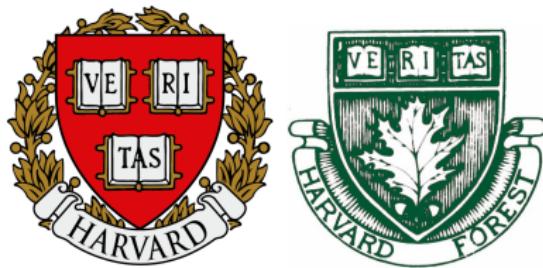


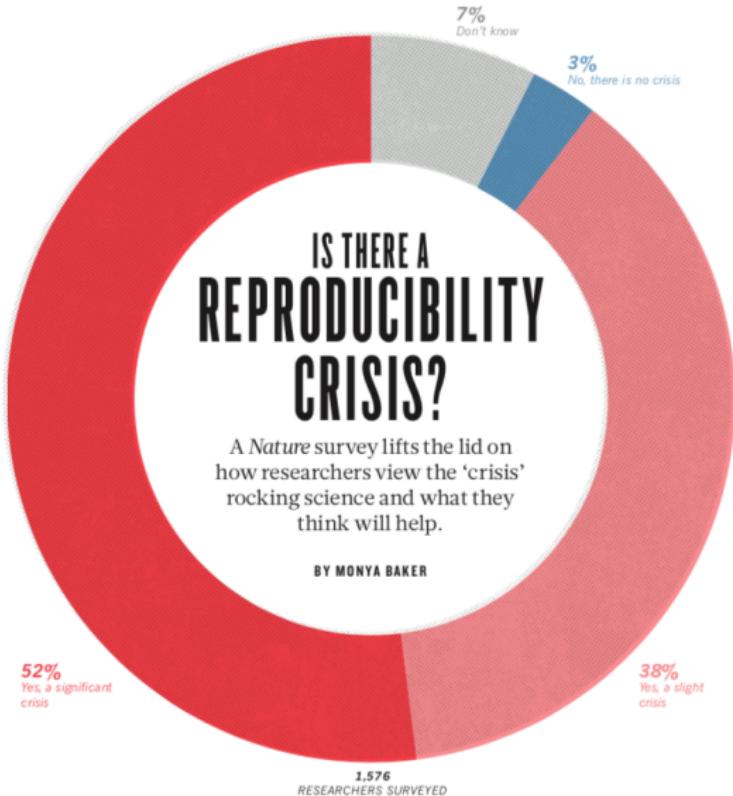
# Opportunity in the Reproducibility Crisis

Computational tools to improve scientific benefaction

Matthew K. Lau, PhD

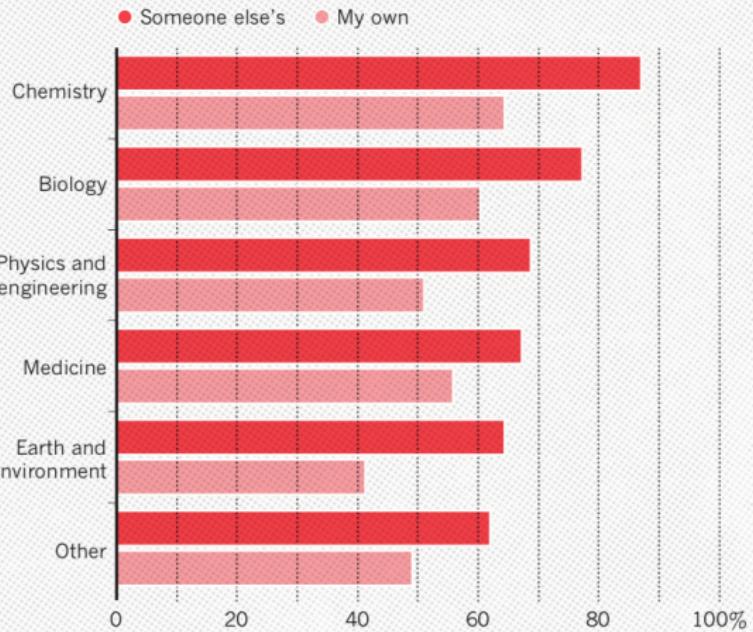


**START**



## HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



# So, how bad is it, really?

## *Harvard Calls for Retraction of Dozens of Studies by Noted Cardiac Researcher*

Some 31 studies by Dr. Piero Anversa contain fabricated or falsified data, officials concluded. Dr. Anversa popularized the idea of stem cell treatment for damaged hearts.



**What are the factors contributing to the crisis?**

# Statistic's Contribution

$$p = 0.05 = 1/20$$

# **Statistic's Contribution**

**One in twenty**

# Hack Your Way To Scientific Glory



You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

## 1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

## 2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

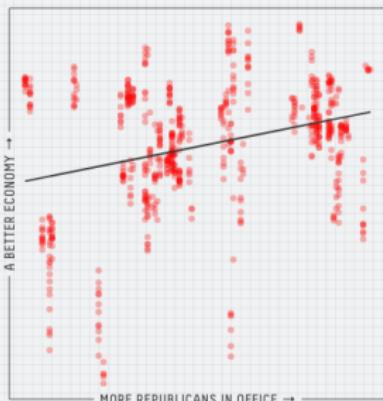
- Employment
- Inflation
- GDP
- Stock prices

Other options

- Factor in power  
Weight more powerful positions more heavily
- Exclude recessions  
Don't include economic recessions

## 3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.



## 4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



## Result: Publishable

You achieved a p-value of less than 0.01 and showed that Republicans have a **positive effect** on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Binder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Aschwanden and King 2015

[fivethirtyeight.com/features/science-isnt-broken](http://fivethirtyeight.com/features/science-isnt-broken)

# Hack Your Way To Scientific Glory



You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

## 1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

## 2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

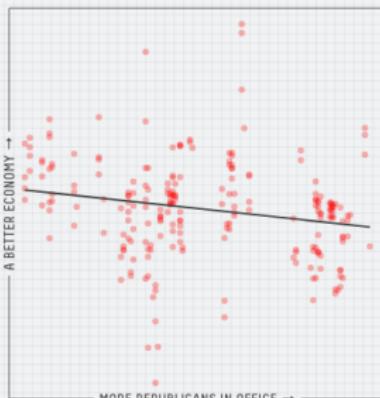
- Employment
- Inflation
- GDP
- Stock prices

Other options

- Factor in power
  - Weight more powerful positions more heavily
- Exclude recessions
  - Don't include economic recessions

## 3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.



## 4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



### Result: Publishable

You achieved a p-value of less than 0.01 and showed that Republicans have a **negative effect** on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Binder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Aschwanden and King 2015

[fivethirtyeight.com/features/science-isnt-broken](http://fivethirtyeight.com/features/science-isnt-broken)

# Hack Your Way To Scientific Glory



You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

## 1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

## 2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

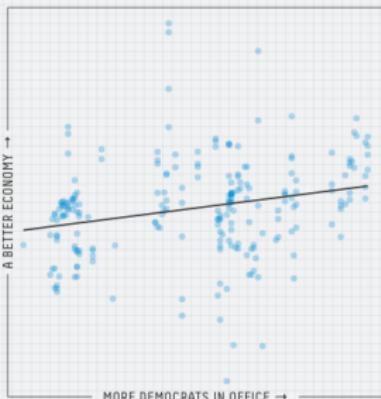
- Employment
- Inflation
- GDP
- Stock prices

Other options

- Factor in power
  - Weight more powerful positions more heavily
- Exclude recessions
  - Don't include economic recessions

## 3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



## 4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



## Result: Publishable

You achieved a p-value of less than 0.01 and showed that **Democrats have a positive effect on the economy**. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Binder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Aschwanden and King 2015

[fivethirtyeight.com/features/science-isnt-broken](http://fivethirtyeight.com/features/science-isnt-broken)

# Hack Your Way To Scientific Glory



You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

## 1 CHOOSE A POLITICAL PARTY

Republicans  Democrats

## 2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

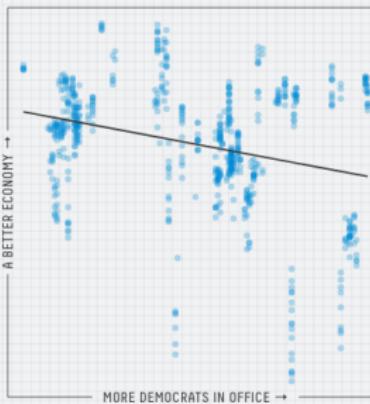
- Employment
- Inflation
- GDP
- Stock prices

Other options

- Factor in power  
Weight more powerful positions more heavily
- Exclude recessions  
Don't include economic recessions

## 3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



## 4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



## Result: Publishable

You achieved a p-value of less than 0.01 and showed that **Democrats have a negative effect on the economy**. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Aschwanden and King 2015

[fivethirtyeight.com/features/science-isnt-broken](http://fivethirtyeight.com/features/science-isnt-broken)

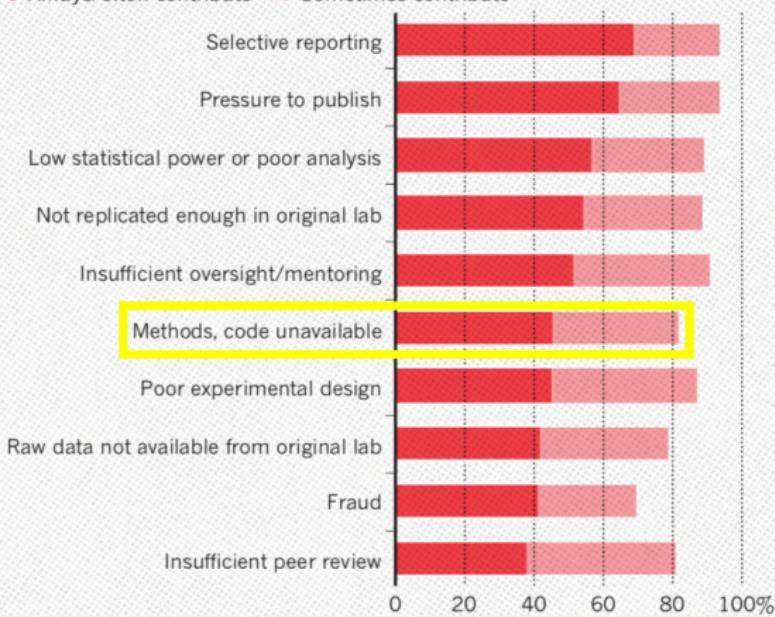
*Please vote!*

# What are the factors contributing to the crisis?

## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

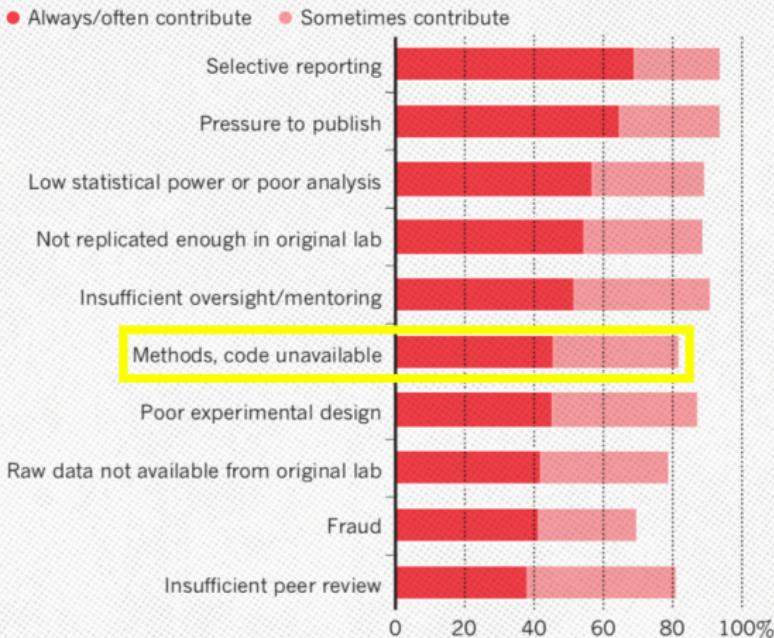
- Always/often contribute
- Sometimes contribute



# Computation is playing a significant role

## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



# Computation is playing a significant role

PNAS



## An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden<sup>a,1</sup>, Jennifer Seiler<sup>b</sup>, and Zhaokun Ma<sup>b</sup>

<sup>a</sup>School of Information Sciences, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and <sup>b</sup>Department of Statistics, Columbia University, New York, NY 10027

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske January 9, 2018

(Received for review July 11, 2017)

A key component of scientific communication is sufficient information for other researchers in the field to reproduce published findings. For computational and data-enabled research, this has often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters. Many journals are revising author guidelines to include data and code availability. This work evaluates the effectiveness of journal policy that requires the data and code necessary for reproducibility be made available postpublication by the authors upon request. We assess the effectiveness of such a policy by (*i*) requesting data and code from authors and (*ii*) attempting replication of the published findings. We chose a random sample of 204 scientific papers published in the journal *Science* after the implementation of their policy in February 2011. We found that we were able to obtain artifacts from 44% of our sample and were able to reproduce the findings for 26%. We find this policy—author remission of data and code postpublication upon request—an improvement over no policy, but currently insufficient for reproducibility.

computational reproducibility of published results. We use a survey instrument to test the availability of data and code for articles published in *Science* in 2011–2012. We then use the scientific communication standards from the 2012 Institute for Computational and Experimental Research in Mathematics (ICERM) workshop report to evaluate the reproducibility of articles for which artifacts were made available (11). We then assess the impact of the policy change directly, by examining articles published in *Science* in 2009–2010 and comparing artifact ability to our postpolicy sample from 2011–2012. Finally, we discuss possible improvements to journal policies for enabling reproducible computational research in light of our results.

### Results

We emailed corresponding authors in our sample to request the data and code associated with their articles and attempted to replicate the findings from a randomly chosen subset of the articles for which we received artifacts. We estimate the artifact recovery rate to be 44% with a 95% bootstrap confidence interval of the proportion [0.36, 0.50], and we estimate the replication rate to be 26% with a 95% bootstrap confidence interval [0.20, 0.32].

reproducible research | data access | code access | reproducibility policy | open science

# Motivation: Journal Policy Impacts

Table 1. Responses to emailed requests ( $n = 180$ )

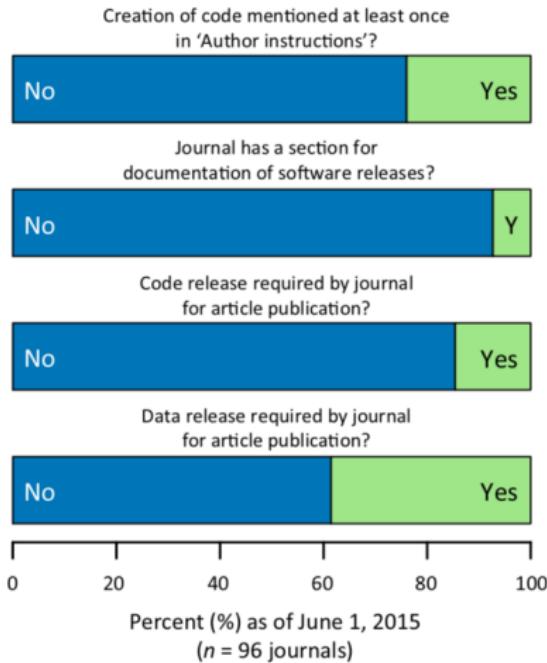
Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	20	11
Asked for reasons	20	11
Refusal to share	12	7
Directed back to supplement	6	3
Unfulfilled promise to follow up	5	3
Impossible to share	3	2
Shared data and code	65	36
Email bounced	3	2
No response	46	26

# Motivation: Journal Policy Impacts

Table 4. Classification of reproducibility effort ( $n = 22$ )

Classification	Percent, %
Impossible to reproduce (missing essential code, data, or methodology)	5
Nearly impossible to reproduce (specialized hardware, intense computation requirements, sensitive data, human study, or other unavoidable reasons)	14
Difficult to reproduce because of unavoidable inherent complexity (e.g., requiring 300 million Markov chain Monte Carlo steps on each dataset, or needing months to do runs)	14
Reproducible with substantial tedious effort (e.g., individual download of a large number of datasets, hand coding of data into a new format, i.e., from an image, many archiving steps required)	5
Reproducible with substantial intellectual effort (e.g., methods well defined but required some knowledge of jargon or understanding of the field; or down the rabbit hole references to past articles required to reproduce; etc.)	5
Could reproduce with fairly substantial skill and knowledge (e.g., required GPU programing abilities to run code that wasn't given; translating complex models into MATLAB code; pseudo code with functions not detailed described in text into code; missing scripts)	23
Reproducible after tweaking (e.g., missing parameters required fiddling to find, missing modified code lines, missing arguments required for differing architecture; missing minor method step)	5
Minor difficulty in reproducing (e.g., installing a specialized library, converting to a different computational system)	18
Straightforward to reproduce with minimal effort	14

# Motivation: Ecology Journal Policies



Meeslan, Heer and White 2016 *Trends in Eco Evo*

# Motivation: Social Science Journal Policies



Crosas et al. 2018 *SocArXiv*

# Crisis and Opportunity

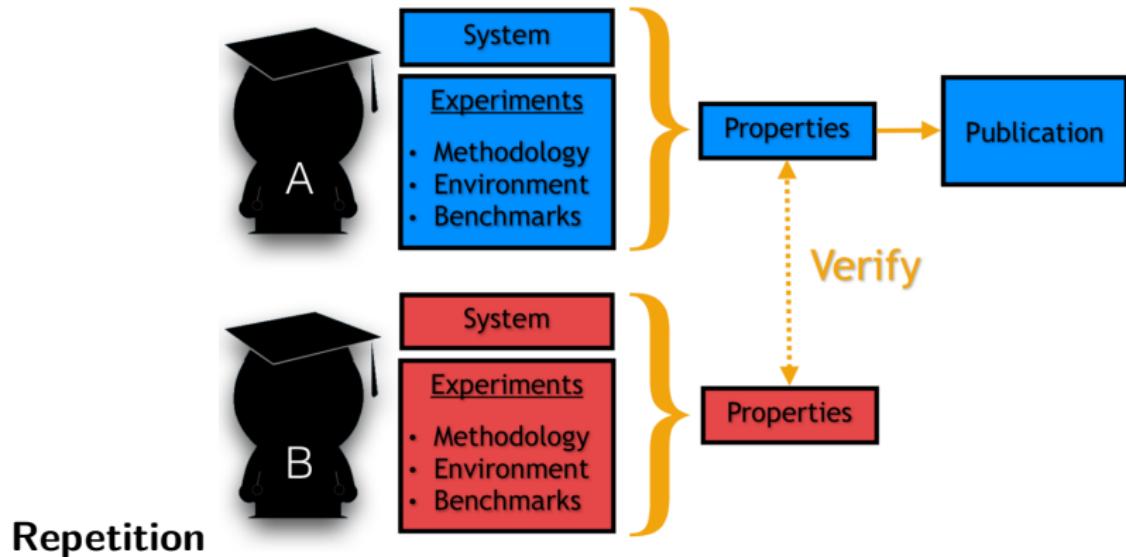
The Chinese use two brush strokes to write the word 'crisis.' One brush stroke stands for danger; the other for opportunity. In a crisis, be aware of the danger—but recognize the opportunity.

— John F. Kennedy

# **Benefaction: the real benefit**

$f(\text{benefaction}) = \text{science}$

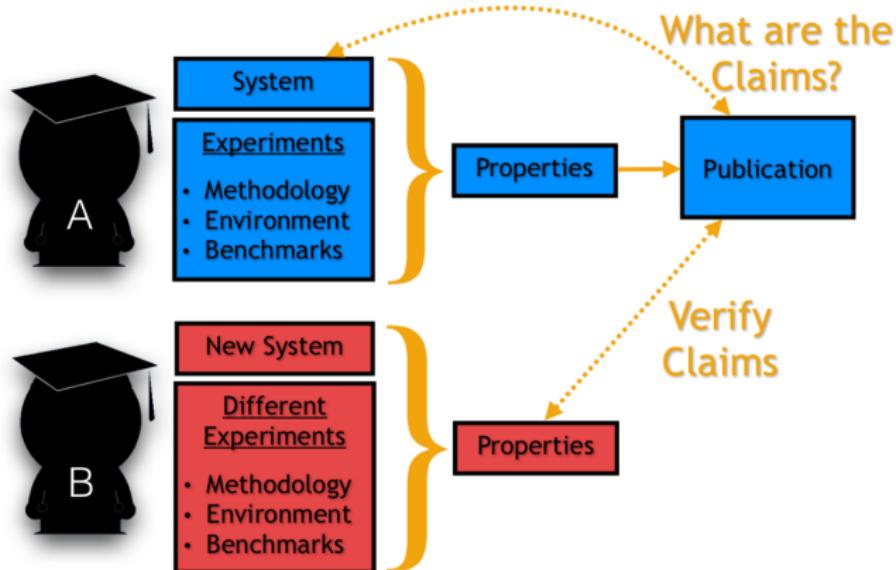
# Opportunity: Benefaction not just reproducibility



*Colberg et al. 2015 Comm ACM*

# Opportunity: Benefaction not just reproducibility

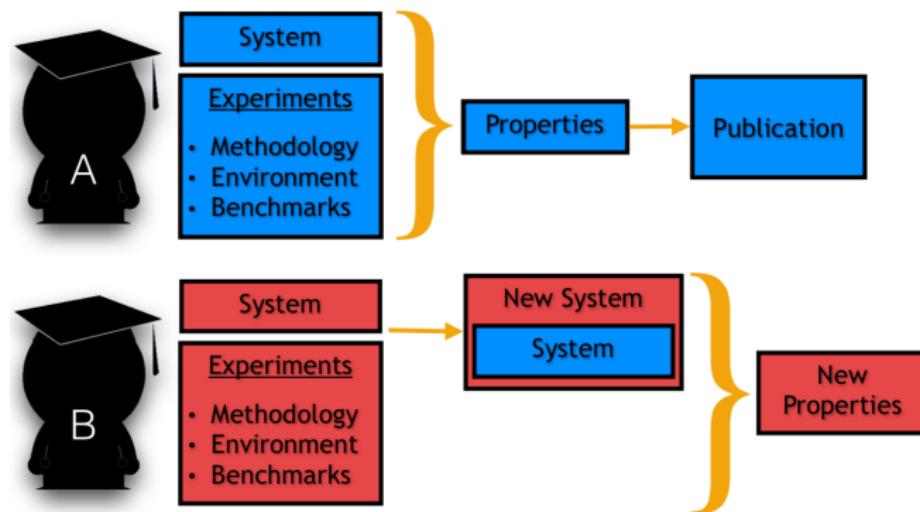
## Reproduction



Colberg et al. 2015 Comm ACM

# Opportunity: Benefaction not just reproducibility

## Benefaction



Colberg et al. 2015 Comm ACM

# What can we do?



# Top Three List

1. Learn a programming language (statistical, command)

## Top Three List

1. Learn a programming language (statistical, command)
2. Make your workflow transparent (data, code and notes)

## Top Three List

1. Learn a programming language (statistical, command)
2. Make your pipelines transparent (data, code and notes)
3. Share and take credit!

# Programming Languages

- ▶ *R*: free, open-source, designed for analysis
- ▶ *python*: also free and open-source, designed for more general computation
- ▶ *BASH*: command language, glues software together

# Programming Languages

- ▶ *R*: free, open-source, designed for analysis
- ▶ *python*: also free and open-source, designed for more general computation
- ▶ *BASH*: command language
- ▶ *Ruby, Java, C++, MatLab, Octave, Stata*, and many more.

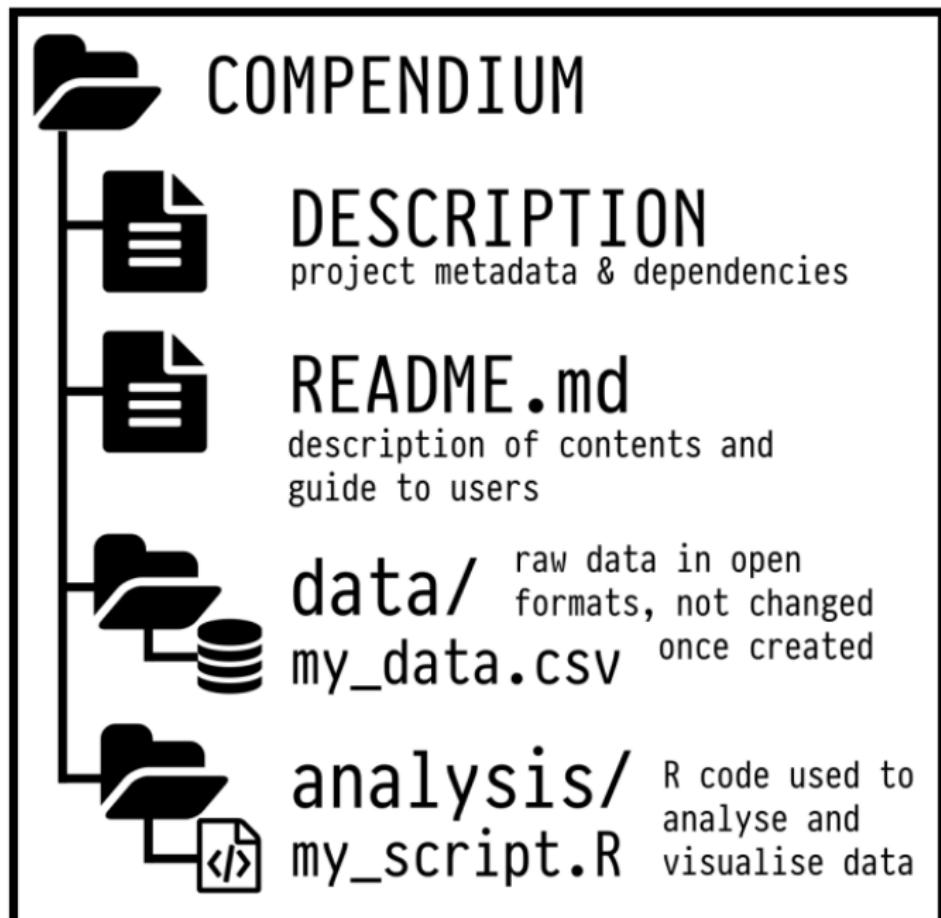
# Transparency

- ▶ Consistent project architecture

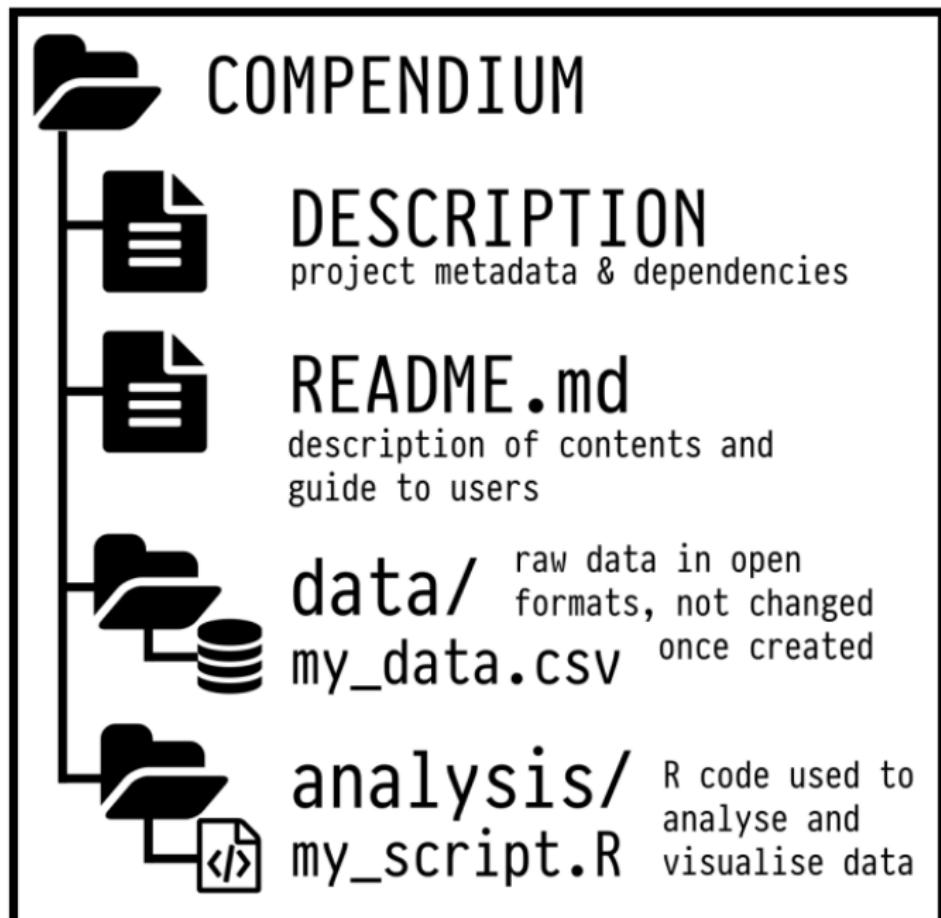
# Transparency

- ▶ Project architecture
- ▶ Version control

# Transparency: Project Architecture



# Transparency: Project Architecture



# Transparency: Version Control

Showing 2 changed files with 173 additions and 19 deletions.

Unified Split

BIN +22.9 KB talk/img/ostrich.jpg

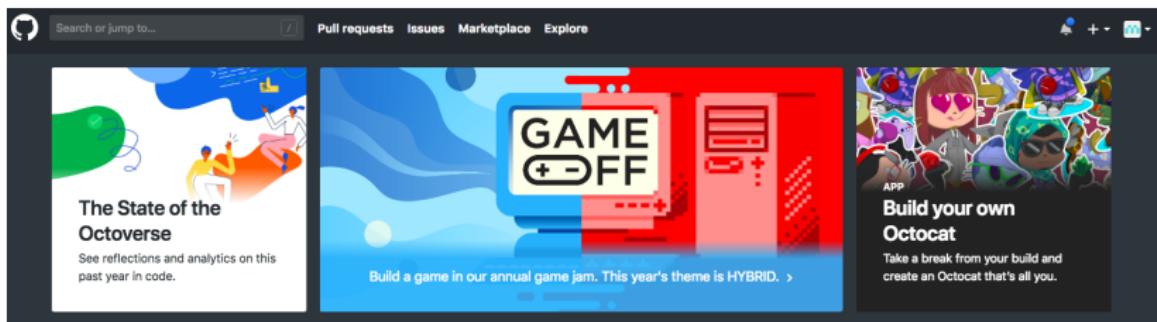
Binary file not shown.

View ▾

```
192  talk/talk.Rmd
@@ -39,7 +39,9 @@ header-includes:
39   - "A scientist, a reviewer and an editor walk into a bar..." 39   - "A scientist, a reviewer and an editor walk into a bar..."
40   - Crosas, White et al., Poisot, Marwick 40   - Crosas, White et al., Poisot, Marwick
41
42 - 3. Tools for improving benefaction not just reproducibility (15 min) 42 + 3. Tools and resources for improving benefaction not just reproducibility (15 min)
43
44 *ReproSci Top 10 Checklist* 45
46 *ReproSci Top 10 Checklist*
47
@@ -65,6 +67,7 @@ header-includes:
65   - Capsules (Code Ocean, ReproZip, encapsulator) 67   - Capsules (Code Ocean, ReproZip, encapsulator)
66   - Literate programming (Rmarkdown, Jupyter, Overleaf, RStudio) 68   - Literate programming (Rmarkdown, Jupyter, OverLeaf, RStudio)
67   - Leaks in the Pipeline: Not recorded information, pseudo-non-determinism 69   - Leaks in the Pipeline: Not recorded information, pseudo-non-determinism
70 + - ??? (ROSA, CodeDepends, others) 70 + - ??? (ROSA, CodeDepends, others)
71
72 4. Let the computers do the work! (10 min) 71
73
74 4. Let the computers do the work! (10 min) 72
75
76 @@ -90,13 +93,158 @@ intelligently and get them to do the right things for us 76
77
78  77
79
78  78
79
76
77
78
79
```

# Transparency: Version Control

[www.github.com](https://www.github.com)



## Based on your interests

### Tencent / rapidjson

A fast JSON parser/generator for C++ with both SAX/DOM style API  
★ 6834 ⚡ 1939

Based on people you follow

### sharkdp / bat

A cat(1) clone with wings.  
★ 9568 ⚡ 172

Based on people you follow

### danburzo / percolate

🌐 → PDF A command-line tool to turn web pages into beautifully formatted PDFs  
★ 2535 ⚡ 83

Based on people you follow

### sinclairtarget / um

Create and maintain your own man pages so you can remember how to do stuff  
★ 1678 ⚡ 36

Based on people you follow

### Kaggle / kaggle

Official Kaggle API  
★ 1678 ⚡ 281

Based on people you follow

# Transparency: Version Control

The screenshot shows the top navigation bar of the Zenodo website. It features the Zenodo logo, a search bar with a magnifying glass icon, an "Upload" button, and a "Communities" link. On the right side, there are "Log in" and "Sign up" buttons.

## Recent uploads

October 30, 2018 (v0.1) Dataset Open Access

Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m

Tomislav Hengl

Soil texture classes (USDA system) for 6 standard soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m. Derived from predicted soil texture fractions using the soiltexture package in R. Processing steps are described in detail here. Antarctica is not included. To access and visualize maps...

Uploaded on November 2, 2018

1 more version(s) exist for this record

View

Zenodo now supports usage statistics!



[Read more](#) about it, in our newest blog post.

November 1, 2018 (v0.1) Software Open Access

"Accelerating and parallelizing Lagrangian simulations of mixing-limited reactive transport—Code Repository"

Nick Engdahl; Michael J. Schmidt

Associated code for: Engdahl, et al., "Accelerating and parallelizing Lagrangian simulations of mixing-limited reactive transport"

Uploaded on November 1, 2018

2 more version(s) exist for this record

View

Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citable.
- **Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- **Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** — because not everything

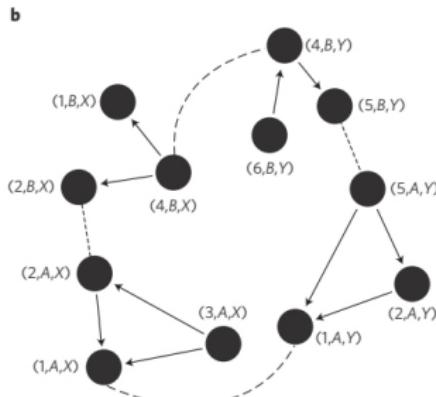
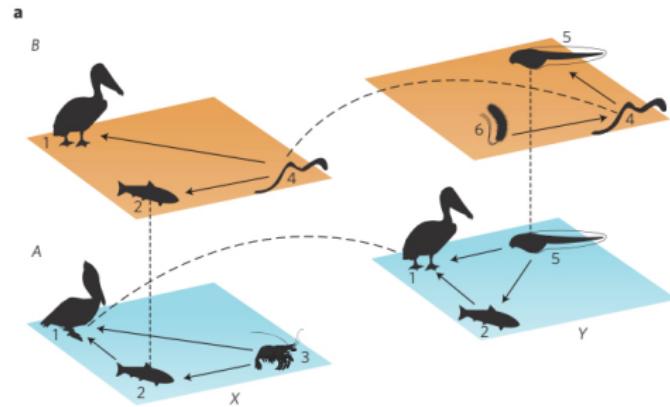
November 1, 2018 (v0.0.5) Software Open Access

ofionnadh/radiowinds: Calculating Thermal Bremsstrahlung Emission from Stellar Winds

Dúalta Ó Flannagáin

View

# Reality



# SCIENTIFIC DATA

A graphic consisting of four rows of binary digits (0s and 1s) in a light blue color, positioned to the right of the journal title.

OPEN

## Comment: If these data could talk

Thomas Pasquier<sup>1</sup>, Matthew K. Lau<sup>2</sup>, Ana Trisovic<sup>3,4</sup>, Emery Boose<sup>2</sup>, Ben Couturier<sup>2</sup>, Mercè Crosas<sup>5</sup>, Aaron M. Ellison<sup>2</sup>, Valerie Gibson<sup>4</sup>, Chris Jones<sup>4</sup> & Margo Seltzer<sup>1</sup>

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through *systematic* and *formal* records of the relationships among data sources, processes, datasets, publications and researchers.

Received: 12 April 2017

Accepted: 24 July 2017

Published: xx xxx 2017

# Reality: Common Ground



# Reality

*Most scientists don't want to produce software, they want to do science.*

# Wild-wild West of Statistical Software



# R-ube Goldberg Coding



# **Code Cleaning and Encapsulation**

*Let's get the computer to work for us and decrease error rates and increase sharing.*

# Code Cleaning and Encapsulation

CODE OCEAN DATA SCIENCE

DASHBOARD EXPLORE HELP + UPLOAD YOUR CODE

Search keyword, research field, title, author, DOI, etc.

More Capsules

**SOCIAL SCIENCES** Oct 2018

Betsy Levy Paluck, Seth Aerial Green, Donald P. Green

The contact hypothesis re-evaluated: code and data

This code reproduces the statistical analyses for 'The contact hypothesis re-evaluated' by Betsy Levy Paluck, ...

Behavioural Public Policy, 2018

**ENGINEERING** Jun 2018

Maninder Chitre

On Writing Reproducible and Interactive Papers

IEEE Journal of Oceanic Engineering, 2018

**MEDICAL SCIENCES** May 2018

James Andrew Watson, Chris Holmes

Exploratory subgroup analysis of the SEAQUAMAT trial using Random Forests:...

This computational notebook (written in R) runs an exploratory subgroup analysis on the SEAQUAMAT trial...

bioRxiv, 2018

**SOCIAL SCIENCES** Aug 2018

Tom Hardwicke, John Ioannidis

Mapping the Universe of Registered Reports

The Registered Reports format offers a substantial departure from traditional publishing models with the...

Nature Human Behaviour, 2018

**MEDICAL SCIENCES** Jul 2018

Restaurant schema

Error Seated Ordering Food

Restaurant at Table

Food

Table

**BIOLOGY** May 2018

**SOCIAL SCIENCES** Jun 2018

tvatobj

**EARTH SCIENCES** May 2018

Pavia University Classification Map GT

A circular icon with a speech bubble and a person icon.

# Code Cleaning and Encapsulation

The screenshot shows the Code Ocean platform interface. At the top, there's a navigation bar with 'CODE OCEAN' logo, 'DASHBOARD', 'EXPLORE', 'HELP', and a 'PREVIEW THE NEW CO' button. Below the navigation is a search bar and a file list.

**File List:**

- Extracting Diurnal Patterns of Real World Activity from Social Media (copy)
- semantic\_expansion.py
- usage\_example.py

**Code Editor:**

```
Code Ocean
Code
usage_example.py
...
1 import semantic_expansion as semexp
2 import sys, cPickle, glob
3 import numpy as np
4 import json as json
5
6 co_occurrences_file = './input/co_occur200plus.pkl'
7 word_df_file = './input/word_df.json'
8
9
10 # read configured number of points from input config file
11
12 print 'loading pickled co_occurrences file'
13 with open(co_occurrences_file, 'rb') as f:
14     co_occurrences = cPickle.load(f)
15
16 print 'loading pickled word document-frequency file'
17 with open(word_df_file, 'rt') as f:
18     word_df = json.load(f)
19
20 print 'generating background language model'
21 random_terms = np.random.choice(co_occurrences.keys(), size=10000, replace=False)
22 bk_pmi = semexp.agg_pmi(random_terms, co_occurrences, sort_results=True, top_n=1000)
23 bk_context_terms, _ = zip(*bk_pmi)
24 bk_context_terms_set = frozenset(bk_context_terms)
25 print 'loading stopwords'
26 stopwords_set = frozenset([l.rstrip() for filename in glob.glob('../input/stopwords_*') for
27     l in open(filename)])
28
29 if __name__ == '__main__':
30     coffee_terms = ['coffee', '#coffee', 'starbucks', '#starbucks', \
31                     'espresso', '#espresso', 'lovecoffee', '#lovecoffee', \
32                     'caffineaddict', '#caffineaddict', 'venti', '#venti', \
33                     'starbucks', '#starbucks', 'mugs', '#mugs', 'latte', \
34                     '#latte', 'caf', '#caf', 'coffeebean', '#coffeebean']
35     sorted_results = dict(semexp.context.pmi(coffee_terms, co_occurrences))
36     print sorted_results
37
38     from wordcloud import WordCloud
39     wordcloud = WordCloud(width=800, height=600, background_color="white").generate_from_f
40     wordcloud.to_image().save("../output/words.png", format='png')
```

**Data:**

File	Size
co_occur200plus.pkl	197.76 MB
stopwords_danish	424 B
stopwords_dutch	453 B
stopwords_english	725 B
stopwords_finnish	1.54 KB
stopwords_forbidden_permutations	16 B
stopwords_french	805 B
stopwords_german	1.31 KB
stopwords_hungarian	1.19 KB
stopwords_indonesian	5.37 KB
stopwords_italian	1.61 KB

**Results:**

- Run 1177939 (Nov 02, 2018 | 12:58)  
Run environment setup failed
- Run 1177900 (Nov 02, 2018 | 12:58)

A blue circular icon with a white 'C' is located in the bottom right corner.

# Code Cleaning and Encapsulation

[About](#) ▾[News](#) ▾[Documentation](#)[GitHub Project](#)[Examples](#)[PyPI Packages](#) ▾[Installers](#) ▾[Tweets about ReproZip](#)

## Automatically pack your research to be run elsewhere!

---

ReproZip allows you to pack your research along with all necessary data files, libraries, environment variables and options.

Then anybody can reproduce the research on a different machine, without tracking down and installing the dependencies, or even having to run the same operating system!

## How It Works

---

ReproZip works by tracing the systems calls used by the experiment to automatically identify which files should be included. You can review and edit this list and the metadata before creating the final package file. Packages can be reproduced in different ways, including chroot environments, [Vagrant](#)-built virtual machines, and [Docker](#) containers; more can be added through plugins.

# Code Cleaning and Encapsulation

## Sharing and Preserving Computational Analyses for Posterity with *encapsulator*

**Thomas Pasquier**

University of Cambridge

**Matthew K. Lau and**

**Xueyuan Han**

Harvard University

**Elizabeth Fong and**

**Barbara S. Lerner**

Mount Holyoke College

**Emery R. Boose, Mercè Crosas, Aaron M. Ellison, and Margo Seltzer**

Harvard University

**Editors:** Lorena A. Barba,  
[labarba@gwu.edu](mailto:labarba@gwu.edu);  
George K. Thiruvathukal,  
[gkt@cs.luc.edu](mailto:gkt@cs.luc.edu)

Reproducibility has become a recurring topic of discussion in many scientific disciplines.<sup>1</sup> Although it might be expected that some studies will be difficult to reproduce, recent conversations highlight important aspects of the scientific endeavor that could be improved to facilitate reproducibility. Open data and open source software are two important parts of a concerted effort to achieve reproducibility.<sup>2</sup> However, multiple publications point out these approaches' shortcomings,<sup>3,4</sup> such as the identification of dependencies, poor documentation of the installation processes, "code rot," failure to capture dynamic inputs, and technical barriers.

In prior work,<sup>5</sup> we pointed out that open data and open source software alone are insufficient to ensure reproducibility, as they do not capture information about the computational execution, that is, the "process" and context that produced the results using the data and code. In keep-

# Code Cleaning and Encapsulation

*RClean*: Simplify code based on specified results.

Lau 2018 CRAN

*Encapsulator*: generate a cleaned capsule.

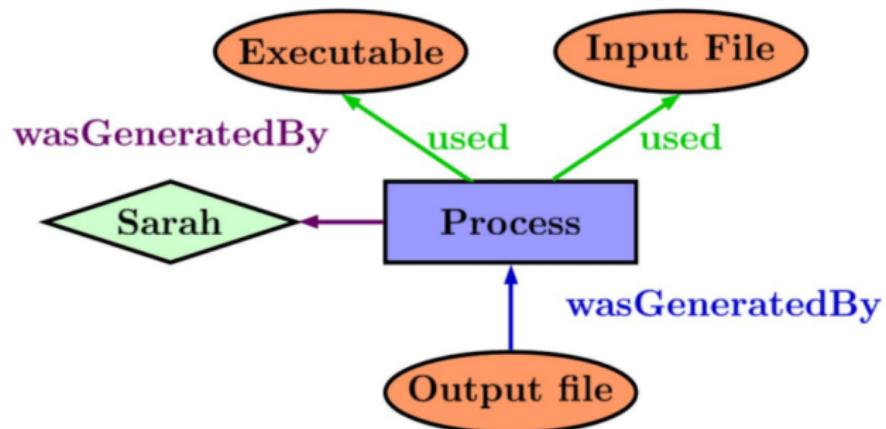
Pasquier et al. 2017 IEEE CISE

- ▶ Capsule = all necessary software and data
- ▶ Cleaned = organize files, remove non-essential code and re-format

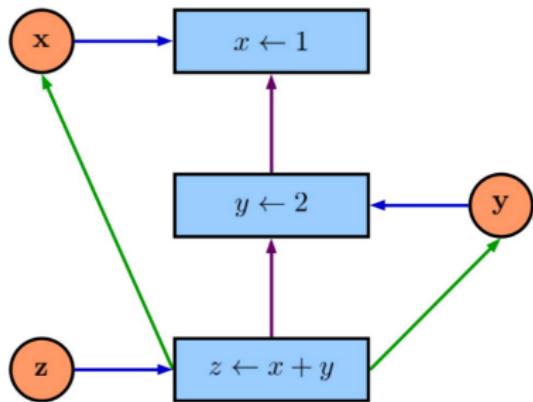
## Tools: Encapsulator

- ▶ near stream-of-consciousness coding that follows a train of thought in script development,
- ▶ output to console that is not written to disk,
- ▶ intermediate objects that are abandoned,
- ▶ library and new data calls throughout the script,
- ▶ output written to disk but not used in final documents,
- ▶ code that is not modularized,
- ▶ code that is syntactically correct but not particularly comprehensible.

# Encapsulation: Under-the-hood



# Rclean and Encapsulator



# Encapsulation: Under-the-hood



Figure 4. Provenance graph corresponding to a small R script (approximately 60 lines of code).

# Encapsulation: Under-the-hood

Example: Messycode

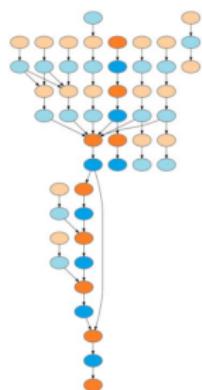


Figure 5. Data dependency transformation of the provenance graph shown in Figure 4.

# Tools: Encapsulator

Basic Usage (current paradigm):

1. Code as usual in your normal environment while recording provenance
2. Run encapsulator from the console
3. List desired results

# Tools: Encapsulator

Basic Usage (current paradigm):

1. Code as usual in your normal environment while recording provenance
2. Run encapsulator from the console
3. List desired results

**Product** = Capsule containing essential code and data organized following project best practices inside a virtual machine

**\*\* Software should not limit science \*\***

*The*  
**HOBBIT**  
or there and back again  
J.R.R. TOLKIEN



THE ENCHANTING PRELUDE TO THE LORD OF THE RINGS

Computation will not replace good scientific thought or practice.

Computation will not replace good scientific thought or practice.  
But hopefully it can help.

# Thanks for listening!

Resources:

- ▶ Databases: *Github Data*, *Data Dryad*, *Figshare*, *Dataverse*
- ▶ Open Source Communities: *ROpenSci*, *RStudio*, *Center for Open Science*, *Transparency in Ecology and Evolution*
- ▶ Tools: *Reprozip*, *CodeOcean*, *Rclean*

Contact Info:

- ▶ Email: [matthewklau@fas.harvard.edu](mailto:matthewklau@fas.harvard.edu)
- ▶ Github: MKLau

*Much of this work was supported by NSF SSI-1450277 (End-to-End Provenance) and ACI-1448123 (Citation++). More details are available at <https://projects.iq.harvard.edu/provenance-at-harvard>*

