Commentary: it is a good idea to have bad ideas in science.

Abstract

There are few truly bad ideas in authentic science. We need to embrace science as a process-driven human endeavour to better understand the world around us. Products are important, but through better transparency, we can leverage ideas, good and bad, ours and others, to do better science. In a brief commentary here inspired by a recent discussion of the topic and previous introspections by other ecologists, it is proposed that whilst it is a good idea to track ideas and all the processes that generate outcomes such as publications, there is inherent merit in all scientific ideas. That said, organizing and framing our ideas into the networks that we already use to examine hypotheses and questions in science is a window into our workflows including ideation, implementation, data analyses, and how we can better map ideas into open science outcomes. Formalizing and describing the linkages between ideas, data, and projects we produce as scientists will enhance and diversify the value of the work we do individually and collectively.

Introduction

A recent editorial suggests that science is all about sorting the wheat from the chaff (Kirwan 2017). To be clear and fair, the author conceptualizes bad ideas using suggestions from eminent scientists on the value and necessity of bad ideas to the advancement of good science. It is concluded that successful science must embrace a pluralism of ideas and that it is not a waste of time to explore ideas that do not work out. The bad-idea science paradigm proposed is a simple dichotomy defining good ideas as those that generate publications and bad as those that do not. This is a functional taxonomy for the purposes of a self-assessment of work done (i.e. effort allocated) and project ideas by the author, and to the defence of this analysis, every effort was made to qualitatively include ideas that were 'built on' for subsequent positive outcomes, i.e. publications. This working definition is an absolutely necessary and convenient short-cut, but it is not the end of the story. The accountability of ideas to the progress of a discipline such as ecology is not a new idea (Grime 1993; Weiner 1995) nor without debate (Aarssen 1997; Weiner 1999). Common ground suggests that we should individually evaluate the merit of our ideas if for no other reason then to better prepare our work for the review process of others. We can even conceptualize some of the less productive ideas as stepping stones to more useful ones or as a counterpoint to frame and anchor the relative merit of those that succeed in whatever capacity we elect to define the positive outcomes. Creation, divine or otherwise, is not a perfect system. Evolution needs mutations. That ideas are bad if they do not either directly or indirectly generate publications is perhaps too limiting. New ideas are novel, and new *and* useful ideas are creative (Runco and Jaeger 2012). The former can become the subtrate for the latter. Useless today can become indispensible tomorrow. Sorting the wheat from the chaff is an exercise in predicting an unknown landscape of discovery. Introspection of our workflows and the relationship between ideation and implementation will nonetheless streamline the mapping of ideas and hypotheses to effective testing through evidence. However, we also need to ensure that ideas do not get lost through self-criticism, lack of data, availability of a mechanism to test today, or through entanglement in the quagmire of of extensive information we process as scientists. In the spirit of replication science (Brandt et al. 2014; Mulkay and Gilbert 1986), I examined my relative idea management and outcomes to explore whether this more generous interpretation of merit and my workflow kept most of the wheat and some of the chaff.

Method of replication and findings

A simplified, direct replication of the concept of idea merit self-assessement from the most recent editorial that inspired this commentary was done. A check of idea efficacy scrapes the files and structures used to support a scientific workflow (hosted locally, but see below on how this landscape is changing). This in and of itself is a superb idea. The need and pressure to publish can encourage one to be less cognizant of the processes that support the final paper particularly if there are many steps or if time to final acceptance is significant (Powell 2016). Many of the simple rules for data and experimental provenance (Kazic 2015) also apply to what can be similarly termed 'scientific idea provenance' including version control, integration, object labelling, and reviewing the semantics. The author in the former idea provenance self-study examined all projects completed in career, recorded initiation date, scored effort for each, and developed an outcome classification scale that included direct and indirect publications from each project (Kirwan 2017). A project was defined as a 'project folder' that likely represented parent directories organized around each independent research thread. The author concluded that 25% of projects generated a publishable result. The assumption of this workflow is that every scientific task is assigned to/nested within a project folder. The folder dimension of my scientific workflows is similar but not a perfect match (and it is fundamentally evolving). I use project folders to organize protocols and ideas, dataset folders for data and some other forms of scientific evidence such as camera trap pictures, an 'idea archive' parent folder to store all ideas, and a 'papers-in-progress' parent folder with subfolders to store ideas that have some development in written form. However, this workflow and structure for organzing the processes that support scientific inquiry is dramatically changing thanks to GitHub and RStudio where I couple code, annotation, and written interpretations with the associated datasets. Scoring effort and time allocated, pre-Github and without effective, time-stamped version control was not viable in terms of reproducibility with my former workflow and its files. Consequently, I restricted analyses primarily to counts. However, I propose that the number of objects within a project, i.e. files, in addition to parent/child folders that support a research thread is an important building block in many instances in the process of turning ideas into more tangible outcomes. This conceptualization of project workflows more fully embraces a provenance paradigm for ideas. Hence, I use a more topological approach to model the merit of ideas herein but expanded the supporting structures associated with ideas and workflows. The diversity of outcomes can also be expanded to recognize open science products published online such as datasets in recognized repositories, slide decks in hosting services, and code that supports data and traditional publications.

To provide context for the idea provenance self-study, I did a scrape of scientific outcomes published online. Papers were recorded as positive outcomes if in print with a peer-reviewed journal, but pre-prints were excluded to avoid non-independence issues. The count of papers was the primary outcome used as the denominator in the previous examination (Kirwan 2017). Datasets published with a DOI in a repository, open slide decks published online with a recognized sharing service, independent repositories with code and data posted to GitHub, and other open science products such as conceptual figures on figshare were also recorded. Total number of independent ideas stored within a parent directory on my local HD were also included in this examination as an estimate of the total available seed pool for all these public outcomes, i.e. the wheat. This more inclusive, open science perspective on outcomes showed that traditional peer-reviewed papers ranked second after the open science products category (Figure 1). This is likely representative of many scientists because we give many talks, and if a slide deck was posted online for many of them, it would exceed the number of publications. Conceptual figures or scientific cartoons are also likely common because many scientists turn predictions or ideas into a theoretical plot to visualize a relationship during the experimental design phase of a project. Many of these visual predictions are unsupported but faciliate sketching out the ideas for the final publication. Finally, the total count of all ideas that I allocated time to record and formally capture as an individual note within the parent folder I termed my 'idea archive' (that is certainly not a Sherlockian mind palace) nearly tripled the count of papers published. In my experiences within working groups, this is within the range of ideas proposed and formally captured in some written capacity. Many more ideas are wildly sown in these collaboratively endeavours, but the capture and germination rate drops off significantly within collective and individual workflows.

At the local level, processes that support these outcomes were scraped from my working machine. These fundamental attributes that reflect my pre-GitHub workflow were comprised of projects (thematic subfolders and files within a parent research projects directory), datasets (thematic subfolders and files within a parent datasets directory), papers-in-progress (subfolders, with files, that in theory will become submissions to journals), and ideas stored within an idea archive parent directory as described above. The total number of subfolders and files were counted using the command line. Subfolders is the closest approximation to the previous idea provenance examination, and for instance, best approximates individual research projects because it aggregated evidence or ideas into outcomes that could become publications. For instance, there are a total of 79 papers-in-progress subfolders and each includes some writing that is the germinant of a potential paper. There were a total of 127 subfolders for datasets with each including sets of experimental data associated with a specific outcome, i.e. survey data on an invasive plant plus a follow-up greenhouse trial within the same subfolder because they were integrated into the same project/research thread and thus subsequent paper. There were a total of 75 project subfolders in my scrape suggesting that I have completed (with colleagues) this number of different research threads. All data and analyses are provided within a **GitHub repository** on this topic. This is not a significant deviation from the workflow of Kirwan, a rheumatologist. An interesting extension is the count of number of files within each of these threads that illuminated the extent that one compiles evidence to explore ideas. Locally stored, there were 4571 datasets, 2005 project files, 804 notes and written documents within papers-in-progress, and 376 annotated ideas (each within a .txt file within that specific category of archiving). To examine the sensitivity of this scale of provenance, both subfolders (replicating the previous analysis) and number of files were used to estimate the proportionate mapping of these scientific elements to published papers (i.e. 126 to date). There was significant variation in the mapping or mobilization of the different building blocks to published papers (Figure 2). The file-level analysis of ideas that estimates individual, annotated ideas in my workflow suggests that 33% of ideas stored locally become papers. Nearly 60% of project subfolders became publications whilst every folder of datasets mapped onto a paper. This is a coarse matching, and the ideal provenance tracking would trace idea-to-data-to-paper, but in reality, the process flows both ways with individual datasets also leading to idea publications without the the data that first inspired the thread. Data and ideas can both propagate laterally to other data and ideas, and individual objects need not directly connect to an outcome to nonetheless promote positive net outcomes. Discovery, even by one individual, is unlikely to be a linear mapping process. It is reasonable to assume from these findings that few ideas can lead to many data that in turn produce fewer publications. The converse is also true that many data can produce fewer ideas, bad or good.

Implications

Without digressing into a theory versus data paradigm clash (Zeller and Carmines 1980), it is legitimate to examine our own scientific workflows, transparently, to ensure that there are few significant impediments in connecting different forms and levels of scientific inquiry or to further ensure that ideas are not lost or unduly discarded. We live in a data deluge wherein we are uninundated with or perhaps compelled to collect data. More intriguingly, it is worthwhile to not only ask can I collect useful data as a scientist, but to revisit the wheat-chaff metaphor, am I a good (i.e. reasonable) idea farmer? Does my process embrace and retain ideas that can support a meaningful, and at times, evidence-based decision process in mobilizing and disseminating research from the insights that I possess as an expert. In this direct and conceptual replication instance, a rheumatologist and an ecologist shared similar levels of career-to-date idea provenance turning 25 and 30% respectively of ideas into publications. This is an experiment of one, twice. It would be ideal to make this process more reproducible and scaleable. My collaborators and I collect a significant volume of independent data files. This strongly suggests that are many objects, i.e. files, that are needed as building blocks between an idea and an outcome. Most folders of data ultimately become a publication (but individual data files are but stepping stones in the scientific process). For me, data best map onto ideas suggesting a cognitive workflow associated with/grounded in ideation from observation. Organizing ideas into project folders is a useful approach and some elements of a wokflow that incorporates aggregation likely increases the capacity for lateral propogation of data and ideas and for increased chances of connecting the dots even within our own individual work. Scientific synthesis is clearly not just a formal method we can use to describe the work of others (Lortie 2014) but a critical tool we need to practice individually. Hence, workflows that stimulate and faciliate connecting ideas and data into networks will promote more creative and more integrated science.

Ideas without context (stored in 'idea archive' on desktop) rarely become more than just that - a noted idea.

Need a venue for ideas that provides both recognition, context via tags, and discoverablility by others. We currently use social media, conference presentations, and a few other channels to share ideas but aggregation and searchability is important. An idea respository like we have for data is just as important, perhaps more so, as a substrate for the advancement of science.

and suggests that different scientists are not entirely dissimilar in how we organize our scientific processes from this perspective.

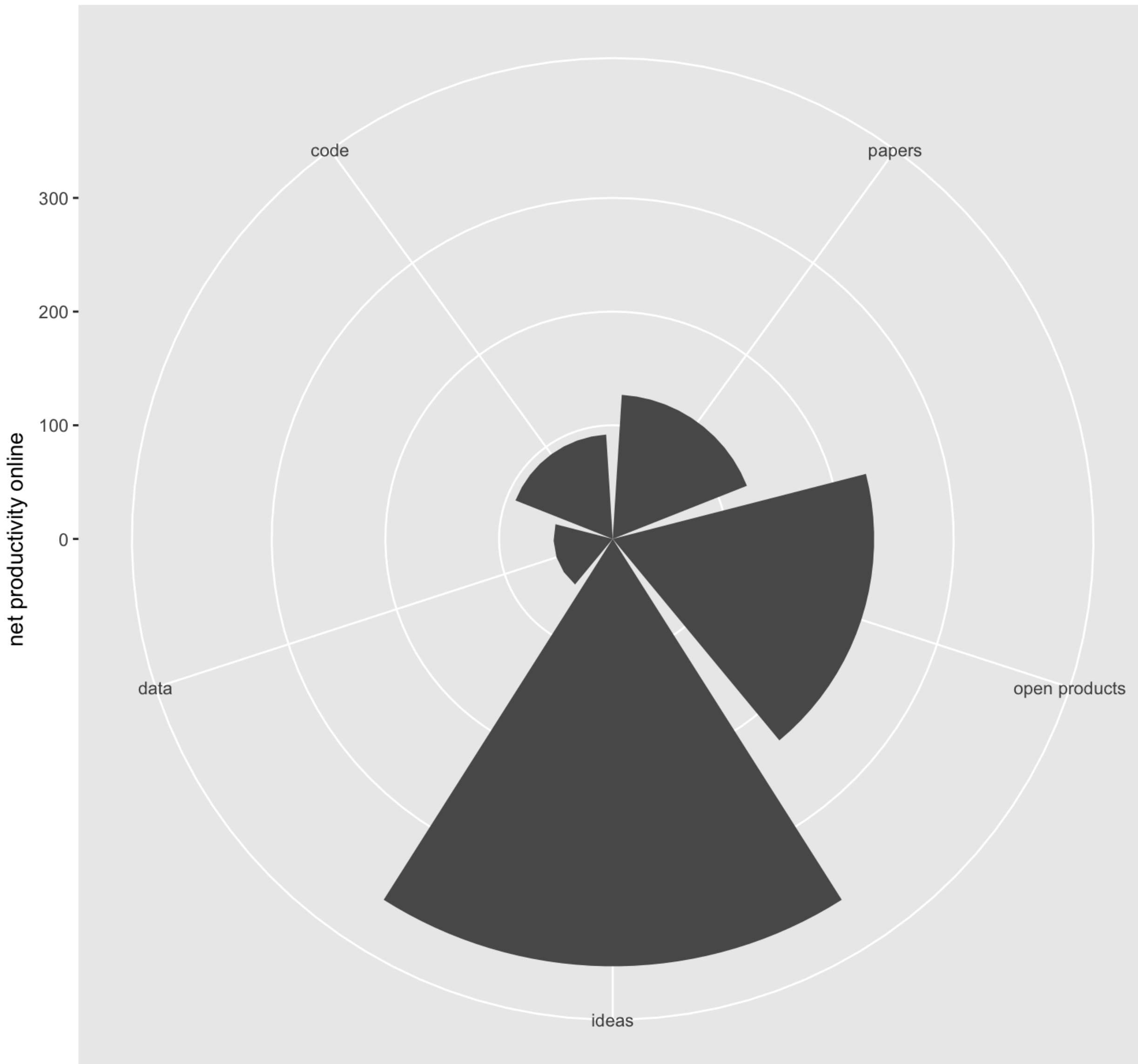

Figure 1: Scientific outcomes published online and ideas captured locally within a collective career-level scrape of productivity for an ecologist. See text for full description of each scientific element class. The open products category includes slide decks and figures posted online in open repositories.

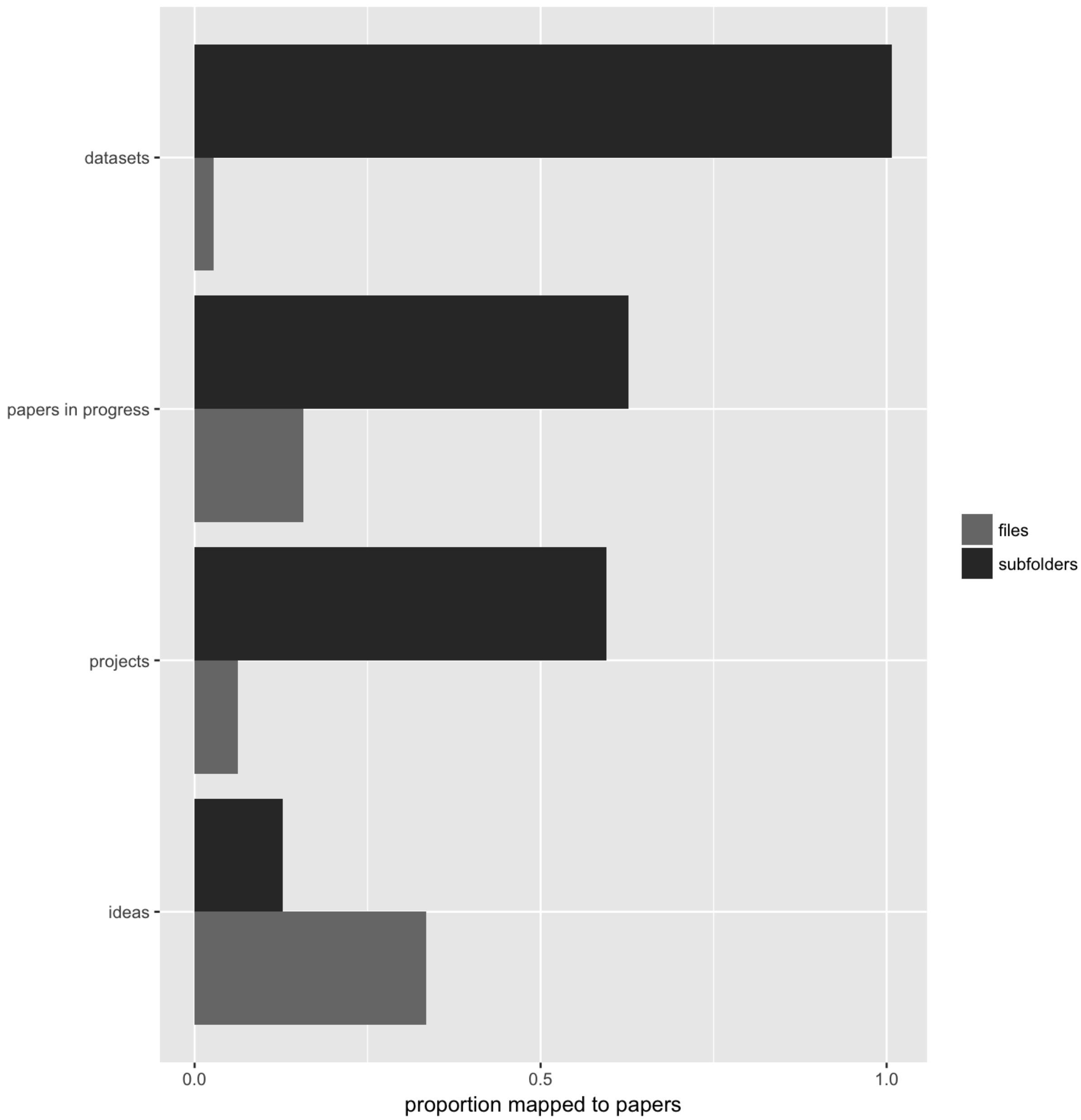

Figure 2: The mapping of ideas and other supporting research processes stored locally to peer-reviewed publications in print or press. The files class represents total number of files stored within a specific category of research thread, and subfolders were the number of folders within parent directories stored on working machine for projects, data, papers-in-progress, and ideas - i.e. the workflow that an ecologist used to organize working files for scientific inquiry and experiments.

Literature cited

Aarssen, L.W. 1997. "On the Progress of Ecology." Journal Article. *Oikos* 80 (1): 177–78.

Brandt, Mark J., Hans Ijzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies, and Anna van 't Veer. 2014. "The Replication Recipe: What Makes for a Convincing Replication?" Journal Article. *Journal of Experimental Social Psychology* 50: 217–24. doi:http://dx.doi.org/10.1016/j.jesp.2013.10.005.

Grime, J.P. 1993. "Ecology Sans Frontieres." Journal Article. *Oikos* 68: 385–92.

Kazic, Toni. 2015. "Ten Simple Rules for Experiments' Provenance." Journal Article. *PLoS Comput Biol* 11 (10): e1004384. doi:10.1371/journal.pcbi.1004384.

Kirwan, J. 2017. "It's Good to Have Lots of Bad Ideas." Journal Article. *Nature* 548: 491. doi:10.1038/nj7668-491a.

Lortie, C. J. 2014. "Formalized Synthesis Opportunities for Ecology: Systematic Reviews and Meta-Analyses." Journal Article. *Oikos* 123: 897–902.

Mulkay, Michael, and G. Nigel Gilbert. 1986. "Replication and Mere Replication." Journal Article. *Philosophy of the Social Sciences* 16 (1): 21–37. doi:10.1177/004839318601600102.

Powell, K. 2016. "The Waiting Vame." Journal Article. *Nature* 530: 148–51.

Runco, M.A., and G.J. Jaeger. 2012. "The Standard Definition of Creativity." Journal Article. *Creativity Research Journal* 24: 92–96.

Weiner, J. 1995. "On the Practice of Ecology." Journal Article. *Journal of Ecology* 83: 153–58.

———. 1999. "On Self-Criticism in Ecology." Journal Article. *Oikos* 85 (2): 373–74.

Zeller, R.A., and E.G. Carmines. 1980. *Measurement in the Social Sciences. the Link Between Theory and Data.* Book. New York: Cambridge University Press.