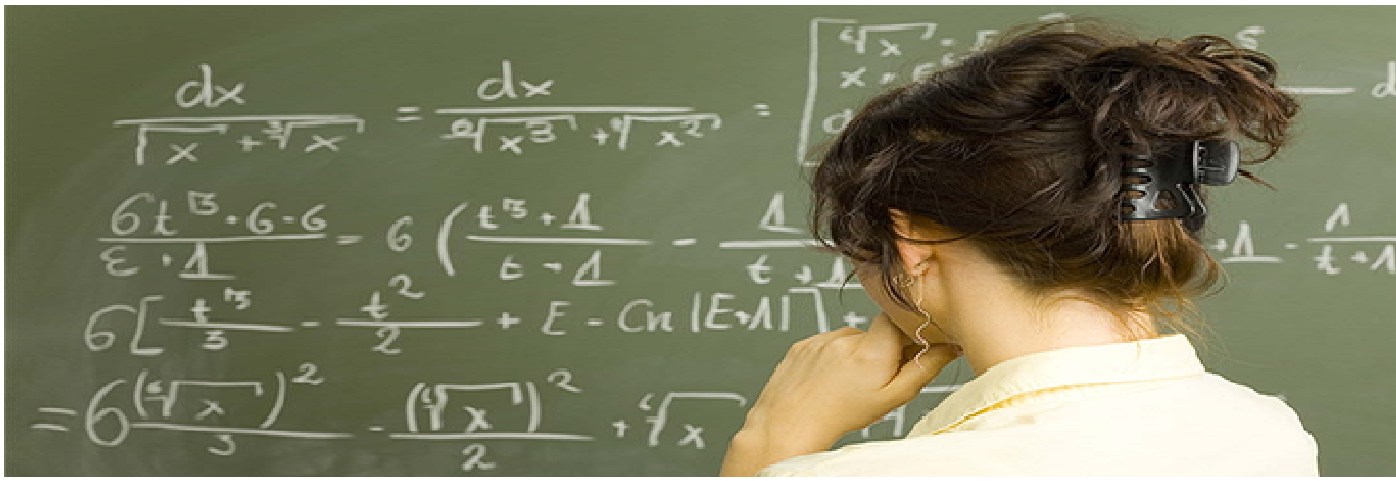
In Data Science, Take Nothing on Faith

Adopting a rigid regimen is required for vetting reproducible computational findings



By James Kobielus | Published May 23, 2014

Taking nothing on faith is a core tenet of true science. Don't even take the word of Nobel Prize–winning scientists on faith. If seeking bedrock confidence in the empirical truths scientists have revealed through their work, that work must be vetted—the assumptions, the methodology, the instrumentation, the data, the logic, and so on—in its entirety. In addition, one would ideally have to reproduce the scientist's findings independently to determine whether they hold up.

Of course, there is no practical need to dig up Galileo's telescope to have utter confidence in his findings. His cosmic discoveries—for example, sunspots, moons of Jupiter, phases of Venus—have been reproduced countless times over the past four centuries by anyone with a powerful enough instrument at their disposal. By contrast, no one has ever provided serious empirical confirmation of the celestial fantasies of his time that Galileo challenged.

If you're a scientist, reproducing somebody else's revolutionary findings is not as career making as, say, announcing those findings in the first place. But science's pioneers recognize that it must happen for their discoveries to join the cumulative foundation of human knowledge. That recognition is why the best of them thoroughly document everything they do, providing a concrete road map for others to follow in their methodological steps.

Providing the means for vetting findings

Computational science cannot claim to be true science if it eschews reproducibility. Fortunately, the data-centric nature of computational science tends to organically generate a detailed documentation trail for others to vet. However, that trail doesn't necessarily ensure ironclad reproducibility, for several steps. For starters, some actions that computational scientists take in their research may not be automated; hence, they're not amenable to automatic logging.

Of those steps that are automated, many may not be logged at every level of detail necessary for later replication. Also, even if every process were logged in detail, the log data may be so voluminous that some of it may later be purged so that limited and costly storage resources can be reallocated to fresh data. And even if it were possible to save exabytes of this data in perpetuity, subsequent researchers may not find it worth their while to dig it all up to re-correlate and reanalyze it all over again.

Nevertheless, computational scientists who need to gain their colleagues' confidence that their discoveries are indeed reproducible should leave a sufficient audit trail—regardless of whether others choose to take them up on that challenge. In this regard, the following list of specific data and analytic governance rules for computational scientists to follow to ensure reproducibility was an interesting find. The rules are quoted here in their entirety from the source article,[1] and the full article offers an in-depth discussion of each:

1. For every result, keep track of how it was produced.
2. Avoid manual data manipulation steps.
3. Archive the exact versions of all external programs used.
4. Version control all custom scripts.
5. Record all intermediate results, when possible, in standardized formats.
6. For analyses that include randomness, note underlying random seeds.
7. Always store raw data behind plots.
8. Generate hierarchical analysis output, allowing layers of increasing detail to be inspected.
9. Connect textual statements to underlying results.
10. Provide public access to scripts, runs, and results.

The authors also spell out exactly why computational scientists should observe these rules scrupulously: "Making reproducibility of your work by peers a realistic possibility sends a strong signal of quality, trustworthiness, and transparency. This could increase the quality and speed of the reviewing process on your work, the chances of your work getting published, and the chances of your work being taken further and cited by other researchers after publication."[2]

Atoning for reproducibility transgressions

But the authors are realistic about this imperative. They also say that amid publication pressures and deadlines, there can be the need to make a trade-off between the ideals of reproducibility and the need to get the research out while it is still relevant.

If one places his or her faith in science, this last statement is equivalent to the tenet of original sin. Professional scientists hope to make a living, and they are always under pressure to publish or perish. So they may transgress when they need to, recognizing that they also may someday need to atone for their occasional dalliance with a necessary evil.

Editor's note: *This article by James Kobielus, big data evangelist and senior program director for product marketing in big data analytics at IBM and editor in chief, IBM Data magazine, is published in association with the Big Data and Enterprise Architecture Conference, June 11–13, 2014, in New York City. In addition to Kobielus, this conference features Bill Inmon, president and CTO, Inmon Consulting, and Krish Krishnan, president and CEO at Sixth Sense Advisors, Inc. Data Management Forum sponsors the conference.*

1,**2** "Ten Simple Rules for Reproducible Computational Research," by Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig, PLOS Computational Biology, October 2013.

Related articles

- Going Beyond Data Science Toward an Analytics Ecosystem: Part 3
- Going Beyond Data Science Toward an Analytics Ecosystem: Part 2
- Going Beyond Data Science Toward an Analytics Ecosystem: Part 1
- It Really Is All About the Data, and Then Some
- Three Data Categories Likely Missing in Your Data Warehouse