SCIENCE · DATA SCIENCE 101 · RESOURCES

RULES FOR CREATING REPRODUCIBLE RESULTS IN DATA ENCE

NDREW TAIT · JULY 3, 2017

In recent years' evidence has been mounting that points to a crisis [illegible] results of scientific research. Reviews of papers in [illegible] and cancer biology found that only 40% and 10%, respectively, of the results, could be reproduced.

Nature published the results of a survey of researchers in 2016 that reported:

- 52% of researchers think there is a significant reproducibility crisis
- 70% of scientists have tried but failed to reproduce another scientist's experiments

In 2013, a team of researchers published a paper describing ten rules for reproducible computational research. These rules, if followed, should lead to more replicable results.

All data science is research. Just because it's not published in an academic paper doesn't alter the fact that we are attempting to draw insights from a jumbled mass of data. Hence, the ten rules in the paper should be of interest to any data scientist doing internal analyses.

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

It's important to know the provenance of your results. Knowing how you went from the raw data to the conclusion allows you to:

- defend the results
- update the results if errors are found
- reproduce the results when data is updated
- submit your results for audit

If you use a programming language (R, Python, Julia, F#, etc) to script your analyses then the path taken should be clear—as long as you avoid any manual steps. Using "point and click" tools (such as Excel) makes it harder to track your steps as you'd need to describe a set of manual activities—which are difficult to both document and re-enact.

RULE #2—AVOID MANUAL DATA MANIPULATION STEPS

There may be a temptation to open data files in an editor and manually clean up a couple of formatting errors or remove an outlier. Also, modern operating systems make it easy to cut and paste been applications. However, the temptation to short-cut your scripting should be resisted. Manual data manipulation is *hidden* manipulation.

RULE #3—ARCHIVE THE EXACT VERSIONS OF ALL EXTERNAL PROGRAMS USED

Ideally, you would set up a virtual machine with all the software used to run your scripts. This allows you to snapshot your analysis ecosystem—making replication of your results trivial.

However, this is not always realistic. For example, if you are using a cloud service, or running your analyses on a big data cluster, it can be hard to circumscribe your entire environment for archiving. Also, the use of commercial tools might make it difficult to share such an environment with others.

At the very least you need to document the edition and version of all the software used—including the operating system. Minor changes to software can impact results.

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

A version control system, such as Git, should be used to track versions of your scripts. You should tag (snapshot) multiple scripts and reference that tag in any results you produce. If you then decide to change your scripts later, as you surely will, it will be possible to go back in time and obtain the exact scripts that were used to produce a given result.

RULE #5—RECORD ALL INTERMEDIATE RESULTS, WHEN POSSIBLE IN STANDARDIZED FORMATS

If you've adhered to Rule #1 it should be possible to recreate any results from the raw data. However, while this might be theoretically possibly, it may be practically limiting. Problems may include:

- lack of resources to run results from scratch (e.g. if considerable cluster computing resources were used)
- lack of licenses for some of the tools, if commercial tools were used
- insufficient technical ability to use some of the tools

In these cases, it can be useful to start from a derived data set that is a few steps downstream from the raw data. Keeping these intermediate datasets (in CSV format, for example), provides more options to build on the analysis and can make it easier to identify where a problematic result when wrong—as there's no need to redo everything.

RULE #6—FOR ANALYSES THAT INCLUDE RANDOMNESS, NOTE UNDERLYING RANDOM SEEDS

One thing that data scientists often fail to do is set the seed values for their analysis. This makes it impossible to exactly recreate machine learning studies. Many machine learning algorithms include a stochastic element and, while robust results might be *statistically* reproducible, there is nothing to compare with the warm glow of matching the *exact* numbers produced by someone else.

If you are using scripts and source code control your seed values can be set in your scripts.

RULE #7—ALWAYS STORE RAW DATA BEHIND PLOTS

If you use a scripting/programming language your charts will often be automatically generated. However, if you are using a tool like Excel to draw your charts, make sure you save the underlying data. This allows the chart to be reproduced, but also allows a more detailed review of the data behind it.

RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

As data scientists, our job is to summarize the data in some form. That is what drawing insights from data involves.

However, summarizing is also an easy way to misuse data so it's important that interested parties can break out the summary into the individual data points. For each summary result, link to the data used to calculate the summary.

RULE #9—CONNECT TEXTUAL STATEMENTS TO UNDERLYING RESULTS

At the end of the day, the results of data analysis are presented as words. And words are imprecise. The link between conclusions and the analysis can sometimes be difficult to pin down. As the report is often the most influential part of a study it's essential that it can be linked back to the results and, because of Rule #1, all the way back to the raw data.

This can be achieved by adding footnotes to the text that reference files or URLs containing the specific data that led to the observation in the report. If you can't make this link you probably haven't documented all the steps sufficiently.

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

In commercial settings, it may not be appropriate to provide public access to all the data. However, it makes sense to provide access to others in your organization. Cloud-based source code control systems, such as Bitbucket and GitHub, allow the creation of private repositories that can be accessed by any authorized colleagues.

Many eyes improve the quality of analysis, so the more you can share, the better your analyses are likely to be.

TAGS: data science Reproducible Results Science Scientific Method

PREVIOUS POST
HOW TO STOP DATA BREACHES FROM RUINING YOUR BUSINESS

NEXT POST
WINNING WITH DATA SCIENCE, GOLDEN STATE WARRIORS STYLE

THE AUTHOR

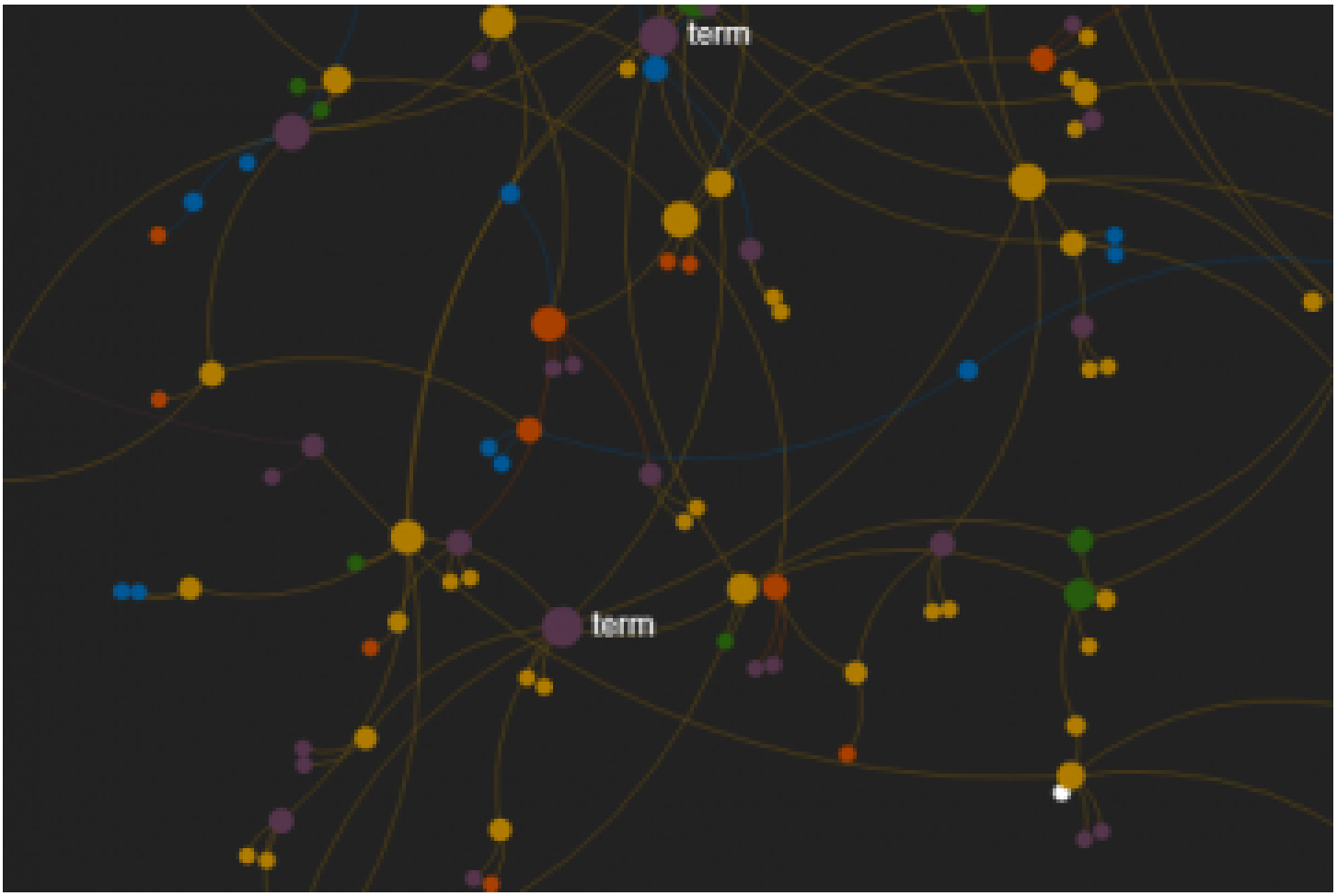
ANDREW TAIT

Andrew Tait is the founder and Chief Technology Officer of Decision Mechanics, a software and consulting firm specializing in the creative use of technology to improve organizational decision-making. He has designed commercial, off-the-shelf, solutions for strategic planning, performance improvement, and conflict management. His work in this area has led to numerous consulting and training relationships with major commercial and government organizations. Andrew has taught for Learning Tree International since 2010 and specializes in the Big Data, .NET, Mobile & Web Development curriculums

RELATED POSTS

DATA SCIENCE · MACHINE LEARNING

SECURING COMPETITIVE ADVANTAGE WITH MACHINE LEARNING



BI & ANALYTICS · BIG DATA · DATA SCIENCE

FOUR STRATEGIC DIFFERENTIATORS OF AN ENTERPRISE KNOWLEDGE GRAPH