

CODECAMP_IASI

/ 26 October 2019

/ Hotel International

Global partners

Cognizant
Softvision



Diamond partners



Platinum partners



REGINA MARIA
RETRAGIA PREȚIA DE SÂNCȘTE

Gold partners



Wellness Partners



Liked By



Media partners





Research Reproducibility - An Opportunity for Software Engineers

Adrian-Tudor Pănescu
Lead Integrations Engineer @ Figshare







Non-reproducibility - anecdotal evidence




THE LANCET

EARLY REPORT | VOLUME 351, ISSUE 9103, P637-641, FEBRUARY 28, 1998

RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Dr A J Wakefield, FRCS  SH Murch, MB  A Anthony, MB  J Linnell, PhD  M Casson, MRCP  M Malik, MRCP  et al.
[Show all authors](#)

Published: February 28, 1998 • DOI: [https://doi.org/10.1016/S0140-6736\(97\)11096-0](https://doi.org/10.1016/S0140-6736(97)11096-0)  [Add to Library](#)

Summary

Introduction

Patients and methods

Results

Discussion

References

Article Info

Figures

Tables

Summary

Background

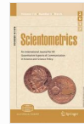
We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods

12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings

Non-reproducibility - scientific evidence





[Scientometrics](#)

March 2017, Volume 110, Issue 3, pp 1471–1493 | [Cite as](#)

Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines

Authors

[Authors and affiliations](#)

Jennifer A. Byrne , Cyril Labbé 

Abstract

Comparing 5 publications from China that described knockdowns of the human *TPD52L2* gene in human cancer cell lines identified unexpected similarities between these publications, flaws in experimental design, and mis-matches between some described experiments and the reported results. Following communications with journal editors, two of these *TPD52L2* publications have been retracted. One retraction notice stated that while the authors claimed that the data were original, the experiments had been out-sourced to a biotechnology company. Using search engine queries, automatic text-analysis, different similarity measures, and further visual inspection, we identified 48 examples of highly similar papers describing single gene knockdowns in 1–2 human cancer cell lines that were all published by investigators from China. The incorrect use of a particular *TPD52L2* shRNA sequence as a negative or non-targeting control was identified in 30/48 (63%) of these publications, using a combination of Google Scholar searches and visual inspection. Overall, these results suggest that some publications describing the effects of single gene knockdowns in human cancer cell lines may include the results of experiments that were not performed by the authors. This has serious implications for the validity of such results, and for their application in future research.



Non-reproducibility - scientific evidence

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund, Thomas E. Nichols, and Hans Knutsson

PNAS July 12, 2016 113 (28) 7900-7905; first published June 28, 2016 <https://doi.org/10.1073/pnas.1602413113>

Significance

Functional MRI (fMRI) is 25 years old, yet surprisingly its most common statistical methods have not been validated using real data. Here, we used resting-state fMRI data from 499 healthy controls to conduct 3 million task group analyses. Using this null data with different experimental designs, we estimate the incidence of significant results. In theory, we should find 5% false positives (for a significance threshold of 5%), but instead we found that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in false-positive rates of up to 70%. These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results.

What about computer science?

Table 6: Summary of results from various studies of repeatability and reproducibility.

Reference	What/Who was studied	What was measured	Results
Kovacevic [17]	15 papers published in the <i>IEEE Transactions on Image Processing</i> .	How well algorithms were explained and whether code and data were available.	All algorithms had proofs, 0% had code available, 33% had data available.
Vandewalle et al. [29]	All the 134 papers published in <i>IEEE Transactions on Image Processing</i> in 2004.	Reproducibility as measured by 2-3 reviewers.	9% of the papers had code available online and 33% had data.
Stodden [26]	Survey responses from 134 registrants affiliated with American universities at NIPS conference.	Proportion of registrants comfortable with sharing post-publication code on the web vs. proportion publishing some code on their web site.	74% are comfortable with sharing, 30% have code on web site.
Table 1 (page 9 of this report)	Artifact evaluation outcomes for the 268 papers that were accepted in the seven conferences for which we have complete information.	Submitted and accepted artifacts as a percentage of accepted papers.	43.3% submitted, 29.5% accepted.
Klein et al. [15]	Formal modeling and mechanized reasoning of nine papers published in ICFP 2009.	Proportion of papers in which no mistakes were found.	0%.
This study	Examination of 402 Computer Systems papers backed by code.	Proportion of papers with shared code, and the extent to which shared code builds successfully in 30 minutes, with extra effort, or by the authors.	56.2% shared, 32.3% builds in 30 minutes, 48.3% builds with extra effort, 54.0% builds by author.

Why is reproducibility important?

nature
cell biology

Editorial | Published: 25 October 2018

The challenge of the post-truth era

Nature Cell Biology |

3305 Accesses | 4

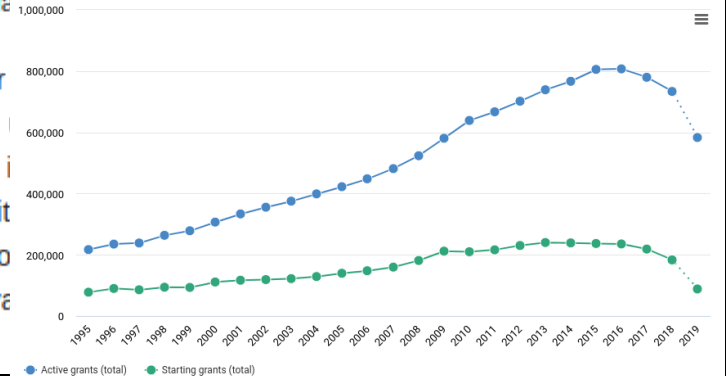
Science denial
conclusions ap
emotion and is
world, scientist
engagement.

Science by press conference

From Wikipedia, the free encyclopedia

Science by press conference (or **science by press conference**) is a practice by which scientists put an emphasis on their research in the media.^[1] The term is intended to associate the target with a lack of questionable scientific merit who they are unlikely to win the approval of the scientific community.

Show years 1995 to 2019 ▾





Opportunities for software engineers

- Develop (**or retrofit**) the technologies required for achieving reproducibility



Opportunities for software engineers

- Develop (**or retrofit**) the technologies required for achieving reproducibility; examples:
 - Repositories
 - Compute environments
 - Preservation

Repositories - today

Explore GitHub


Science

Scientists around the world are working together to solve some of the biggest questions in research. Take a look at some of the examples featured here to find out more.


📁 12 repositories 🔗 4 languages 🕒 Last updated on Nov 30, 2016

Stars

Language




dfm / emcee




emcee - seriously kick-ass MCMC... it's big in astrophysics. It has been used to study [planets like our own Earth](#), [binary stars](#), the [Milky Way](#), and even [astrophotographers](#).

Python ★ 955 🗄 354 Updated 4 hours ago




cms-sw / cmssw




In 2012 researchers at [CERN](#) discovered the [Higgs Boson](#), the elementary particle responsible for particles having mass. One of the experiments that made this discovery, [CMS](#), is a collaboration of more than 3,800 people from 42 different countries. [cmssw](#) is the software that powers the data-management pipeline of CMS and was built by 200 different people on GitHub.

C++ ★ 702 🗄 3,058 Updated 9 minutes ago



astropy / astropy



Astropy is a community effort in the astrophysics community to develop a single core package for Astronomy in Python. Three years after the [first commit](#) more than [75 researchers](#) have contributed over



Repositories - challenges

- Ease of use - not all users in the scientific community have the required technical skills, how do we cater to them?
- Applicability - research outputs are diverse (data - figshare.com, preprints - arxiv.org, MRIs - openfmri.org etc.); repositories should handle all, and link between the sources.
- **Discovery** - how do we find all the above outputs (i.e. Google Scholar for *all* research)?



Detour - Reproducibility vs. Replicability

- Reproducibility - same code, same data, different **analyst**
- Replicability - same code, different **data**

Reproducibility - code

```
fits_clinical = fits.join(data.groupby('subj_idx').mean()[clinical_rating_cols]).dropna()
#fits_clinical = fits_clinical.query('a_diff > -.5')
x = fits_clinical['motorscore - Total motor score (TMS)']
fits_clinical['log(motorscore)'] = np.log(x+np.sqrt(x**2+1))
```

```
In [2]: fits_orig = pd.read_csv('fits_hd.csv', index_col=0)
fits_orig.head()
#accumodel.models.WaldAntiPDA
#accumodel.estimators.estimate_sample(x[1], depends_on='a': 'cond', init_vals=x[2], name=x[0])
```

Out[2]:

	t	a(cong)	a(incong)	v_pro	v_stop	v_anti	t_anti
378155	7.802374e-02	1.039099	1.158032	3.494729	2.239062	3.975630	1.402957e-01
560225	5.392383e-14	0.938888	0.853591	3.265282	1.806058	5.662315	2.427171e-01
797677	2.238457e-01	0.605942	1.081538	2.657450	4.091814	4.722220	3.479237e-13
1146341	3.669387e-12	1.589054	1.941488	4.059399	0.833342	8.835723	1.452986e-01
1374832	1.069551e-01	1.837055	2.275932	4.506693	5.224003	6.540063	2.822893e-13

```
In [3]: data_hd = pd.read_csv('hd_antisaccade.csv')
data_hd = data_hd.loc[data_hd.cond != 'prosaccade-only']
data_hd.loc[data_hd.cond == 'antisaccade', 'cond'] = 'incong'
data_hd.loc[data_hd.cond == 'prosaccade', 'cond'] = 'cong'
data_hd = hddm.utils.flip_errors(data_hd)
data_hd.loc[data_hd.cond == 'incong', 'rt'] *= -1
```

```
In [4]: import accumodel
accumodel.set_estimator(accumodel.estimators.OptimizeEstimator(accumodel.models.WaldAntiStop))

Initializing random numbers of (10000, 2)
```

Cite

Download (2.17 MB)

Share

Embed

+ Collect ...

DataCite

Select your citation style and then place

your mouse over the citation text to
select it.

Wiecki, Thomas (2015): Notebook file containing all analyses to reproduce plots and statistics.
figshare. Software.

<https://doi.org/10.6084/m9.figshare.2008401.v1>

Notebook file containing all analyses to reproduce
plots and statistics

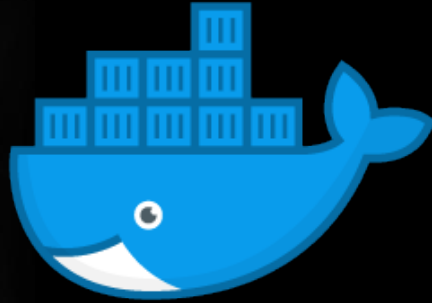
Software posted on 11.12.2015, 12:37 by Thomas Wiecki

913
views

63
downloads

0
citations

Reproducibility - environment



docker



**ANACONDA
CLOUD**

Reproducibility - paper

This is a Reproducible document. See the original article or source.

Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

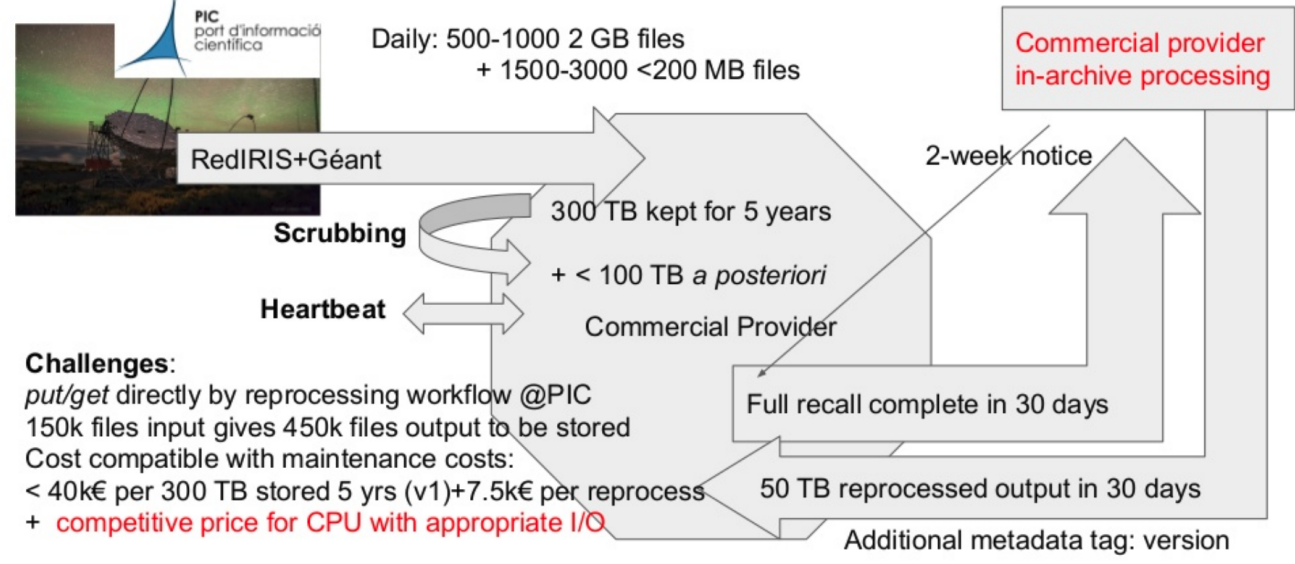
L Michelle Lewis, Meredith C Edwards, Zachary R Meyers, C Conover Talbot Jr, Haiping Hao, David Blum, Reproducibility Project: Cancer Biology

As part of the [Reproducibility Project: Cancer Biology](#), we published a Registered Report (Blum et al., 2015), that described how we intended to replicate selected experiments from the paper 'Transcriptional amplification in tumor cells with elevated c-Myc' (Lin et al., 2012). Here we report the results. We found overexpression of c-Myc increased total levels of RNA in P493-6 Burkitt's lymphoma cells; however, while the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication. Digital gene expression analysis for a set of genes was also performed on P493-6 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression, similar to the original study (Figure 3F; Lin et al., 2012). Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012). Treating the data as paired, we found a statistically significant increase in gene expression for both active and silent genes upon c-Myc induction, with the change in gene expression greater for active genes compared to silent genes. Finally, we report meta-analyses for each result.

NOTE: This is a demonstration of a reproducible view of an existing eLife article. You can inspect the code that was used to generate the figures, make changes and re-run the code. For technical reasons the article differs slightly from the [original article](#). The reference list is missing, references are external links and figure supplements are missing.

Preservation and continuity

+ In-archive processing scenario workflow/scenario



Preservation and continuity

We are building the universal software archive



Collect
Preserve
Share

We **collect** and **preserve** software in source code form, because software embodies our technical and scientific knowledge and humanity cannot afford the risk of losing it.

Software is a precious part of our cultural heritage. We curate and make accessible all the software we collect, because only by **sharing** it we can guarantee its preservation in the very long term.

[Discover our mission](#)

Do we already have your code?

We harvest publicly available source code from many software projects and keep up with development happening there. As of today our archive already contains and keeps safe for you:

Source files
6,034,792,688

Commits
1,343,174,362

Projects
89,421,958

Preservation is not just storage

US military will no longer use floppy disks to coordinate nuke launches

It now has a "highly-secure solid state digital storage solution."



Steve Dent, @stevetdent
10.18.19 in Politics

48

Comments

114832

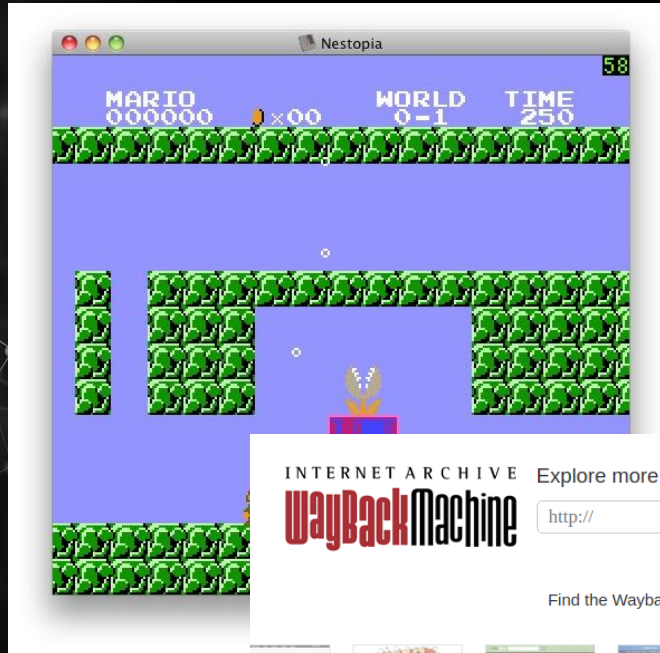
Shares



Robert Gauthier via Getty Images

<https://www.engadget.com/2019/10/18/us-military-nuclear-missiles-floppy-disks>

Preservation is not just storage



```
uint32_t chunkLen=[fh readUInt32BE];
off_t nextChunk=[fh offsetInFile]+(chunkLen*3)&~3;

// At this point, I'd like to take a moment to speak to you about the Adobe PSD format.
// PSD is not a good format. PSD is not even a bad format. Calling it such would be an
// insult to other bad formats, such as PCX or JPEG. No, PSD is an abysmal format. Having
// worked on this code for several weeks now, my hate for PSD has grown to a raging fire
// that burns with the fierce passion of a million suns.
// If there are two different ways of doing something, PSD will do both, in different
// places. It will then make up three more ways no sane human would think of, and do those
// too. PSD makes inconsistency an art form. Why, for instance, did it suddenly decide
// that *these* particular chunks should be aligned to four bytes, and that this alignment
// should *not* be included in the size? Other chunks in other places are either unaligned,
// or aligned with the alignment included in the size. Here, though, it is not included.
// Either one of these three behaviours would be fine. A sane format would pick one. PSD,
// of course, uses all three, and more.
// Trying to get data out of a PSD file is like trying to find something in the attic of
// your eccentric old uncle who died in a freak freshwater shark attack on his 58th
// birthday. That last detail may not be important for the purposes of the simile, but
// at this point I am spending a lot of time imagining amusing fates for the people
// responsible for this Rube Goldberg of a file format.
// Earlier, I tried to get a hold of the latest specs for the PSD file format. To do this,
// I had to apply to them for permission to apply to them to have them consider sending
// if some document or
// this process so
// s abomination. I
// e, but if I had done
// t them all on fire.
// specs, and launch
```

INTERNET ARCHIVE

WayBackMachine

Explore more than 388 billion web pages saved over time

http://

BROWSE HISTORY

Find the Wayback Machine useful?

DONATE



Opportunities for software engineers

- Develop (**or retrofit**) the technologies required for achieving reproducibility
- Lead by example





Lead by example, please.

[Comment](#) | [Open Access](#) | [Published: 23 August 2016](#)

Gene name errors are widespread in the scientific literature

[Mark Ziemann](#), [Yotam Eren](#) & [Assam El-Osta](#) 

[Genome Biology](#) **17**, Article number: 177 (2016) | [Download Citation](#) 

102k Accesses | **34** Citations | **1828** Altmetric | [Metrics](#) 

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Lead by example, please.

The dumb reason your fancy Computer Vision app isn't working: Exif Orientation



Adam Geitgey [Follow](#)

Oct 9 · 6 min read



Camera

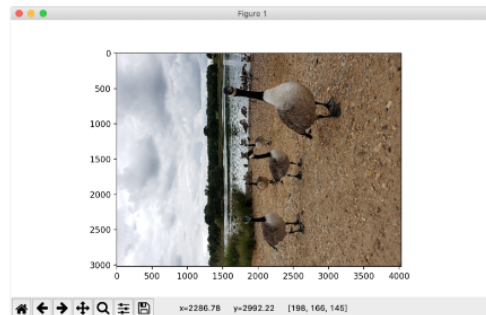


Image In Memory



Lead by example

- Ensure that CS research is reproducible!
- Have more software engineers specialize in scientific computing, emphasize the reproducibility aspect
 - Should we take the lead on *teaching* reproducibility? See software-carpentry.org
 - Transfer engineering principles to scientific software: testing, versioning, etc.
- Apply new *trends* to reproducibility - distributed systems, microservices, blockchain, linked data



Lead by example

Programming as a profession is only moderately interesting. It can be a good job, but you could make about the same money and be happier running a fast food joint. You're much better off using code as your secret weapon in another profession.

People who can code in the world of technology companies are a dime a dozen and get no respect. People who can code in biology, medicine, government, sociology, physics, history, and mathematics are respected and can do amazing things to advance those disciplines.

CODECAMP_❤️ FEEDBACK



codecamp.ro/feedback

tudor@figshare.com