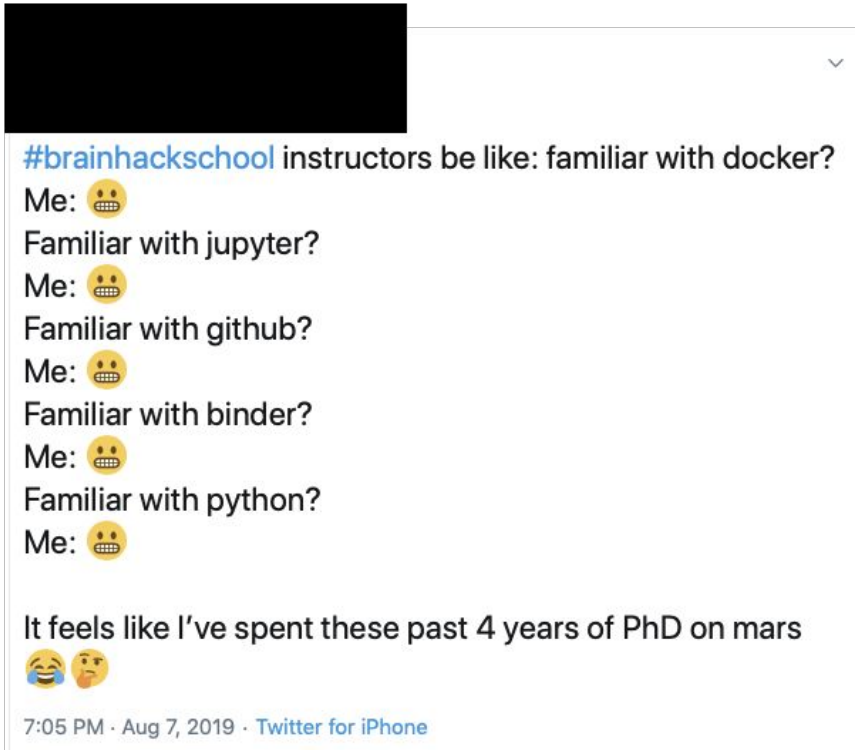
How do you adopt Open and Reproducible Research Practices?

Dr Andrew J. Stewart

Fellow of the Software Sustainability Institute

E: drandrewjstewart@gmail.com
T: @ajstewart_lang
G: ajstewartlang

The world of research is rapidly changing...



Open Research is recognised by the G7…

Focus: Incentives and the researcher ecosystem

Ambition: Foster a research environment in which career advancement takes into account Open Science activities, through incentives and rewards for researchers, and valuing the skills and capabilities in the Open Science workforce.

Recommendations:

At national levels: G7 nations should each engage with research stakeholders to identify and implement enhancements to research evaluation and reward systems that take into consideration the Open Science activities carried out by researchers and research institutions. Topics that could be discussed include:

- Recognizing Open Science practices during evaluation of research funding proposals, and research outcomes;
- Recognizing and rewarding research productivity and impact that reflect open science activities by researchers during career advancement reviews;
- Including credit for service activities such as reviewing, evaluating, and curation and management of research data; and,
- Developing metrics of Open Science practices.

...by the REF...

200. The sub-panels welcome research practice that supports reproducible science and the application of best practice. Examples include registered reports, pre-registration, publication of data sets, experimental materials, analytic code, and use of reporting checklists for publication purposes and those relating to the use of animals in research. These contribute to the evaluation of rigour for submitted outputs. Replication studies may be submitted as outputs and will be evaluated on the extent to which they contribute significant new knowledge, improved methods, or advance theory or practice1.

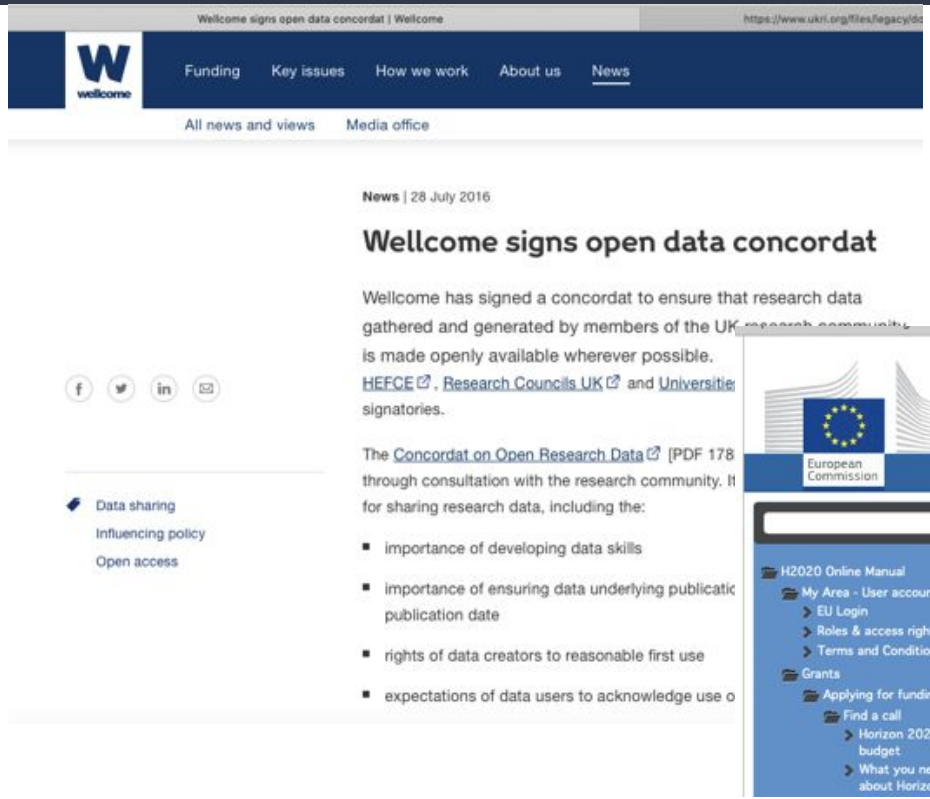
346...
Within the context of the institution's strategy, how the submitting unit is progressing towards an open research environment, including where this goes above and beyond the REF open access policy requirements, and wider activity to encourage the effective sharing and management of research data, as appropriate to the discipline. Consideration of reproducibility should also be included where relevant to the discipline.

…is forming part of University's teaching manifestos…

Teaching with Open Science commitment:
To teach the practices and skills of open research and science in our undergraduate and postgraduate degree programmes

a. Promote open science in our teaching.
b. Design a Research Methods curriculum that teaches skills for open science and uses open science to enhance teaching (for example: teach R and use open data to practice analysis skills).
c. Learn about and adopt open educational practices in our teaching.
d. Produce and promote tools for helping student researchers adopt open practices, including training and guidance suitable to their level of study.
e. Author, share and use open educational resources to promote teaching with open science beyond our School and Institution.
f. Support our colleagues to learn the skills of teaching Open Science.

...and is required by many funders.



Concordat on Open Research Data - Nine Principles

- Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice.
- There are sound reasons why the openness of research data may need to be restricted but any restrictions must be justified and justifiable.
- Open access to research data carries a significant cost, which should be respected by all parties.
- The right of the creators of research data to reasonable first use is recognised.
- Use of others' data should always conform to legal, ethical and regulatory frameworks including appropriate acknowledgement.

https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/

Concordat on Open Research Data – Nine Principles

- Good data management is fundamental to all stages of the research process and should be established at the outset.
- Data curation is vital to make data useful for others and for long-term preservation of data.
- Data supporting publications should be accessible by the publication date and should be in a citeable form.
- Support for the development of appropriate data skills is recognised as a responsibility for all stakeholders.

https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/

TOP GUIDELINES

TRANSPARENCY AND OPENNESS PROMOTION

Transparency, open sharing, and reproducibility are core values of science, but not always part of daily practice. Journals, funders, and societies can increase research reproducibility by adopting the TOP Guidelines.

8 MODULAR STANDARDS

CITATION STANDARDS Cite shared data to incentivize their publication	**DATA TRANSPARENCY** Disclose, require, or verify shared data
ANALYTICAL METHODS TRANSPARENCY Disclose, require, or verify shared code	**RESEARCH MATERIALS TRANSPARENCY** Disclose, require, or verify shared materials
DESIGN AND ANALYSIS TRANSPARENCY Sets standards for research design disclosures	**PREREGISTRATION OF STUDIES** Specification of study details before data collection
PREREGISTRATION OF ANALYSIS PLANS Specification of analytical details before data collection	**REPLICATION** Encourages publication of replication studies

ACROSS 3 TIERS

DISCLOSURE:
The article must disclose whether or not materials are available.

REQUIREMENT:
The article must share materials when possible.

VERIFICATION:
Third party must verify that the standard is being met.

HOW TOP IS IMPLEMENTED

TOP Statements are standardized tools for disclosing research outputs such as datasets.

Open Science Badges signal transparent research.

Registered Reports protect research against biased analysis and publication.

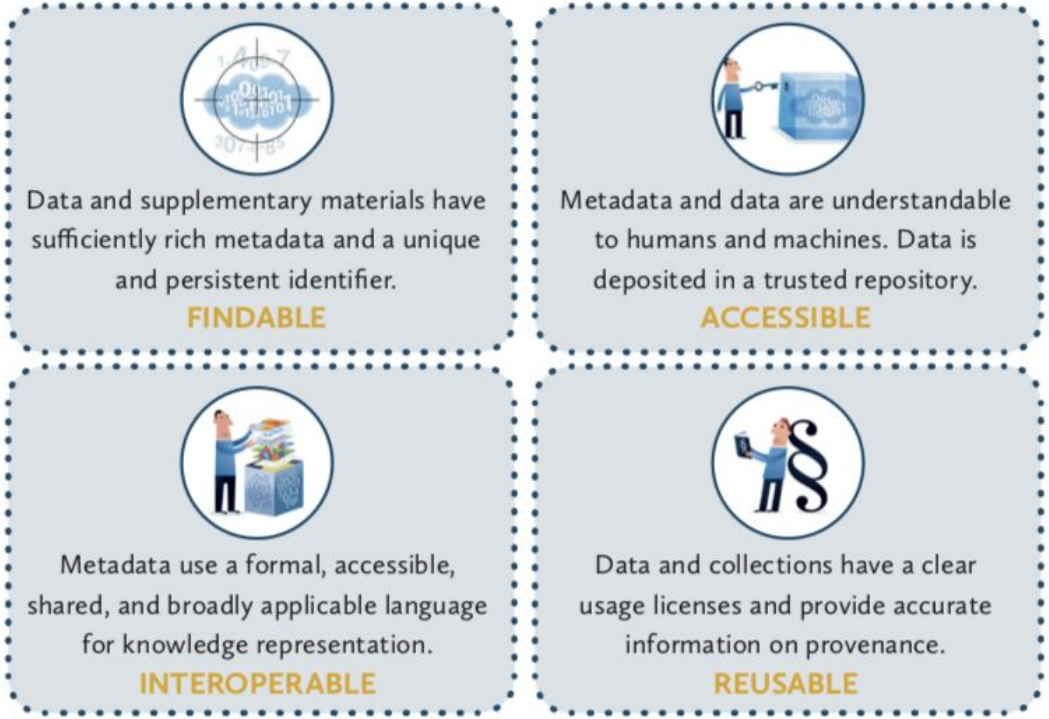
OVER 5,000 JOURNAL SIGNATORIES

LEARN MORE AT COS.IO/TOP

The Center for Open Science is a non-profit organization with the mission of improving openness, integrity, and reproducibility in scientific research.

COS: cos.io | **OSF:** osf.io | **Email:** contact@cos.io | **CenterForOpenScience** | **@OSFramework**

Your data need to be FAIR



https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

And you need to remember to add a license – such as CC BY

Attribution
CC BY

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

View License Deed | View Legal Code

Attribution-ShareAlike
CC BY-SA

This license lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms. This license is often compared to "copyleft" free and open source software licenses. All new works based on yours will carry the same license, so any derivatives will also allow commercial use. This is the license used by Wikipedia, and is recommended for materials that would benefit from incorporating content from Wikipedia and similarly licensed projects.

View License Deed | View Legal Code

Attribution-NoDerivs
CC BY-ND

This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.

View License Deed | View Legal Code

Attribution-NonCommercial
CC BY-NC

This license lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.

View License Deed | View Legal Code

Attribution-NonCommercial-ShareAlike
CC BY-NC-SA

This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.

View License Deed | View Legal Code

Attribution-NonCommercial-NoDerivs
CC BY-NC-ND

This license is the most restrictive of our six main licenses, only allowing others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially.

View License Deed | View Legal Code

Choosing a License

Which of the following best describes your situation?

I need to work in a community.

Use the **license preferred by the community** you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to **add a license**.

I want it simple and permissive.

The **MIT License** is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, **.NET Core**, and **Rails** use the MIT License.

I care about sharing improvements.

The **GNU GPLv3** also lets people do almost anything they want with your project, *except* distributing closed source versions.

Ansible, **Bash**, and **GIMP** use the GNU GPLv3.

https://choosealicense.com

STM – 2020 is Research Data Year

STM is working with publishers and other partners to boost effective sharing of research data:

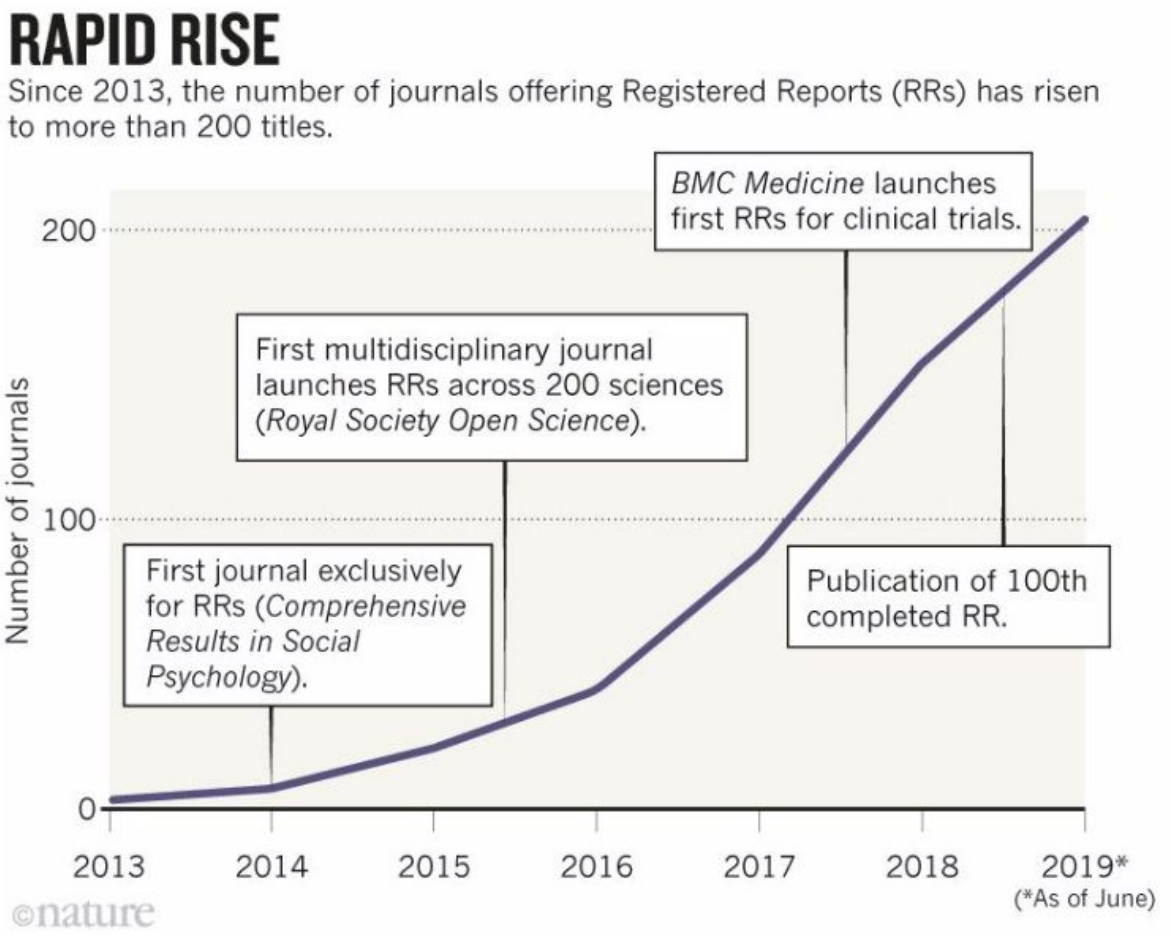
- SHARE: Increase the number of journals with data policies and articles with Data Availability Statements (DAS)
- LINK: Increase the number of journals that deposit the data links to the SCHOLIX framework
- CITE: Increase the citations to datasets along the Joint Declaration of Data Citation Principles

Before data collection

- Specify your hypotheses and analysis plan.
- Pre-register your hypotheses and analysis plan at osf.io
- Consider data simulation so that you can write your analysis script **before** you have your real data.
- Consider submitting as a **registered report** - currently more than 200 journals now support this route. This involves acceptance in principle before you have even started collecting your data - so you are **guaranteed** a publication regardless of what you find.

Registered Reports

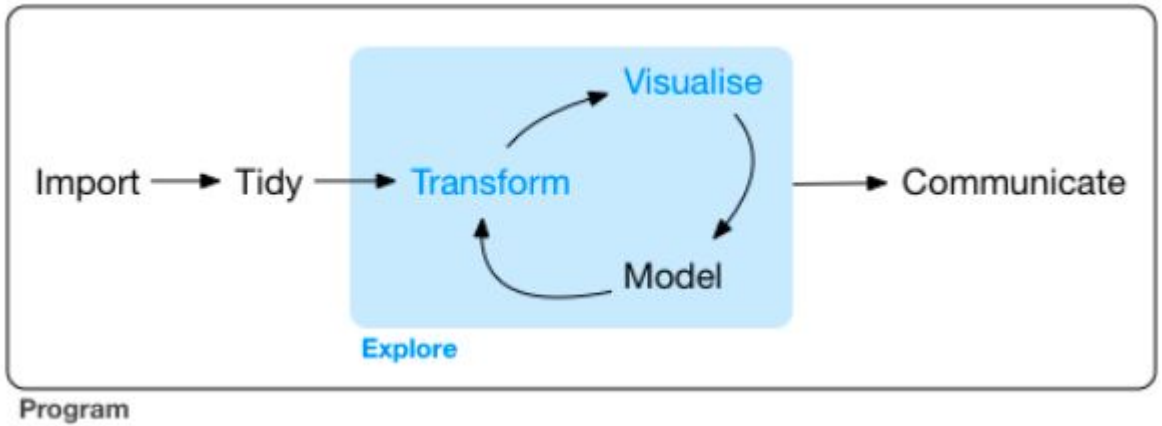
https://www.cos.io/rr



https://www.nature.com/articles/d41586-019-02674-6

After data collection

You need to use analysis software that allows for open sharing and reproducibility of the entire data wrangling/analysis/write-up workflow.



Hadley Wickham and Garrett Grolemund, R4DS

Lukas Schlögl @LukasSchlogl · Sep 7

Just priceless. An estimated 20% of **genetic** research papers contain errors because **Excel** converted some gene names into calendar **dates**.

PLOS ONE PHYLOGENY/FLICKR (CC BY 2.0)

One in five genetics papers contains errors thanks to Microsoft Excel

By **Jessica Boddy** | Aug. 29, 2016 , 1:45 PM

Andrew Whitby @EconAndrew · Sep 7

This is top shelf trolling, because thanks to Excel "1 in 5" genetics papers contain errors in gene names. sciencemag.org/news/2016/08/o ... twitter.com/msexcel/status...

Show this thread

186 4.6K 7.3K

Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander
BBC News

20 April 2013 Share

This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

Codifying your Analysis Workflow

You can’t use proprietary (closed) software like SPSS, GraphPad, MATLAB etc.

You need to use open source analysis software such as R, Octave or Python.

You need to think about your entire analysis pipeline from data importing, data wrangling, visualisation, statistical modelling, and report generation.

Scripts are fine for small tasks but what about the multiple tasks needed in a research project? How do you connect them together so the entire workflow can be reproduced?

What happens when you want to share your workflow with colleagues elsewhere or in another lab (with a different infrastructure)?

"It worked on my machine!" - but it’s not good when it doesn't work on your collaborators' machines (or on your new machine!)

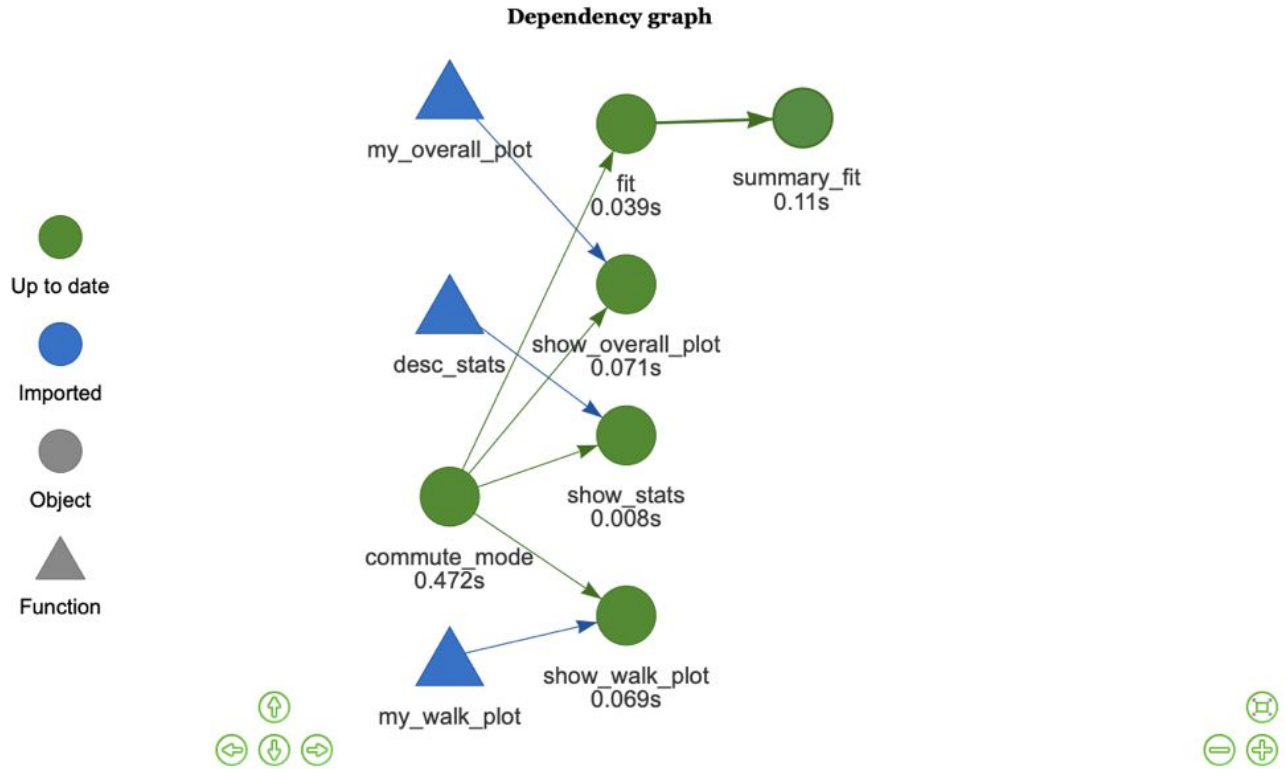
The {drake} package for R

"Data analysis can be slow. A round of scientific computation can take several minutes, hours, or even days to complete. After it finishes, if you update your code or data, your hard-earned results may no longer be valid. How much of that valuable output can you keep, and how much do you need to update? How much runtime must you endure all over again?

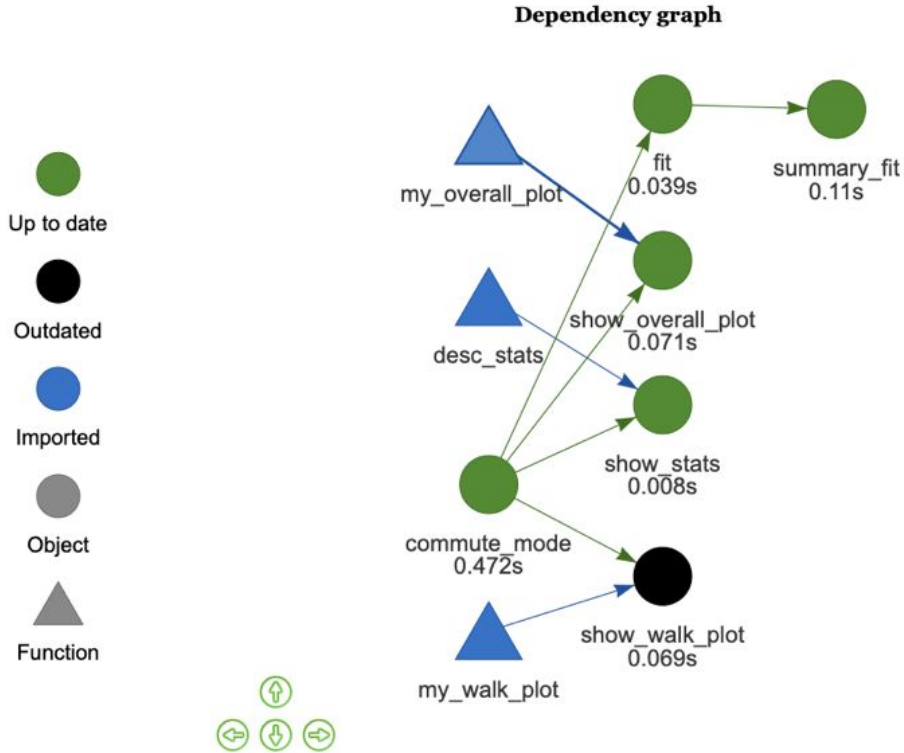
For projects in R, the drake package can help. It analyzes your workflow, skips steps with up-to-date results, and orchestrates the rest with optional distributed computing. At the end, drake provides evidence that your results match the underlying code and data, which increases your ability to trust your research."

https://ropenscilabs.github.io/drake-manual/index.html

Dependency graphs in {drake}



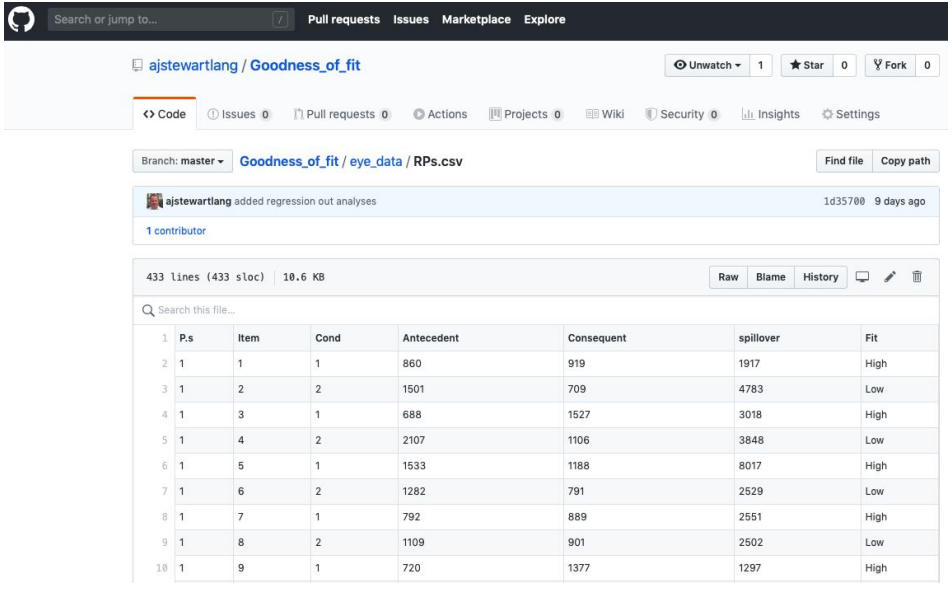
Dependency graphs in {drake}



Adopt literate programming principles

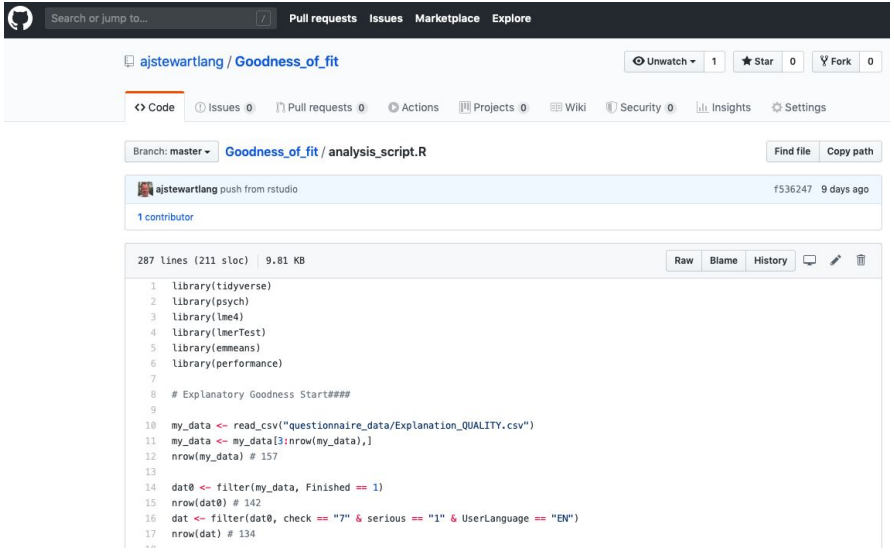
- Write analysis code that's easy for others (and future you!) to understand - have lots of commenting - and break your code down into manageable chunks or separate scripts/functions controlled by a master script.
- Later in this unit we'll explore one example of literate programming using R Markdown where you can generate documents containing a blend of narrative, code, and output 'knitted' together (in lots of possible document formats - html, pdf, doc etc.)
- The workshops in this unit are all written in R Markdown - you can even examine the source code behind each of these by looking at the .Rmd file in each workshop repository.

Sharing your data and code (and add a licence!)



	P.s	Item	Cond	Antecedent	Consequent	spillover	Fit
2	1	1	1	860	919	1917	High
3	1	2	2	1501	709	4783	Low
4	1	3	1	688	1527	3018	High
5	1	4	2	2107	1106	3848	Low
6	1	5	1	1533	1188	8017	High
7	1	6	2	1282	791	2529	Low
8	1	7	1	792	889	2551	High
9	1	8	2	1109	901	2502	Low
10	1	9	1	720	1377	1297	High

GitHub



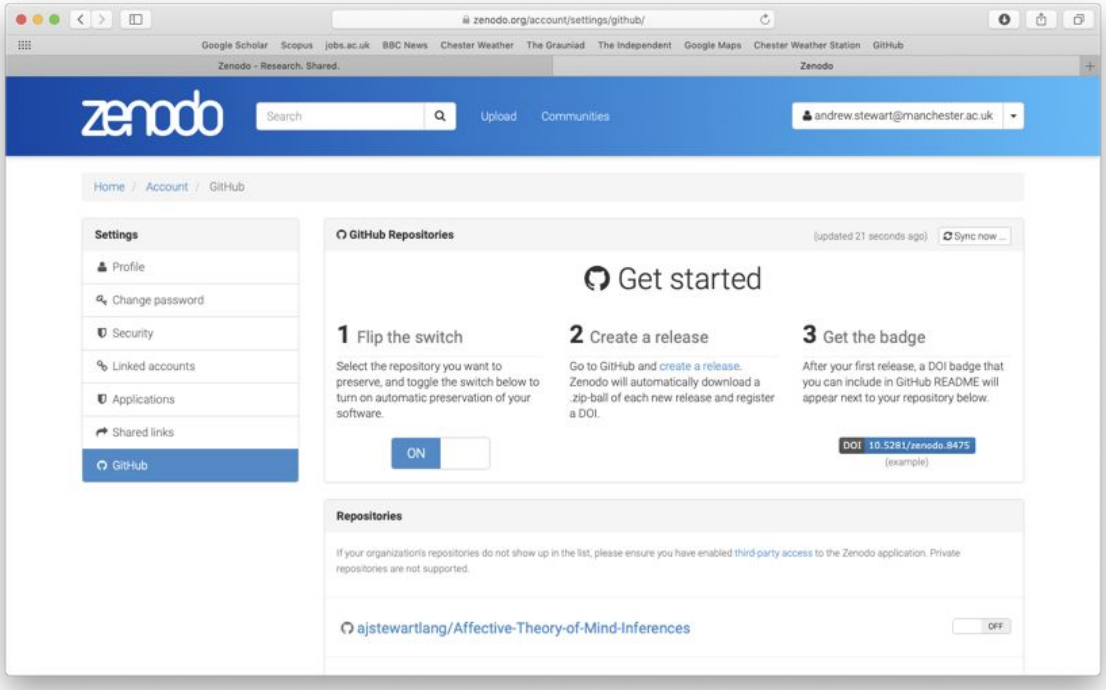

```
library(tidyverse)
library(psych)
library(lme4)
library(lmerTest)
library(emmeans)
library(performance)

# Explanatory Goodness Start####

my_data <- read_csv("questionnaire_data/Explanation_QUALITY.csv")
my_data <- my_data[3:nrow(my_data),]
nrow(my_data) # 157

dat0 <- filter(my_data, Finished == 1)
nrow(dat0) # 142
dat <- filter(dat0, check == "7" & serious == "1" & UserLanguage == "EN")
nrow(dat) # 134
```

Make your data and code citeable - preserve it with a doi on Zenodo



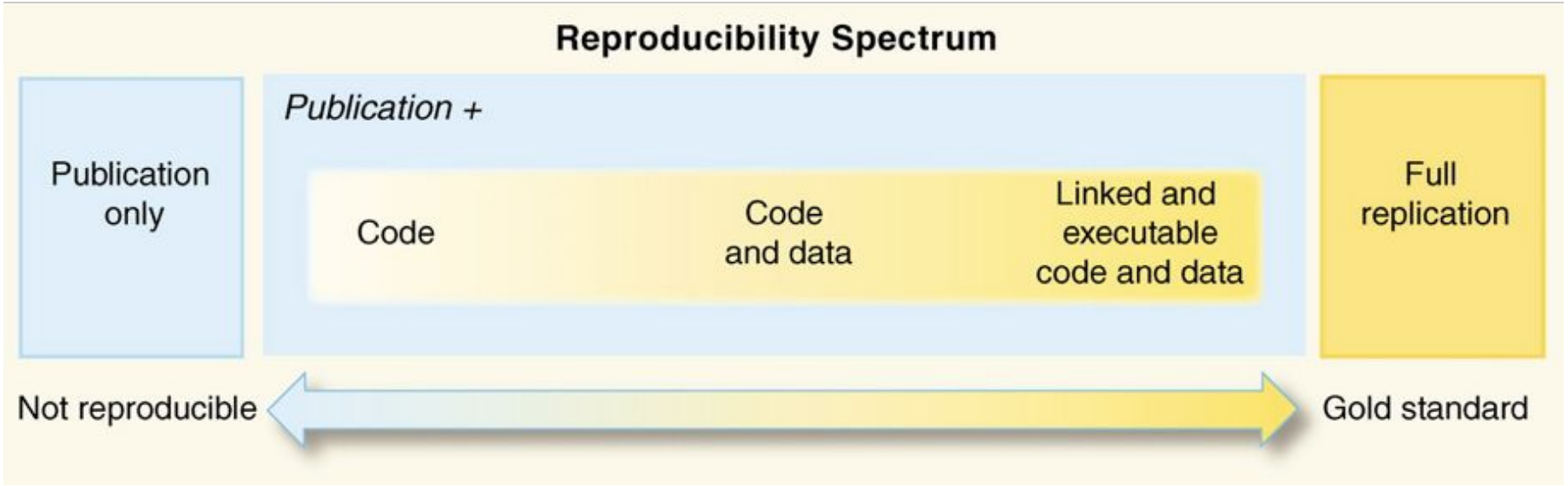
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



Why do we need to reproduce the computational environment?

Quite often analysis code ‘breaks’ - often in one of two ways:

Code that worked previously now doesn’t - maybe a function in an R package was updated (e.g., `lsmeans` became `emmeans` so old code using `lsmeans` wouldn’t now run).

Code that worked previously still works - but produces a slightly different result or now throws a warning where it didn’t previousy (e.g., convergence/singular fit warnings in `lme4` version 1.1-19 vs. version 1.1-20).

When R moved from version 3.5 to 3.6, the way in which random numbers were generated using the `sample()` function changed - so even with the same randomisation seed, *different* random numbers were produced.

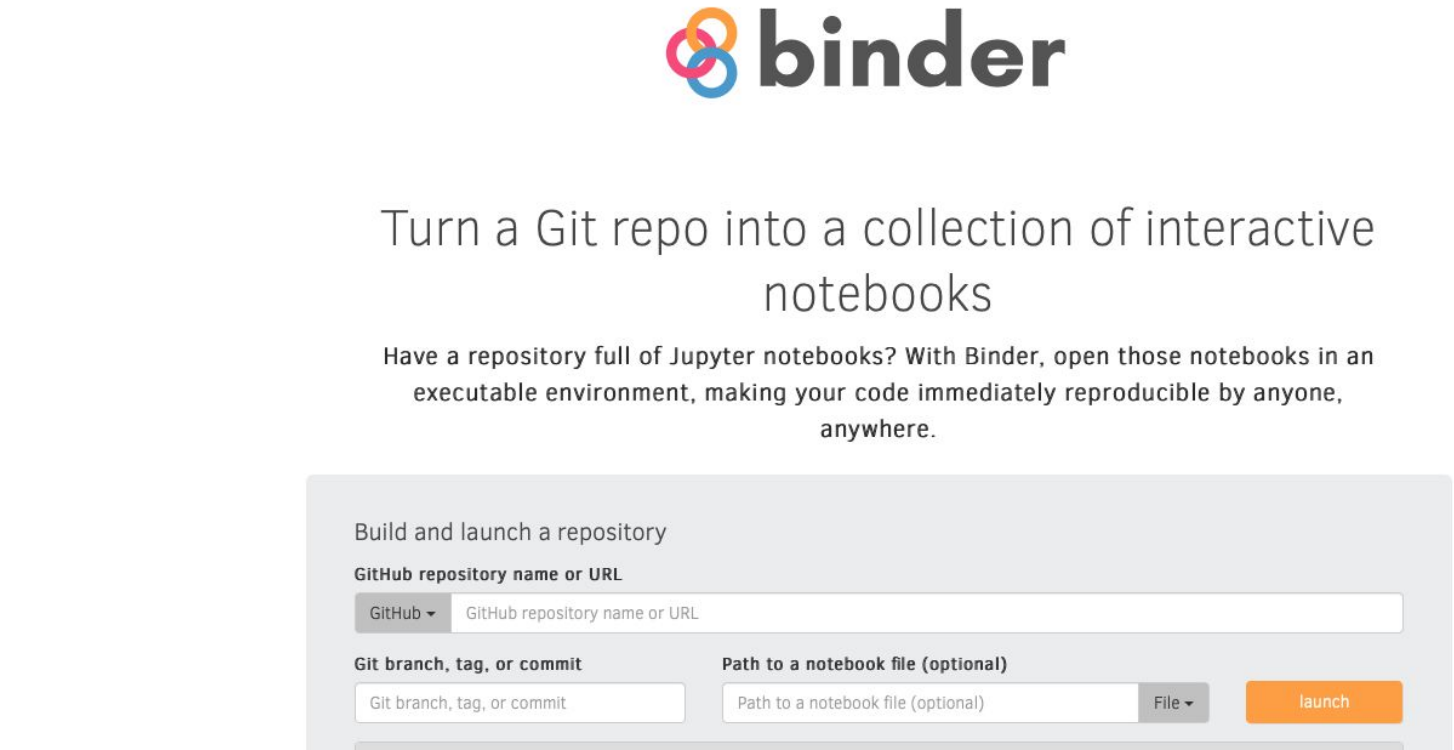
Much Twitter confusion ensued!

Capturing your local computational environment

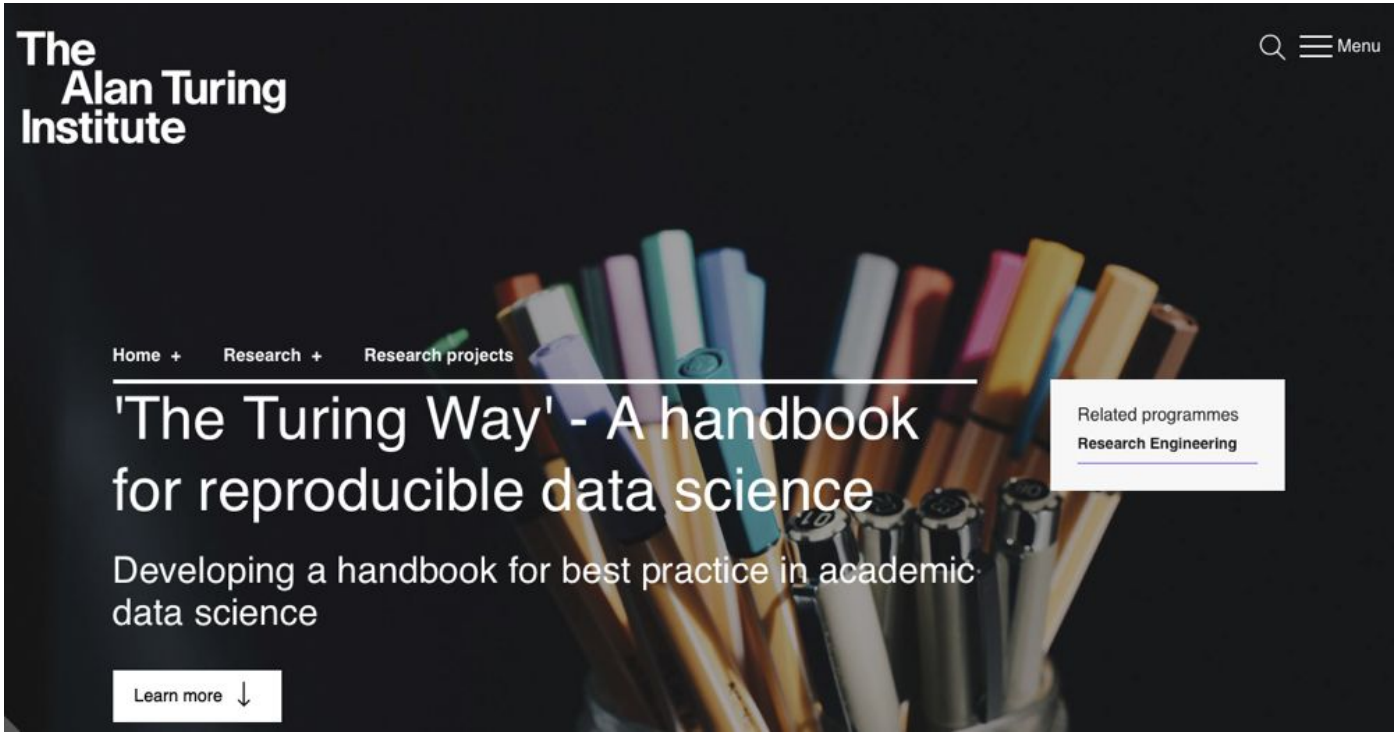
- You need to capture the versions of the different your software packages (plus their dependencies incl. system-level ones).

- This may sound trivial but trying running some old analysis code and be amazed at how many things now don’t work as they once did!

Binder to the rescue!



The Turing Way



https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science

Summary

- Open data and open code are the future - remember to make your data FAIR and add a license to both your data and code.
- Pre-registration allows you to capture your predictions - time-stamped at a point in time so when you come to write up your work you can use the pre-reg as evidence that you really did make your predictions before data collection commenced.
- In many cases conducting research in a reproducible manner is easy - it requires a bit of planning and organisation up-front, but the pay-off is huge.
- Not only will others be able to reproduce your results, but so will you at some future point in time.
- Working in an open and reproducible manner also makes large-scale collaborations easier - with the extra computational skills that you acquire, you'll be a more effective researcher.