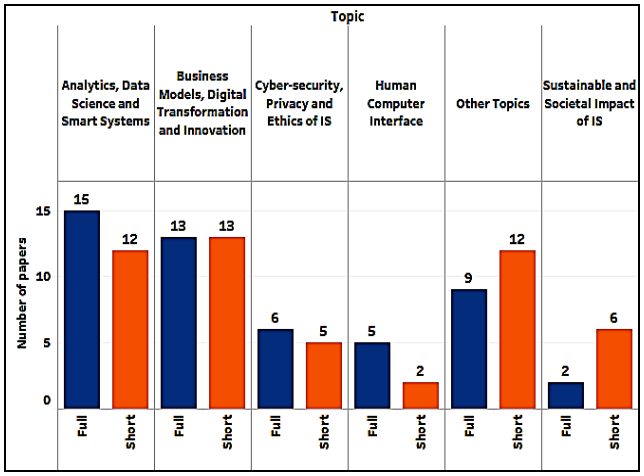

FIGURE 2. Distribution of papers by paper length and topic.

For example, while reviewing the variable 'Goal', each paper was reviewed manually for an explicit mention of the research goal, such as "Our research goal is. . ." or "The goal of the research is to. . .". Similarly, all variable codes were manually assessed by each researcher for all papers. The codes for each paper were then compared, and any resulting discrepancies were resolved by a combined re-evaluation of the paper in question until a consensus was reached. In this way, an inter-rater reliability of 90% was achieved. To reiterate, we used the reproducibility metrics from [31], [32] to quantify whether a paper is R1D, R2D, or R3D reproducible, and to what degree.

IV. RESULTS AND ANALYSIS

The data was analyzed using Python for its data preprocessing, descriptive statistics, and correlation analysis capabilities. Tableau, the business intelligence tool, was used to visualize the reproducibility outcomes. We initially present below the descriptive statistics for the metrics and factors.

Table 4 presents the descriptive statistics for the three composite reproducibility metrics. R1D is a composite score that covers Method, Data and Experiment, while R2D covers Method and Data and R3D represents the value of Method only. The mean for R3D (0.6657) is the highest, followed by R2D (0.5634) and R1D (0.4256). These outcomes demonstrate that most papers tend to share the documentation for Method only, rather than for all three (including Data and Experiment).

Table 5 below presents the descriptive statistics for the three factors measuring reproducibility. The average of Method (0.6657) is the highest, followed by Data (0.4611) and Experiment (0.15). Again, these outcomes suggest that there is a trend for sharing the methodology, which makes methodology more reproducible. Some papers, though deemed empirical, did not conduct an experiment (e.g., an analysis) involving data, which may a least partially explain why Data and Experiment are less reproducible. In addition, data sharing is still challenging for several reasons, including ownership, confidentiality, copyright and competitive advantage. Finally, the experiments may not be sufficiently standardized.

TABLE 3. Method, data and experiment and the variables that specify them.

Factor		Variable	Description
Method (14)	1	Problem statement	Is there an explicit mention of the problem the research seeks to address?
	2	Goal	Is the research goal explicitly mentioned?
	3	Research question	Is there an explicit mention of the research question(s) addressed?
	4	Research method	Is there an explicit mention of the research method used?
	5	Algorithm	Is there an explicit mention of the algorithm(s) the research used?
	6	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	7	Prediction	Is there an explicit mention of prediction related to the hypotheses?
	8	Experiment setup	Are the variable settings shared, such as hyperparameters?
	9	Contributions	Does the paper state the contributions or implications of the research?
	10	Related study	Does the paper explicitly mention related literature?
	11	Scope and limitations	Is there an explicit mention of scope and limitation of the research?
	12	Machine learning	Is there an explicit mention of using machine learning for analysis?
	13	Statistical analysis	Does the paper conduct any statistical analysis?
	14	Conclusion	Is there an explicit outcome concluded in the paper?
Data (9)	15	Model results	Is the output of the model constructed shared?
	16	Training data	Is the training set shared?
	17	Validation data	Is the validation set shared?
	18	Testing data	Is the test set shared?
	19	Evaluation criteria	Is the evaluation metrics (e.g., Accuracy, R-squared, etc.) of the model shared?

TABLE 3. *(Continued.)* **Method, data and experiment and the variables that specify them.**

	20	Data preprocessing	Is the method of data preprocessing of the analysis shared, including data merging or feature engineering?
	21	Publicly available data	Is the data available to the public?
	22	Time series data	Does the paper use time series data for analysis (e.g., across multiple years or months)?
	23	Data Source	Does the paper explicitly state the data source?
Experiment (2)	24	Method source code	Is the system code available as open source?
	25	Software used	Is the software used in the research explicitly mentioned (Python, R, etc.)?

TABLE 4. **Descriptive statistics for the three composite reproducibility metrics.**

Metric	Count	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
R1D	100	0.4256	0.1451	0.1561	0.3142	0.4087	0.4861	0.8307
R2D	100	0.5634	0.1435	0.2341	0.4524	0.5794	0.6706	0.873
R3D	100	0.6657	0.1201	0.3571	0.5714	0.7143	0.7143	0.9286

TABLE 5. **Descriptive statistics for the three reproducibility factors.**

Factor	Count	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
Method	100	0.6657	0.1201	0.3571	0.5714	0.7143	0.7143	0.9286
Data	100	0.4611	0.2173	0.1111	0.3333	0.4444	0.6667	1
Experiment	100	0.15	0.2513	0	0	0	0.5	1

Table 6 displays the descriptive statistics for the absolute scores, the sums of variables listed under the three factors Method, Data and Experiment. Similarly, the overall absolute score (Abs Overall) represents the sum of variables across the three factors. On average, each paper has approximately 13.77 reproducibility variables, in which 9.32 are Method, 4.15 are Data, and only 0.3 are Experiment. It is noticed that overall, there are 25 variables, in which 14 are for Method, 9 for Data, but only 2 for Experiment. Each paper has at least 5 variables for Method, and more than 50% of the papers have more than 10 variables for Method. Each paper also has at least one Data variable, and about 25% of the papers have three or fewer Data variables. More than half of the papers do not show reproducibility for the Experiment factor.

TABLE 6. **Descriptive statistics for absolute scores.**

Absolute Score	Count	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
Abs Method	100	9.32	1.681	5	8	10	10	13
Abs Data	100	4.15	1.956	1	3	4	6	9
Abs Experiment	100	0.3	0.5025	0	0	0	1	2
Abs Overall	100	13.77	3.2188	6	11	14	16	21

Table 7 presents the descriptive statistics for the variables comprising the factor Method for the 100 empirical papers. The frequency count indicates the number of papers that explicitly mentioned the variable. For example, the frequency count of 86 for 'Goal' indicates that 86 papers mentioned the research goal. Over 90% of the documentation surveyed mentioned the problem statement (97%), research method (93%) and conclusion (94%).

Table 8 presents the descriptive statistics of the sample of 100 empirical papers for the variables making up the Data factor. The frequency count, again, represents the number of papers with the specific variable. All 100 papers surveyed mentioned the source of data, whether primary or secondary. More than half of the documentation surveyed provided the model results (65%) and evaluation criteria (57%).

Table 9 presents the descriptive statistics for the two variables comprising the factor Experiment for the 100 empirical papers. Only 7% of the documentation shared the method's source code, and only 23% identified the software used for analysis.

Table 10 shows the mean score of the three reproducibility factors in each topic. Cyber-security, Privacy and Ethics of IS as paper topics have the highest average Method score (0.7143), Analytics, Data Science and Smart Systems papers score highest in Data (0.5062). Papers in Business Models, Digital Transformation and Innovation provide the highest score in Experiment (0.2115).

Table 11 shows the mean value of R1D, R2D and R3D by topics. Analytics, Data Science and Smart Systems (0.4439), Business Models, Digital Transformation and Innovation (0.4462) and Other Topics (0.4478) have the highest R1D score. Analytics, Data Science and Smart Systems have the

TABLE 7. Descriptive statistics for the 14 variables of method.

Variable	Count	Percentage	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
Problem Statement	97	97%	0.97	0.1714	0	1	1	1	1
Goal	86	86%	0.86	0.3487	0	1	1	1	1
Research question	76	76%	0.76	0.4292	0	1	1	1	1
Research method	93	93%	0.93	0.2564	0	1	1	1	1
Algorithm	35	35%	0.35	0.4794	0	0	0	1	1
Hypothesis	46	46%	0.46	0.5009	0	0	0	1	1
Prediction	24	24%	0.24	0.4292	0	0	0	0	1
Experiment setup	70	70%	0.7	0.4606	0	0	1	1	1
Contributions	90	90%	0.9	0.3015	0	1	1	1	1
Related study	81	81%	0.81	0.3943	0	1	1	1	1
Scope limitations	68	68%	0.68	0.4688	0	0	1	1	1
Machine learning	15	15%	0.15	0.3589	0	0	0	0	1
Statistical analysis	57	57%	0.57	0.4976	0	0	1	1	1
Conclusion	94	94%	0.94	0.2387	0	1	1	1	1

highest R2D score (0.5917) while Cyber-security, Privacy and Ethics of IS have the highest R3D score (0.7143).

A. FACTORS

Figure 3 depicts three diagrams that spider plot the means for the variables in each of the three factors of Method, Data and Experiment for the sample of empirical papers. Under the Method factor, the problem statement, research method, and conclusion have the highest scores; more than 90 percent of the papers contain these variables. Algorithm, machine learning, and prediction appeared least often. Under the Data factor, data source, evaluation criteria and model results are mentioned most often, and data preprocessing is barely discussed at all. Under the Experiment factor, even though there are only two variables, it appears that the frequency of method source code and software used are below 30 percent, indicating that most papers do not give sufficient details about the experiments to support reproducibility. Comparing the spider plots reveals that the business computing research papers we examined pay more attention to the Method factors, with many variables scoring above 80.

TABLE 8. Descriptive statistics for nine variables of data.

Variable	Count	Percentage	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
Model results	65	65%	0.65	0.4794	0	0	1	1	1
Training data	45	45%	0.45	0.5	0	0	0	1	1
Validation data	21	21%	0.21	0.4094	0	0	0	0	1
Testing data	23	23%	0.23	0.423	0	0	0	0	1
Evaluation criteria	57	57%	0.57	0.4976	0	0	1	1	1
Data preprocessing	34	34%	0.34	0.4761	0	0	0	1	1
Publicly available data	40	40%	0.4	0.4924	0	0	0	1	1
Time series data	30	30%	0.3	0.4606	0	0	0	1	1
Data Source	100	100%	1	0	1	1	1	1	1

TABLE 9. Descriptive statistics for two variables of experiment.

Variable	Count	Percentage	Mean	Standard deviation	Min	0.25	0.5 Median	0.75	Max
Method source code	7	7%	0.07	0.2564	0	0	0	0	1
Software used	23	23%	0.23	0.423	0	0	0	0	1

TABLE 10. Average method, data and experiment scores in each topic.

	Analytics, Data Science and Smart Systems	Business Models, Digital Transformation and Innovation	Cyber-security, Privacy and Ethics of IS	Human Computer Interface	Other Topics	Sustainable and Societal Impact of IS
Method (Mean)	0.6772	0.6484	0.7143	0.6122	0.6769	0.6339
Data (Mean)	0.5062	0.4786	0.4343	0.3810	0.476	0.3194
Experiment (Mean)	0.1481	0.2115	0.0909	0	0.1905	0.0625

Variables such as problem statement, research method and conclusions, which have scores over 90, are typically given priority in these papers. In contrast, the Experiment variables

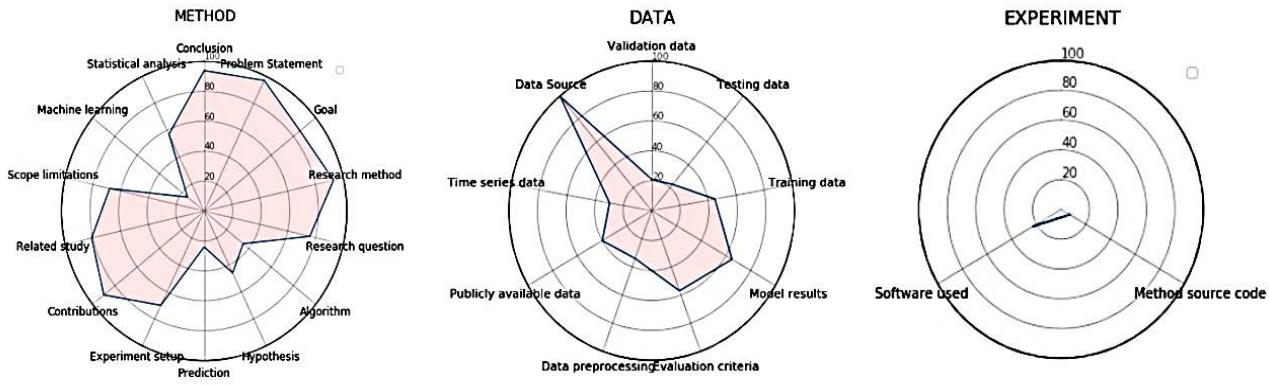

FIGURE 3. Spider plot with variables of method, data, and experiment for the papers.

TABLE 11. Average R1D, R2D and R3D of topics.

	Analytics, Data Science and Smart Systems	Business Models, Digital Transformation and Innovation	Cyber-security, Privacy and Ethics of IS	Human Computer Interface	Other Topics	Sustainable and Societal Impact of IS
R1D (Mean)	0.4439	0.4462	0.4132	0.3311	0.4478	0.3386
R2D (Mean)	0.5917	0.5635	0.5743	0.4966	0.5765	0.4767
R3D (Mean)	0.6772	0.6484	0.7143	0.6122	0.6769	0.6339

score at 20 or less, indicating that experiment details are scant or absent. These findings are understandable: it is relatively more difficult to explain the details of software and code than the details of other aspects of the research. Likewise, typical empirical papers in business computing research are more data-driven, and focus on association or correlation rather than on causality, for which experiments are more appropriate.

B. REPRODUCIBILITY METRICS

The results for the reproducibility metrics appear in Figure 4. These bar charts show the distribution of scores for Method, Data and Experiment, and none of them follow a normal distribution. The charts show the mean values for variables for each of the factors described in Table 3. For example, Figure 4 shows papers usually have a better score in the Method factor, indicated by the range of the scores between 0.6 and 0.8. Papers in the Experiment factor typically have a lower score. The scores of many papers fall in the range of 0.1 to 0.75 for Experiment, indicating that the papers mentioned very little about their experiments. Relevant information, such as source code or details about the software used for analysis, while important for artifact design, matters little to data-driven research. The Conclusion variable score is in line with the descriptive statistics provided earlier. Generally, business computing researchers have a strong awareness of the details in regard to Method, while reproducibility can be improved further by providing more details about Data and Experiment.

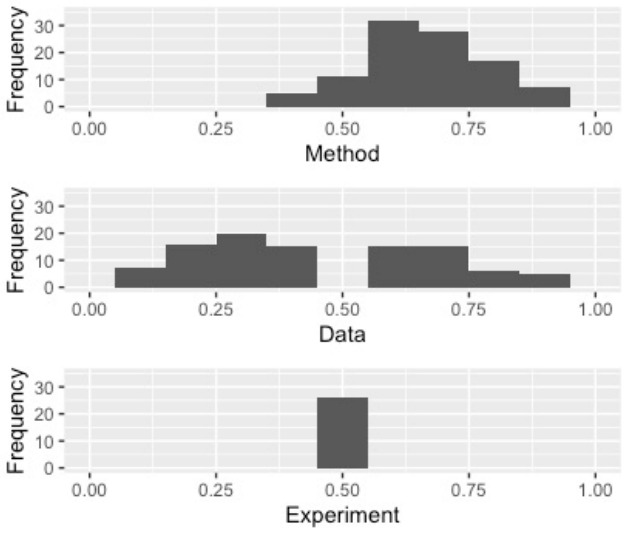

FIGURE 4. Distribution of papers scores.

Figure 5a depicts bar charts of the frequency of the composite reproducibility metrics R1D, R2D and R3D for all papers. Figure 5b depicts the distribution for the six groups of topics. To recap, R1D is a composite score that covers all the values of Method, Data and Experiment; R2D covers Method and Data; and R3D covers Method only. Figure 5(a) shows distinct variations in the frequency distributions. Most of the

papers have an R1D of 0.2 to 0.5, while a few have an R1D in the range of 0.6 to 0.8. According to the analysis by topic for R1D, Figure 5(b) shows that papers in analytics, data science and smart systems, as well as business models, digital transformation and innovation have the highest R1D score, at over 0.44. As indicated by the composite reproducibility score, reproducibility for R1D is not high. The bar charts for R2D (figure 5a) show that the highest frequency ranges are from 0.3 to 0.5, and no papers have an R2D below 0.1. This finding shows that reproducibility performance is much higher when Experiment is not included. For R3D, Figure 5a shows that the highest frequency falls in the interval of 0.6 to 0.8, and almost no papers have an R3D measuring below 0.25. In terms of distribution by topic (Figure 5b) papers in analytics, data science and smart systems have the highest average R2D score (over 0.59), and papers in each topic have a mean score of over 0.47. Cyber-security, privacy and ethics score the highest in R3D (0.71), followed by analytics, data science and smart systems papers (0.68). This means that overall, the reproducibility performance of Method is better than that of Data and Experiment. And analytics, data science, and smart systems papers usually produce better reproducibility levels than papers in other topics, although the difference is not significant.

Figure 6 shows the analysis by paper length (full vs. short). The scatter plot with trend lines shows blue squares representing short papers and orange crosses representing full papers. The X- and Y-axis depict the average scores of Method and Data respectively. The chart shows a high correlation between Data and Method in both paper lengths: also, if a paper performs well for Data, it is likely to perform well for Method, too ($p<0.05$). Thus, the quality of reproducibility in terms of referring to the Method and Data metrics, shows a significantly positive relationship with the coefficient estimates (0.77 for short VS 0.57 for full) greater than 1. The R-squared for short papers (0.1613) is slightly higher than that for a full paper (0.1066), indicating that for short papers a larger variation in Data scores can be explained by the Method scores. Papers that share details on their method are highly likely to share details of their data, especially for short papers.

Figure 7 is a scatter plot that shows the linear association between overall reproducibility - R1D (which is the weighted average of Data, Method and Experiment) and the reproducibility of method and data - R2D (which is the weighted average of Data and Method). The blue square represents short papers, and the orange cross represents full papers, with the size of the square representing the composite reproducibility R1D score. The chart shows that R1D increases as R2D increases for both types of papers, with R2D significantly accounting for more than 67% ($p<0.0001$, $R^2 = 0.6722$) of the variation in R1D. In other words, overall reproducibility is largely determined by the disclosure in the data and method sections. Compared to short papers (0.74), R2D of the long papers tend to have a stronger impact on R1D indicated by a higher coefficient estimate (0.93). Most of the highest-scored papers at the top-right corner are long papers. Therefore, papers with high R2D scores always have higher R1D scores, and long papers generally reflect higher reproducibility.

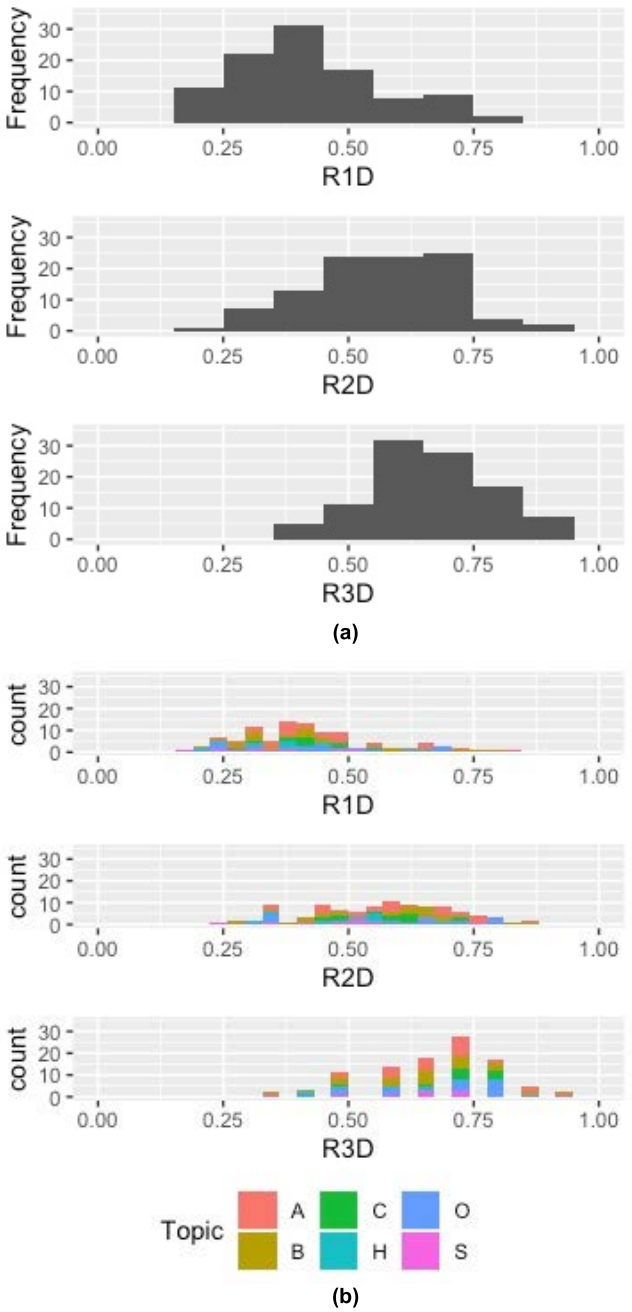

FIGURE 5. (a): Distribution of R1D, R2D & R3D for all papers (b): Distribution of R1D, R2D & R3D by topic Note: A: Analytics, data science and smart systems, B: Business models, digital transformation and innovation, C: Cyber-security privacy and ethics of IS, H: Human computer interface, O: Other topics, S: Sustainable and societal impact of IS.

Figure 8 is a scatter plot that shows the linear association between method reproducibility R3D (method score only) and overall reproducibility R1D (weighted score of all three). The orange crosses stand for papers with an experiment

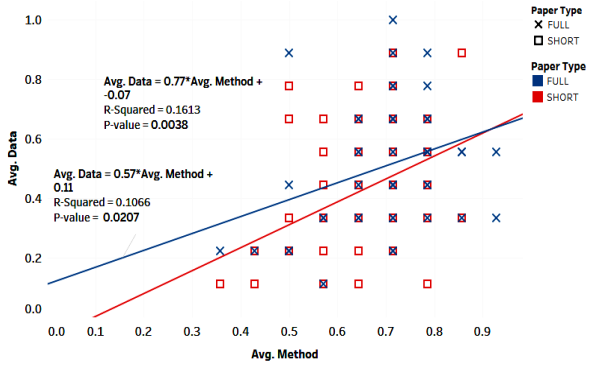

FIGURE 6. Correlation of paper scores in data and method.

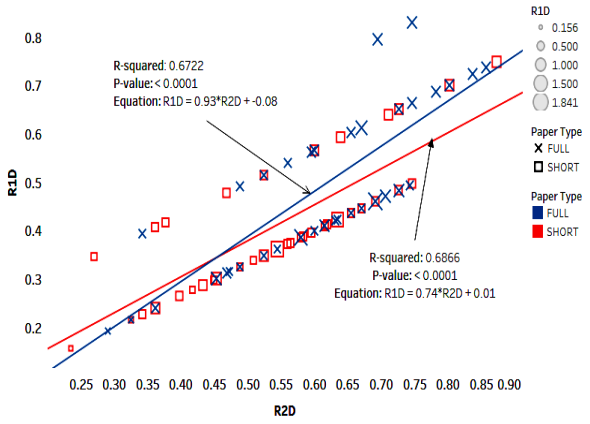

FIGURE 7. Association between R2D and R1D scores by paper type.

setup and the blue circles stand for papers without an experiment setup. The size of the point represents the R1D score. Regardless of experiment setup, R1D goes up as R3D goes up. However, for papers with an experiment setup, the relationship is statistically significant ($p<0.0001$) and R3D can explain 31% of the variation in R1D ($R^2 = 0.3108$).

On the other hand, papers without an experiment setup do not have a statistically significant relationship between R3D and R1D ($p>0.05$). Most of the highest-scoring papers at the right corner are papers with experiment setups. Therefore, papers with high R3D (method) scores tend to have high R1D (overall reproducibility) scores, and papers with experiment setups tend to be more reproducible.

Figure 9 is a scatter plot that shows the linear association between R2D (the weighted average score of Data and Method) and R3D (score of Method). The blue circles represent papers without data preprocessing, and the orange cross represents papers with data preprocessing. The chart shows that R2D tends to increase as R3D grows, regardless of whether the data preprocessing is shared or not. The coefficient estimates for both with and without data preprocessing are statistically significant ($p<0.05$). The R-squared for papers without data preprocessing (0.6621) is higher, about 66% of the variation in R2D can be explained by R3D.

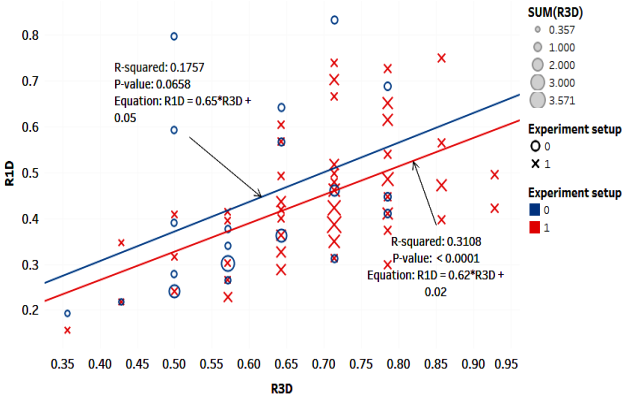

FIGURE 8. Association between R3D and R1D scores by experiment.

The coefficient estimate for papers without data preprocessing (0.78) is also higher, indicating that each unit increase in R3D will result in a greater increase in R2D. Therefore, papers without data preprocessing can be made more reproducible by having a more rigorous methodology.

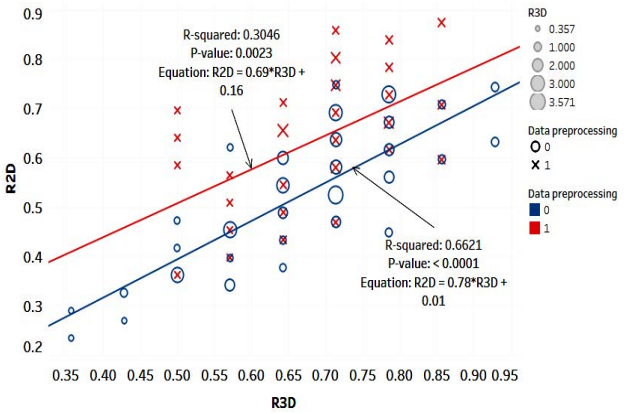

FIGURE 9. Association between R3D and R2D scores by data preprocessing.

Figure 10 shows a series of box plots for the six groups of research paper topics analyzed. Papers on topics such as human computer interface, as well as the sustainable and social impact of business computing have a lower average R1D (reproducibility for method, data and experiment) and R2D (reproducibility for method and data) scores. The results imply that, for these papers, the data availability is poor and little to no source code or details on methodology are provided in the research literature. Hence, it makes reproducing the experiments harder for some of the papers under these topics. But given the empirical nature of papers in this conference, it is likely most of the research did not require experiments.

Figure 11 is a quadrant chart that maps the relationship between the Method and Data scores. The color of each dot represents the average composite R1D score, and the dot size represents the number of papers surveyed for each topic.

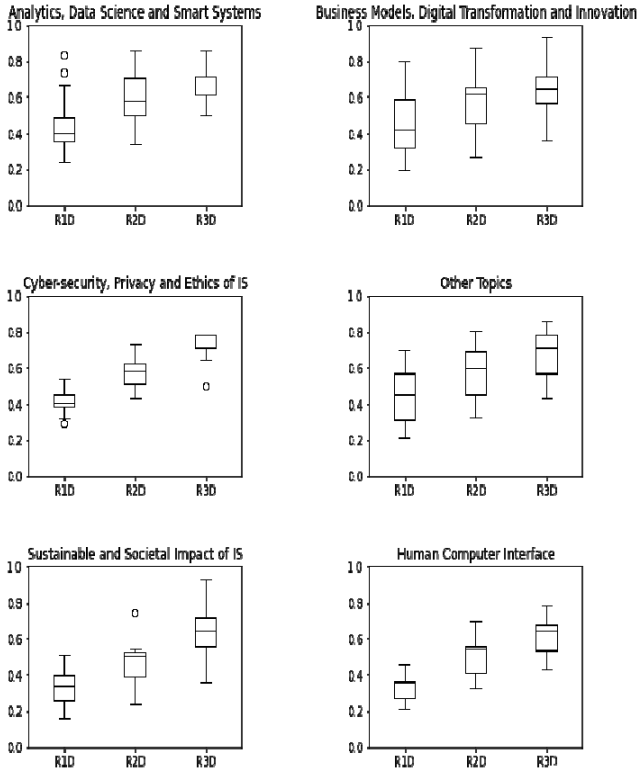

FIGURE 10. Reproducibility metrics for the groups of six topics.

The trend line shows a positive association between the average Method score and the average Data score. Documentation of topics with a higher Method score tend to have a higher Data score. Examples include analytics, data science and smart systems. Analytics, data science and smart systems lead in Data, while cyber-security, privacy, and ethics of business computing lead in Method. On the other hand, sustainable and societal impact as well as human computer interface have a below-average Method and Data score, and thereby have a relatively low R1D score. However, business models, digital transformation and innovation are the only topics that tend to share more about data and less about method, while still gaining a high average composite reproducibility score for R1D. Therefore, when publishing their research, researchers should consider sharing more specifics regarding the method and data of their studies to increase the reproducibility. Doing so will no doubt enhance the overall quality of the research.

Figure 12 is a quadrant diagram that maps the relationship between Experiment and R1D (Method, Data and Experiment) scores. The color of the dots represents the average composite R1D score, and the size of each dot represents the number of papers surveyed for each topic. The trend line shows a positive association between the average Experiment score and the R1D. Topics such as analytics, data science and smart systems outperformed for both scores. In fact, Other Topics (see Table 2 above) is dominant in R1D, while business models, digital transformation, and innovation are notable in Experiment. Sustainable and societal impact as

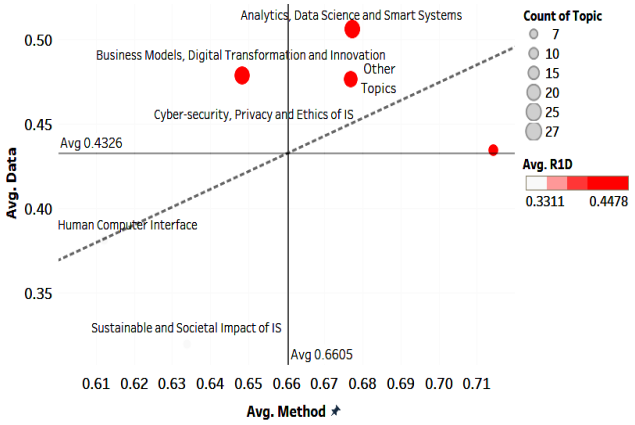

FIGURE 11. Average of method and data scores by topic.

well as human computer interface remain below-average for Experiment and R1D, and thereby are the least reproducible. To increase the overall reproducibility of the business computing research, a disclosure of the experiment process is very important to consider when publishing the research findings.

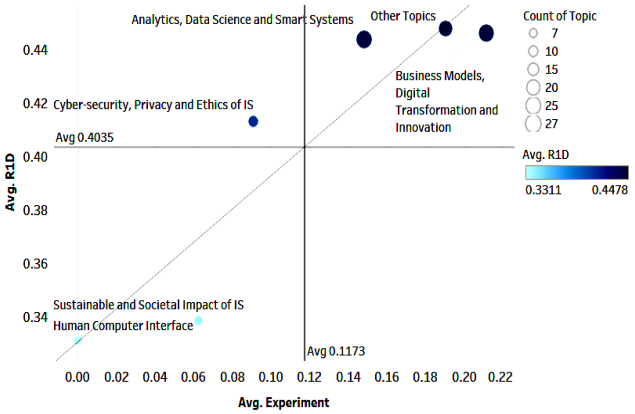

FIGURE 12. Average of experiment and R1D scores by topic.

Figure 13 depicts a pair of boxplots for the reproducibility metrics R1D (Method, Data and Experiment), R2D (Data and Method) and R3D (Data) for all the papers. Compared to the R1D of short papers (with the mean below 0.4), the R1D of full papers has a higher mean value (above 0.4). The mean values for R2D and R3D are also slightly higher for full papers. The implication here is that full papers tend to have more detailed explanations than short papers and are likely to include more details for Method, Data and Experiment. Hence, full papers are generally more reproducible than short papers. To encourage reproducibility, the authors should consider publishing their papers with fuller content.

Figure 14 is a bar chart showing the count analysis for Conclusion, grouped by paper length (full vs. short). According to the figure, the number of papers (with and without conclusions) for full and short papers is the same. Almost all the papers, regardless of length, provide a conclusion for the study.

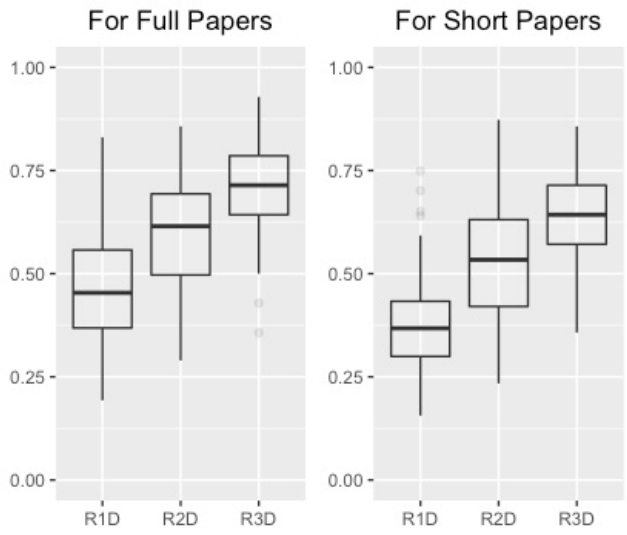

FIGURE 13. Reproducibility metrics for full and short paper.

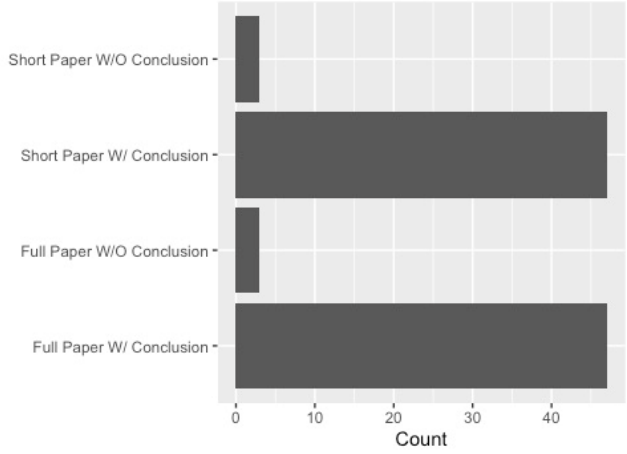

FIGURE 14. Distribution of full and short papers with and without conclusions.

Figure 15 is a bar chart comparing the average scores of composite R1D and the three factors for full and short papers. The chart shows that full papers outperformed short papers in the composite reproducibility score R1D (0.464 vs. 0.387) as well as for the three factors: Method (0.696 vs. 0.636), Data (0.507 vs. 0.416) and Experiment (0.19 vs. 0.11). Short papers are less likely to share, especially for the Data (0.416) and Experiment (0.11) factors, and therefore are less reproducible. Short papers do not typically elaborate the details of the data or experiments, nor do they provide the tools for analysis. To increase reproducibility, researchers or publishers are encouraged to publish full papers, including details on the data, method and experiments. Conferences may consider reevaluating the option to submit short papers.

Figure 16 is a set of bar charts showing the distribution of the absolute scores for Method, Data and Experiment. The absolute score represents the sum of the variables listed under

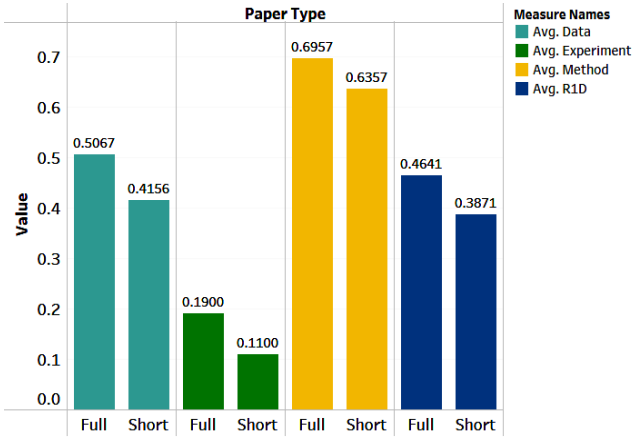

FIGURE 15. R1D Score and reproducibility factors by paper type.

each of these three factors. It is notable that there are, in total, 14 variables for Method, 9 for Data, and only 2 for Experiment. To examine individual papers more closely, we apply a random rule of thumb. We assume that a factor, to be designated as reproducible, has at least half of the variables present. A paper is method reproducible when it has seven or more variables, data reproducible for more than four variables, and experiment reproducible with at least one variable. For Method, only 5% (2+3) of the papers are not reproducible, and 70% (18+28+17+5+2) of the papers have more than eight variables. Forty-two percent (15+15+6+5+1) of the papers are Data reproducible. Interestingly, 72% of the papers have neither of the two variables in the Experiment factor, indicating that only 28% of the papers have Experiment reproducibility. This finding could be attributed to the fact that most of the papers in the conference are data-driven, not experiment-driven.

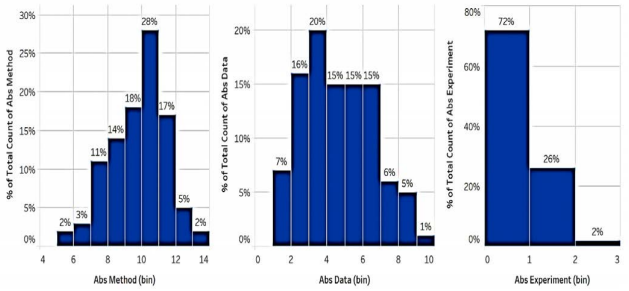

FIGURE 16. Distribution of Absolute Scores.

Figure 17 is a bar chart showing the distribution of the overall absolute score, which is the sum of all the variables. There are 25 variables in total representing the reproducibility performance. A paper is defined as reproducible if more than 12 variables across the three factors are present. Sixty-seven percent (12+10+14+10+10+4+4+2+1) of the papers have more than 12 reproducibility variables, and 11% (4+4+2+1) have 18 or more variables. Typically, a majority

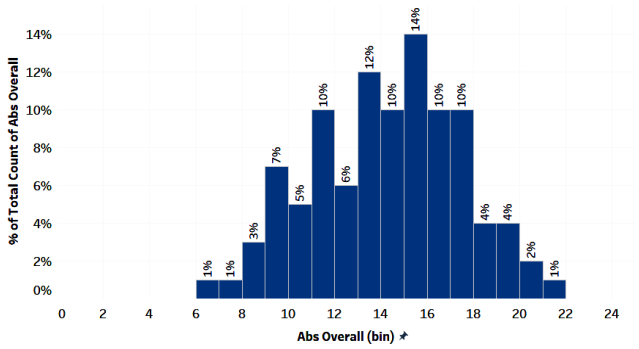

FIGURE 17. Distribution of overall absolute score.

of the variables shared are under Method, while very few share the variables under Experiment.

V. DISCUSSION

Analyzing the reproducibility for the 100 papers, we found that 67 papers, or 67%, are reproducible. As many as 95% of the papers are Method reproducible, 42% are Data reproducible, while only 28% are Experiment reproducible. Many papers in the field of business computing performed well for Method, but further improvements of reproducibility performance can be made for Data and Experiment.

The findings indicate that full papers generally score higher in all the reproducibility metrics, and in all the three factors. This outcome stems from the fact that short papers inevitably cannot provide details in Method, Data and Experiment. To encourage reproducibility, academic researchers should prioritize publishing research documentation in full context, with details explaining their method, data and experiments. Reproducibility varies by topic. Other topics, including healthcare, economics and design science, have high reproducibility. It is also evident that topics such as the sustainable and societal impact of IS as well as human computer interface are the least reproducible and received the lowest scores for all the reproducibility metrics. These topics are emerging, and data availability is limited. It is likely the papers are more case-driven or based on interviews and the like, resulting in qualitative data as sources. In terms of reproducibility by factor, cyber-security, privacy and ethics perform the best in Method; analytics, data science and smart systems lead in Data; while business models, digital transformation and innovation are the topics that lead in Experiment. It must be noted while we adapted the methodology from [31], [32], there are several key differences. While [31] compares academic papers to industry papers published in the period 2013-2016 and is a panel study, this paper focuses only on academic papers that were published during one year. While their studies focus on research in artificial intelligence, this study focuses on business computing/information systems research. Additionally, this study analyzes the data based on more specific topics and delineates the reproducibility differences among the topics. We also developed numerous additional charts to shed light on this rich dataset.

VI. LIMITATIONS

Our research has a few limitations. First, the sample data selected for review was limited because less than 30% of the papers published in ICIS 2019 were reviewed. The papers used for this research likely do not fully represent the entire population of papers, thereby impacting generalizability. In addition to the limited number of papers surveyed, this study is a snapshot in time. Future studies could examine conference papers over time and thereby identify trends. Additional limitations include the validity of the data. Although we cross-validated the results, human errors do occur when conducting manual data collections and survey type analyses. By and large these errors are minimized when multiple teams cross-check individual paper classifications. To reiterate, each paper was read and the reproducibility features were coded by four coders. The four coders were trained in the methodology and checklist template. Any differences in coding were reconciled through discussion and consensus. While the reading itself has elements of subjectivity, this approach is typically used in this type of study. It should also be noted that certain variables, such as algorithm, machine learning, prediction and source code may not be relevant to the overall theme of the conference and papers. Likewise, considering they are data driven and associative, certain studies may not require experiments. It can also be argued that the reproducibility score may depend on the research topic itself. Research topics that are data-based and quantitative in nature, would likely score better on Method and Data. Furthermore, these papers are mostly data-driven research and not about the design of computer artifacts. Therefore, code is not a prominent issue in this study. This study is also a descriptive analytic study of 'data as is' and determining the relative absence or presence of reproducibility. This is not a predictive study attempting to *predict* the absence or presence of reproducibility. In the future, reproducibility models for qualitative research may be developed. This study looked at reproducibility through a documentarian's lens. However, there are other methods that can independently assess reproducibility, or serve to complement similar studies. Finally, one size does not fit all. In the future, more sophisticated frameworks may be developed to suit the conferences and journals of individual disciplines.

VII. CONCLUSION

Using visualization and descriptive analytics, this exploratory study offers a panoramic view of the state of reproducibility of business computing research. The study paints a mixed picture. While 67% of the surveyed papers appear to be reproducible, this outcome indicates there is significant room for improvement in publishing reproducible papers. Among the three factors of Method, Data and Experiment, none of the papers meet all 25 criteria, leaving much room for improvement. Data and Method are closely associated to

one another, as expected, since data is typically utilized in the analysis process. Experiment falls short, but it must be acknowledged that the topics and nature of the conference do not lend themselves well to experiments. Emerging and sharply-focused topics—such as the economics of computing, health information technology, design science and future of work, appear to have better quality reproducibility compared to such other topics as sustainable and societal impact, and human computer interfaces. Because they impact the reproducibility mode one may apply, further research is warranted to help delineate the differences among topics. Also, research in several of the topics is more slanted towards the conceptual. Additionally, we found that paper length (full vs. short) also matters in terms of reproducibility. Full papers with greater documentation are likely to provide more details about the method, data and experiments (R1D), and they are generally more reproducible. It seems an obvious suggestion, but we offer advice to conferences that they accept only full papers and peer reviewed papers; this would improve the reproducibility of the findings, such as those in our study. It is conceivable that the review process would also evolve over time to include more reproducibility-related criteria for evaluation.

From a prescriptive perspective, we offer several recommendations to enhance the reproducibility of computing research. Our framework and check list are starting points as they can be applied both to assess reproducibility before a research study is carried out, and to evaluate a paper or report arising from the research. A significant benefit is the mitigation of the risk of carrying out a research project only to discover it is not reproducible at a later stage. We suggest prospective authors ask themselves the questions given in the check list for their study area. This would be a major departure from the traditional approach of merely making code or data available at journal sites, repositories such as GitHub, validating code post facto, etc. In addition, we suggest that reproducibility analyses be conducted in the context of data governance, ethics, awareness of intellectual property issues, privacy, security, transparency and other issues. There is also an urgent need to continue to build methods, models and tools to conduct studies, both at the paper/project level and at a large-scale macro level, for example, to assess the reproducibility of entire sets of conference papers. These are dauting tasks since we know from the literature review there is an eclectic group of models and tools across the broader scientific disciplines and the more specific sub-fields of computing, and at the same time, one model may not fit all research situations. We would be remiss if we did not mention the need for additional research into the validation of the reproducibility methods themselves. While studies, including this one, are emerging to examine the presence of reproducibility, there is a dearth of 'how-to' mechanisms. This gap must be addressed. Though there is an increased awareness for the need for reproducibility, better communication of the benefits of rigor in computing research, and the risks and consequences of a failure to reproduce or repeat/replicate research findings, is needed so the larger benefits of computing and technology research can be harnessed.

Reproducibility in general, and in business computing research specifically, is at a critical stage of development but increased awareness and advances in reproducibility methods and tools can accelerate the maturing process.

REFERENCES

[1] D. B. Allison, A. W. Brown, B. J. George, and K. A. Kaiser, "Reproducibility: A tragedy of errors," *Nature*, vol. 530, no. 7588, pp. 27–29, Feb. 2016.
[2] D. B. Allison, R. M. Shiffrin, and V. Stodden, "Reproducibility of research: Issues and proposed remedies," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 11, pp. 2561–2562, Mar. 2018.
[3] H. Fineberg, V. Stodden, and X. L. Meng, "Highlights of the U.S. National Academies report on 'reproducibility and replicability in science,'" *Harvard Data Sci. Rev.*, vol. 2, no. 4, 2020, doi: 10.1162/99608f92.cb310198.
[4] M. Krafczyk, A. Shi, A. Bhaskar, D. Marinov, and V. Stodden, "Scientific tests and continuous integration strategies to enhance reproducibility in the scientific software context," in *Proc. 2nd Int. Workshop Practical Reproducible Eval. Comput. Syst. (P-RECS)*, 2019, pp. 23–28.
[5] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research," *Commun. ACM*, vol. 59, no. 3, pp. 62–69, Feb. 2016.
[6] M. Flittner, R. Bauer, A. Rizk, S. Geißler, T. Zinner, and M. Zitterbart, "Taming the complexity of artifact reproducibility," in *Proc. Reproducibility Workshop*, Aug. 2017, pp. 14–16.
[7] D. Ghoshal, D. Paine, G. Pastorello, A. Elbashandy, D. Gunter, O. Amusat, and L. Ramakrishnan, "Experiences with reproducibility: Case studies from scientific workflows," in *Proc. 4th Int. Workshop Practical Reproducible Eval. Comput. Syst.*, Jun. 2020, pp. 3–8.
[8] F. S. Collins and L. A. Tabak, "Policy: NIH plans to enhance reproducibility," *Nature*, vol. 505, no. 7485, pp. 612–613, Jan. 2014.
[9] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Med.*, vol. 2, no. 8, p. e124, Aug. 2005.
[10] H. Pashler and E. J. Wagenmakers, "Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?" *Perspect. Psychol. Sci.*, vol. 7, no. 6, pp. 528–530, 2012.
[11] M. Baker, "Reproducibility crisis," *Nature*, vol. 533, no. 26, pp. 452–454, 2016.
[12] N. C. Nelson, K. Ichikawa, J. Chung, and M. M. Malik, "Mapping the discursive dimensions of the reproducibility crisis: A mixed methods analysis," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254090.
[13] G. Naik, "Scientists' elusive goal: Reproducing study results," *Wall Street J.*, vol. 258, no. 130, p. A1, 2011.
[14] The Economist. (2013). *Trouble at the Lab*. Accessed: Jan. 28, 2022. [Online]. Available: https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab
[15] D. H. Freedman, "Lies, damned lies, and medical science," *Atlantic*, vol. 306, no. 4, pp. 76–84, 2010. [Online]. Available: https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/
[16] E. Yong. (Aug. 27, 2015). *How Reliable are Psychology Studies? Atlantic*. Accessed: Jan. 28, 2022. [Online]. Available: https://www.theatlantic.com/science/archive/2015/08/psychology-studies-reliability-reproducability-nosek/402466/
[17] E. Yong. (Jan. 18, 2017). *How Reliable are Cancer Studies? Atlantic*. Accessed: Jan. 28, 2022. [Online]. Available: https://www.theatlantic.com/science/archive/2017/01/what-proportion-of-cancer-studies-are-reliable/513485/
[18] O. Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, Aug. 2015.
[19] K. A. Baggerly and K. R. Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology," *Ann. Appl. Statist.*, vol. 3, no. 4, pp. 1309–1334, 2009.
[20] F. Sayre and A. Riegelman, "The reproducibility crisis and academic libraries," *College Res. Libraries*, vol. 79, no. 1, p. 2, Jan. 2018.
[21] R. O. Gilmore, "Progress toward openness, transparency, and reproducibility in cognitive neuroscience," *Ann. New York Acad. Sci.*, vol. 1396, no. 1, pp. 5–18, 2017.
[22] C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, pp. 531–533, 2012.

[23] B. C. Gibb, "Reproducibility," *Nature Chem.*, vol. 6, no. 8, 2014.
[24] R. E. Benestad, "Learning from mistakes in climate research," *Theor. Appl. Climatol.*, vol. 126, pp. 699–703, Aug. 2016.
[25] T. Herndon, M. Ash, and R. Pollin, "Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff," in *Proc. Political Economy Res. Inst. Work. Paper*, 2013. [Online]. Available: https://peri.umass.edu/images/WP322.pdf
[26] M. C. Makel and J. A. Plucker, "Facts are more important than novelty: Replication in the education sciences," *Educ. Researcher*, vol. 43, no. 6, pp. 304–316, 2014.
[27] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, "The economics of reproducibility in preclinical research," *PLOS Biol.*, vol. 13, no. 6, Jun. 2015, Art. no. e1002165.
[28] P. Hunter, "The reproducibility 'crisis' Reaction to replication crisis should not stifle innovation," *EMBO Rep.*, vol. 18, no. 9, pp. 1493–1496, 2017.
[29] R. F. Boisvert, "Incentivizing reproducibility," *Commun. ACM*, vol. 59, no. 10, p. 5, Sep. 2016.
[30] J. Freire, N. Fuhr, and A. Rauber, "Reproducibility of data-oriented experiments in e-science (dagstuhl seminar 16041)," Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Wadern, Germany, Dagstuhl Rep. 6-1, 2016.
[31] O. E. Gundersen, "Standing on the feet of giants—Reproducibility in AI," *AI Mag.*, vol. 40, no. 4, pp. 9–23, 2019.
[32] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," in *Proc. AAAI*, Feb. 2018, pp. 1644–1651.
[33] O. E. Gundersen, Y. Gil, and D. W. Aha, "On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications," *AI Mag.*, vol. 39, no. 3, pp. 56–68, Sep. 2018.
[34] National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*, National Academies, Washington, DC, USA, 2019.
[35] Y. Yang, W. Youyou, and B. Uzzi, "Estimating the deep replicability of scientific findings using human and artificial intelligence," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 20, pp. 10762–10768, May 2020.
[36] A. A. Aarts, C. J. Anderson, J. Anderson, M. A. van Assen, P. R. Attridge, A. S. Attwood, and E. Baranski. (2015). *Reproducibility project: Psychology*. [Online]. Available: https://www.osf.io/ezcuj
[37] C. G. Begley and J. P. A. Ioannidis, "Reproducibility in science: Improving the standard for basic and preclinical research," *Circulat. Res.*, vol. 116, no. 1, pp. 116–126, 2014.
[38] P. Ivie and D. Thain, "Reproducibility in scientific computing," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–36, May 2019.
[39] V. Bajpai, M. Kühlewind, J. Ott, J. Schönwälder, A. Sperotto, and B. Trammell, "Challenges with reproducibility," in *Proc. Reproducibility Workshop*, Aug. 2017, pp. 1–4.
[40] S. Greengard, "An inability to reproduce," *Commun. ACM*, vol. 62, no. 9, pp. 13–15, Aug. 2019.
[41] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," *PLoS Comput. Biol.*, vol. 9, no. 10, Oct. 2013, Art. no. e1003285.
[42] Q. Scheitle, M. Wählisch, O. Gasser, T. C. Schmidt, and G. Carle, "Towards an ecosystem for reproducible research in computer networking," in *Proc. Reproducibility Workshop*, Aug. 2017, pp. 5–8.
[43] R. Sinha and P. S. Sudhish, "A principled approach to reproducible research: A comparative review towards scientific integrity in computational research," in *Proc. IEEE Int. Symp. Ethics Eng., Sci. Technol. (ETHICS)*, May 2016, pp. 1–9.
[44] J. Vitek and T. Kalibera, "Repeatability, reproducibility and rigor in systems research," in *Proc. 9th ACM Int. Conf. Embedded Softw. (EMSOFT)*, Oct. 2011, pp. 33–38.
[45] J. Cito, V. Ferme, and H. C. Gall, "Using Docker containers to improve reproducibility in software and web engineering research," in *Proc. Int. Conf. Web Eng.* Cham, Switzerland: Springer, Jun. 2016, pp. 609–612.
[46] V. Bajpai, O. Bonaventure, K. Claffy, and D. Karrenberg, "Encouraging reproducibility in scientific research of the internet," *Dagstuhl Rep.*, vol. 8, no. 10, pp. 1–22, 2019.
[47] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, and K. Turner, "Computing environments for reproducibility: Capturing the 'whole tale,'" *Future Gener. Comput. Syst.*, vol. 94, pp. 854–867, May 2019.
[48] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, Dec. 2011.
[49] M. A. Heroux, L. Barba, M. Parashar, V. Stodden, and M. Taufer, "Toward a compatible reproducibility taxonomy for computational and computing sciences," Sandia National Lab. (SNL-NM), Albuquerque, NM, USA, Tech. Rep. SAND2018-11186, 2018.
[50] A. A. Alsheikh-Ali, W. Qureshi, M. H. Al-Mallah, and J. P. A. Ioannidis, "Public availability of published research data in high-impact journals," *PLoS ONE*, vol. 6, no. 9, Sep. 2011, Art. no. e24357.
[51] V. Stodden and S. Miguez, "Best practices for computational science: Software infrastructure and environments for reproducible and extensible research," *J. Open Res. Softw.*, vol. 2, no. 1, p. e21, Jul. 2014.
[52] V. Stodden and D. H. W. Bailey, B. LeVeque, and R. Stein. (2017). *Setting the default to reproducible. Reproducibility in Computational and Experimental Mathematics*. Accessed: Jan. 28, 2022. [Online]. Available: http://icerm.brown.edu/tw12-5-rcem
[53] J. F. Claerbout and M. Karrenbach, "Electronic documents give reproducible research a new meaning," in *Proc. SEG Tech. Program Expanded Abstr.*, 1992, pp. 601–604.
[54] D. H. Bailey, J. M. Borwein, and V. Stodden, "Facilitating reproducibility in scientific computing: Principles and practice," in *Reproducibility: Principles, Problems, Practices, and Prospects*. New York, NY, USA: Wiley, 2016, pp. 205–232.
[55] ACM. (2016). *Result and Artifact Review and Badging*. Accessed: Jan. 28, 2022. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-badging
[56] V. Stodden, P. Guo, and Z. Ma, "Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e67111.
[57] J. Freire, P. Bonnet, and D. Shasha, "Computational reproducibility: State-of-the-art, challenges, and database research opportunities," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2012, pp. 593–596.
[58] N. Ferro, "Reproducibility challenges in information retrieval evaluation," *J. Data Inf. Qual.*, vol. 8, no. 2, pp. 1–4, Feb. 2017.
[59] IEEE. (2020). *IEEE Reproducibility Practices Survey—Summary of Findings*. Accessed: Jan. 28, 2022. [Online]. Available: https://ieeecs-media.computer.org/media/tech-news/ieee-reproducibility-practices-survey-summary-of-findings-1.pdf
[60] J. Goodrich. (2021). *Study Shows Ensuring Reproducibility in Research Is Needed The IEEE Computer Society Suggests Improvements*. Accessed: Jan. 27, 2022. [Online]. Available: https://spectrum.ieee.org/study-shows-ensuring-reproducibility-in-research-is-needed
[61] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: The incredibles," *ACM SIGMOBILE*, vol. 9, no. 4, pp. 50–61, 2005.
[62] P. Vandewalle, J. Kovacević, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 37–47, May 2009.
[63] V. Stodden, J. Seiler, and Z. Ma, "An empirical analysis of journal policy effectiveness for computational reproducibility," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 11, pp. 2584–2589, Mar. 2018.
[64] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis, "What does research reproducibility mean?" *Sci. Transl. Med.*, vol. 8, no. 341, Jun. 2016, Art. no. 341ps12.
[65] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden, "Reproducible research in computational harmonic analysis," *Comput. Sci. Eng.*, vol. 11, no. 1, pp. 8–18, Jan./Feb. 2009.
[66] D. Delling, C. Demetrescu, D. S. Johnson, and J. Vitek, "Rethinking experimental methods in computing (dagstuhl seminar 16111)," Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Wadern, Germany, Dagstuhl Rep. 6-3, 2016.
[67] S. Krishnamurthi and J. Vitek, "The real software crisis: Repeatability as a core value," *Commun. ACM*, vol. 58, no. 3, pp. 34–36, 2015.
[68] B. A. Nosek, J. R. Spies, and M. Motyl, "Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability," *Perspect. Psychol. Sci.*, vol. 7, no. 6, pp. 615–631, 2012.
[69] O. Bonaventure, "The January 2017 issue," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 47, no. 1, pp. 1–3, 2017.
[70] L. Oliveira, D. Wilkinson, D. Mossé, and B. Childers, "Supporting long-term reproducible software execution," in *Proc. 1st Int. Workshop Practical Reproducible Eval. Comput. Syst.*, Jun. 2018, pp. 1–6.
[71] B. J. Oates, *Researching Information Systems and Computing*. Newbury Park, CA, USA: Sage, 2006.

[72] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson, "Using prediction markets to estimate the reproducibility of scientific research," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 50, pp. 15343–15347, Dec. 2015.
[73] L. K. John, G. Loewenstein, and D. Prelec, "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychol. Sci.*, vol. 23, no. 5, pp. 524–532, 2012.
[74] V. Bajpai, A. W. Berger, P. Eardley, J. Ott, and J. Schönwälder, "Global measurements: Practice and experience (report on dagstuhl seminar# 16012)," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 46, no. 2, pp. 32–39, 2016.
[75] M. M. McGill, "Discovering empirically-based best practices in computing education through replication, reproducibility, and meta-analysis studies," in *Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res.*, Nov. 2019, pp. 1–5.
[76] V. Bajpai, O. Bonaventure, K. Claffy, and D. Karrenberg, "Encouraging reproducibility in scientific research of the internet UCL-Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium, Tech. Rep. Dagstuhl Seminar 18412, 2018.
[77] A. A. Ali, M. El-Kalioby, and M. Abouelhoda, "The case for Docker in multicloud enabled bioinformatics applications," in *Proc. Int. Conf. Bioinf. Biomed. Eng.* Cham, Switzerland: Springer, Apr. 2016, pp. 587–601.
[78] D. G. Feitelson, "From repeatability to reproducibility and corroboration," *ACM SIGOPS Operating Syst. Rev.*, vol. 49, no. 1, pp. 3–11, Jan. 2015.
[79] J. Freire, D. Koop, F. Chirigati, and C. T. Silva, "Reproducibility using vistrails," in *Implementing Reproducible Research*. Boca Raton, FL, USA: CRC Press, 2018, pp. 33–56.
[80] Seven Bridges. 2015. *Docker-Based Solutions to Reproducibility in Science*. [Online]. Available: https://blog.sbgenomics. com/docker-based-solutions-to-reproducibility-in-science/
[81] V. V. Sochat, C. J. Prybol, and G. M. Kurtzer, "Enhancing reproducibility in scientific computing: Metrics and registry for singularity containers," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0188511.
[82] S. Hunold, "A survey on reproducibility in parallel computing," 2015, *arXiv:1511.04217*.
[83] V. E. Johnson, "Revised standards for statistical evidence," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 48, pp. 19313–19317, 2013.
[84] S. Hunold and J. L. Traff, "On the state and importance of reproducible experimental research in parallel computing," 2013, *arXiv:1308.3648*.
[85] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, Feb. 2018.
[86] A. Abedi, A. Heard, and T. Brecht, "Conducting repeatable experiments and fair comparisons using 802.11n MIMO networks," *ACM SIGOPS Operating Syst. Rev.*, vol. 49, no. 1, pp. 41–50, 2015.
[87] J. B. Buckheit and D. L. Donoho, "Wavelab and reproducible research," in *Wavelets and Statistics*. New York, NY, USA: Springer, 1995, pp. 55–81.
[88] K. Börner, A. Bueckle, and M. Ginda, "Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 6, pp. 1857–1864, Feb. 2019.
[89] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansman. (2010). *Solving Problems with Visual Analytics*. [Online]. Available: http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf
[90] D. A. Keim, "Visual exploration of large data sets," *Commun. ACM*, vol. 44, no. 8, pp. 38–44, 2001.
[91] V. Raghupathi, J. Ren, and W. Raghupathi, "Understanding the nature and dimensions of litigation crowdfunding: A visual analytics approach," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0250522.
[92] P. C. Wong and J. Thomas, "Guest editors' introduction–visual analytics," *IEEE Comput. Graph. Appl.*, vol. 24, no. 5, pp. 20–21, Sep. 2004.
[93] J. Kohlhammer, D. Keim, M. Pohl, G. Santucci, and G. Andrienko, "Solving problems with visual analytics," *Proc. Comput. Sci.*, vol. 7, pp. 117–120, Jan. 2011, doi: 10.1016/j.procs.2011.12.035.
[94] J. Thomas and K. Cook, "Illuminating the path: Research and department agenda for visual analytics," Dept. Homeland Secur. Washington, DC, USA, Tech. Rep., 2005.

WULLIANALLUR RAGHUPATHI (Member, IEEE) is currently a Professor of information systems with the School of Business, Fordham University, New York, NY, USA, and the Director of the Center for Digital Transformation (http://www.fordhamcdt.org/). He has published over 60 refereed journal articles, including in IEEE Access, *Communications of the ACM*, and others. He has over 4500 Google citations. He has also written papers in refereed conference proceedings, abstracts in international conferences, book chapters, editorials, and reviews. His research interests include artificial intelligence, big data, crowdfunding, deep learning, entrepreneurship, healthcare IT, sustainability, and others. He was the Founding Editor of the *International Journal of Computational Intelligence and Organizations* from 1995 to 1997. He is a Co-Editor of North America of the *International Journal of Health Information Systems and Informatics*.

VIJU RAGHUPATHI received the Ph.D. degree in information systems from The Graduate Center, City University of New York. She is currently an Associate Professor at the Koppelman School of Business, Brooklyn College, City University of New York. She has published in academic journals in the information systems and healthcare areas, such as *Communications of the AIS*, *Technological Forecasting and Social Change*, IEEE Access, *Journal of Business Venturing Insights*, *PLOS ONE*, *International Journal of Environmental Research and Public Health*, *Health Policy and Technology*, and *Journal of Medical Internet Research*. Her research interests include business analytics, big data, crowdfunding, sharing platforms, sustainability, innovation and entrepreneurship, and healthcare IT.

JIE REN is currently an Assistant Professor in information, technology, and operations area at the Gabelli School of Business, Fordham University. She is interested in exploring the business impact of collective online behaviors. She studies crowdsourcing, social media, and online reviews. Her publications occur in major *Information Systems* journals, such as the *European Journal of Information Systems*, *Decision Support System*, and *Journal of the Association for Information Science and Technology*. Due to the interdisciplinary nature of her research, her publications also occur in other major journals, such as the *Journal of Medical Internet Research*, the *International Journal of Hospitality Management*, and *PLOS ONE*. She received the Best Paper Award in the track of social computing from the Americas Conference on Information Systems (AMCIS) 2020 and the Best Paper Second Runner-Up from AMCIS 2018.