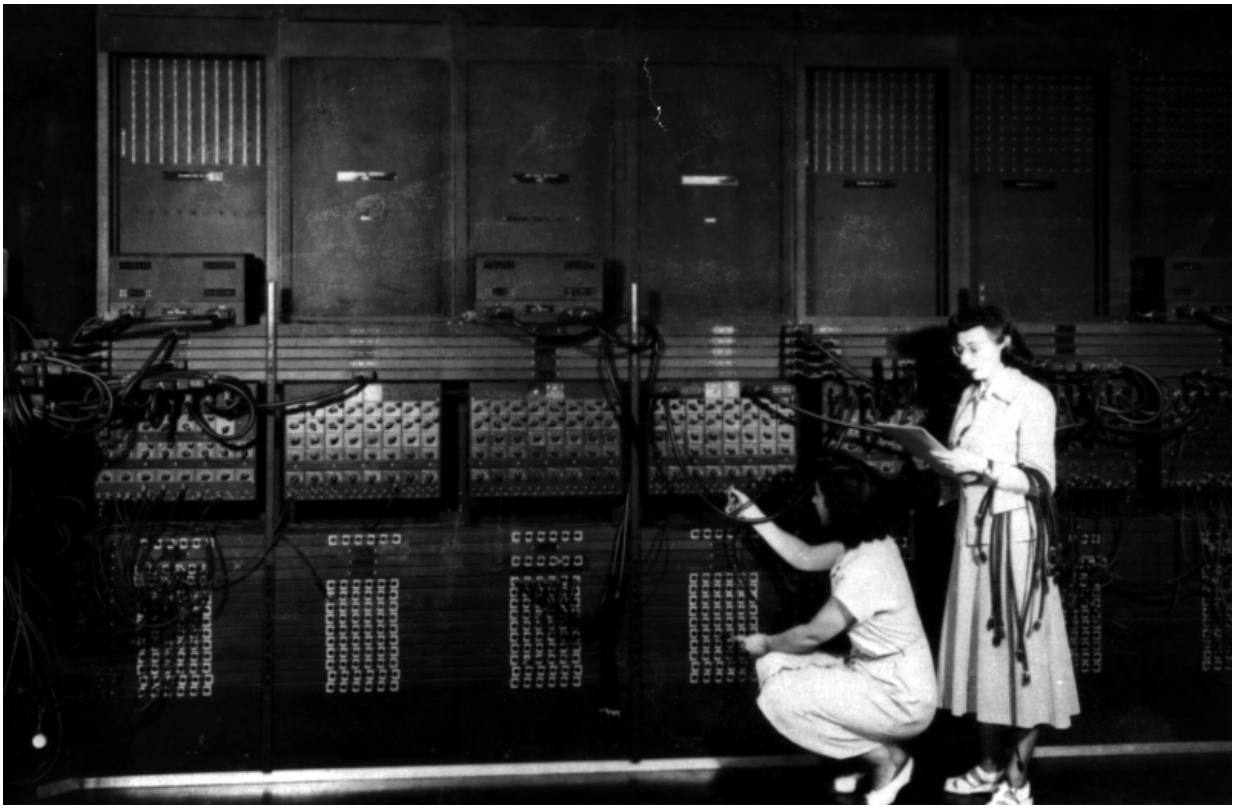
# How computers broke science – and what we can do to fix it

November 9 2015, by Ben Marwick



Computer… or black box for data? Credit: US Army

Reproducibility is one of the cornerstones of science. Made popular by British scientist Robert Boyle in the 1660s, the idea is that a discovery should be reproducible before being accepted as scientific knowledge.

In essence, you should be able to produce the same results I did if you follow the method I describe when announcing my discovery in a scholarly publication. For example, if researchers can reproduce the effectiveness of a new drug at treating a disease, that's a good sign it could work for all sufferers of the disease. If not, we're left wondering what accident or mistake produced the original favorable result, and would doubt the drug's usefulness.

For most of the history of science, researchers have reported their methods in a way that enabled independent reproduction of their results. But, since the introduction of the personal computer – and the point-and-click software programs that have evolved to make it more user-friendly – reproducibility of much research has become questionable, if not impossible. Too much of the research process is now shrouded by the opaque use of computers that many researchers have come to depend on. This makes it almost impossible for an outsider to recreate their results.

Recently, several groups have proposed similar solutions to this problem. Together they would break scientific data out of the black box of unrecorded computer manipulations so independent readers can again critically assess and reproduce results. Researchers, the public, and science itself would benefit.

## Computers wrangle the data, but also obscure it

Statistician Victoria Stodden has described the unique place personal computers hold in the history of science. They're not just an instrument – like a telescope or microscope – that enables new research. The computer is revolutionary in a different way; it's a tiny factory for producing all kinds of new "scopes" to see new patterns in scientific data.

It's hard to find a modern researcher who works without a computer,

even in fields that aren't intensely quantitative. Ecologists use computers to simulate the effect of disasters on animal populations. Biologists use computers to search massive amounts of DNA data. Astronomers use computers to control vast arrays of telescopes, and then process the collected data. Oceanographers use computers to combine data from satellites, ships and buoys to predict global climates. Social scientists use computers to discover and predict the effects of policy or to analyze interview transcripts. Computers help researchers in almost every discipline identify what's interesting within their data.

Computers also tend to be personal instruments. We typically have exclusive use of our own, and the files and folders it contains are generally considered a private space, hidden from public view. Preparing data, analyzing it, visualizing the results – these are tasks done on the computer, in private. Only at the very end of the pipeline comes a publicly visible journal article summarizing all the private tasks.

The problem is that most modern science is so complicated, and most journal articles so brief, it's impossible for the article to include details of many important methods and decisions made by the researcher as he analyzed his data on his computer. How, then, can another researcher judge the reliability of the results, or reproduce the analysis?

## How much transparency do scientists owe?

Stanford statisticians Jonathan Buckheit and David Donoho described this issue as early as 1995, when the personal computer was still a fairly new idea.

*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.*

They make a radical claim. It means all those private files on our personal computers, and the private analysis tasks we do as we work toward preparing for publication should be made public along with the journal article.

This would be a huge change in the way scientists work. We'd need to prepare from the start for everything we do on the computer to eventually be made available for others to see. For many researchers, that's an overwhelming thought. Victoria Stodden has found the biggest objection to sharing files is the time it takes to prepare them by writing documentation and cleaning them up. The second biggest concern is the risk of not receiving credit for the files if someone else uses them.

## A new toolbox to enhance reproducibility

Recently, several different groups of scientists have converged on recommendations for tools and methods to make it easier to keep track of files and analyses done on computers. These groups include biologists, ecologists, nuclear engineers, neuroscientists, economists and political scientists. Manifesto-like papers lay out their recommendations. When researchers from such different fields converge on a common course of action, it's a sign a major watershed in doing science might be under way.

One major recommendation: minimize and replace point-and-click procedures during data analysis as much as possible by using scripts that contain instructions for the computer to carry out. This solves the problem of recording ephemeral mouse movements that leave few traces, are difficult to communicate to other people, and hard to automate. They're common during data cleaning and organizing tasks using a spreadsheet program like Microsoft Excel. A script, on the other hand, contains unambiguous instructions that can be read by its author far into the future (when the specific details have been forgotten) and by

other researchers. It can also be included within a journal article, since they aren't big files. And scripts can easily be adapted to automate research tasks, saving time and reducing the potential for human error.

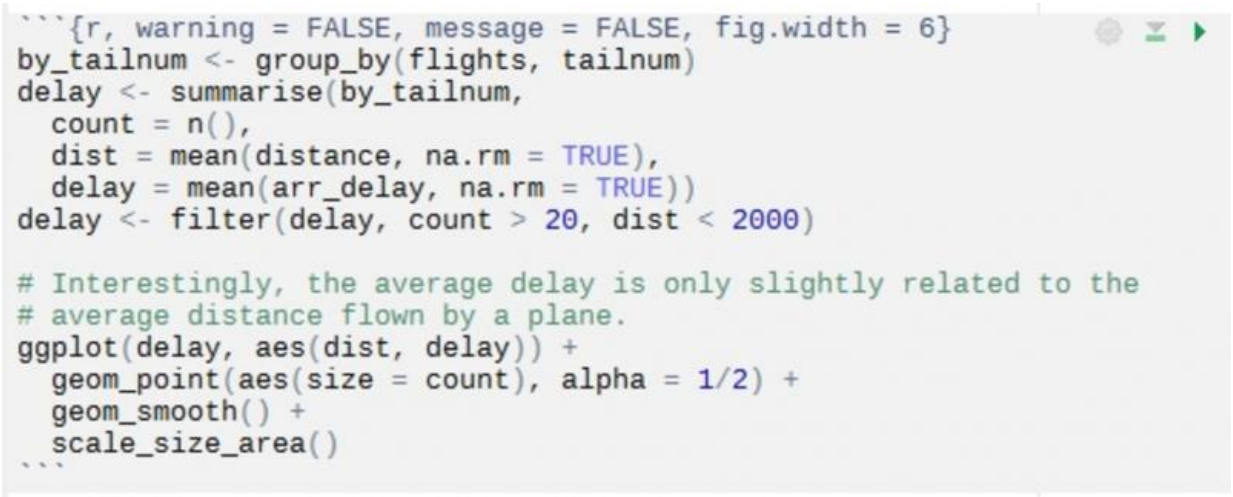
We can see examples of this in microbiology, ecology, political science and archaeology. Instead of mousing around menus and buttons, manually editing cells in a spreadsheet and dragging files between several different software programs to obtain results, these researchers wrote scripts. Their scripts automate the movement of files, the cleaning of the data, the statistical analysis, and the creation of graphs, figures and tables. This saves a lot of time when checking the analysis and redoing it to explore different options. And by looking at the code in the script file, which becomes part of the publication, anyone can see the exact steps that produced the published results.

Other recommendations include the use of common, nonproprietary file formats for storing files (such as CSV, or comma separated variables, for tables of data) and simple rubrics for systematically organizing files into folders to make it easy for others to understand how the information is structured. They recommend free software that is available for all computer systems (eg. Windows, Mac, and Linux) for analyzing and visualizing data (such as R and Python). For collaboration, they recommend a free program called Git, that helps to track changes when many people are editing the same document.

Currently, these are the tools and methods of the avant-garde, and many midcareer and senior researchers have only a vague awareness of them. But many undergraduates are learning them now. Many graduate students, seeing personal advantages to getting organized, using open formats, free software and streamlined collaboration, are seeking out training and tools from volunteer organizations such as Software Carpentry, Data Carpentry and rOpenSci to fill the gaps in their formal training. My university recently created an eScience Institute, where we

help researchers adopt these recommendations. Our institute is part of a bigger movement that includes similar institutes at Berkeley and New York University.

As students learning these skills graduate and progress into positions of influence, we'll see these standards become the new normal in science. Scholarly journals will require code and data files to accompany publications. Funding agencies will require they be placed in publicly accessible online repositories.



```r
```{r, warning = FALSE, message = FALSE, fig.width = 6}
by_tailnum <- group_by(flights, tailnum)
delay <- summarise(by_tailnum,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE))
delay <- filter(delay, count > 20, dist < 2000)

# Interestingly, the average delay is only slightly related to the
# average distance flown by a plane.
ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  geom_smooth() +
  scale_size_area()
```
```

Example of a script used to analyze data. Credit: Author provided

## Open formats and free software are a win/win

This change in the way researchers use computers will be beneficial for public engagement with science. As researchers become more comfortable sharing more of their files and methods, members of the public will have much better access to scientific research. For example, a high school teacher will be able to show students raw data from a

recently published discovery and walk the students through the main parts of the analysis, because all of these files will be available with the journal article.

Similarly, as researchers increasingly use free software, members of the public will be able to use the same software to remix and extend results published in journal articles. Currently many researchers use expensive commercial software programs, the cost of which makes them inaccessible to people outside of universities or large corporations.

Of course, the personal computer is not the sole cause of problems with reproducibility in science. Poor experimental design, inappropriate statistical methods, a highly competitive research environment and the high value placed on novelty and publication in high-profile journals are all to blame.

What's unique about the role of the computer is that we have a solution to the problem. We have clear recommendations for mature tools and well-tested methods borrowed from computer science research to improve the reproducibility of research done by any kind of scientist on a computer. With a small investment of time to learn these tools, we can help restore this cornerstone of science.



Source: The Conversation