Reproducibility is process

Matthew Brett

Computational reproducibility

An analysis is computationally reproducible} when someone other than the original author of an analysis can produce on their own computer the reported results using the authors' data, code, and instructions (Buckheit and Donoho 1995).

Why computational reproducibility?

> *The scientific method's central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error.*

(Donoho et al. 2009)

Why computational reproducibility?

> *Computing results are now being presented in a very loose, "breezy" way—in journal articles, in conferences, and in books. All too often one simply takes computations at face value. This is spectacularly against the evidence of my own experience. I would much rather that at talks and in referee reports, the possibility of such error were seriously examined.*

(Donoho 2010)

Otherwise

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely* ***advertising*** *of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

(Buckheit and Donoho 1995).

Towards reproducibility

- structured, commented code
- code review
- tests
- version control
- records of discussion and decisions
- all analysis by code.

Put otherwise, You can't do data science in a GUI

This is too much, what about baby steps?

With a joke.

This is too much, reprise

Reproducible and Collaborative Statistical Data Science.

- Autumn 2015, UC Berkeley.
- ~40 students
- undergraduates, some graduates, mainly statistics
- neuroimaging as example scientific problem
- large open-ended project
- projects had to be fully reproducible.

(Millman et al. 2018),
https://www.frontiersin.org/articles/10.3389/fnins.2018.00727/full

The projects

- https://github.com/berkeley-stat159
- A simple project
- An heroic project

Why?

- massive increase in efficiency, reduction in error

Why not?

- I don't make many errors
- The errors I make aren't very important

How shall this be done?

- Factors in Berkeley.
- What we need here.

Is this the end?

Yes, it's the end of the talk.

All material for these slides at https://github.com/matthew-brett/open-science-seminar

Handout at https://github.com/matthew-brett/open-science-seminar/blob/master/data_confuse_deceive_handout.pdf