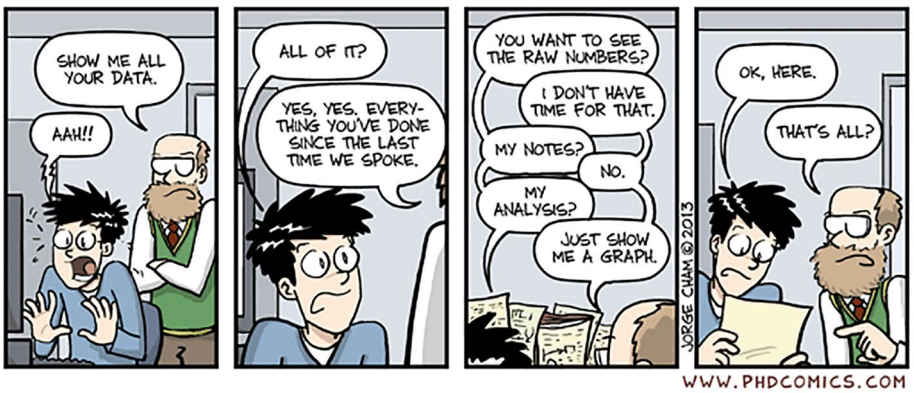
Make the data available

Obtaining valuable datasets is often arduous, costly and time-consuming. Give them away for wider reuse.

We publish papers to share knowledge, technical insight, hypotheses, methodology, implications, data and algorithms. But for most scientific disciplines, raw and processed data, and custom data-generating or data-processing code are most often not freely available or accessible. Why? Cloud space is cheap (and free for many researchers), and network bandwidth is rarely an impediment. Depositing the datasets and scripts in public repositories is much less laborious than obtaining the data and writing code (*Nature* **584**, 656–658; 2020). And researchers are keen to help others build on their work. Why isn't data deposition widespread then? Are researchers weary of data misuse, of being scooped on future uses of the data, or of wider scrutiny? Do old habits die hard? Possibly. Yet misalignments of incentives, rewards and penalties may play a bigger role.

Researchers, especially when funded by taxpayer money, are morally and culturally prompted to share data widely after patenting or publishing their work (after all, science thrives when open collaboration and constructive criticism are nurtured). Funders of research and research institutions increasingly encourage — and in some cases, mandate — data-sharing plans and practices. Many journals require authors to report how their data and custom code can be accessed, and are experimenting on how to best encourage (and help) their deposition in public repositories (in particular, through the use of badges). There are therefore plenty of incentives (both carrots and sticks) for authors to present their data on a silver platter.

Rewards are harder to come by. Contributions to new knowledge from data reuse, to reproducibility and to trust in the scientific endeavour do not lend themselves to quantification. More citations, new collaborations (*Nature* **569**, 445–447; 2019) and instances of reuse by others can be counted, but it is difficult to do so, especially for individual projects or papers. Having curated, organized, interoperable and annotated datasets available for reuse (as per the FAIR principles) requires time and resources that are not always available, yet aids the training of young researchers and new team members, and speeds up related projects. However, these benefits do not necessarily accrue to those who need to sort and label the datasets.



Credit: "Piled Higher and Deeper" by Jorge Cham www.phdcomics.com

Penalties are, in practice, inexistent or unenforced. Most funders and journals only mandate data deposition for specific datasets (such as nucleic acid sequences and protein sequences) and may not consistently track whether published research data are freely available. And interested researchers can obtain data through collaborations, or by simply asking authors or data owners (success is not guaranteed, however (*Proc. Natl Acad. Sci. USA* **115**, 2584–2589; 2020)). But a lack of public accessibility to specific research datasets poses implicit barriers: it signals that the datasets may not be readily usable; requesting the data might be seen as a nuisance or as a favour to be returned; and limiting data sharing to parties that need to identify themselves can be perceived as a way of exerting control.

Despite the incentives, the uncertain long-term rewards and the insignificant penalties do not make for the best outcomes. How, then, can incentives, rewards and penalties be aligned toward increasing wider data availability? There may not be easy wins. Data deposition across the board cannot be swiftly enforced or monitored. Requiring authors to deposit their datasets will hamper reuse if done hastily or carelessly. Instituting future penalties for non-compliant researchers requires proper (and costly) oversight.

It is better to offer tangible incentives at the right time, elevate the rewards, and use penalties only as a backstop. For instance, grant funders could demand proof of previous data deposition and of data reuse by the community. Research institutions could make it easier for their researchers to compile, structure and annotate data for public deposition, either via data-expert officers or by recruiting professional help from the marketplace. Journals should make it easier for authors too, by identifying useful datasets in the papers they publish and by requiring authors to report how the different datasets can be accessed, and whether there are any access restrictions. *Nature Biomedical Engineering* will now encourage authors submitting an invited revised manuscript — rather than when promising acceptance, as we have been doing — to provide links to the data used to make the graphs (*Nat. Biomed. Eng.* **1**, 0079; 2017), to any necessary custom code and, for some types of dataset and when feasible, to the annotated raw data in machine-readable format and with descriptive headings. (Thus far, we have been requesting this only for a small subset of manuscripts for which we deemed it important that the data or code be available to the peer reviewers.) We will highlight to authors the short-term rewards of doing so, such as increased visibility of the work (for example, via direct citations to the datasets), and will provide our feedback on the appropriateness of a potential data-descriptor paper (whose authorship can particularly reward those who acquired and processed the data). Moreover, when commissioning News & Views pieces for papers with publicly available raw or processed datasets, we will encourage the News & Views author to cite and discuss the usefulness of the (to be) published dataset or code.

Naturally, not all new research datasets can be made publicly available. Identifiable patient information or personal data that are difficult to de-identify; data not entirely owned by the researchers themselves; constraints from private funders and from equipment or software providers; and huge datasets — all impose legal or practical constraints on public sharing. These should be made clear in the manuscript's data availability statement. Sometimes, a representative subset of the data, 'demo data' or synthetic data can be made openly available instead.

The most expected datasets today are those from the developers of COVID-19 vaccines. Press releases touting high antibody titres and a lack of adverse events generate publicity and false hope, and undermine trust. Before claims about vaccine safety and efficacy are made publicly, clinical trial data must be made available for public scrutiny. Although most research studies are not as widely impactful or time-pressing, the moral principles for data sharing are the same. Think of the upsides. Make the data available. ❐

Published online: 10 September 2020
https://doi.org/10.1038/s41551-020-00620-w