Reproducible Computing @ JSM 2019

Colin Rundel

July 27, 2019

Reproducible Computing

Schedule

Time	Activity
08:30 - 10:15	Literate programming and organization
10:15 - 10:30	:coffee:
10:30 - 12:30	Version control with Git and GitHub
12:30 - 14:00	:fork_and_knife:
14:00 - 15:15	Scaling reproducible projects, make
15:15 - 15:30	:coffee:
15:30 - 17:00	More make, wrapup

Reproducibility:

Who cares?

Science retracts gay marriage paper without agreement of lead author

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months earlier.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results. + Survey incentives misrepresented. + Sponsorship statement false.
- Two Berkeley grad students who attempted to replicate the study quickly discovered that the data must have been faked.
- Methods we'll discuss today can't prevent this, but they can make it easier to discover issues.

Source: http://news.sciencemag.org/policy/2015/05/science-retracts-gay-marriage-paper-without-lead-author-s-consent

Bad spreadsheet merge kills depression paper, quick fix resurrects it

- **Original conclusion:** Lower levels of CSF IL-6 were associated with current depression and with future depression [...].

- **Revised conclusion:** Higher levels of CSF IL-6 and IL-8 were associated with current depression [...].

Source: http://retractionwatch.com/2014/07/01/bad-spreadsheet-merge-kills-depression-paper-quick-fix-resurrects-it/

Divorce study felled by a coding error gets a second chance

- **Original conclusion:** The risk of divorce in a heterosexual marriage increases when the wife falls ill, but not the husband.

- **Corrected conclusion:** Based on the corrected analysis, we conclude that there are not gender differences in the relationship between gender, pooled illness onset, and divorce.

Source: http://retractionwatch.com/2015/09/10/divorce-study-felled-by-a-coding-error-gets-a-second-chance/#more-32151

Divorce study retraction: Editor's note

- "The research environment is fast-paced given the ethos to “publish or perish"."
- "[...] research is becoming increasingly complex, with greater calls for transdisciplinary collaborations, “big data,” and more sophisticated research questions and methods [...] data sets often have multiple files that require merging, change the wording of questions over time, provide incomplete codebooks, and have unclear and sometimes duplicative variables."
- "Given these issues, I would not be surprised if coding errors were fairly common [...]"

Source: http://retractionwatch.com/2015/09/10/divorce-study-felled-by-a-coding-error-gets-a-second-chance/#more-32151

One in five genetics papers contains errors thanks to Microsoft Excel

- "Autoformatting in Microsoft Excel has caused many a headache—but now, a new study shows that one in five genetics papers in top scientific journals contains errors from the program, The Washington Post reports. The errors often arose when gene names in a spreadsheet were automatically changed to calendar dates or numerical values."

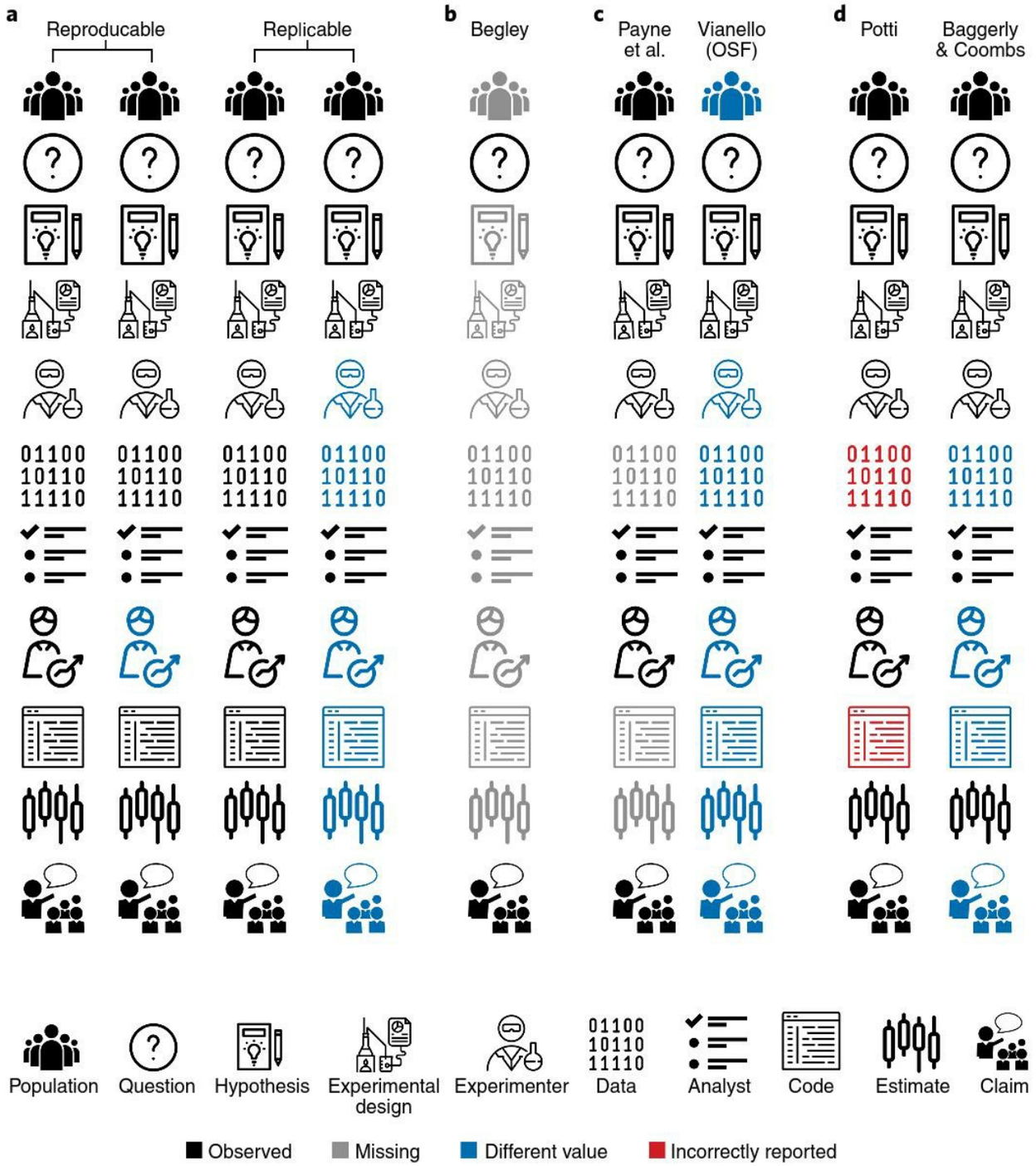
- "For example, one gene called Septin-2 is commonly shortened to SEPT2, but is changed to 2-SEP and stored as the date 2 September 2016 by Excel."

Source: https://www.sciencemag.org/news/2016/08/one-five-genetics-papers-contains-errors-thanks-microsoft-excel

Reproducibility:

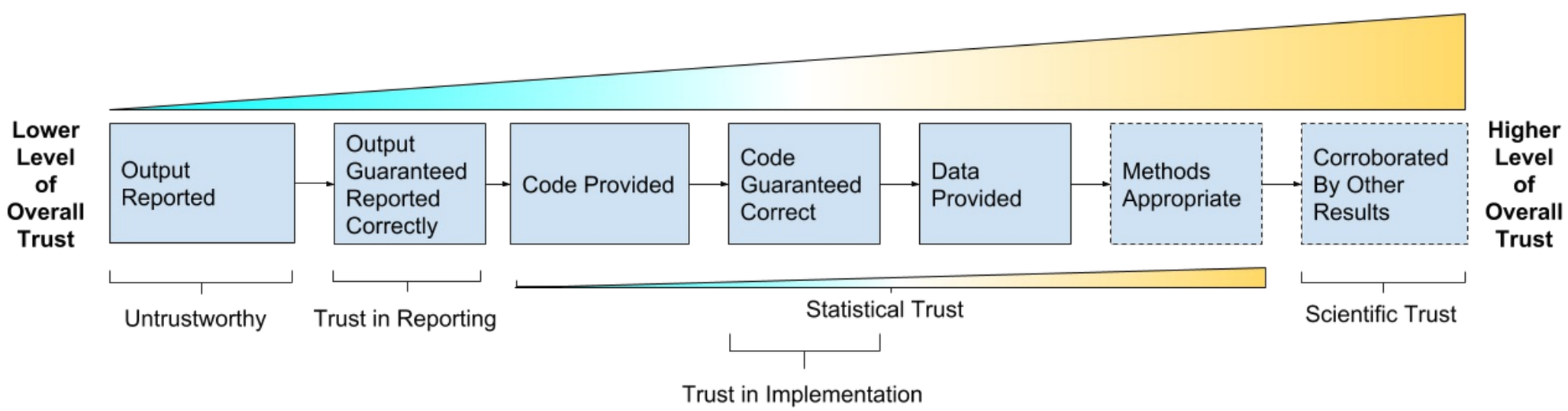
Why should you care?

Reproducible vs Replicable



Source: Patil, Peng, Leek (2019) A visual tool for defining reproducibility and replicability. Nature Human Behaviour

Reproducibility as a trust scale



Source: Gabriel Becker - Keynote - Advanced R Course - May Institute for Computational Proteomics 2019

Think back to every time...

- The results in Table 1 don't seem to correspond to those in Figure 2.
- In what order do I run these scripts?
- Where did we get this data file?
- Why did I omit those samples?
- How did I make that figure?
- "Your script is now giving an error."
- "The attached is similar to the code we used."

Source: Karl Broman - [steps to reproducible research](https://kbroman.org/steps2rr/)

No collaborators?

> Your closest collaborator is you six months ago,
> but you don’t reply to emails.

- Mark Holder

Reproducibility: How?

Reproducibility checklist

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear why it was done? (e.g. how were hyper / tuning parameters chosen?)
- Can the code be used for other data?
- Can you extend the code to do other things?

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

1 Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology, Duke University, Durham, North Carolina, United States of America, **4** Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.
* gvwilson@software-carpentry.org

Author summary

Computers are now essential in all branches of science, but most researchers are never taught the equivalent of basic lab skills for research computing. As a result, data can get lost, analyses can take much longer than necessary, and researchers are limited in how effectively they can work with software and data. Computing workflows need to follow the same practices as lab projects and notebooks, with organized data, documented steps, and the project structured for reproducibility, but researchers new to computing often don't know where to start. This paper presents a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill. These practices, which encompass data management, programming, collaborating with colleagues, organizing projects, tracking work, and writing manuscripts, are drawn from a wide variety of published sources from our daily lives and from our work with volunteer organizations that have delivered workshops to over 11,000 people since 2010.

Ambitious goal + other concerns

We need an environment where

- data, analysis, and results are tightly connected, or better yet, inseparable
- reproducibility is built in
    - the original data remains untouched
    - all data manipulations and analyses are inherently documented
    - documentation is human readable and syntax is minimal

Toolkit

Roadmap

Scriptability \(\rightarrow\) R

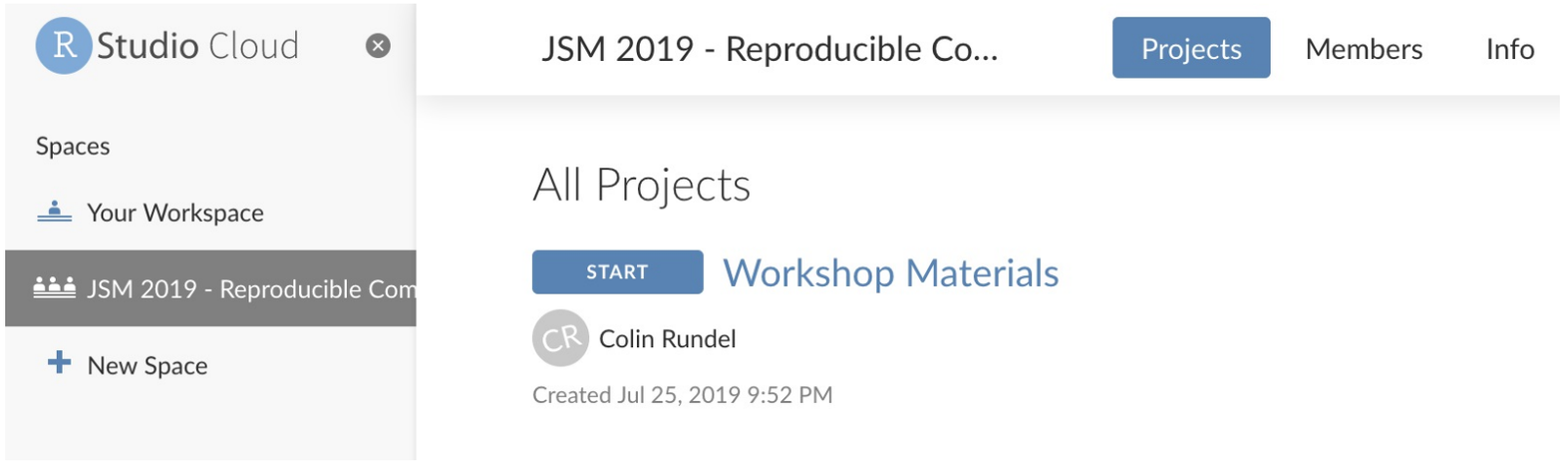
Literate programming \(\rightarrow\) R Markdown

Version control \(\rightarrow\) `git` / GitHub

Scaling and automation \(\rightarrow\) `make`

Computing access

- Go to http://bit.ly/jsm2019-repro-comp
- Log in by creating an Account or using your Google / GitHub credentials.
- Click the Start Button next to the Workshop Materials project



- You should now be inside an RStudio Cloud Session that contains all of the workshop files