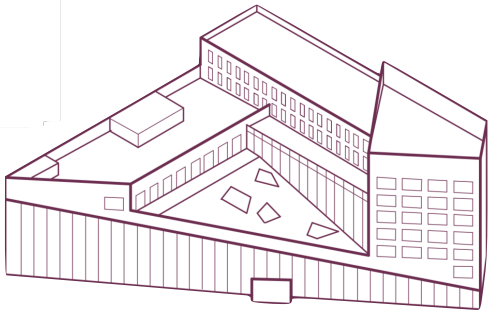
Best Practice in Reproducible Data Science

Matt Forshaw
Lecturer, Newcastle University.
Data Skills Lead, The Alan Turing Institute

@mattforshaw



Data Science North East
21st January 2020

The Alan Turing Institute

What do you understand about
the meaning of “Reproducibility” or “Open Science”?

Discussion

“Open science is at a stage where no-one is quite sure what it is, but they think it’s a good idea.”

–Martyn Rittman, Editor at MDPI

Science

This article is more than **10 months old**

Attempt to replicate major social scientific findings of past decade fails

Scientists and the design of experiments under scrutiny after a major project fails to reproduce results of high profile studies

Hannah Devlin
Science correspondent
@hannahdev

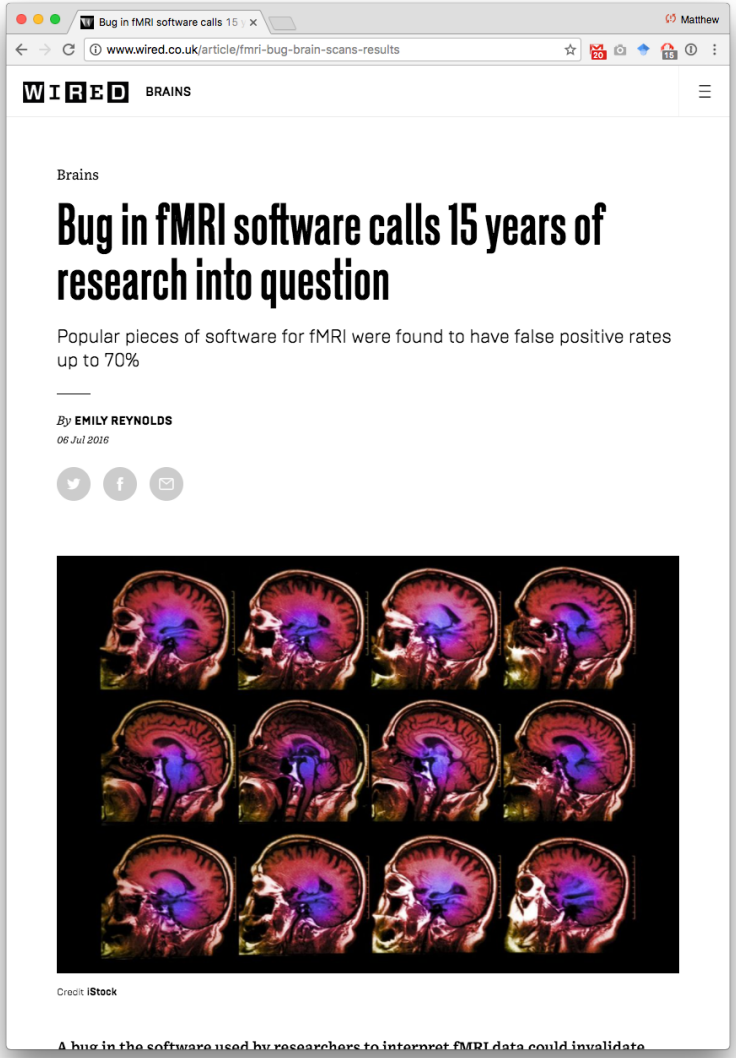
Mon 27 Aug 2018
16.00 BST

▲ One finding which this study was unable to replicate was that people who viewed a picture of Rodin's sculpture The Thinker subsequently reported weaker religious beliefs. Photograph: Alamy

Some of the most high profile findings in social sciences of the past decade do not stand up to replication, a major investigation has found.

The project, which aimed to repeat 21 experiments that had been published in Science or Nature - science's two preeminent journals - found that only 13 of the original findings could be reproduced.

The research, which follows similar efforts in psychology and biomedical



Bug in fMRI software calls 15 ×

Matthew

www.wired.co.uk/article/fmri-bug-brain-scans-results

WIRED BRAINS

Brains

Bug in fMRI software calls 15 years of research into question

Popular pieces of software for fMRI were found to have false positive rates up to 70%

By EMILY REYNOLDS

06 Jul 2016

Credit iStock

A bug in the software used by researchers to interpret fMRI data could invalidate

*"A bug in the software used by researchers to interpret fMRI data could invalidate **fifteen years** worth of neuroscientific research, a paper claims.*

*Three of the most popular pieces of software for fMRI – SPM, FSL and AFNI – were all found to have **false positive rates of up to 70 per cent**. These findings could invalidate "up to **40,000 papers**", researchers claim.*

*"Though fMRI is 25 years old, surprisingly its most common statistical methods have **not been validated** using real data," said Anders Eklund.*

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 201602413. http://www.pnas.org/content/113/28/7900.abstract

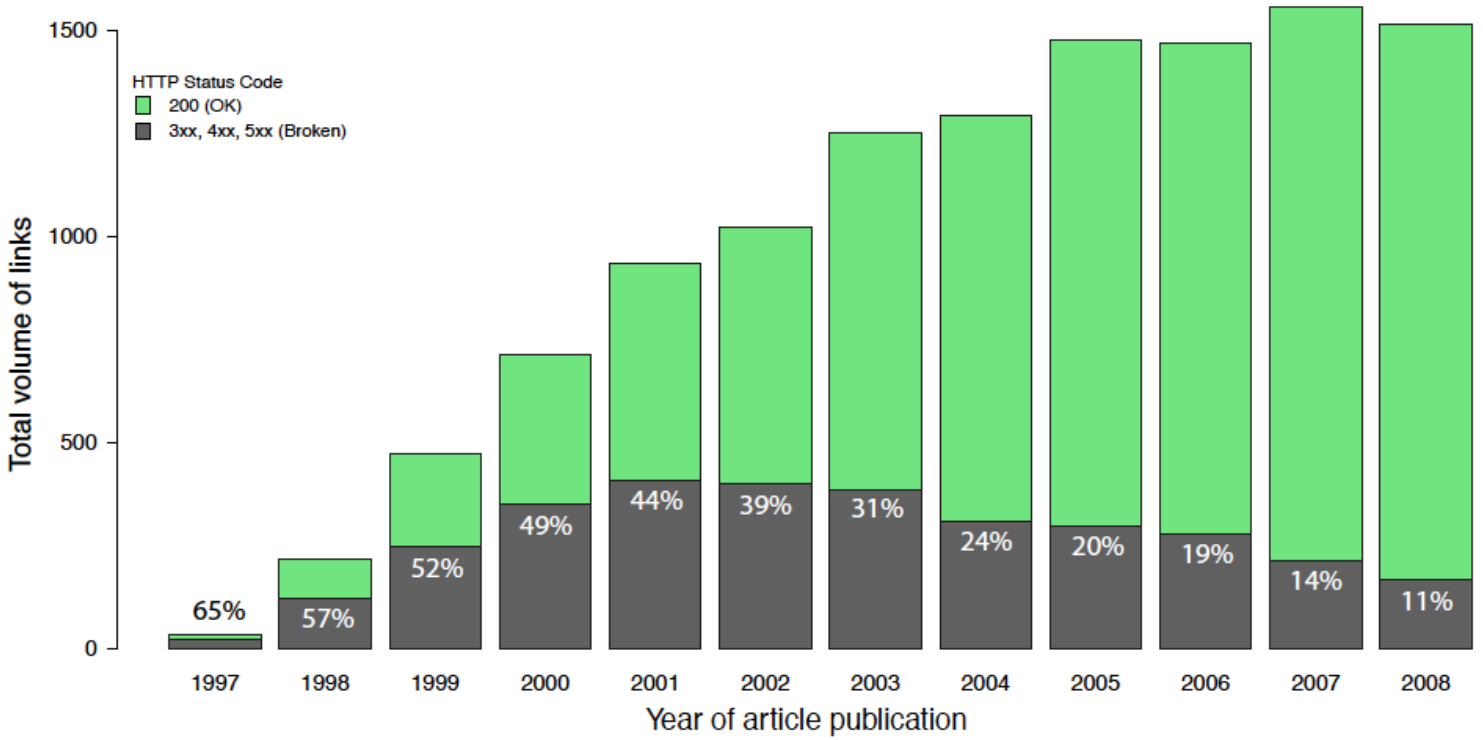


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

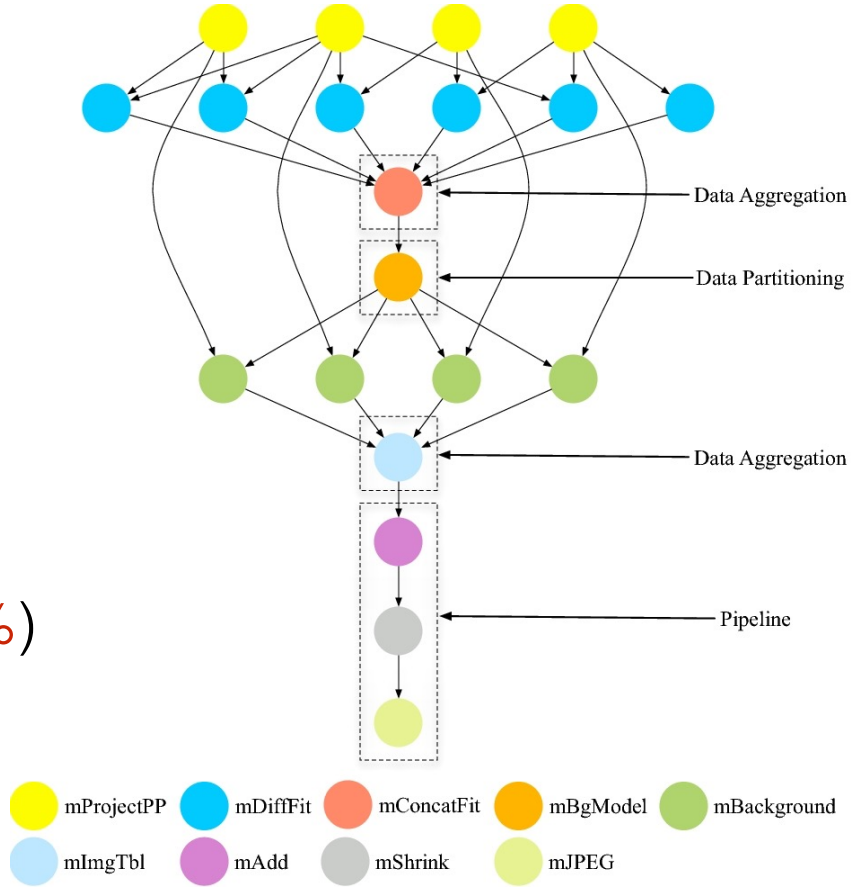
Pepe A, Goodman A, Muench A, Crosas M, Erdmann C (2014) How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE 9(8): e104798. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104798

Zhao, *et al*:

18 of 92 (19.57%)

Mayer, *et al:*

341 of 1443 (23.63%)

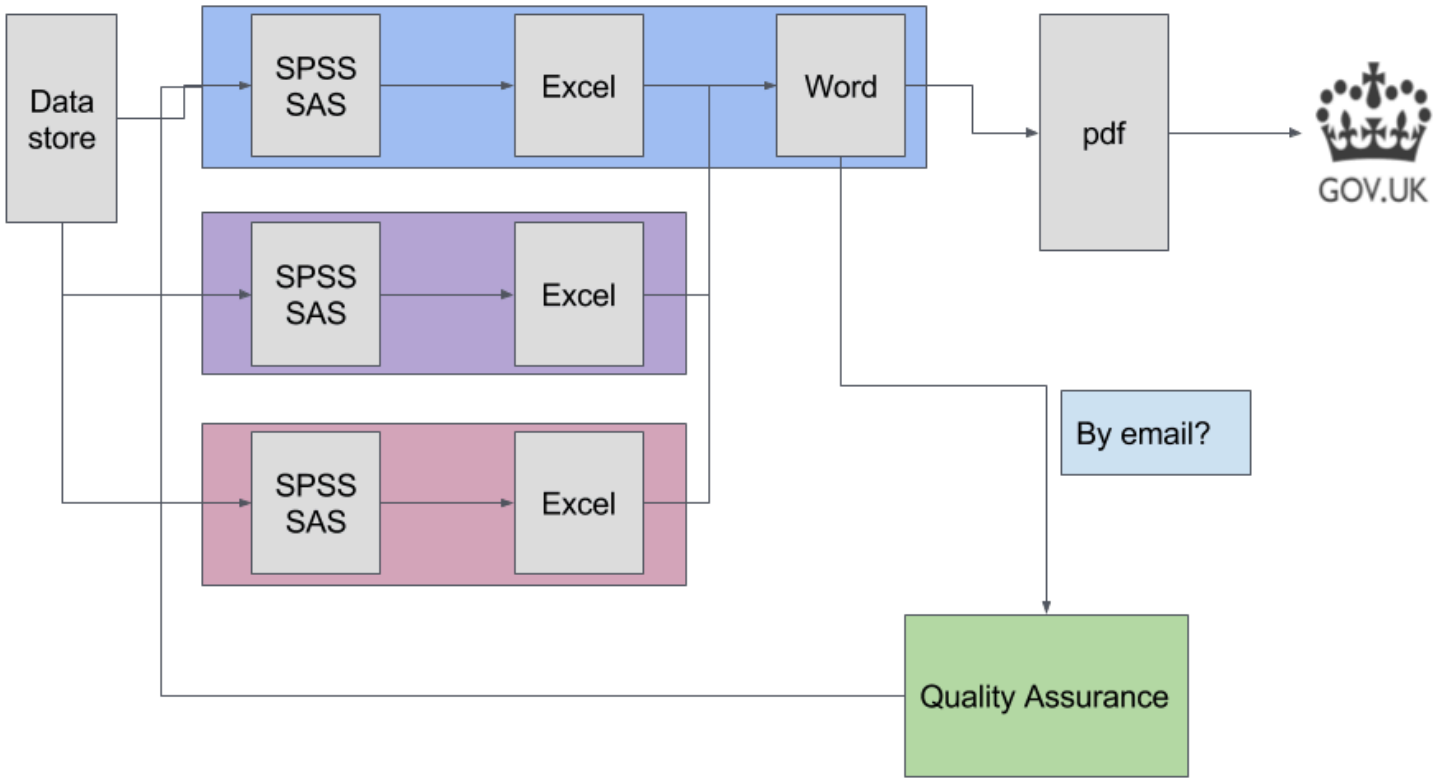


Zhao, Jun, et al. "Why workflows break—Understanding and combating decay in Taverna workflows." E-Science (e-Science), 2012 IEEE 8th International Conference on. IEEE, 2012.
http://ieeexplore.ieee.org/document/6404482/

Mayer, Rudolf, and Andreas Rauber. "A quantitative study on the re-executability of publicly shared scientific workflows." e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, 2015.
http://ieeexplore.ieee.org/abstract/document/7304314/

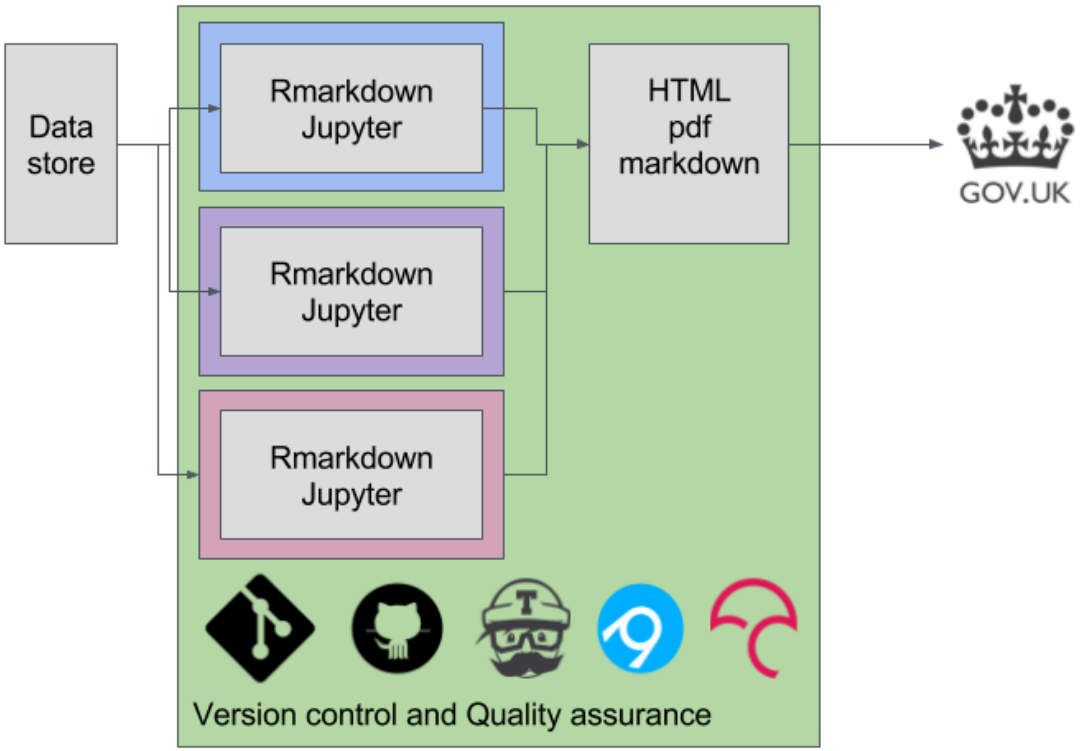

Review code snippet (3 mins)

A case study of gov.uk

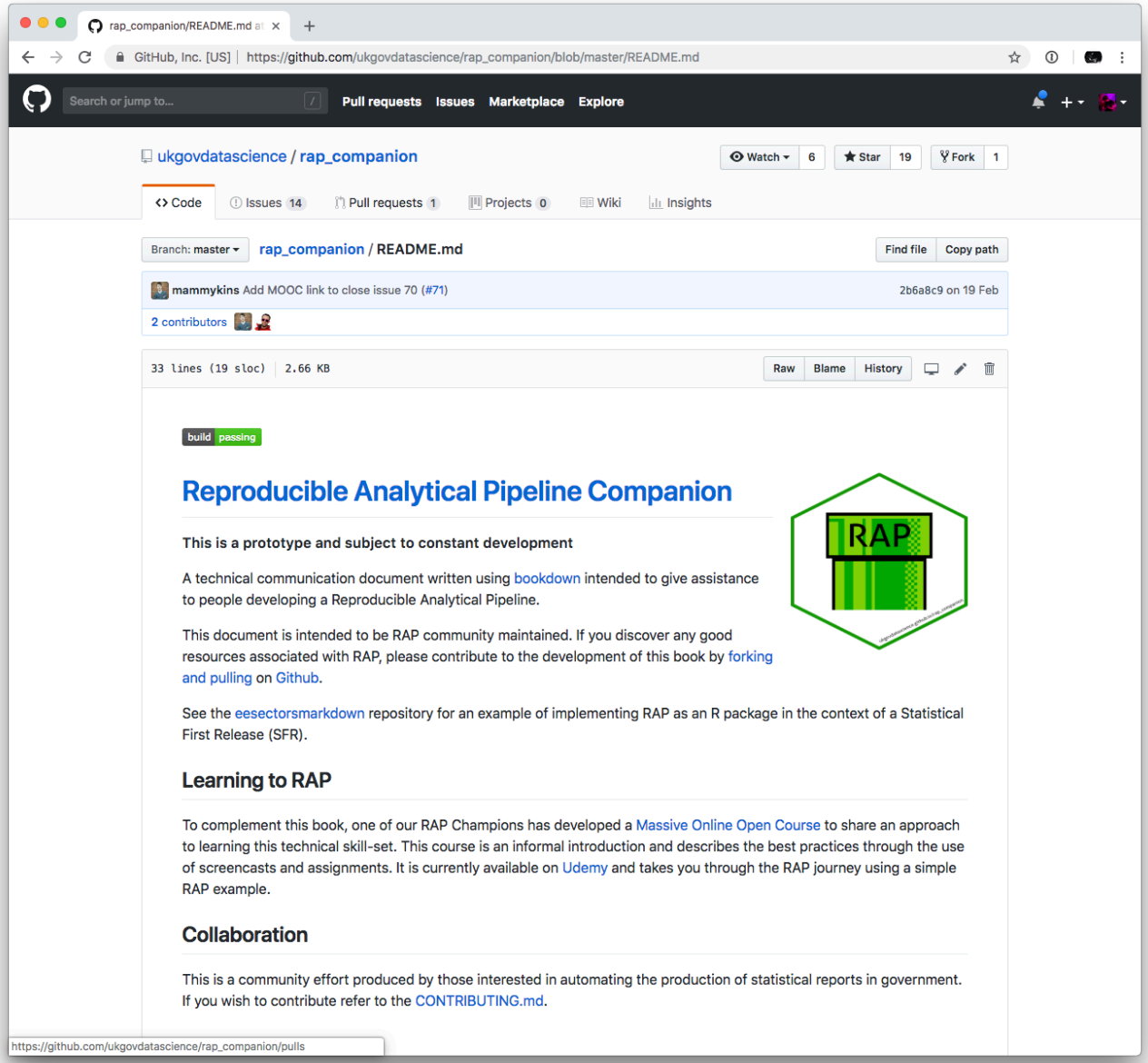


https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process

A case study of gov.uk



https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process

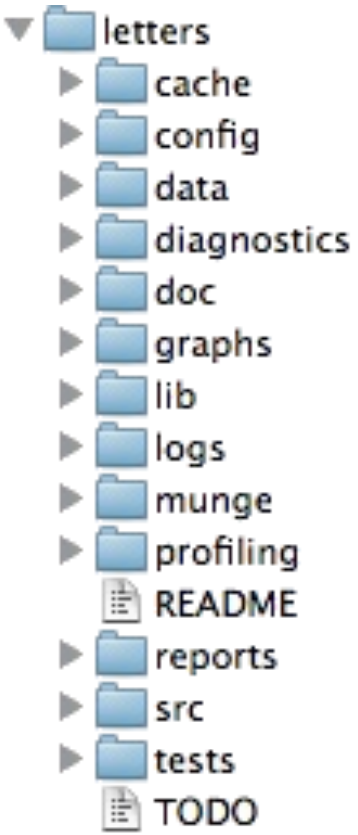


https://github.com/ukgovdatascience/rap_companion


Live Demo

What is ProjectTemplate?

- ProjectTemplate is a system for automating the thoughtless parts of a data analysis project:
  - Organising the files in your project.
  - Loading all the R packages you'll use.
  - Loading all of your data sets into memory.
  - Munging and preprocessing your data into a form that's suitable for analysis.



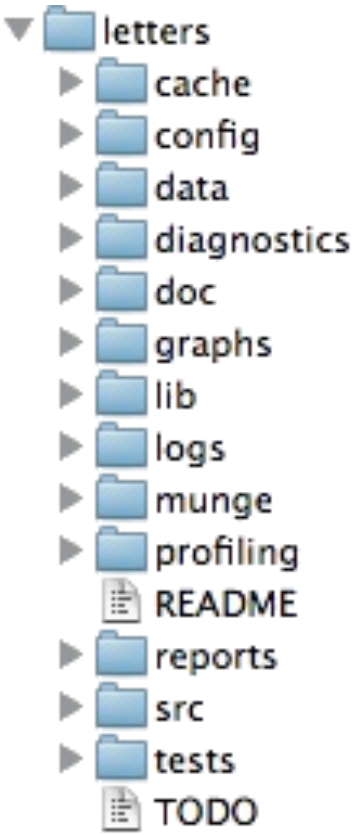
Benefits

1. Documents and automatically install any dependencies used by your analysis.

2. Caches intermediate results

3. Gives you a common structure to locate your files.

4. Avoids drift between code and workspace contents.



Creating a Project

```
library('ProjectTemplate')
create.project('new_project')
```

Loading a Project

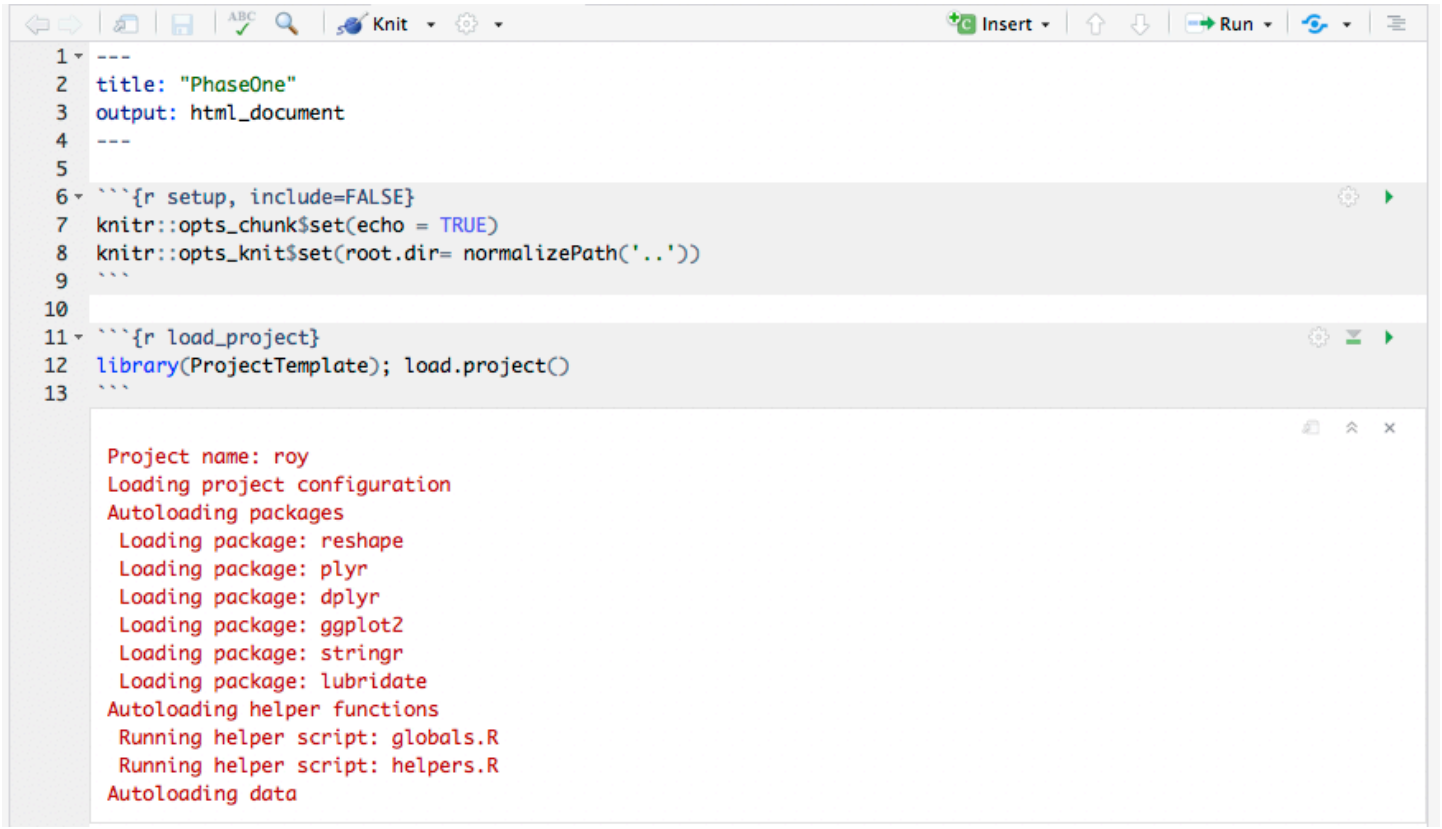
```
library('ProjectTemplate')
setwd('~/projects/new_project')
load.project()
```

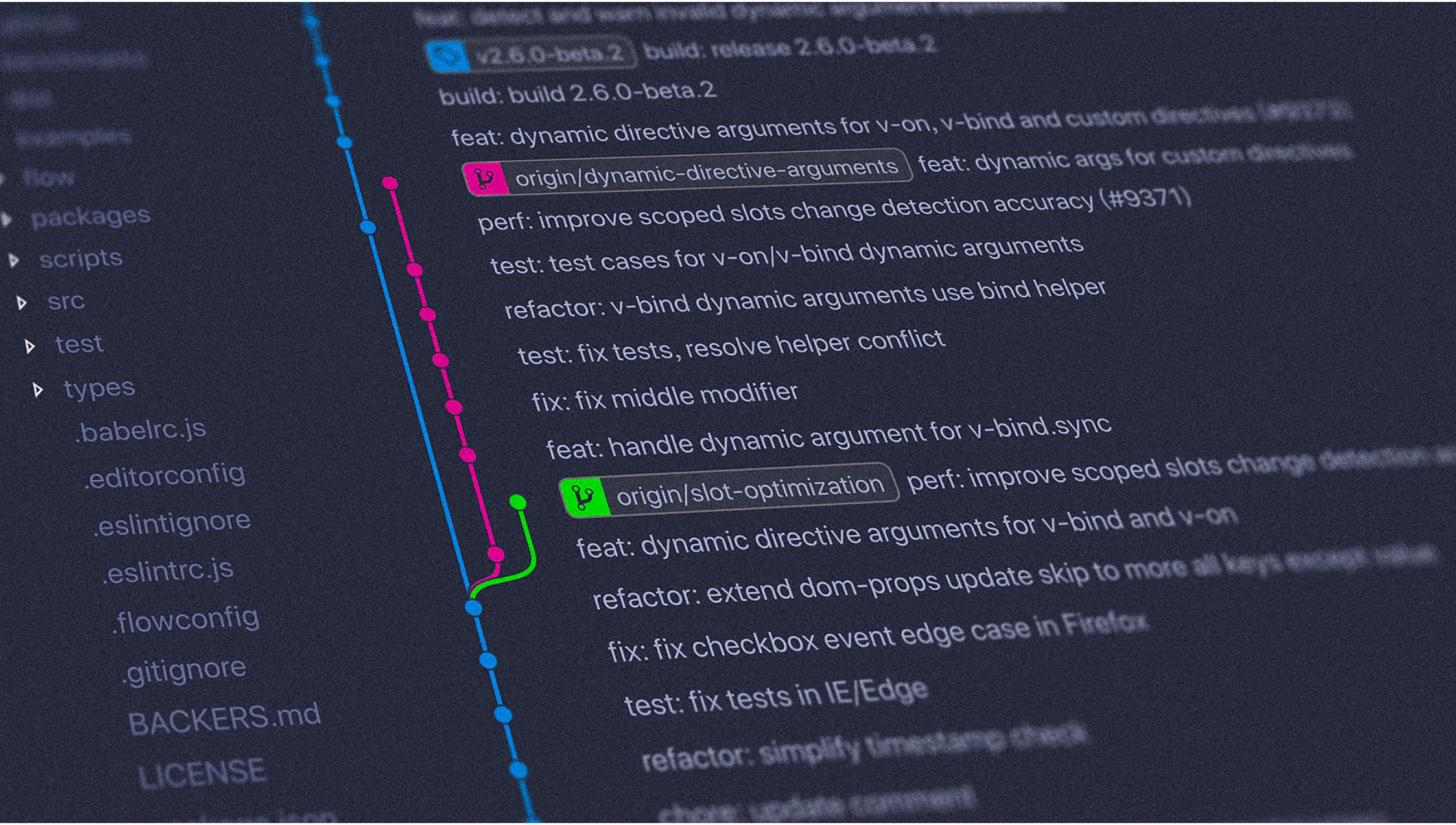
```
> library('ProjectTemplate')
Loading required package: testthat
> load.project()
Loading project configuration
Autoloading utility functions
Autoloading data
 Loading data set: letters
Munging data
 Running preprocessing script: 01-A.R
>
```

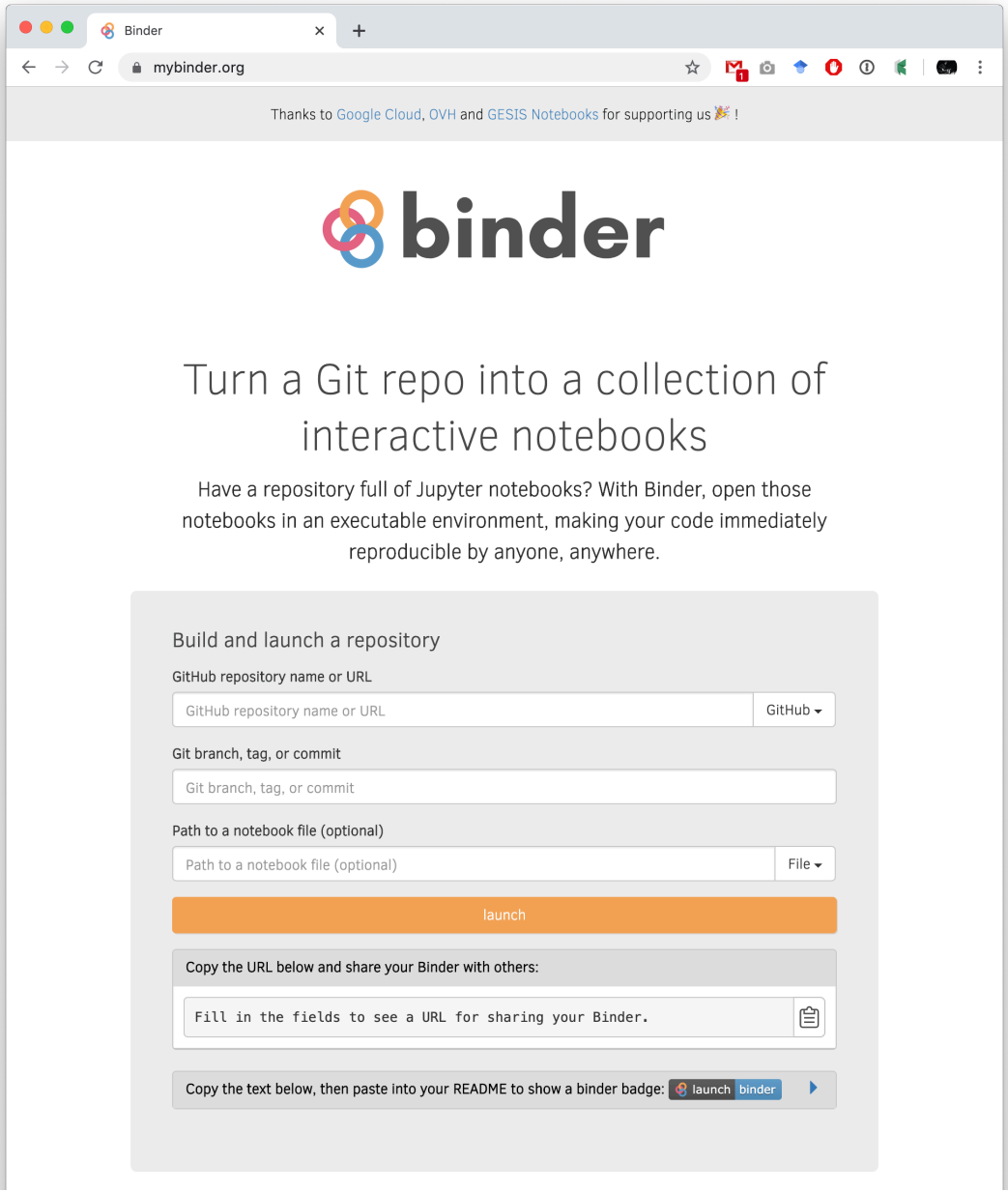
Hints and Tips

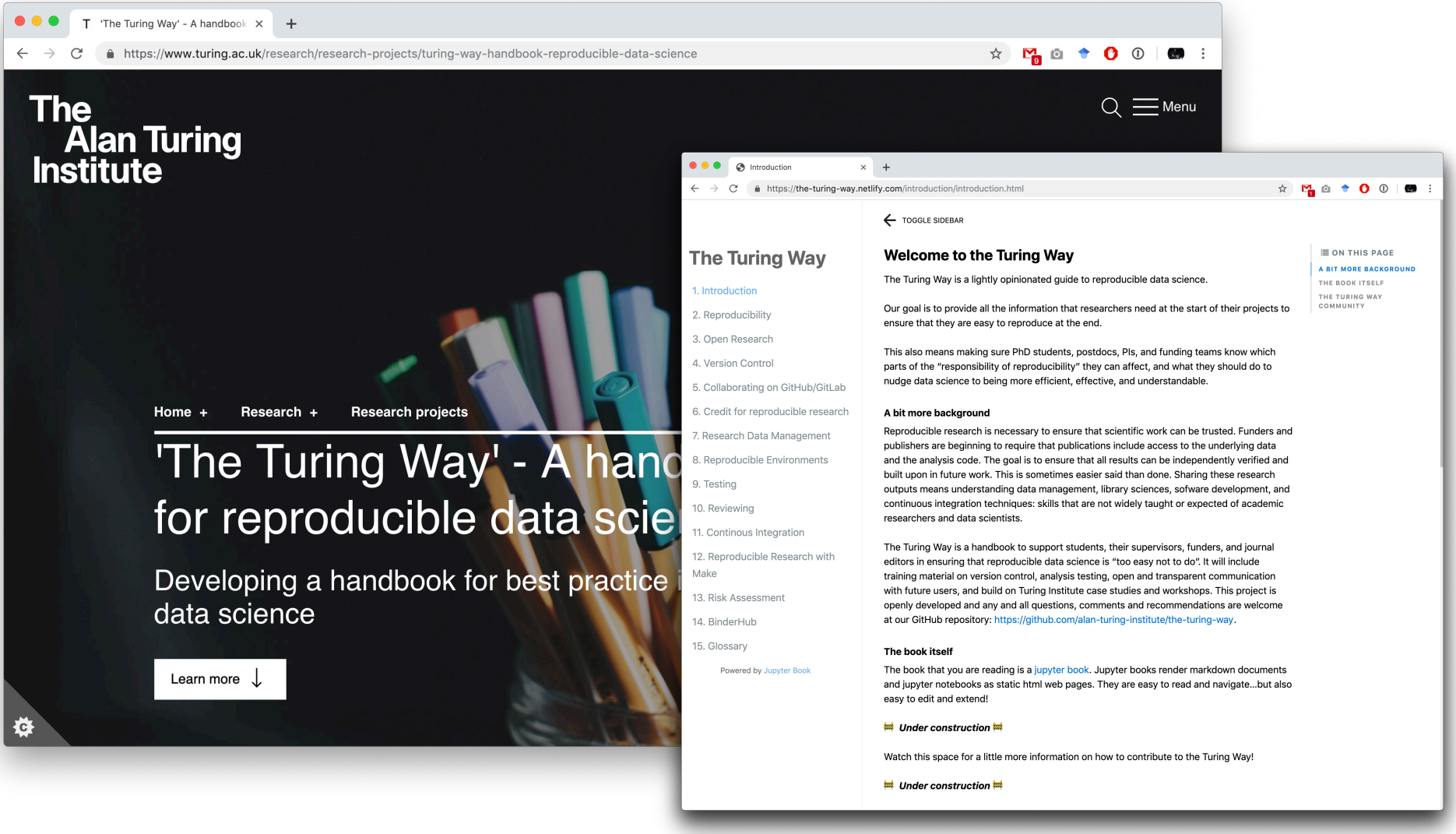
- Append your '`munge`' script names with a number, so they run in a predictable order.
- Do not edit your raw data!
- Specify all dependencies within `config/global.dcf`

knitr::opts_knit$set(root.dir= normalizePath('..'))




v2.6.0-beta.2
build: release 2.6.0-beta.2
build: build 2.6.0-beta.2
feat: dynamic directive arguments for v-on, v-bind and custom directives
origin/dynamic-directive-arguments
feat: dynamic args for custom directives
perf: improve scoped slots change detection accuracy (#9371)
test: test cases for v-on/v-bind dynamic arguments
refactor: v-bind dynamic arguments use bind helper
test: fix tests, resolve helper conflict
fix: fix middle modifier
feat: handle dynamic argument for v-bind.sync
origin/slot-optimization
perf: improve scoped slots change detection
feat: dynamic directive arguments for v-bind and v-on
refactor: extend dom-props update skip to more all keys except value
fix: fix checkbox event edge case in Firefox
test: fix tests in IE/Edge
refactor: simplify timestamp check
chore: update comment
flow
packages
scripts
src
test
types
.babelrc.js
.editorconfig
.eslintignore
.eslintrc.js
.flowconfig
.gitignore
BACKERS.md
LICENSE


Binder
mybinder.org
Thanks to Google Cloud, OVH and GESIS Notebooks for supporting us 🎉 !
binder
Turn a Git repo into a collection of interactive notebooks
Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
Build and launch a repository
GitHub repository name or URL
GitHub repository name or URL
GitHub
Git branch, tag, or commit
Git branch, tag, or commit
Path to a notebook file (optional)
Path to a notebook file (optional)
File
launch
Copy the URL below and share your Binder with others:
Fill in the fields to see a URL for sharing your Binder.
Copy the text below, then paste into your README to show a binder badge:
launch
binder


'The Turing Way' - A handbook
https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science
The Alan Turing Institute
Menu
Home +
Research +
Research projects
'The Turing Way' - A hand
for reproducible data scie
Developing a handbook for best practice
data science
Learn more
Introduction
https://the-turing-way.netlify.com/introduction/introduction.html
TOGGLE SIDEBAR
The Turing Way
1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Credit for reproducible research
7. Research Data Management
8. Reproducible Environments
9. Testing
10. Reviewing
11. Continous Integration
12. Reproducible Research with Make
13. Risk Assessment
14. BinderHub
15. Glossary
Powered by Jupyter Book
Welcome to the Turing Way
The Turing Way is a lightly opinionated guide to reproducible data science.
Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.
This also means making sure PhD students, postdocs, PIs, and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective, and understandable.
A bit more background
Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.
The Turing Way is a handbook to support students, their supervisors, funders, and journal editors in ensuring that reproducible data science is "too easy not to do". It will include training material on version control, analysis testing, open and transparent communication with future users, and build on Turing Institute case studies and workshops. This project is openly developed and any and all questions, comments and recommendations are welcome at our GitHub repository: https://github.com/alan-turing-institute/the-turing-way.
The book itself
The book that you are reading is a jupyter book. Jupyter books render markdown documents and jupyter notebooks as static html web pages. They are easy to read and navigate...but also easy to edit and extend!
Under construction
Watch this space for a little more information on how to contribute to the Turing Way!
Under construction
ON THIS PAGE
A BIT MORE BACKGROUND
THE BOOK ITSELF
THE TURING WAY COMMUNITY

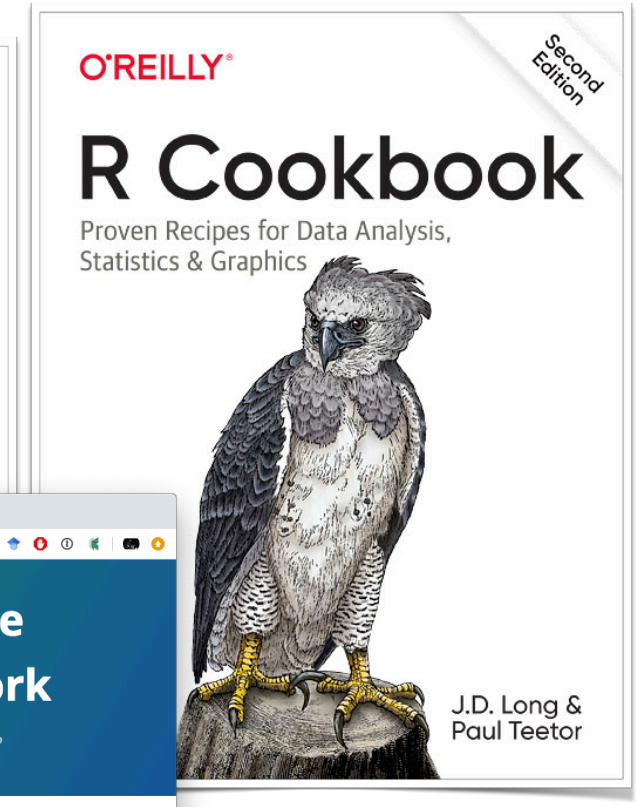
JD Long
@CMastication

Reproducible actuarial workflows with #rstats ... I'm glad to see actuarial science adopting some good data science workflow methods.

https://philipdarke.com/reproducible-actuarial-work/

10:21 AM · Mar 2, 2019 · Twitter for iPad

15 Retweets **64** Likes



Reproducible data science techniques in actuarial work

What can actuaries learn from open science and other professions?

How can actuaries ensure that workflows are efficient, minimise the risk of errors and allow complex work to be reproduced by others in their organisation?

Actuaries exploring the use of data science have the opportunity to revisit existing ways of working and consider whether they remain appropriate. These challenges are also being faced in science and by other professions.

This presentation looks at why the concept of *reproducible work* is key and how it can help address the challenges of working in data intensive fields.

Institute and Faculty of Actuaries

Reproducible data science techniques in actuarial work

What can actuaries learn from open science and other professions?

“Reproducibility is like brushing your teeth. It is good for you, but it takes time and effort. Once you learn it, it becomes a habit. ”

Irakli Loladze, mathematical biologist

Source: Nature 2016, 533, 531–533,; Picture: http://cu.ac.kr