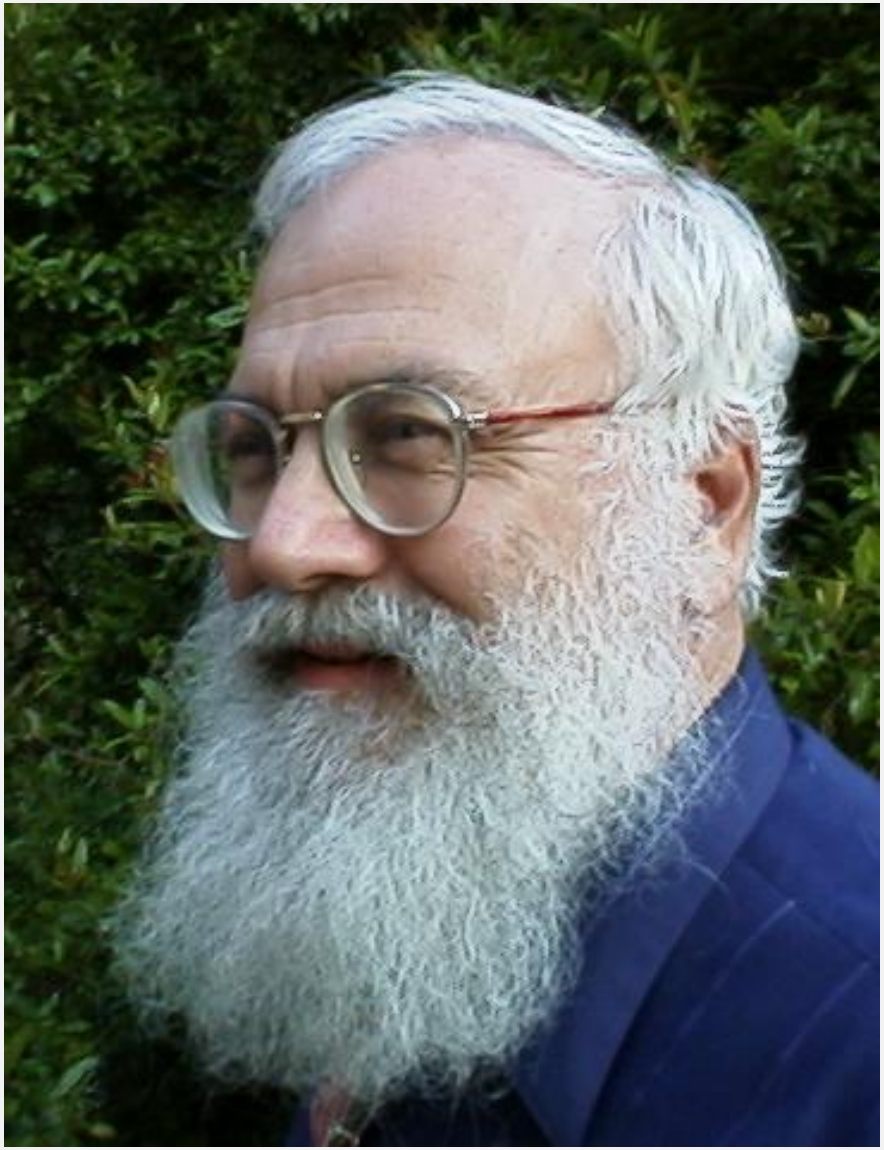
Research Software & Open Reproducible Research

Karthik Ram
@_inundata



Jon Clarebout

Electronic Documents Give Reproducible Research a New Meaning

Jon F. Claerbout and Martin Karrenbach, Stanford Univ. **1992**

“

A revolution in education...

marriage of word processing

and **software command scripts**

Electronic Documents Give Reproducible Research a New Meaning

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

> “In this marriage, an author attaches to every figure caption **a push button** to **recalculate** the **figures** from all its **data**, **parameters** and **programs**

Electronic Documents Give Reproducible Research a New Meaning

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

“ This provides a **concrete definition of reproducibility** in computationally oriented research

Verification of reproducibility

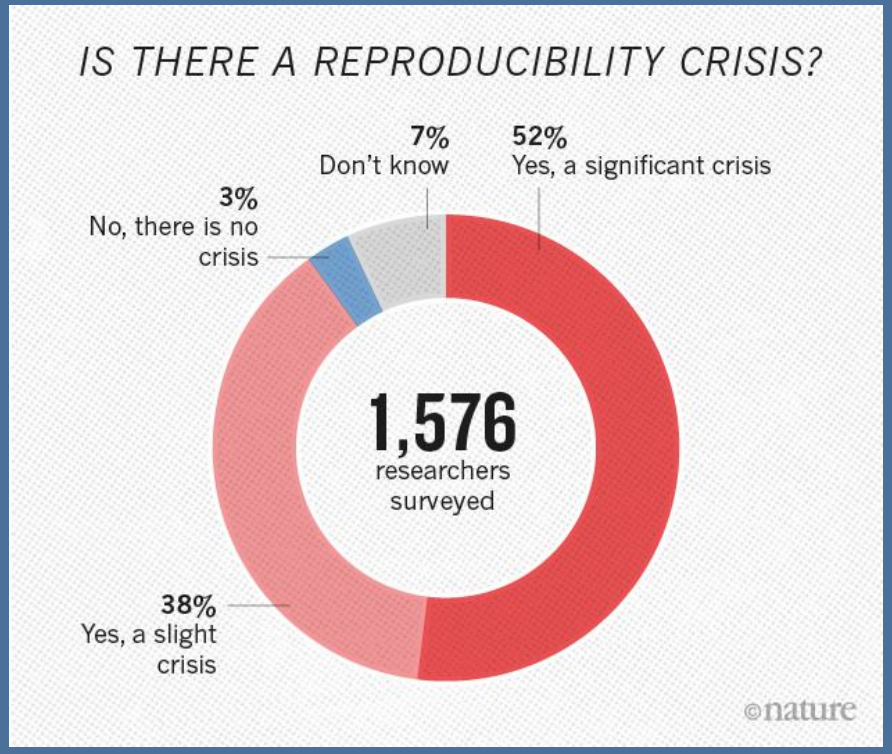
> “Judgement of the reproducibility of computationally oriented research **no longer requires an expert - a clerk can do it**

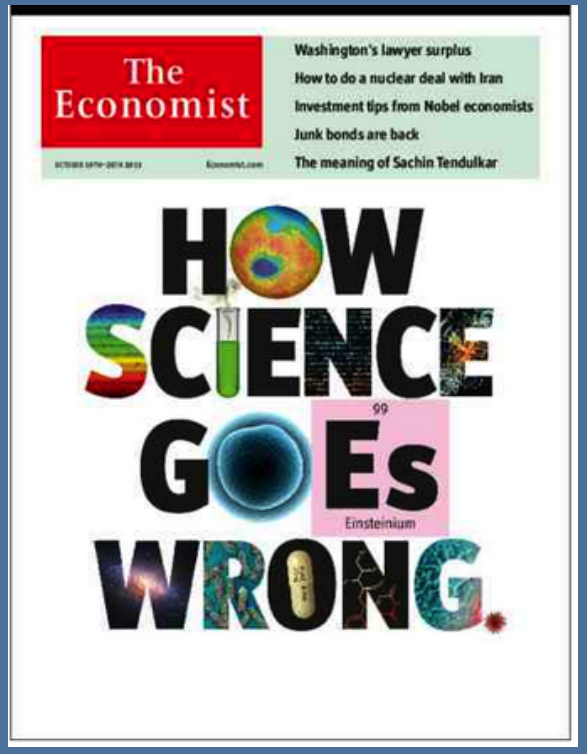
*" The actual scholarship is in the **full software environment**, **code**, & **data** that produce the result"*

David Donoho

10.1007/978-1-4612-2544-

The **reproducibility crisis** is widespread

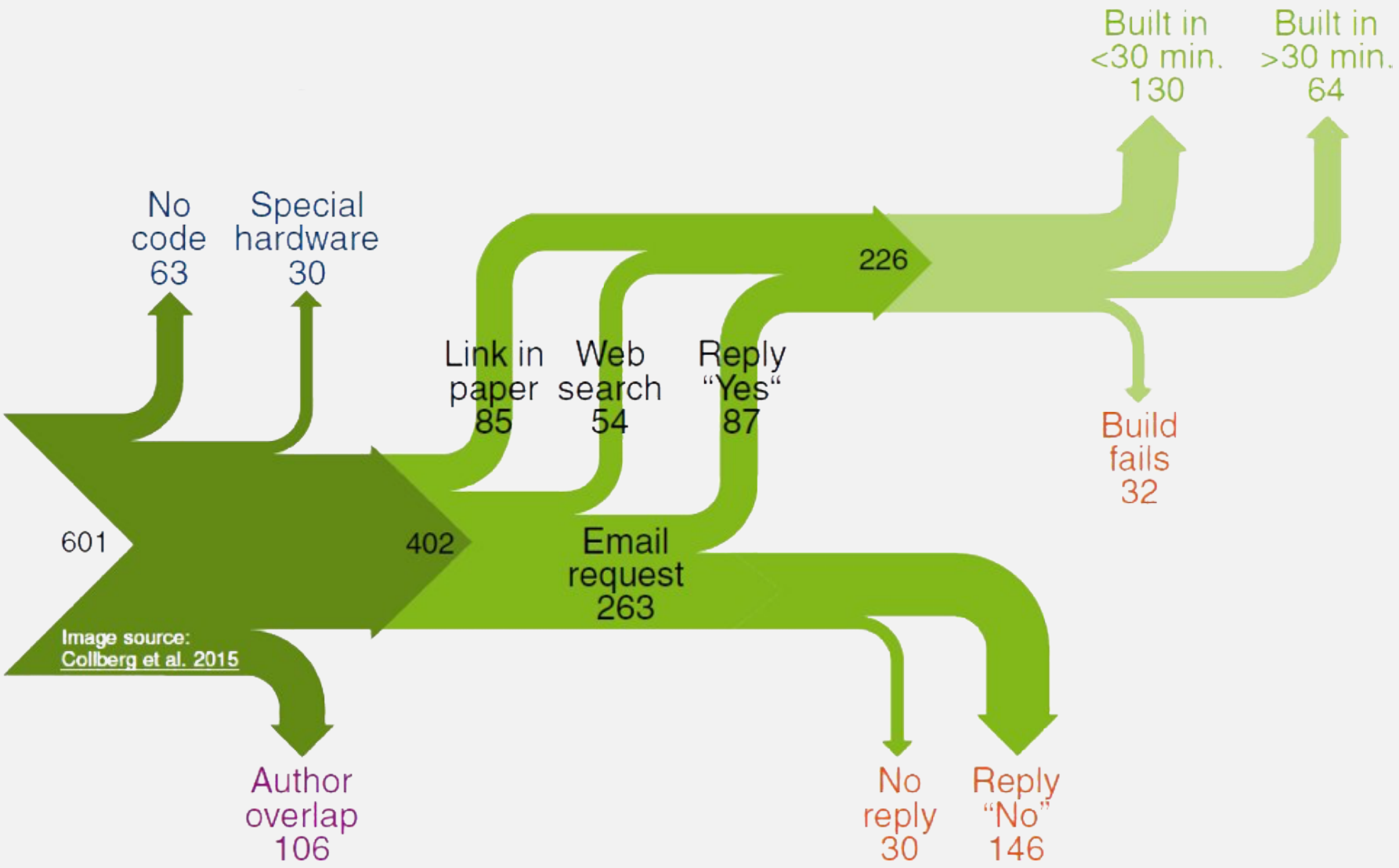




Baker, 2015. Baker & Dolgin 2017, Aschwanden, C. 2016, Casadevall & Fang 2010

Lack of reproducibility is quite widespread even in applied computational research

Collberg et al 2014


No code 63
Special hardware 30
601
Image source: Collberg et al. 2015
Author overlap 106
Link in paper 85
Web search 54
Reply "Yes" 87
402
Email request 263
No reply 30
Reply "No" 146
226
Build fails 32
Built in <30 min. 130
Built in >30 min. 64

The **extent to which code would actually build** with reasonable effort is quite low

< 20%

Collberg et al 2014

Software is critical for research but **we don't value it as scholarship**


Prof. Daniel
Bolnick

Recently, Dr. Tony Wilson from CUNY Brooklyn tried to recreate my analysis, so that he could figure out how it worked and apply it to his own data ... ***he couldn't quite recreate some of my core results.***

“

I dug up my original code, sent it to him, and after a couple of back-and-forth emails ***we found my error****.*

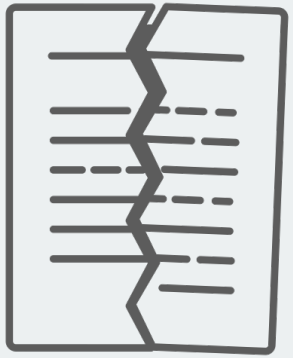
*“I immediately sent a retraction email to the journal (Evolutionary Ecology Research), which will be appearing soon in the print version. So let me say this clearly, **I was wrong**.*

“*So: how many results, negative or positive, that enter the published literature are tainted by a coding mistake as mine was. We just don’t know.* ***Which raises an important question: why don’t we review code (or other custom software) as part of the peer-review process?***

> *I suspect that I am not the only biologist out there to make a **small mistake** in my code that has a **big impact**.*

When software is not visible, it is often **excluded from peer review**

Computational science has **culture problems**



no **verification**,

no **transparency**

no **efficiency**

Training

Training in computational skills is one of the **largest unmet needs**

Barone et al, 2017

Training

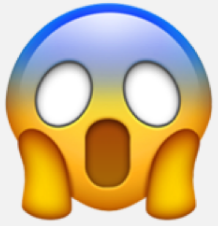
Expected output

Research Software

Use	90%	95%
Can't continue without	70%	63%

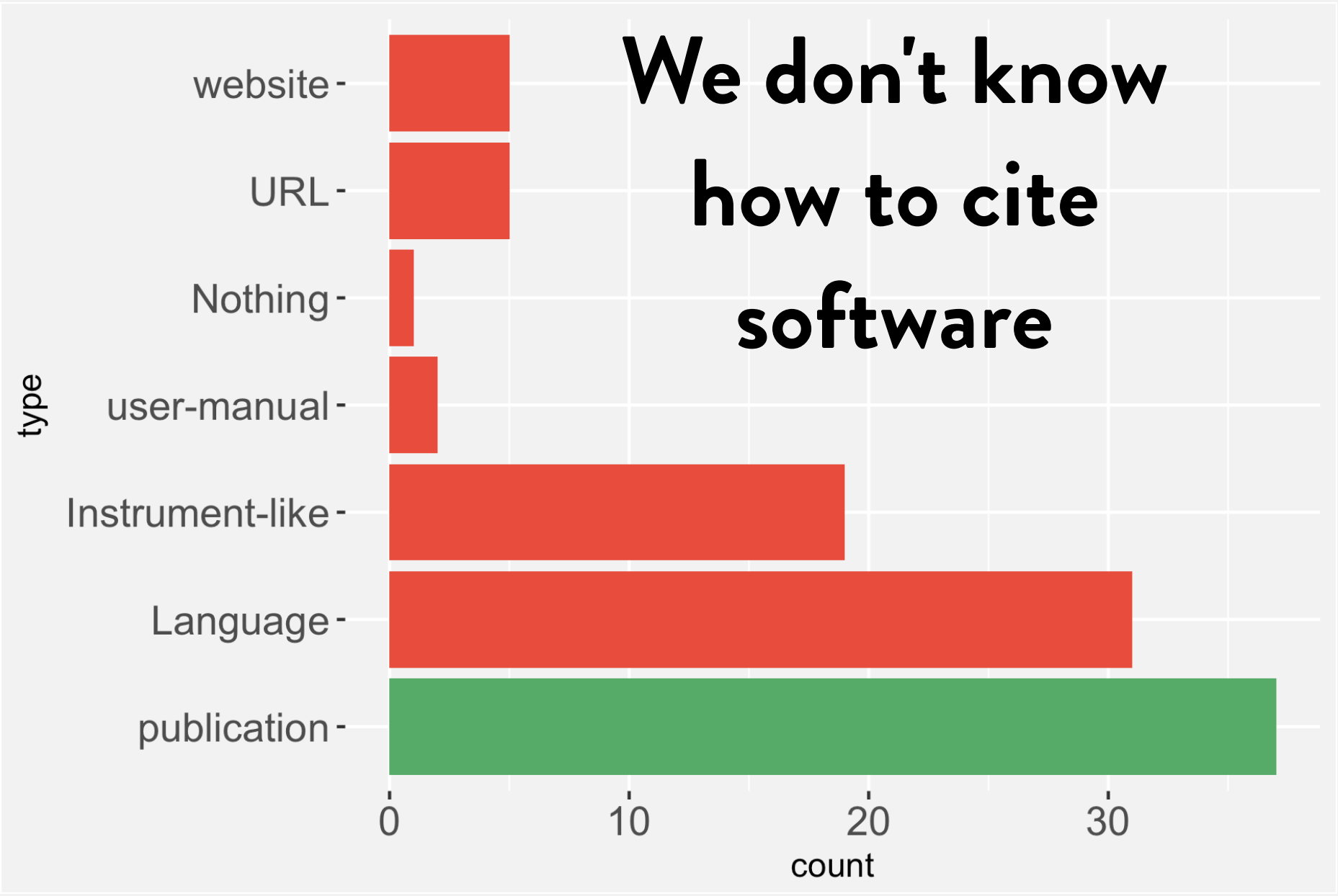
19%

of researcher-contributed packages have unit tests



Keyes et al 2015

Credit



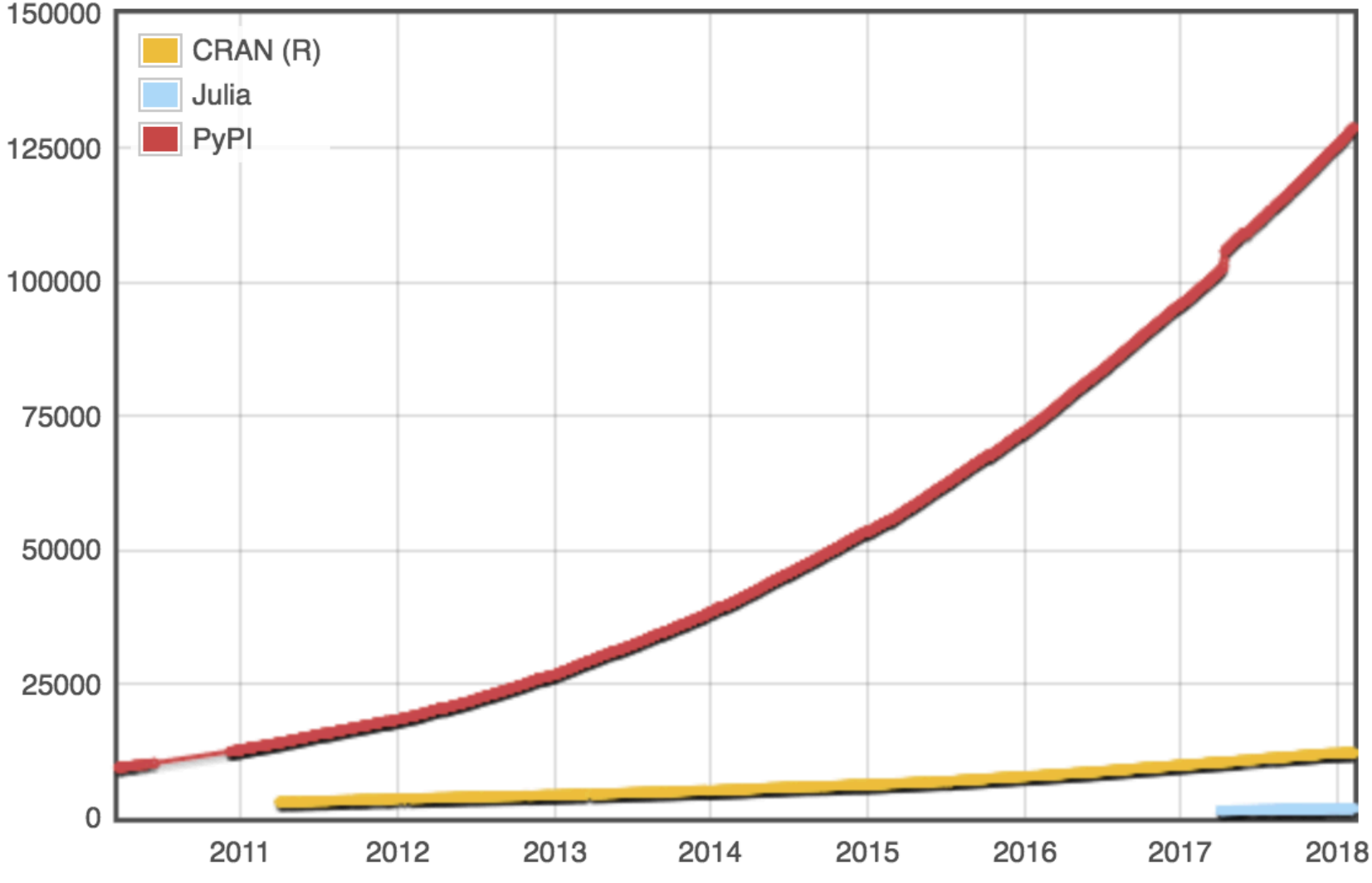
Howison & Bullard 2016

Lack of visibility means that **incentives** to produce high-quality, widely shared, and collaboratively developed software **are lacking**

Formal citations: 31% - 43%

Informal mentions are the norm, even in high impact journals

Software is frequently inaccessible (15 - 29%)


150000
125000
100000
75000
50000
25000
0
CRAN (R)
Julia
PyPI
2011
2012
2013
2014
2015
2016
2017
2018

Sustainability

Academic data science

Universities rush to add data science majors as demand explodes

Yale establishes biomedical data science center

Yale Daily News (blog) · Feb 9, 2018

London universities join national drive to transform health through data science

Imperial College London · Feb 8, 2018

California University of Pennsylvania adds new data science degree

Tribune-Review · Jan 26, 2018

'Memphis: data science capital' gains steam as university adds new SAS training center

The Commercial Appeal · Jan 27, 2018

Glasgow University wins share of £14m data science award

The Scotsman · Jan 30, 2018

David Donoho

50 years of data science

Journal Of Computational And Graphical Statistics, 2017

Academic data science is the big tent

ROpenSci

binder

jupyter

scikit-image
image processing in python

What will data science look like in 2065?

Open Science Takes Over

What will data science look like in 2065?

> **Reproducible computation** is finally being recognized today by many scientific leaders as a **central requirement for valid scientific publication.**

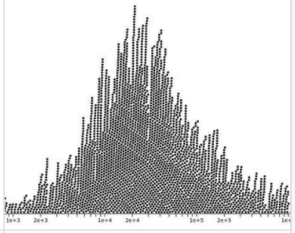
What will data science look like in 2065?

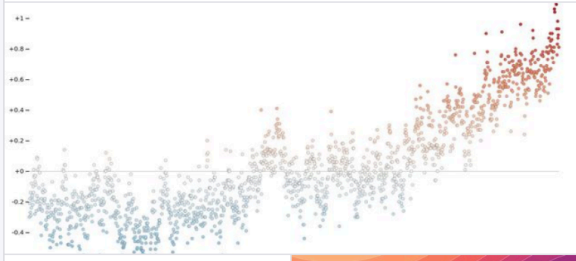
> To work reproducibly in today's computational environment, one **constructs automated workflows** that **generate all the computations** and all the analyses in a project.

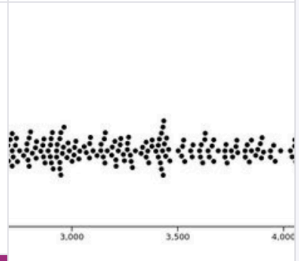
We are currently in the golden age that Clarebout talks about

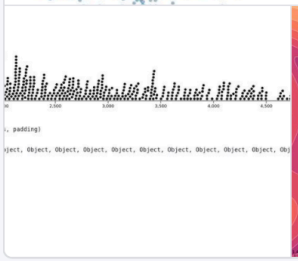
Visualization

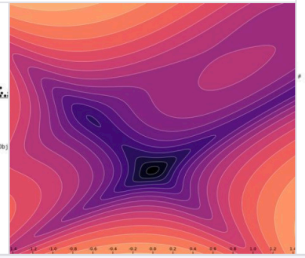
Explore and explain patterns in quantitative data using D3 and Vega.

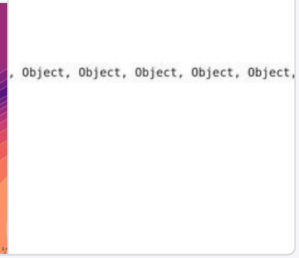






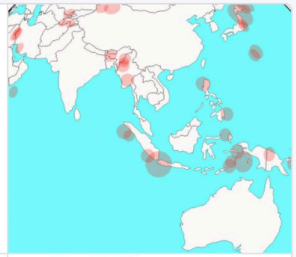


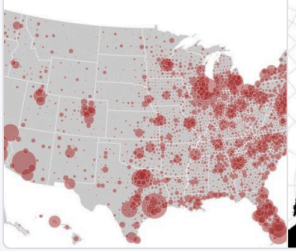


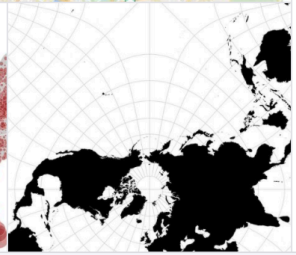


Maps

Embrace your inner shapefile. Or GeoJSON or TopoJSON. 🌍







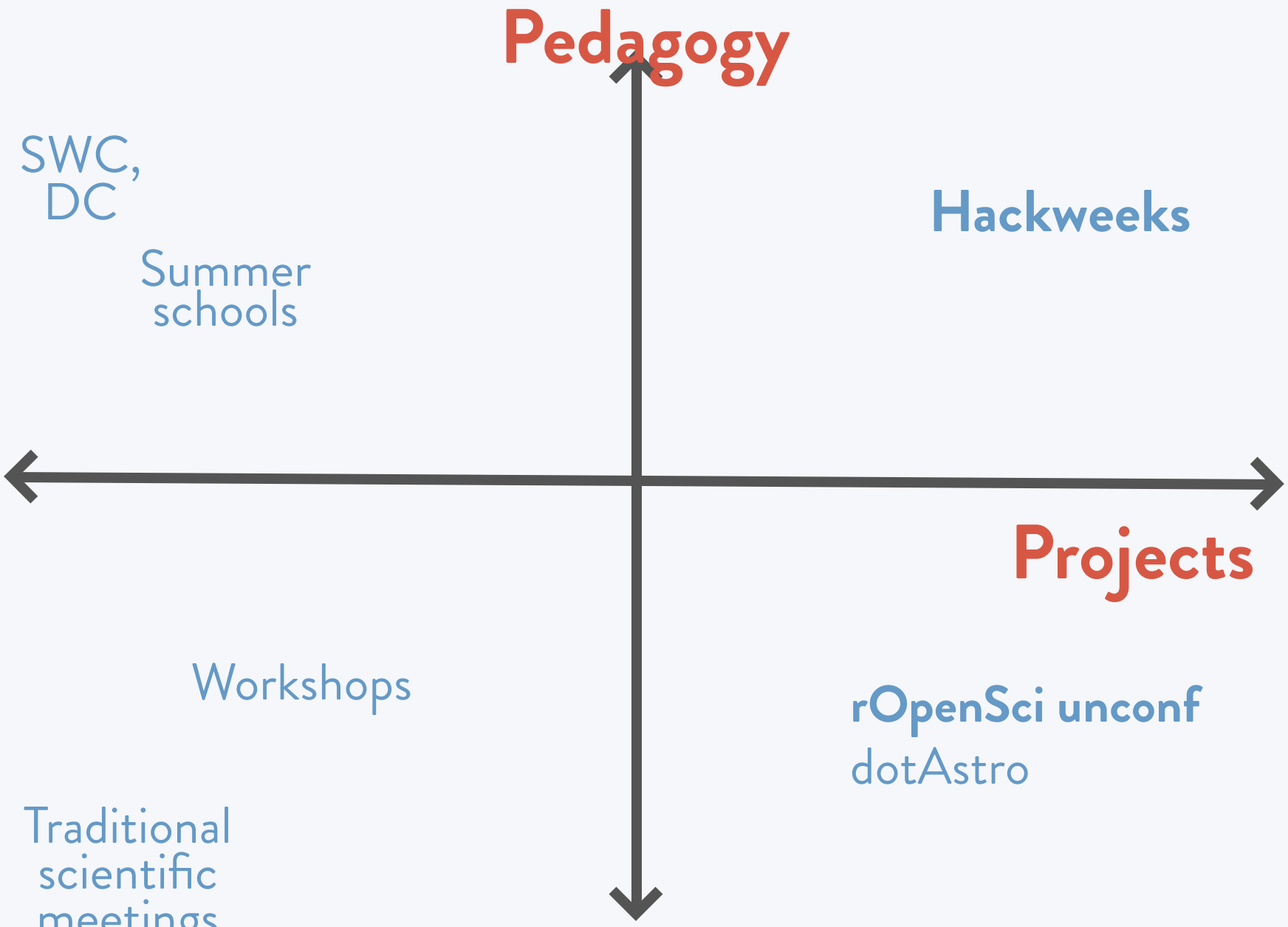
Notebooks are finally becoming the lingua franca of computational science

Fluent interfaces

The best technology is something you don't know you're using

Training

The carpentries, hackweeks, project based learning, data science university courses


Pedagogy
SWC,
DC
Summer
schools
Hackweeks
Projects
Workshops
rOpenSci unconf
dotAstro
Traditional
scientific
meetings

Hackweeks

Hack Weeks fill the gaps between **pedagogically focused** and **project focused** models.

arxiv.org/abs/1711.00028

Developing **a community of research software engineers**, and the next generation of data science mentors.

tidy text

Making text analysis easier and reproducible

textworkshop17.ropensci.org

Credit

We are developing new journals aimed at developers, and highlighting reproducibility as scholarship

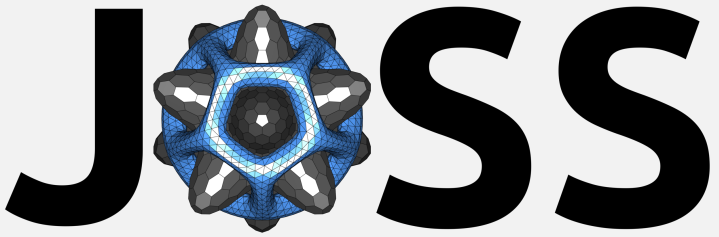
Journal of Open Source Software

joss.theoj.org

Arfon Smith

Data Science Mission Office

(DSMO) Head, STSCI



Journal of Open Source Software

joss.theoj.org

A mechanism for research software developers to get **credit** *within the current merit system of science*

JOSS

Submission only require a **Github repository URL**, an **ORCID**, and a succinct high-level description of the software

The Journal of Open Source Software

tidytext: Text Mining and Analysis Using Tidy Data Principles in R

Julia Silge[1] **and David Robinson**[2]

1 Datassist **2** Stack Overflow

DOI: 10.21105/joss.00037

Software

- Review
- Repository
- Archive

Summary

The tidytext package (Silge, Robinson, and Hester 2016) is an R package (R Core Team 2016) for text mining using tidy data principles. As described by Hadley Wickham (Wickham 2014), tidy data has a specific structure:

- each variable is a column
- each observation is a row
- each type of observational unit is a table

joss.theoj.org

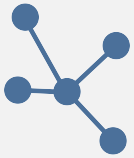
ROpenSci

ropensci.org • @ropensci

Founded in 2011 by Carl Boettiger, Scott Chamberlain and myself.

Early motivation was to make data access easier and reproducible

100+ software packages to support data science. e.g. **spatial data**, biodiversity informatics & climate change, **glue for workflows**.



Community

Developing a **community of research software engineers**, and the next generation of data science mentors.

Culture change

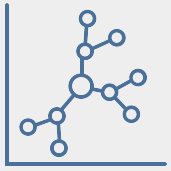
Bringing the best parts of academic peer-review to research software

Incentivize scientists who engage in reproducible research (50k fellowships)

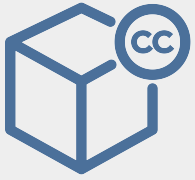
Glue software

ROpenSci

Data retrieval (APIs, data storage services, journals)



Data visualization (plot.ly, magick)



Data sharing (figshare, Zenodo)

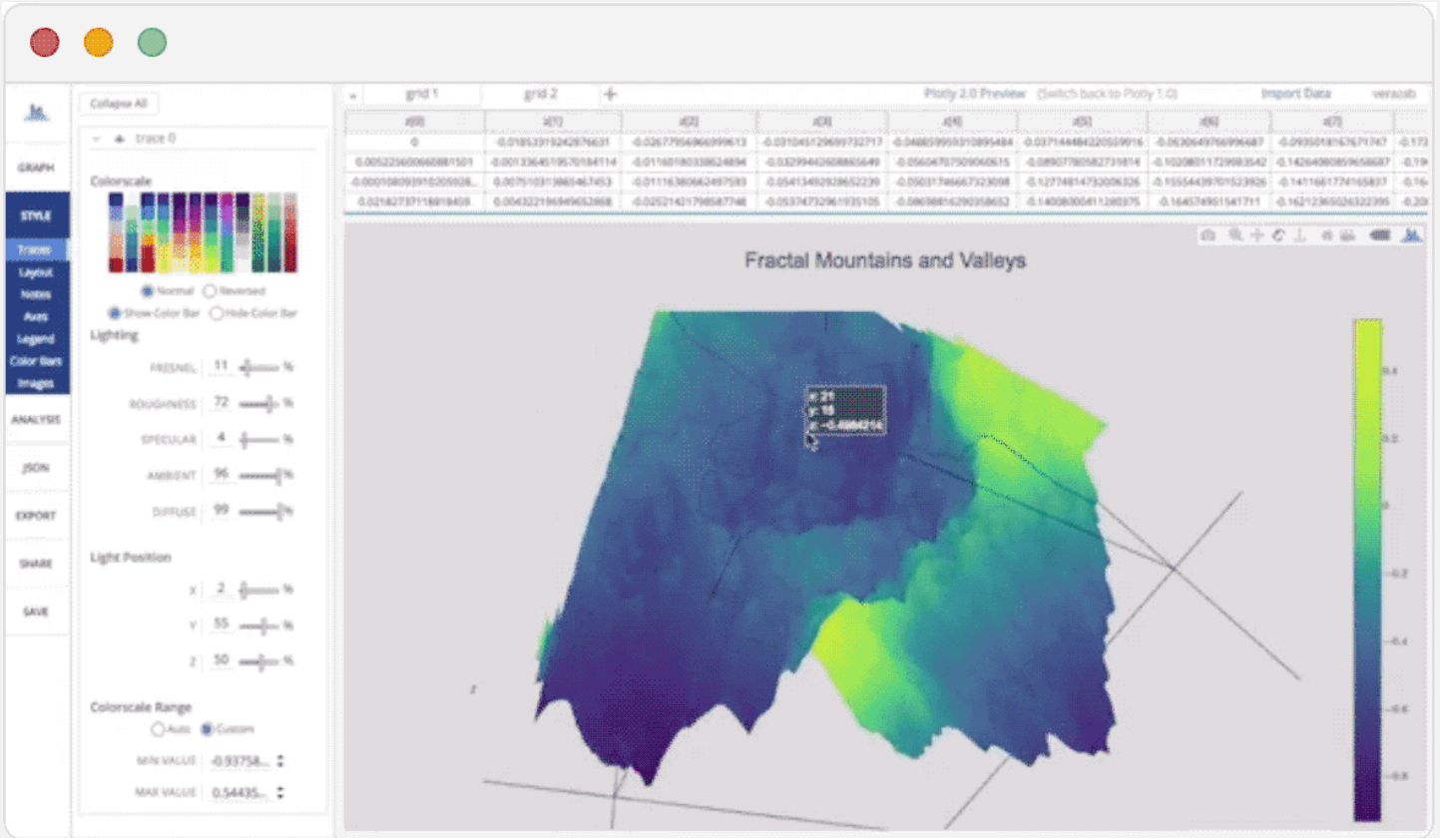
Reproducibility

text mining

Enabling retrieval of full text from **journals** (PLOS, Biomed Central, eLife, Springer IEEE arXiv preprints bioarxiv preprints) and **PDF**s (tabularizer, tesseract, pdftools) along with **tools for text analysis and NLP** (text reuse, tokenizers)

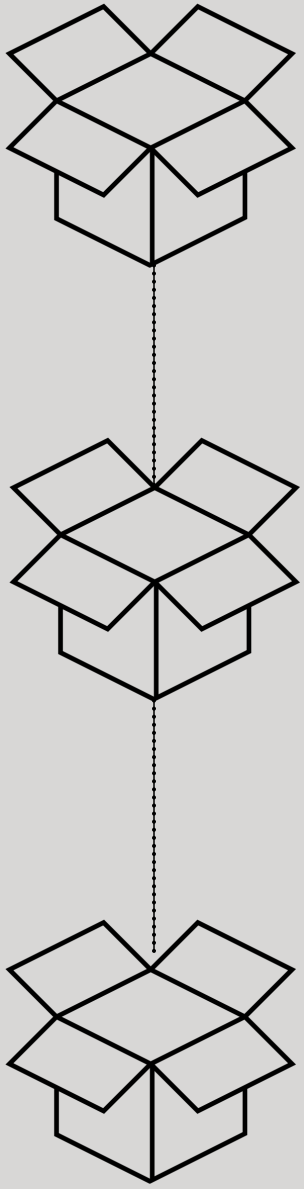
Data visualization

Computing resources of R + interactivity of JavaScript



Reviewing software without a publication

Even without software pubs, we need to create a culture around peer-reviewing our research software



🤔 Pre-submission inquiry

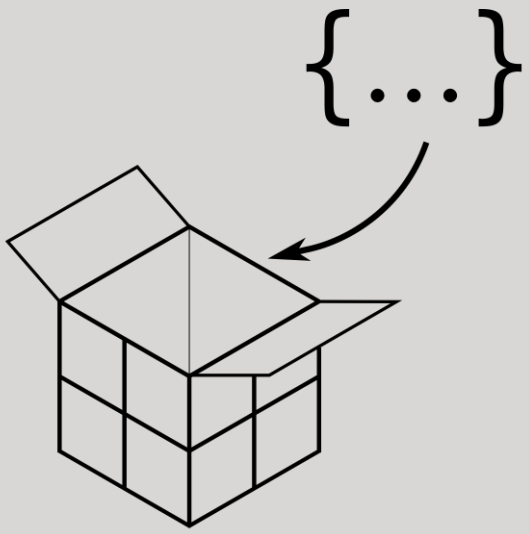
Fit based on our criteria

Peer-review 😒

evaluate the package for usability, quality, and style based on our guidelines

Acceptance 💯

Packages are badged and added to our system



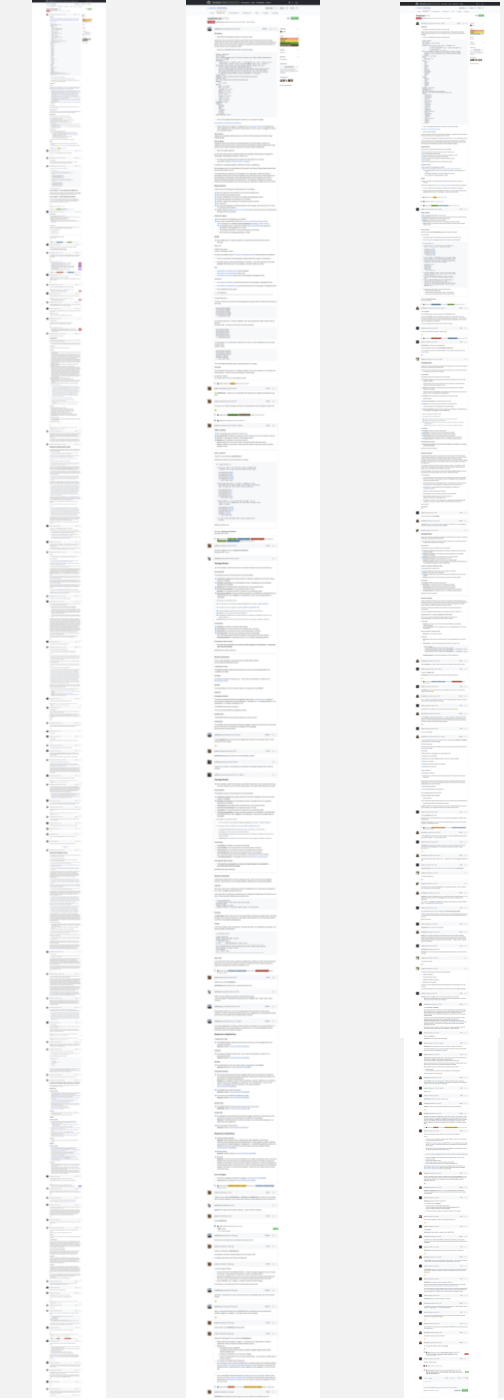
Open & **non-adverserial**

No rejections

Makes the process constructive for everyone involved

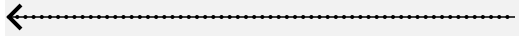
Software Review

- OSI compatible license
- Complete documentation
- High test coverage
- Readable code
- Usability



Nov 2017

A typical software review thread



Feb 2018

benmarwick commented 15 days ago — Member

Yes, thanks, I'm happy. You've done lots of work to address the concerns in my review, that's excellent to see. The package is much more accessible now, and it's easier to see how to get started using it.

🎉 1

wlandau-lilly commented 15 days ago — Member

Thank you, **@benmarwick**! I am so glad you think the changes are making a difference.

By the way, I just updated the review summaries in light of wlandau-lilly/drake#195.

gothub commented 15 days ago — Member

All of the points I mentioned have been fully addressed with clear and complete documentation, and I have no other issues, so thumbs up from me.

I have one remaining question that is not really an issue for the review, but I'm just curious about. If i were to develop a workflow with drake, what drake specific artifacts would need to be preserved for a researcher to reproduce my analysis (other than the R scripts and data that may have existed before even starting to build the drake workflow)? I'm assuming that it's simply a CSV from the workflow itself, e.g. 'my_plan.csv' from the basic example. If more than that is required then explaining that in the docs or having an export function would be a good idea.

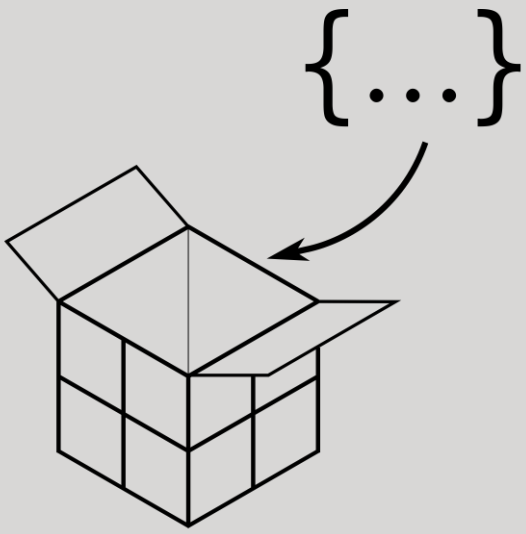
🎉 1

wlandau commented 11 days ago

Member

I sincerely thank you all for helping me take my favorite project to the next level. Your advice and coaching were eye-opening. I feel much more able to help people do their work.

👍 1



"I don't really see myself writing another serious package without having it go through code review."

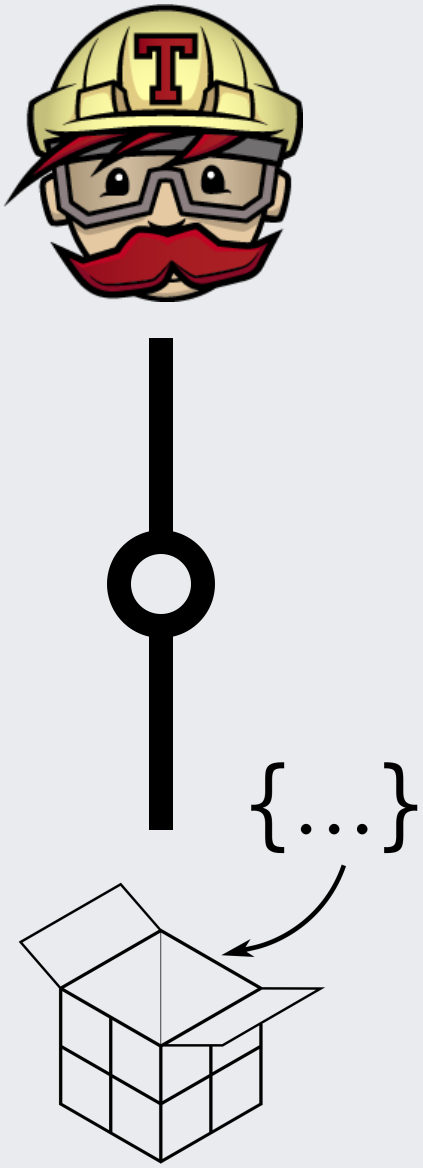
Methods in Ecology and Evolution

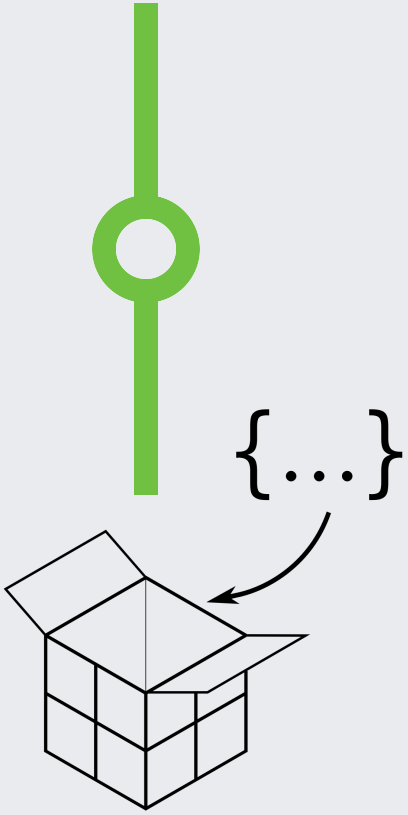
BRITISH ECOLOGICAL SOCIETY

Sign up as a reviewer

ropensci.org/onboarding


TRAVIS + TIC

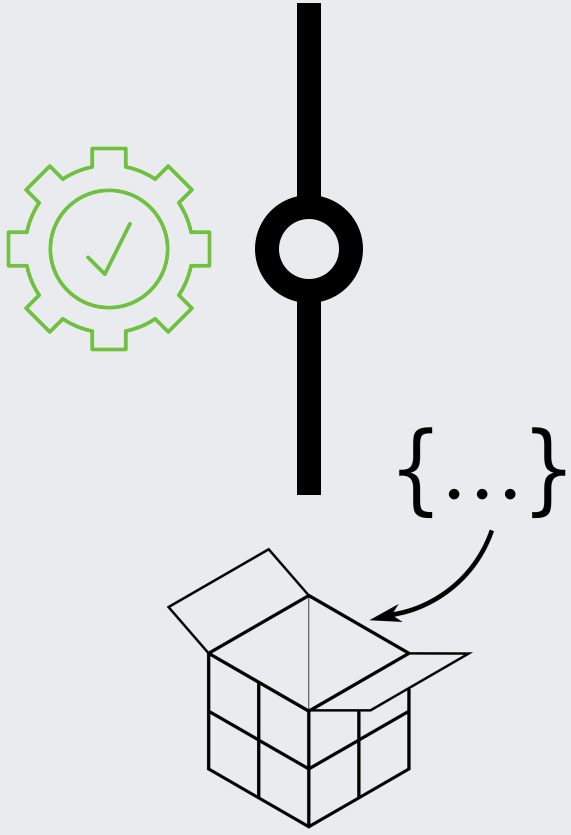




All checks have passed

1 successful check

Travis

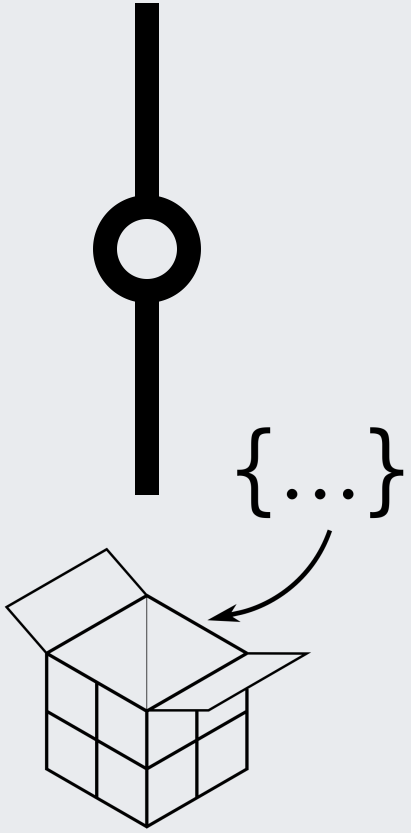



```
travis_enable()
#Create deploy keys
use_travis_deploy()
```

Tasks Integrated Continuously

Tic

Works with **Travis CI**, **AppVeyor**, or the CI tool of your choice

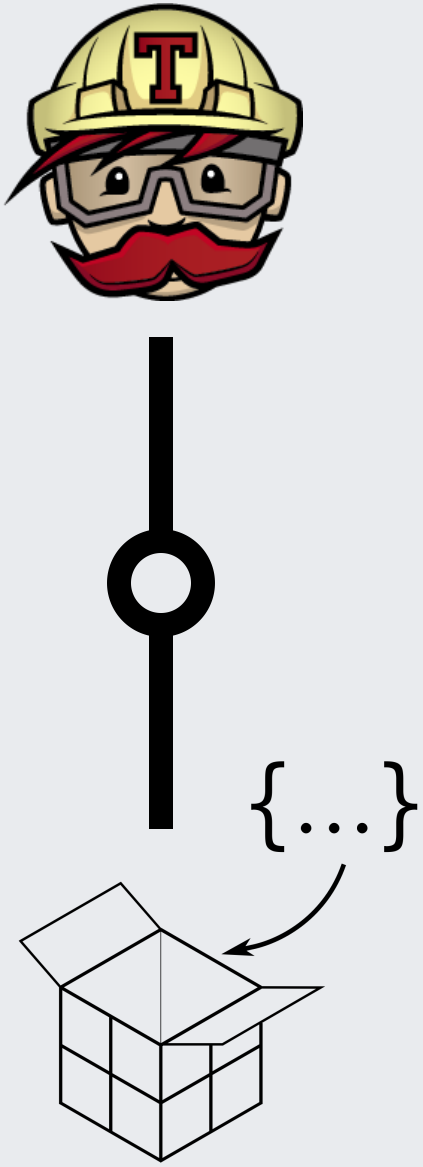


`tic`

→ Checks with steps need to run

→ Prepares everything for that step

→ Runs it at the right time



Tic

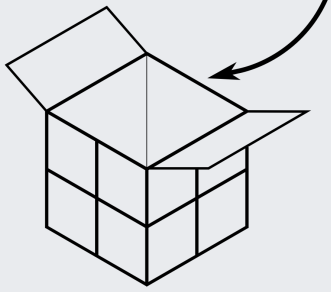
Specify steps to be run at each **stage**.
Using a **simple** DSL.

Stages

```
after_success
deploy
```

Preinstalled steps

{...}



step_hello_world:

step_run_covr:

step_install_ssh_key: *make available a private SSH key (which has been added before to your project by travis::use_travis_deploy())*

step_test_ssh

step_build_pkgdown

Preinstalled steps

step_push_deploy: deploy to GitHub with arguments:

path: which path to deploy
branch: which branch to deploy to
remote_url: the remote URL to push to,
commit_message: adds a [ci skip] to avoid a loop

Configuring tic

```
#before_script
before_script:
- R -q -e 'devtools::install_github("ropenscilabs/tic"); tic::prepare_all_stages()'

#after_success
after_success:
- R -q -e 'tic::after_success()'

#deploy
deploy:
  provider: script
  script: R -q -e 'tic::deploy()'
  on:
    all_branches: true
```

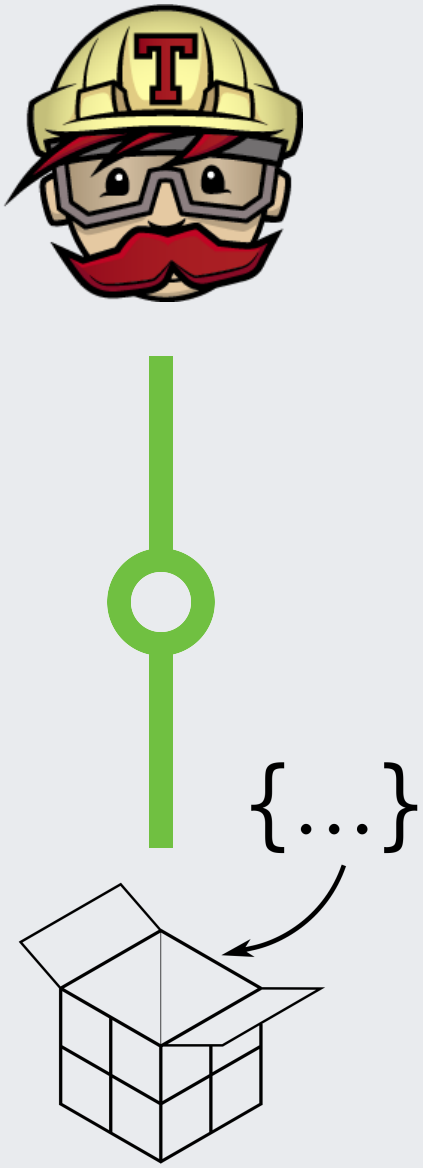
Add tic.R

```
get_stage("after_success") %>%
  add_step(step_hello_world()) %>%
  add_step(step_run_covr())

get_stage("deploy") %>%
  add_step(step_install_ssh_keys()) %>%
  add_step(step_test_ssh())
```

Write your own steps!

github.com/ropenscilabs/tic#how-steps-are-run



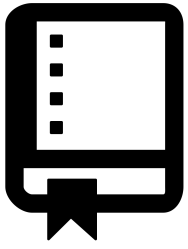
blogdown

pkgdown

packagedocs

data release

drat repo

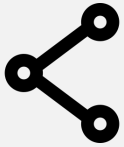


ropenscilabs.github.io/tic/

Get involved

 discuss.ropensci.org

 onboarding.ropensci.org

 ropensci.org/community

 @ropensci