NATURE | NEWS: Q&A

Why scientists must share their research code

'Reproducibility editor' Victoria Stodden explains the growing movement to make code and data available to others.

Monya Baker

13 September 2016

George Dyson

Many scientists worry over the reproducibility of wet-lab experiments, but data scientist Victoria Stodden's focus is on how to validate computational research: analyses that can involve thousands of lines of code and complex data sets.

Beginning this month, Stodden — who works at the University of Illinois at Urbana-Champaign — becomes one of three 'reproducibility editors' appointed to look over code and data sets submitted by authors to the Applications and Case Studies (ACS) section of the *Journal of the American Statistical Association* (*JASA*). Other journals including *Nature* have established guidelines for accommodating data requests after publication, but they rarely consider the availability of code and data during the review of a manuscript. *JASA* ACS will now insist that — with a few exceptions for privacy — authors submit this information as a condition of publication.

Nature spoke to Stodden about computational reproducibility and the emerging norms of sharing data and code.

Why is *JASA* ACS taking this step?

This is really about what it means to do science. We have publication processes to root out error for research that is done without a computer. Once you introduce a computer, the materials section in a typical scientific paper doesn't come close to providing the information that you need to verify the results. Analysing complicated data by computer requires instructions consisting of script and code. Hence we need the code, and we need the data. The reproducibility editors will gather the code and gather data and gather workflow information, and we'll enforce the requirement that the data and code that support the claims in an article are made available.

What does computational reproducibility mean?

It means that all details of computation — code and data — are made routinely available to others. If I can run your code on your data, then I can understand what you did. We need to expose all the steps that went into any discovery that relies on a computer.

What's the scientific value of running the same data with the same code and getting the same result?

It's true that running the same code on the same data is not advancing science, but it is necessary to advance science. I can

guarantee you that two independent implementations of a computational experiment — that is, two people asking the same questions of the same data set — will not give the exact same output. What's important is that we are able to reconcile the differences. And the only way you are going to do that is if you can see the code and the data.

Don't statisticians and scientists already share their data and code online?

If you open up any journal to see if you can get the data, for most articles you can't. Researchers who are sharing are using resources like CRAN (the Comprehensive R Archive Network), but that doesn't link up to publications and doesn't say how they used their code to get the published results.

If we have something like journal standards, then sharing becomes part of the incentive structure. The idea is to make it the routine and the default. A lot of scientists are sharing code and data, but they are acting as valiant warriors; they risk hurting their careers because it takes time and effort to release the data.

How should researchers make their computational results more reproducible?

Get used to the idea of sharing. First, throw stuff out there — in repositories such as GitHub or Dryad, or in the Harvard Dataverse Network. Even sharing on your website is better than leaving data on your hard drive. Then start thinking about the details. Do you have sufficient metadata? Is it in a discoverable place? All of those things can come in time.

Do you see open data becoming the status quo?

The idea that people should be able to get hold of code and data as the general default, that's where we are moving. We're leaving the world of a narrative that depends on computational work without any supporting digital artefacts.

Nature | doi:10.1038/nature.2016.20504