Using big data to study and improve scientific software

Váleri N. Vásquez*[1] and Richard Barnes*[1]

[1]Energy and Resources Group, University of California, Berkeley

1 Challenge

Computation has emerged as a "third leg" of science. The traditional legs—theory and experimentation—require quality instruments and communication to function well. Similarly, computation requires high-quality software that is correct, publicly available, and documented. The process of designing, developing, and refining such software should—like the discoveries it is being used to elucidate—be treated empirically, and include iterative experimentation guided by data. A primary obstacle in building the evidence base that is required to apply the scientific method to scientific software is the *observability challenge*: representative data is required, but has been historically difficult to obtain. **Here, we propose to use public archives of open source software and scientific publications to help solve the observability challenge.**

Previous work on the *observability challenge* has used surveys to characterize how scientists develop and use software in their work; however, these efforts have been largely qualitative and restricted to small sample sizes, such as $N \leq 2,000$ in [1]. This may introduce bias while limiting generalization across scientific domains. Surveys, for instance, may be unable to quantify the usability, quality, or sustainability of software and, unless they are conducted continuously, cannot track the evolution of discipline-specific software applications nor trends in the best workflows and community practices for software development.

2 Opportunity

As detailed below, information concerning the development and use of scientific software is publicly accessible at an unprecedented scale and ever-increasing rate via platforms such as GitHub, arXiv and its discipline-specific sub-archives[1], and open access academic journals. Likewise, citation metadata has become available from services like Web of Science, a global citation database. The recent proliferation of these digital venues and others like them is thanks in part to the **growing appreciation of transparency and accessibility in scientific disciplines [2]: a cultural shift in expectation and practice that presents a prime opportunity** for addressing the observability challenge.

With 73M developer-users and 200M code repositories (of which 28M are public), GitHub is the largest source code host, though services such as Code Ocean are growing in popularity. GitHub offers both the current and historical code of a project along with a rich database of metadata including the frequency and size of proposed changes to the code (pull requests), the length of time pull requests and issues remain unaddressed, counts of unique and regular contributors, cross-contributions by particular developers, and the number of users following a project. This data has been used to investigate the effects of gender bias by linking user names to aspects of the software development cycle [3], the impact of issue tracking techniques on project success [4], and how team characteristics and the choice of programming language affects software quality [5].

Over the past 30 years the number of preprints on services like arXiv has increased 63-fold [6]. While this accounts for only 4% of all research articles [6], in some fields such as theoretical computer science and machine learning over 60% of published papers are on arXiv, with usage rising in other areas [7]. Of these preprints, 41–56% go on to be published in academic journals [6, 7], with some disciplines exceeding even that: in the case of bioRxiv, nearly 38,000 preprints were uploaded in its first five years, two-thirds of which were later published in peer-reviewed journals [8]. arXiv has made science more observable by making scientific results available sooner and without paywalls in machine-readable LaTeX format. As a result, anyone can observe and analyze scientific progress in real-time. For instance, analysis of arXiv full-text

*Contributed equally.

[1] e.g., bioRxiv, ChemRxiv, EarthArXiv, medRxiv, SocArXiv

corpuses has been used to study the co-evolution of scientific topics and collaboration networks [9], text reuse and its effects on citations [10], and to predict future directions in AI [11].

The Web of Science catalogues 182M records including journals, books, and conference proceedings, spanning literature from the year 1800 onward and patents dating back to 1963. This data can be used to understand how scientific works relate to each other and what their impact is. For instance, citation metadata from Web of Science has been used to show that progress ossifies in large fields of science due to citation network effects [12] and that the deaths of preeminent scientists lead to reorganizations of citation networks and the emergence of new leaders and ideas [13]. This data has also been used to conclude that 28% of the scientific literature is open access and that open access articles receive 18% more citations [14].

3 Timeliness

The foregoing demonstrates that open access to scientific texts, source code, and citation metadata allows for myriad analyses of each of these data sources on their own. **But only the fusion of these data sources will allow us to fully characterize the development and use of scientific software**, drawing a more direct line between innovation and the practices that drive it. Some preliminary work in this direction does exist. Combining citation and code data indicates that adding software to package management systems can increase a paper's citation count by 280% [15], while combining textual and code data indicates that high-profile projects tend to use statically-typed languages and tend to have fewer female contributors [16].

However, in our review of the literature we find that most "fusion" work to date focuses on fairly simple, descriptive conclusions. For instance, a representative paper reviews all code-paper pairs in the Microsoft Academic Graph only to conclude that repository stars follow a power distribution [17]. **Emerging efforts in social and cognitive science, as well as human-computer interaction (HCI) [18], can help bridge the gap between challenge and opportunity** in scientific software. By fusing the big data sources described here with social science methods, researchers can overcome the observability challenge to answer timely questions such as:

- Do software design choices influence subsequent scientific innovation or research question creativity?
- Can the development of "foundational" software be causative to subsequent innovation?
- Do co-author demographics predict software quality or longevity, or features of the software's community?
- What methods are most successful in supporting long-term software projects (e.g., intermediate publications, citations garnered, or cyber-infrastructure grants)?
- Do good software practices (such as those described in [19]) spread via collaboration networks or are they driven by outside factors?

The impact of answering questions such as those exemplified above is an improved understanding of how to design and maintain scientific software, reward and incentivize developers, build communities, and, ultimately, drive new scientific advances.

[1] Hannay et al. 2009. DOI: 10.1109/SECSE.2009.5069155. **[2]** Wilkinson et al. 2016. DOI: 10.1038/sdata.2016.18. **[3]** Imtiaz et al. 2019. DOI: 10.1109/ICSE.2019.00079. **[4]** Bissyandé et al. 2013. DOI: 10.1109/ISSRE.2013.6698918. **[5]** Ray et al. 2017. DOI: 10.1145/3126905. **[6]** Xie et al. 2021. arXiv: 2102.09066. **[7]** Sutton et al. 2017. arXiv: 1710.05225. **[8]** Abdill et al. 2019. DOI: 10.7554/eLife.45133. **[9]** Citron et al. 2018. DOI: 10.1016/j.joi.2017.12.008. **[10]** Citron et al. 2015. DOI: 10.1073/pnas.1415135111. **[11]** Hao. 2019. URL: https://bit.ly/2YXUJlb. **[12]** Chu et al. 2021. DOI: 10.1073/pnas.2021636118. **[13]** Azoulay et al. 2019. DOI: 10.1257/aer.20161574. **[14]** Piwowar et al. 2018. DOI: 10.7717/peerj.4375. **[15]** Jalili et al. 2020. DOI: 10.1101/2020.11.16.385211. **[16]** Russell et al. 2018. DOI: 10.1371/journal.pone.0205898. **[17]** Färber. 2020. DOI: 10.1145/3383583.3398578. **[18]** Myers et al. 2016. DOI: 10.1109/MC.2016.200. **[19]** Hunter-Zinck et al. 2021. DOI: 10.1371/journal.pcbi.1009481.