NATURE ECOLOGY AND EVOLUTION EMBARGO
– MAY 23, 2017, 10:00 EDT –

Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes[1]*, Benjamin D. Best[2], Courtney Scarborough[1], Jamie C. Afflerbach[1], Melanie R. Frazier[1], Casey C. O'Hara[1], Ning Jiang[1], Benjamin S. Halpern[1,3,4]

[1]National Center for Ecological Analysis and Synthesis, University of California at Santa Barbara, Santa Barbara, CA, United States

[2]EcoQuants.com, Santa Barbara, CA, United States

[3]Bren School for Environmental Science and Management, University of California, Santa Barbara, CA, United States

[4]Silwood Park Campus, Imperial College London, Ascot, United Kingdom

*corresponding author: lowndes@nceas.ucsb.edu

Preface

Reproducibility has long been a tenet of science but has been challenging to achieve — we learned this the hard way when our old approaches proved inadequate to efficiently reproduce our own work. Here we describe how several free software tools have fundamentally upgraded our approach to collaborative research, making our entire workflow more transparent and streamlined. By describing specific tools and how we incrementally began using them for the Ocean Health Index project, we hope to encourage others in the scientific community to do the same — so we can all produce better science in less time.

Keywords

collaboration, data science, Ocean Health Index, open science, reproducibility, transparency

Scientists need data science

Science, now more than ever, demands reproducibility, collaboration, and effective communication to strengthen public trust and effectively inform policy. Recent high-profile difficulties in reproducing and repeating scientific studies have put the spotlight on psychology and cancer biology[1–3], but it is widely acknowledged that reproducibility challenges persist across scientific disciplines[4–6]. Environmental scientists face potentially unique challenges in achieving goals of transparency and reproducibility because they rely on vast amounts of data spanning natural, economic, and social sciences that create semantic and synthesis issues exceeding those for most other disciplines[7–9]. Furthermore, proposed environmental solutions can be complex, controversial, and resource intensive, increasing the need for scientists to work transparently and efficiently with data to foster understanding and trust.

Environmental scientists are expected to work effectively with ever-increasing quantities of highly heterogeneous data even though they are seldom formally trained to do so[10–14]. This was recently highlighted by a survey of 704 US National Science Foundation principle investigators in the biological sciences that found training in data skills to be the largest unmet need[15]. Without training, scientists tend to develop their own bespoke workarounds to keep pace, but with this comes wasted time struggling to create their own conventions for managing, wrangling, and versioning data. If done haphazardly or without a clear protocol, these efforts are likely to result in work that is not reproducible — by the scientist's own 'future self' or by anyone else[12]. As a team of environmental scientists tasked with reproducing our own science annually, we experienced this struggle first-hand. When we began our project, we worked with data in the same way as we always had, taking extra care to make our methods reproducible for planned future re-use. But when we began to reproduce our workflow a second time and repeat our methods with updated data, we found our approaches to reproducibility were insufficient. However, by borrowing philosophies, tools, and workflows primarily created for software development, we have been able to dramatically improve the ability for ourselves and others to reproduce our science, while also reducing the time involved to do so: the result is better science in less time (Fig. 1).

Here we share a tangible narrative of our transformation to better science in less time — meaning more transparent, reproducible, collaborative, and openly shared and communicated science — with an aim of inspiring others. Our story is only one potential path because there are many ways to upgrade scientific practices — whether collaborating only with your 'future self' or as a team — and they depend on the shared commitment of individuals, institutions, and publishers[6,16,17]. We do not review the important, ongoing work regarding data management architecture and archiving[8,18], workflows[11,19–21], sharing and publishing data[22–25] and code[25–27], or how to tackle reproducibility and openness in science[28–32]. Instead, we focus on our experience, because it required changing the way we had always worked, which was extraordinarily intimidating. We give concrete examples of how we use tools and practices from data science, the discipline of 'turning raw data into understanding'[33]. It was out of necessity that we began to engage in data science, which we did incrementally by introducing new tools, learning new skills, and creating deliberate workflows — all while maintaining annual deadlines. Through our work with academics, governments, and non-profit groups around the world, we have seen that the need to improve practices is common if not ubiquitous. In this narrative we describe specific software tools, why we use them, how we use them in our workflow, and how we work openly as a collaborative team. In doing so we underscore two key lessons we learned that we hope encourage others to incorporate these practices into their own research. The first is that powerful tools exist and are freely available to use; the barriers to entry seem to be exposure to relevant tools and building confidence using them. The second is that engagement may best be approached as an evolution rather than as a revolution that may never come.

Improving reproducibility and collaboration

From then to now

The Ocean Health Index (OHI) operates at the interface of data-intensive marine science, coastal management and policy, and now, data science[34,35]. It is a scientific framework to quantify ocean-derived benefits to humans and to help inform sustainable ocean management using the best available information [36,37]. Assessments using the OHI framework require synthesising heterogeneous data from nearly one hundred different sources, ranging from categorical tabular data to high-resolution rasters. Methods must be reproducible, so that others can produce the same results, and also repeatable, so that newly available data can be incorporated in subsequent assessments. Repeated assessments using the same methods enable quantifiable comparison of changes in ocean health through time, which can be used to inform policy and track progress[34].

Using the OHI framework, we lead annual global assessments of 220 coastal nations and territories, completing our first assessment in 2012[36]. Despite our best efforts, we struggled to efficiently repeat our own work during the second assessment in 2013 because of our approaches to data preparation[37]. Data preparation is a critical aspect of making science reproducible but is seldom explicitly reported in research publications; we thought we had documented our methods sufficiently in 130-pages of published supplemental materials[36], but we had not.

However, by adopting data science principles and freely available tools that we describe below, we began building an OHI 'Toolbox' and fundamentally changed our approach to science (Figure 1). The OHI Toolbox provides a file structure, data, code, and instruction, operates across computer operating systems, and is shared online for free so that anyone can begin building directly from previous OHI assessments without reinventing the wheel[34]. While these changes required an investment of our team's time to learn and develop the necessary skills, the pay-off has been substantial. Most significantly we are now able to share and extend our workflow with a growing community of government, non-profit, and academic collaborations around the world that use the OHI for science-driven marine management. There are currently two dozen OHI assessments underway, most of which are led by independent groups[34], and the Toolbox has helped lower the barriers to entry. Further, our own team has just released the fifth annual global OHI assessment[38] and continues to lead assessments at smaller spatial scales, including the Northeastern United States, where the OHI is included in President Obama's first Ocean Plan[39].

We thought we were doing reproducible science

For the first global OHI assessment in 2012 we employed an approach to reproducibility that is standard to our field, which focused on scientific methods, not data science methods[36]. Data from nearly one hundred sources were prepared manually — i.e. without coding, typically in Microsoft Excel — which included organising, transforming, rescaling, gap-filling, and formatting data. Processing decisions were documented primarily within the Excel files themselves, emails, and Microsoft Word documents. We programmatically coded models and meticulously documented their development, (resulting in the 130-page supplemental materials)[36], and upon publication, we also made the model inputs (i.e., prepared data and metadata) freely available to download. This level of documentation and transparency is beyond the norm for environmental science[16,40].

We also worked collaboratively in the same ways we always had. Our team included scientists and analysts with diverse skill sets and disciplines, and we had distinct, domain-specific roles assigned to scientists and to a single analytical programmer. Scientists were responsible for developing the models conceptually, preparing data, and interpreting modeled results, and the programmer was responsible for coding the models. We communicated and shared files frequently, with long, often-forwarded, and vaguely-titled email chains (e.g. `Re: Fwd: data question`) with manually versioned data files (e.g. `data_final_updated2.xls`). All team members were responsible for organising those files with their own conventions on their local computers. Final versions of prepared files were stored on the servers and used in models, but records of the data processing itself were scattered.

Upon beginning the second annual assessment in 2013, we realised that our approach was insufficient since it took too much time and relied heavily on individuals' data organisation, email chains, and memory — particularly problematic as original team members moved on and new team members joined. We quickly realised we needed a nimble and robust approach to sharing data, methods, and results within and outside our team — we needed to completely upgrade our workflow.

Actually doing reproducible science

As we began the second global OHI assessment in 2013 we faced challenges across three main fronts: 1) *reproducibility*, including transparency and repeatability, particularly in data preparation; 2) *collaboration*, including team record keeping and internal collaboration; and 3) *communication* with scientific and broader communities. We knew that environmental scientists are increasingly using `R`[11] because it is free, cross-platform, and open source, and also because of the training and support provided by developers[33] and independent groups[12,41] alike. We decided to base our work in `R`[42] and RStudio[43] for coding and visualisation, Git[44] for version control, GitHub[45] for collaboration, and a combination of GitHub and RStudio for organisation, documentation, project management, online publishing, distribution, and communication (Table 1). These tools can help scientists organise, document, version, and easily share data and methods, thus not only increasing reproducibility but also reducing the amount of time involved to do so[14,46,47]. Many available tools are free so long as work is shared publicly online, which enables open science, defined by Hampton et al.[40] as "the concept of transparency at all stages of the research process, coupled with free and open access to data, code, and papers". When integrated into the scientific process, data science tools that enable open science — let's call them "open data science" tools — can help realise reproducibility in collaborative scientific research[6,16,40,48,49].

Open data science tools helped us upgrade our approach to reproducible, collaborative, and transparent science, but they did require a substantial investment to learn, which we did incrementally over time (Figure 1; Box 1). Previous to this evolution, most team members with any coding experience — not necessarily in `R` — had learned just enough to accomplish whatever task had been before them using their own unique conventions. Given the complexity of the OHI project, we needed to learn to code collaboratively and incorporate best[50,51] or good enough practices[12,52] into our coding, so that our methods could be co-developed and vetted by multiple team members. Using a version control system not only improved our file and data management, but allowed individuals to feel less inhibited about their coding contributions, since files could always be reverted back to previous versions if there were problems. We built confidence using these tools by sharing our imperfect code, discussing our challenges, and learning as a team. These tools quickly became the keystone of how we work, and have overhauled our approach to science, perhaps as much as email did in decades prior. They have changed the way we think about science and about what is possible. The following describes how we have been using open data science practices and tools to overcome the biggest challenges we encountered to reproducibility, collaboration, and communication.

Reproducibility

Data preparation - coding and documenting

Our first priority was to code all data preparation, create a standard format for final data layers, and do so using a single programmatic language, `R`[42]. Code enables us to reproduce the full process of data preparation, from data download to final model inputs[37,53], and a single language makes it more practical for our team to learn and contribute collaboratively. We code in `R` and use RStudio[43] to power our workflow because it has a user-friendly interface and built-in tools useful for coders of all skill levels, and, importantly, it can be configured with Git to directly sync with GitHub online (See Collaboration section). We have succeeded in transitioning to `R` as our primary coding language for data preparation, including for spatial data, although some operations still require additional languages and tools such as ArcGIS, QGIS, and Python[54–56].

All our code is underpinned by the principles of tidy data, the grammar of data manipulation, and the `tidyverse` `R` packages developed by Wickham[33,57–59]. This deliberate philosophy for thinking about data helped bridge our scientific questions with the data processing required to get there, and the readability and conciseness of `tidyverse` operations makes our data analysis read more as a story arc. Operations require less syntax — which can mean fewer potential errors that are easier to identify — and they can be chained together, minimising intermediate steps and data objects that can cause clutter and confusion[33,60]. `tidyverse` tools for wrangling data have expedited our transformation as coders and made `R` less intimidating to learn. We heavily rely on a few packages for data wrangling and visualisation that are bundled in the `tidyverse` package[58,59] — particularly `dplyr`, `tidyr`, and `ggplot2` — as well as accompanying books, cheatsheets, and archived webinars (Box 1).

We keep detailed documentation describing metadata (e.g., source, date of access, links) and data processing decisions — trying to capture not only the processing we decided to do, but what we decided against. We started with small plain text files accompanying each `R` file, but have transitioned to documenting with R Markdown[61,62] because it combines plain text and executable chunks of `R` code within the same file and serves as a living lab notebook. Every time R Markdown output files are regenerated the `R` code is rerun so the text and figures will also be regenerated and reflect any updates to the code or underlying data. R Markdown files increase our reproducibility and efficiency by streamlining documentation and eliminating the need to constantly paste updated figures into reports as they are developed.

Modeling – `R` functions and packages

Once the data are prepared, we develop assessment-specific models to calculate OHI scores. Models were originally coded in multiple languages to accommodate disparate data types and formatting. By standardising our approach to data preparation and final data layer format, we have been able to translate all models into `R`. In addition to assessment-specific models, the OHI framework includes core analytical operations that are used by all OHI assessments[34], and thus we created an `R` package called `ohicore`[63], which was greatly facilitated by the `devtools` and `roxygen2` packages[64–66]. The `ohicore` package is maintained in and installed from a dedicated GitHub repository — using `devtools::install_github('ohi-science/ohicore')` — from any computer with `R` and an internet connection, enabling groups leading independent OHI assessments to use it for their own work[34].

Version control

We use Git[44] as a version control system. Version control systems track changes within files and allow you to examine or rewind to previous versions. This saves time that would otherwise be spent duplicating, renaming, and organising files to preserve past versions. It also makes folders easier to navigate since they are no longer overcrowded with multiple files suffixed with dates or initials (e.g., `final_JL-2012-02-26.csv`)[67–69]. Once Git is configured on each team member's machine, they work as before but frequently commit to saving a snapshot of their files, along with a human-readable "commit message"[67,68]. Any line modified in a file tracked by Git will then be attributed to that user.

We interface with Git primarily through RStudio, using the command line for infrequently encountered tasks. Using RStudio to interact with Git was key for our team's uptake of a version control system, since the command line can be an intimidating hurdle or even a barrier for beginners to get onboard with using version control. We were less resistant because we could use a familiar interface, and as we gained fluency in Git's operations through RStudio we translated that confidence to the command line.

Organisation

Our team developed conventions to standardise the structure and names of files to improve consistency and organisation. Along with the GitHub workflow (see Collaboration section below), having a structured approach to file organisation and naming has helped those within and outside our team navigate our methods more easily. We organise parts of the project in folders that are both RStudio "projects" and GitHub "repositories", which has also helped us collaborate using shared conventions rather than each team member spending time duplicating and organising files.

Collaboration within our team

Coding collaboratively

We transitioned from a team of distinct roles (scientists-and-programmer) to becoming a team with overlapping skill sets (scientists-as-programmers, or simply, data scientists). Having both environmental expertise and coding skills in the same person increases project efficiency, enables us to vet code as a team, and reduces the bottleneck of relying on a single programmer. We, like Duhigg[70], have found that "groups tend to innovate faster, see mistakes more quickly and find better solutions to problems". Developing these skills and creating the team culture around them requires leadership with the understanding that fostering more efficient and productive scientists is worth the long-term investment. Our team had the freedom to experiment with available tools and their value was recognised with a commitment that we, as a team, would adopt and pursue these methods further. In addition to supportive leadership, having a "champion" with experience of how tools can be introduced over time and interoperate can expedite the process, but is not the only path (Box 2). Taking the time to experiment and invest in learning data science principles, tools, and skills enabled our team to establish a system of best practices for developing, using, and teaching the OHI Toolbox.

Our (simplified) GitHub workflow

GitHub is one of many web-based platforms that enables files tracked with Git to be collaboratively shared online so contributors can keep their work synchronised[45,68,69], and it is increasingly being adopted by scientific communities for project management[71]. Versioned files are synced online with GitHub similar to the way Dropbox operates, except syncs require a committed, human-readable message and reflect deliberate snapshots of changes made that are attributed to the user, line-by-line, through time. Built for large, distributed teams of software developers, GitHub provides many features that we as a scientific team, new to data science, do not immediately need, and thus we mostly ignore features such as branching, forking, and pull requests. Our team uses a simplified GitHub workflow whereby all members have administrative privileges to the repositories within our `ohi-science` organisation. Each team member is able to sync their local work to GitHub.com, making it easier to attribute contribution, as well as identify to whom to direct questions.

GitHub is now central to many facets of our collaboration as a team and with other communities — we use it along with screensharing to teach and troubleshoot with groups leading independent OHI assessments, as well as to communicate our ongoing work and final results (see Communication section). Now there are very few files emailed back and forth within our team since we all have access to all repositories within the `ohi-science` organisation, and can navigate to and edit whatever we need. Additionally, these organised files are always found with the same file path, whether on GitHub.com or on someone's local computer; this, along with RStudio `.Rproj` files, eases the file path problems that can plague collaborative coding and frustrate new coders.

Internal communication

We use a feature of GitHub called 'Issues' in place of email for discussions about data preparation and analysis. We use Issues in a separate private repository to keep our conservations private but our work public. All team members can see and contribute to all conversations, which are a record of all our decisions and discussions across the project and are searchable in a single place. Team members can communicate clearly by linking to specific lines of code in current or past versions of specific files since they are stored on GitHub and thus have a URL, as well as paste images and screenshots, link to other websites, and send an email to specific team members directly by mentioning their GitHub username. In addition to discussing analytical options, we use Issues to track ongoing tasks, tricks we have learned, and future ideas. Issues provide a written reference of institutional memory so new team members can get up to speed more easily. Most importantly, GitHub Issues have helped us move past the never-ending forwarded email chains and instead to conversations available to any current or future team member.

Communication outside the project

Sharing data and code

Our code is online in GitHub repositories, publicly available for any researcher or interested person to see and access (github.com/ohi-science). As we work, GitHub renders code, text, images, and tabular and spatial data and displays differences between versions, essentially creating webpages that can be easily shared with collaborators, whether or not they use GitHub. Additionally, we create 'Releases' for each global assessment[36,37] so the code and data we use for peer-reviewed publication are preserved while we continue our work (https://github.com/OHI-Science/ohi-global/releases).

Sharing methods and instruction

We use R Markdown not only for data preparation but also for broader communication. R Markdown files can be generated into a wide variety of formatted outputs, including PDFs, slides, Microsoft Word documents, HTML files, books, or full websites[61,62]. These can all be published online for free through GitHub using the same RStudio-GitHub workflow that we use for our analyses, which has made communication an ongoing part of our work, instead of a final step in completed analyses.

We built a website using GitHub and RStudio publishing tools: ohi-science.org. Team members can update content directly, and using the same workflow makes it easier for us to keep it current. Our website is intended for scientists interested in our methods as well as those leading their own assessments[34]. Thus, the website provides scientific methods, publications, data, and code, as well as instruction, news, blog posts, and a map displaying where all ongoing OHI assessments are taking place so that groups can learn directly from and build off of each other's code. ohi-science.org provides technical information to complement oceanhealthindex.org, our overview website intended for more general audiences.

Meeting scientists where they are

We are environmental scientists whose impetus for upgrading approaches to collaborative, data-intensive science was driven by our great difficulty reproducing our own methods. Many researchers do not attempt to reproduce their own work[17,72] — ourselves included before 2013 — and thus may not realise that there could be reproducibility issues in their own approaches. But they can likely identify inefficiencies. Integrating open data science practices and tools into science can save time, while also improving reproducibility for our most important collaborator: our 'future selves'. We have found this as individuals and as a team: We could not be as productive[34,35] without open data science practices and tools. We would also not be able to efficiently share and communicate our work while it is ongoing rather than only post-publication, which is particularly important for bridging science and policy. As environmental scientists who are still learning, we hope sharing our experiences will empower other scientists to upgrade their own approaches, helping further shift the scientific culture to value transparency and openness as a benefit to all instead of as a vulnerability[16,40,48].

From our own experience and from teaching other academic, non-profit, and government groups through the Ocean Health Index project[34], we find that the main barriers to engagement boil down to exposure and confidence: first knowing which tools exist that can be directly useful to one's research, and then having the confidence to develop the skills to use them. These two points are simple but critical. We are among the many environmental scientists who were never formally trained to work deliberately with data. Thus, we were unaware of how significantly open data science tools could directly benefit our research[11,73], and upon learning about them we were hesitant, or even resistant, to engage. However, we were able to develop confidence in large part because of the open, inclusive, and encouraging online developer community that builds tools and creates tutorials that meet scientists where they are (Box 1, Box 2). It takes motivation, patience, diligence, and time to overcome the conceptual and technical challenges involved in developing computing skills but resources are available to help scientists get started[11,51,73]. Coding is "as important to modern scientific research as telescopes and test tubes"[50], but it is critical to "dispel the misconception that these skills are intuitive, obvious, or in any way inherent"[41].

There is ongoing and important work by the informatics community on the architecture and systems for data management and archiving[7,8,18,74], as well as efforts to enable scientists to publish the code that they do have[26,31,52]. This work is critical, but comes with the *a priori* assumption that scientists are already thinking about data and coding in a way that they would seek out further resources. In reality, this is not always the case, and without visible examples of how to use these tools within their scientific fields, common stumbling blocks will be continually combatted with individual workarounds instead of addressed with intention. These workarounds can greatly delay focusing on actual scientific research, particularly when scientific questions that may not yet have answers — e.g., how the behavior of X changes with Y — are conflated with data science questions that have many existing answers — e.g., how to operate on only criteria X and Y.

Scientific advancement comes from building off the past work of others; scientists can also embrace this principle for using software tools to tackle some of the challenges encountered in modern scientific research. In a recent survey in *Nature*, 90% of the 1,500 respondents across scientific fields agreed that there was a reproducibility crisis in science, and one third of the respondents reported not having their own "established procedures for reproducibility"[4]. While reproducibility means distinct things within the protocols of each sub-discipline or specialty, underpinning reproducibility across all disciplines in modern science is working effectively and collaboratively with data, including wrangling, formatting, and other tasks that can take 50-80% of a data scientist's time[75]. While reaching full reproducibility is extremely difficult[5,76], incrementally incorporating open data science practices and tools into scientific workflows has the potential to alleviate many of the troubles plaguing science, including collaboration and preserving institutional memory[12]. Further, sharing openly is fundamental to truly expediting scientific progress because others can build directly off previous work if well-documented, re-usable code are available[16,47,48,77]. Until quite recently, making research open required a great deal of extra work for researchers and was less likely to be done. Now, with available tools, the benefits of openness can be a byproduct of time-saving efficiencies, because tools that reduce data headaches also result in science that is more transparent, reproducible, collaborative, and freely accessible to others.

Ecologists and environmental scientists arguably have a heightened responsibility for transparency and openness, as data products provide important snapshots of systems that may be forever altered due to climate change and other human pressures[16,18]. There is particular urgency for efficiency and transparency, as well as opportunity to democratise science in fields that operate at the interface of science and policy. Individuals play an important part by promoting good practices and creating supportive communities[16,41,48]. But it is also critical for the broader science community to build a culture where openness and reproducibility are valued, formally taught, and practiced, where we all agree that they are worth the investment.

References

1. Baker, M. Over half of psychology studies fail reproducibility test. *Nature* (2015). doi:10.1038/nature.2015.18248

2. Baker, M. & Dolgin, E. Cancer reproducibility project releases first results. *Nature News* **541,** 269 (2017). doi:10.1038/541269a

3. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015). doi:10.1126/science.aac4716

4. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (2016). (2016). doi:10.1038/533452a

5. Aschwanden, C. Science isn't broken. FiveThirtyEight. (2015). Available at: https://fivethirtyeight.com/features/science-isnt-broken/. (Accessed: 24th January 2017)

6. Buck, S. Solving reproducibility. *Science* **348,** 1403–1403 (2015). doi:10.1126/science.aac8041

7. Frew, J. & Dozier, J. Environmental informatics. *Annual Review of Environment and Resources* **37,** 449–472 (2012). doi:10.1146/annurev-environ-042711-121244

8. Jones, M. B., Schildhauer, M. P., Reichman, O. J. & Bowers, S. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics* **37,** 519–544 (2006).

9. Michener, W. K. & Jones, M. B. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution* **27,** 85–93 (2012). doi:10.1016/j.tree.2011.11.016

10. Check Hayden, E. Mozilla plan seeks to debug scientific code. *Nature News* **501,** 472 (2013). doi:10.1038/501472a

11. Boettiger, C., Chamberlain, S., Hart, E. & Ram, K. Building software, building community: Lessons from the rOpenSci project. *Journal of Open Research Software* **3,** (2015). doi:10.5334/jors.bu

12. Wilson, G. *et al.* Good enough practices in scientific computing. *arXiv:1609.00037 [cs]* (2016).

13. Wilson, G. V. Where's the real bottleneck in scientific computing? *Sigma Xi, The Scientific Research Society* **Jan-Feb,** 5–6 (2006).

14. Baker, M. Scientific computing: Code alert. *Nature* **541,** 563–565 (2017). doi:10.1038/nj7638-563a

15. Barone, L., Williams, J. & Micklos, D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *bioRxiv* 108555 (2017). doi:10.1101/108555

16. Wolkovich, E. M., Regetz, J. & O'Connor, M. I. Advances in global change research require open science by individual researchers. *Global Change Biology* **18,** 2102–2110 (2012). doi:10.1111/j.1365-2486.2012.02693.x

17. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348,** 1422–1425 (2015). doi:10.1126/science.aab2374

18. Reichman, O. J., Jones, M. B. & Schildhauer, M. P. Challenges and opportunities of open data in ecology. *Science* **331,** 703–705 (2011). doi:10.1126/science.1197962

19. Shade, A. & Teal, T. K. Computing workflows for biologists: A roadmap. *PLOS Biol* **13,** e1002303 (2015). doi:10.1371/journal.pbio.1002303

20. Goodman, A. *et al.* Ten simple rules for the care and feeding of scientific data. *PLOS Computational Biology* **10,** e1003542 (2014). doi:10.1371/journal.pcbi.1003542

21. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLOS Computational Biology* **9,** e1003285 (2013). doi:10.1371/journal.pcbi.1003285

22. White, E. P. *et al.* Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution* **6,** (2013).

23. Kervin, K., Michener, W. & Cook, R. Common errors in ecological data sharing. *Journal of eScience Librarianship* **2,** (2013). doi:10.7191/jeslib.2013.1024

24. Lewandowsky, S. & Bishop, D. Research integrity: Don't let transparency damage science. *Nature News* **529,** 459 (2016). doi:10.1038/529459a

25. Michener, W. K. Ten simple rules for creating a good data management plan. *PLOS Computational Biology* **11,** e1004525 (2015). doi:10.1371/journal.pcbi.1004525

26. Mislan, K. A. S., Heer, J. M. & White, E. P. Elevating the status of code in ecology. *Trends in Ecology & Evolution* **31,** 4–7 (2016). doi:10.1016/j.tree.2015.11.006

27. Kratz, J. & Strasser, C. Data publication consensus and controversies. *F1000Research* (2014). doi:10.12688/f1000research.3979.3

28. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1,** 0021 (2017). doi:10.1038/s41562-016-0021

29. Martinez, C. *et al. Reproducibility in science: A guide to enhancing reproducibility in scientific results and writing*. (2014). Available at: http://ropensci.github.io/reproducibility-guide/

30. Tuyl, S. V. & Whitmire, A. L. Water, water, everywhere: Defining and assessing data sharing in academia. *PLOS ONE* **11,** e0147942 (2016). doi:10.1371/journal.pone.0147942

31. Baker, M. Why scientists must share their research code. *Nature* (2016). doi:10.1038/nature.2016.20504

32. Kidwell, M. C. *et al.* Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology* **14,** e1002456 (2016). doi:10.1371/journal.pbio.1002456

33. Wickham, H. & Grolemund, G. *R for data science*. (O'Reilly, 2016). Available at: http://r4ds.had.co.nz/

34. Lowndes, J. S. S. *et al.* Best practices for assessing ocean health in multiple contexts using tailorable frameworks. *PeerJ* **3,** e1503 (2015). doi:10.7717/peerj.1503

35. Lowndes, J. S. S. A biography of the ocean health index. OHI-science. (2017). Available at: http://ohi-science.org/news/Biography-OHI. (Accessed: 26th January 2017)

36. Halpern, B. S. *et al.* An index to assess the health and benefits of the global ocean. *Nature* (2012). doi:10.1038/nature11397

37. Halpern, B. S. *et al.* Patterns and emerging trends in global ocean health. *PLoS ONE* **10,** e0117863 (2015). doi:10.1371/journal.pone.0117863

38. Ocean Health Index Team. Five years of global ocean health index assessments. (2016). Available at: http://ohi-science.org/ohi-global/.

39. Goldfuss, C. & Holdren. The nation's first ocean plans. The white house of president barack obama. (2016). Available at: https://obamawhitehouse.archives.gov/blog/2016/12/07/nations-first-ocean-plans. (Accessed: 24th January 2017)

40. Hampton, S. E. *et al.* The tao of open science for ecology. *Ecosphere* **6,** 1–13 (2015). doi:10.1890/ES14-00402.1

41. Mills, B. Introducing mozilla science study groups. Mozilla science. (2015). Available at: https://science.mozilla.org/blog/introducing-mozilla-science-study-groups. (Accessed: 2nd August 2016)

42. R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing (2016). https://www.R-project.org/

43. RStudio Team. *RStudio: Integrated development for R*. RStudio, Inc, (2016). www.rstudio.com

44. Git Team. *Git version control system*. (Git, 2016). https://git-scm.com/

45. GitHub. *GitHub: A collaborative online platform to build software.* (GitHub, 2016). https://github.com

46. Wilson, G. V. Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science and Engineering* 66–69 (2006).

47. Broman, K. Initial steps toward reproducible research. (2016). Available at: http://kbroman.org/steps2rr/.

48. McKiernan, E. C. *et al.* How open science helps researchers succeed. *eLife* **5,** e16800 (2016). doi:10.7554/eLife.16800

49. Seltenrich, N. Scaling the heights of data science. *Breakthroughs: The magazine of the UC Berkeley College of Natural Resources* **Fall,** (2016). Available at: https://nature.berkeley.edu/breakthroughs/opensci-data

50. Wilson, G. *et al.* Best practices for scientific computing. *PLOS Biol* **12,** e1001745 (2014). doi:10.1371/journal.pbio.1001745

51. Haddock, S. H. & Dunn, C. W. *Practical computing for biologists*. (Sinauer Associates, Inc., 2011). http://practicalcomputing.org/

52. Barnes, N. Publish your computer code: It is good enough. *Nature* **467,** 753 (2010).

53. Frazier, M., Longo, C. & Halpern, B. S. Mapping uncertainty due to missing data in the global ocean health index. *PLOS ONE* **11,** e0160377 (2016). doi:10.1371/journal.pone.0160377

54. ESRI. *ArcGIS platform*. (2016). http://www.esri.com/software/arcgis

55. The QGIS Team. *QGIS project*. (2016). http://www.qgis.org

56. The Python Team. *Python.org*. (2016).

57. Wickham, H. Tidy data. *Journal of Statistical Software* **59,** 1–23 (2014). doi:10.18637/jss.v059.i10

58. Wickham, H. Tidyverse tidyweb. (2017). Available at: http://tidyverse.org/. (Accessed: 29th January 2017)

59. Wickham, H. *Tidyverse: Easily install and load 'tidyverse' packages*. (2016). Available at: https://CRAN.R-project.org/package=tidyverse

60. Fischetti, T. How dplyr replaced my most common r idioms. StatsBlogs. (2014). Available at: http://www.onthelambda.com/2014/02/10/how-dplyr-replaced-my-most-common-r-idioms/. (Accessed: 22nd September 2016)

61. RStudio Team. R Markdown. (2016). Available at: http://rmarkdown.rstudio.com/.

62. Allaire, J.J. *et al. R Markdown: Dynamic documents for R*. (2016). Available at: https://CRAN.R-project.org/package=rmarkdown

63. Ocean Health Index Team. *Ocean Health Index ohicore package*. (2016).

64. Wickham, H. *R packages*. (O'Reilly, 2015). Available at: http://r-pkgs.had.co.nz/

65. Wickham, H. & Chang, W. *Devtools: Tools to make developing R packages easier.* (2016). Available at: https://CRAN.R-project.org/package=devtools

66. Wickham, H., Danenberg, P. & Eugster, M. *Roxygen2: In-source documentation for R*. (2015). Available at: https://CRAN.R-project.org/package=roxygen2

67. Ram, K. Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine* **8,** 7 (2013). doi:10.1186/1751-0473-8-7

68. Blischak, J. D., Davenport, E. R. & Wilson, G. A quick introduction to version control with git and GitHub. *PLOS Comput Biol* **12,** e1004668 (2016). doi:10.1371/journal.pcbi.1004668

69. Perez-Riverol, Y. *et al.* Ten simple rules for taking advantage of git and GitHub. *PLOS Comput Biol* **12,** e1004947 (2016). doi:10.1371/journal.pcbi.1004947

70. Duhigg, C. What google learned from its quest to build the perfect team. *The New York Times* (2016). Available at: http://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html

71. Perkel, J. Democratic databases: Science on GitHub. *Nature* **538,** 127–128 (2016). doi:10.1038/514127a

72. Casadevall, A. & Fang, F. C. Reproducible science. *Infection and Immunity* **78,** 4972–4975 (2010).

73. Wilson, G. Software carpentry: Lessons learned. *F1000Research* (2016). doi:10.12688/f1000research.3-62.v2

74. Hampton, S. E. *et al.* Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11,** 156–162 (2013). doi:10.1890/120103

75. Lohr, S. For big-data scientists, 'janitor work' is key hurdle to insights. *The New York Times* (2014). Available at: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

76. FitzJohn, R., Pennell, M., Zanne, A. & Cornell, W. Reproducible research is still a challenge. ROpenSci. (2014). Available at: /blog/2014/06/09/reproducibility. (Accessed: 2nd August 2016)

77. Boland, M. R., Karczewski, K. J. & Tatonetti, N. P. Ten simple rules to enable multi-site collaborations through data sharing. *PLOS Computational Biology* **13,** e1005278 (2017). doi:10.1371/journal.pcbi.1005278

78. Perkel, J. M. Scientific writing: The online cooperative. *Nature* **514,** 127–128 (2014). doi:10.1038/514127a

79. Rbloggers. How twitter improved my ecological model. Rbloggers. (2015). Available at: https://www.r-bloggers.com/how-twitter-improved-my-ecological-model/. (Accessed: 26th January 2017)

Acknowledgements

The Ocean Health Index is a collaboration between Conservation International and the National Center for Ecological Analysis and Synthesis at the University of California at Santa Barbara. We thank Johanna Polsenberg, Steve Katona, Erich Pacheco, and Lindsay Mosher who are our partners at Conservation International. We thank all past contributors and funders that have supported the Ocean Health Index, including Beau and Heather Wrigley and The Pacific Life Foundation. We also thank all the individuals and groups that openly make their data, tools, and tutorials freely available to others. Finally, we thank Hadley Wickham, Karthik Ram, Kara Woo, and Mark Schildhauer for friendly review of the developing manuscript.

See ohi-science.org/betterscienceinlesstime as an example of a website built with RMarkdown and the RStudio-GitHub workflow, and for links and resources referenced in the paper.

Author Contributions

All authors developed concepts and wrote the paper.

Competing interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to Julia Lowndes: lowndes@nceas.ucsb.edu

Figures

Figure 1

Better science in less time, illustrated by the Ocean Health Index project. Every year since 2012 we have repeated Ocean Health Index (OHI) methods to track change in global ocean health[35,36]. Increased reproducibility and collaboration has reduced the amount of time required to repeat methods (size of bubbles) with updated data annually, allowing us to focus on improving methods each year (biggest innovations written as text). The original assessment in 2012 focused solely on scientific methods (e.g., obtaining and analyzing data; developing models; calculating and presenting results (dark shading)). In 2013, by necessity we gave more focus to data science (e.g., data organisation and wrangling; coding; versioning; documentation (light shading)), using open data science tools. We established R as the main language for all data preparation and modeling (using RStudio), which drastically decreased the time involved to complete the assessment. In 2014, we adopted Git and GitHub for version control, project management, and collaboration. This further decreased the time required to repeat the assessment. We also created the OHI Toolbox, which includes our R package ohicore for core analytical operations used in all OHI assessments. In subsequent years we have continued (and plan to continue) this trajectory towards better science in less time by improving code with principles of tidy data[33]; standardising file and data structure; and focusing more on communication, in part by creating websites with the same open data science tools and workflow. See text and Table 1 for more details.

Tables

Table 1

Summary of the primary open data science tools used to upgrade reproducibility, collaboration, and communication, by task. The transition to using open data science tools was incremental (see Figure 1). All tasks are accomplished with the RStudio–GitHub workflow that is underpinned by R and Git. This workflow streamlines collaboration by capturing each individual's contribution to the project – thus taking care of bookkeeping – for tasks from data processing and analysis to creating documents and websites with embedded results that are updatable. Note that collaboration is not only for labs and teams, but also for each individual's 'future self'.

	Task	Then	Now	Primary open data science tools
	data preparation	manually (i.e., Excel)	coded in R	R packages: `tidyverse` (`dplyr`, `tidyr`, `ggplot2`). Documentation: R Markdown
Reproducibility	modeling	multiple programming languages	R functions and `ohicore` package	R packages: `tidyverse`, `devtools`, `roxygen2`, `git2r`
	version control	file duplication and renaming	Git	Git; interface with Git and GitHub primarily through RStudio
	organisation	individual conventions	standardised team convention	RStudio projects, GitHub repositories. File structure protocols
	coding	separate languages and conventions	R; standardised team convention	Principles of tidy data; `tidyverse`
Collaboration	workflow and project management	individual conventions	(simplified) GitHub workflow	GitHub, RStudio
	internal	email	centralised,	GitHub issues

	collaboration		archived conversations	
	sharing data	ftp download	all versions and Releases available online	ohi-science.org/ohi-global
Communication	sharing methods	published manuscript and supplementary material	published on our website	ohi-science.org website, with linked R Markdown outputs (webpages, presentations, etc)

Boxes

Box 1

Resources to learn open data science tools. These are some of the free, online resources that we used to learn and develop a workflow with R, RStudio, Git, and GitHub. These resources exposed us to what was possible, and helped us build skills to incorporate concepts and tools into our own workflow. This is by no means an exhaustive list. See also Box 2 for strategies of how to get started.

Primarily R

- **R for Data Science** book by Hadley Wickham and Garrett Grolemund (r4ds.had.co.nz)
- **RStudio's webinars** on-demand videos by RStudio (rstudio.com/resources/webinars)
- **RStudio's cheatsheets** PDFs by RStudio (rstudio.com/resources/cheatsheets)
- **CRAN Task Views** to identify useful packages by category of task (cran.r-project.org/web/views)
- **R Packages** book by Hadley Wickham (r-pkgs.had.co.nz)

Combination RStudio-GitHub

- **Happy Git With R** short-course by Jenny Bryan (happygitwithr.com)
- **UBC Stats545: Data Wrangling, Exploration, and Analysis with R** university course by Jenny Bryan (stat545-ubc.com)
- **Software Carpentry** workshops, teaching and learning communities (software-carpentry.org)
  - example 2-day course: "Reproducible Science with RStudio and GitHub" jules32.github.io/2016-07-12-Oxford/overview

Community discussion

- **#rstats on Twitter** online discussions (twitter.com/search?q=%23rstats&src=typd)
- **Not So Standard Deviations** podcast by Roger Peng and Hilary Parker (soundcloud.com/nssd-podcast)
- **R-Bloggers** blog (r-bloggers.com)

RStudio blog
(blog.rstudio.org)
Data Carpentry blog
(datacarpentry.org/blog)

Box 2

Strategies to learn in an intentional way. The resources listed in Box 1 have helped us learn open data science principles and tools in an intentional way: We felt empowered (vs. panicked), we learned to think ahead (vs. quick fixes for single purposes), and we learned with a community (vs. in isolation). There is a whole ecosystem of open data science principles, practices, and tools (including R, RStudio, Git, and GitHub) and no single way to begin learning. These are a few strategies you can consider as you get engaged.

Self-paced learning
Box 1 lists resources to learn open data science principles and tools that you can use at your own pace. The books and courses provide in-depth philosophies and are good for initial learning as well as for reference later on. Webinars and podcasts are generally under an hour.

Join and/or create communities
Learning together and supporting each other peer-to-peer can be more fun and rewarding. You can become a "champion" for others by showing leadership as you learn. Start off by watching a webinar with a friend or group during lunch or a happy hour. Learn enough about a useful R package to share in your lab meetings; you learn best by teaching. In traditional journal clubs or lab meetings, discuss an academic article on importance of reproducibility, collaboration, and coding[14,22,69,78]. Search if your institution or city has local Meetup.com groups, or create your own.

Additionally, join or keep tabs on communities online. Mozilla Study Groups are a network of 'journal-clubs' where scientists teach scientists computing skills (science.mozilla.org/programs/studygroups/join). rOpenSci is a developer collective building R-based tools to facilitate open science (ropensci.org). Also look on Twitter for #rstats discussions and then follow individuals from those conversations.

Ask for help
Local and online communities are a great resource to ask when you need help. Expecting that someone has already asked your question can help you both articulate the problem clearly and identify useful answers. Often, pasting error messages directly into Google will get you to the best answers quickly. Many answers come from online forums, including

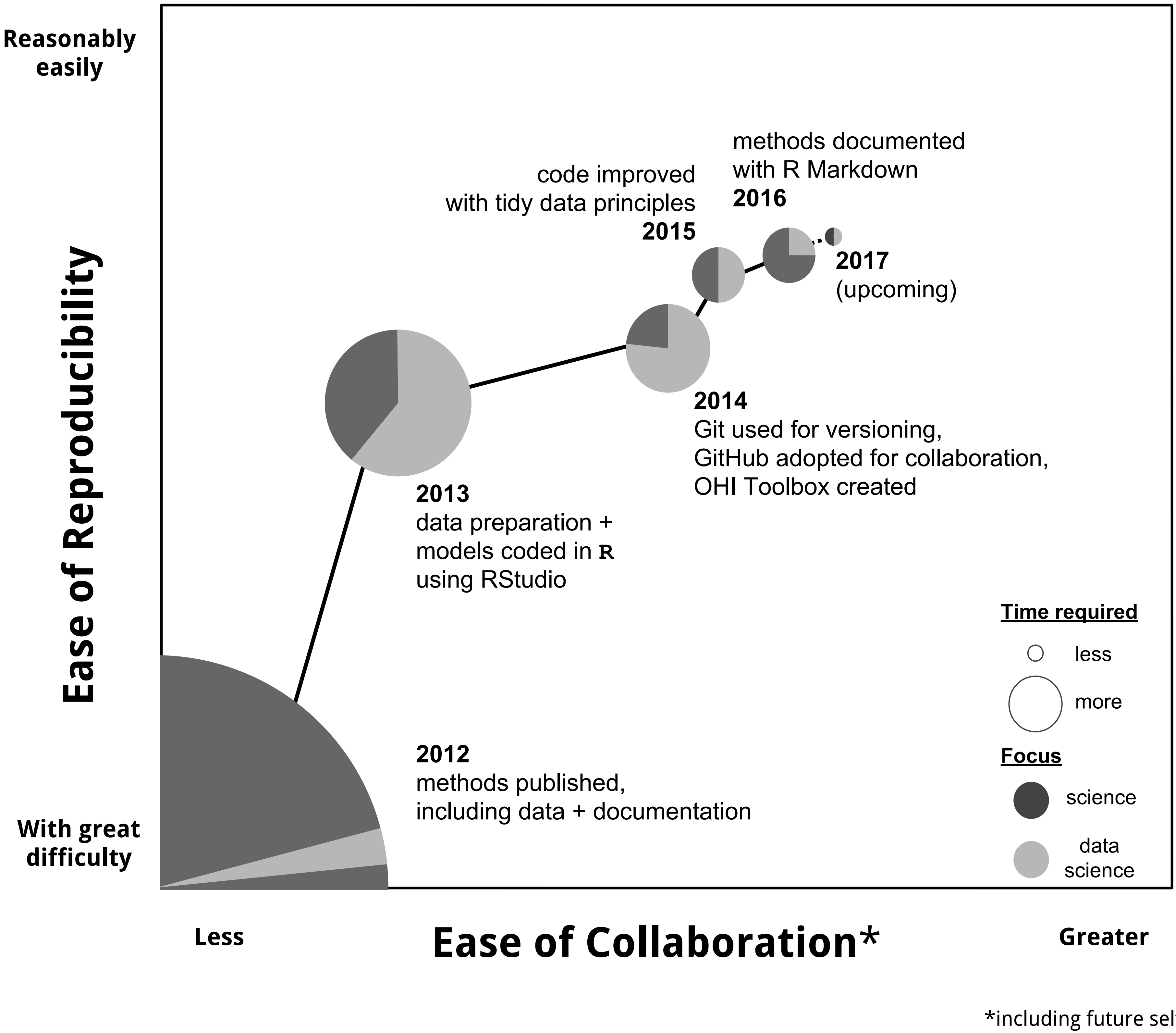
StackOverflow.com[14], or even Twitter itself (e.g., 'How Twitter Improved My Ecological Model')[79].

Attend in-person workshops and conferences
In-person workshops can be extremely valuable and give you an opportunity to get direct help from instructors and helpers. Software Carpentry and Data Carpentry run 2-day bootcamps that teach skills for research computing; you can attend a scheduled workshop or request your own (software-carpentry.org; datacarpentry.org). Attend conferences like useR (example: user2017.brussels) both for skill-building and to learn how others are using these tools.

Watch presentations from past conferences
More and more, slide decks and videos of presentations are appearing online. For example, you can see presentations from the the 2016 useR conference (user2016.org) and the 2017 RStudio conference (rstudio.com/conference).

Read Blogs

There are many individuals who blog about open data science concepts, R packages, workflows, etc. Try Googling a package you're using, or going to the website of someone you are following on Twitter.



Reasonably easily
With great difficulty
Ease of Reproducibility
methods documented with R Markdown
2016
code improved with tidy data principles
2015
2017 (upcoming)
2014
Git used for versioning, GitHub adopted for collaboration, OHI Toolbox created
2013
data preparation + models coded in R using RStudio
Time required
less
more
Focus
science
data science
2012
methods published, including data + documentation
Less
Ease of Collaboration*
Greater
*including future self