The essence of reproducibility for Data Scientists

What is reproducibility?

Why does reproducibility matter in Data Science?

- Modern data analysis pipelines are often complex – they involve many steps and many choices (from preprocessing to data aggregation) and dependencies.
- These unfold often over long periods of time – during which the available tools (e.g. versions of Python or other software) change.
- If you are unable to reproduce the same result given the same data, you might be in trouble (or at least on treacherous ground).
- You will likely collaborate on these projects with many people.
- You might want to deploy your project to the world (e.g. on Github), all who should get the same – predictable – results.
- And yet, this is often an afterthought or not done in a principled way.

This is a whole field of research and study

A tragedy of errors

Mistakes in peer-reviewed papers are easy to find but hard to fix, report **David B. Allison** and colleagues.

Just how error-prone and self-correcting is science? We have spent the past 18 months getting a sense of that.

We are a group of researchers working on obesity, nutrition and energetics. In the summer of 2014, one of us (D.B.A.) read a research paper in a well-regarded journal estimating how a change in fast-food consumption would affect children's weight, and he noted that the analysis applied a mathematical model that over-estimated effects by more than tenfold. We and others submitted a letter[1] to the editor explaining the problem. Months later, we were gratified to learn that the authors had elected to retract their paper. In the face of popular articles proclaiming that science is stumbling, this episode was an affirmation that science is self-correcting.

Sadly, in our experience, the case is not representative. In the course of assembling weekly lists of articles in our field, we began noticing more peer-reviewed articles containing what we call substantial or invalidating errors. These involve factual mistakes or veer substantially from clearly accepted procedures in ways that, if corrected, might alter a paper's conclusions.

After attempting to address more than 25 of these errors with letters to authors or journals, and identifying at least a dozen more, we had to stop — the work took too much of our time. Our efforts revealed invalidating practices that occur repeat… and s… react … correc…

We

NATURE.COM
For *Nature*'s special collection on reproducibility, see:
go.nature.com/huhbyr

4 FEBRUAR…

Vicky Rampin

Research Data Management & Reproducibility Librarian

Liaison Relationship
Arts & Science: Data Science. Courant Institute of Mathematical Sciences: Computer Science
Departments Data Services

Elmer Holmes Bobst Library, 70 Washington Square South, New York, NY 10012
vs77@nyu.edu +1 (212) 992-6269 Subject Specialist

Request an Appointment

Best practice recommendations

- I'll share a presentation from Vicky about best practices to a previous class on Brightspace. Here are some highlights:
- The **3-2-1** principle: Be sure to make 3 copies of each essential data file, have 2 backups, 1 of them off-site.
- Each project folder should be standardized and have the following structure: "**Data**", "**Code**", "**Results**" and "**Documentation**"
- Contextual information about the data files (e.g. column headers) and the analysis (e.g. dependencies) should be placed in a **readme** file in the project folder.
- Sensitive data should be stored in a secure repository with additional protections, if privacy is a concern, e.g. "**NYU Box**".
- Version control systems like **Git** are recommended to avoid proliferation of slightly edited versions of the same file