ReScience

KONRAD HINSEN
NICOLAS P. ROUGIER
GITHUB.RESCIENCE.IO

MAY 2021

Reproducible science is good.
Replicated science is better.

SCIENCE TODAY

~~Science~~ - Broken Science

Paywalled, proprietary tools & software, non shared data, non-reproducible, anonymous peer-review, publish or perish (alone)!

PAYWALLED (32$)

Nature 171 (1953)
Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid
J.D. WATSON & F. H. C. CRICK.

I have heard from graduate students opting out of academia, assistant professors afraid to come up for tenure, mid-career people wondering how to protect their labs, and senior faculty retiring early, all because of methodological terrorism.

APS Observer (2016)

METHODOLOGICAL TERRORISM

A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited.There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as research parasites

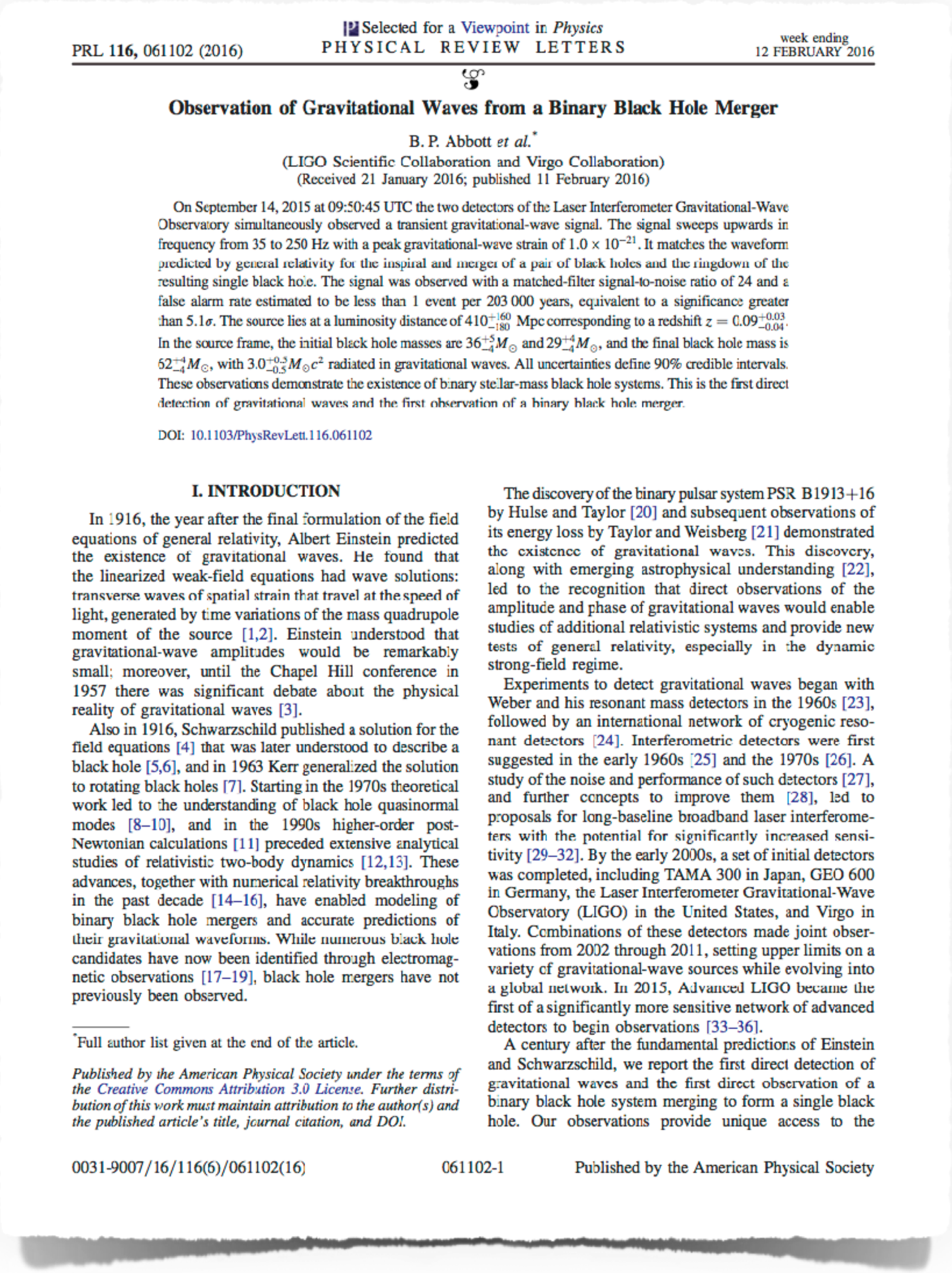
The New England Journal of Medicine (2016)

RESEARCH PARASITES

SCIENCE TOMORROW

~~Open Science~~ - Science

Open Access, Open Source, Open Data, Open Methodology, Open Education, Open Peer-review, Much more fun & efficient!



Selected for a Viewpoint in *Physics*

PRL **116,** 061102 (2016) PHYSICAL REVIEW LETTERS week ending 12 FEBRUARY 2016

Observation of Gravitational Waves from a Binary Black Hole Merger

B. P. Abbott *et al.*[*]

(LIGO Scientific Collaboration and Virgo Collaboration)

(Received 21 January 2016; published 11 February 2016)

On September 14, 2015 at 09:50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal sweeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a matched-filter signal-to-noise ratio of 24 and a false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a significance greater than 5.1σ. The source lies at a luminosity distance of 410^{+160}_{-180} Mpc corresponding to a redshift $z = 0.09^{+0.03}_{-0.04}$. In the source frame, the initial black hole masses are $36^{+5}_{-4} M_\odot$ and $29^{+4}_{-4} M_\odot$, and the final black hole mass is $62^{+4}_{-4} M_\odot$, with $3.0^{+0.5}_{-0.5} M_\odot c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

DOI: 10.1103/PhysRevLett.116.061102

I. INTRODUCTION

In 1916, the year after the final formulation of the field equations of general relativity, Albert Einstein predicted the existence of gravitational waves. He found that the linearized weak-field equations had wave solutions: transverse waves of spatial strain that travel at the speed of light, generated by time variations of the mass quadrupole moment of the source [1,2]. Einstein understood that gravitational-wave amplitudes would be remarkably small; moreover, until the Chapel Hill conference in 1957 there was significant debate about the physical reality of gravitational waves [3].

Also in 1916, Schwarzschild published a solution for the field equations [4] that was later understood to describe a black hole [5,6], and in 1963 Kerr generalized the solution to rotating black holes [7]. Starting in the 1970s theoretical work led to the understanding of black hole quasinormal modes [8–10], and in the 1990s higher-order post-Newtonian calculations [11] preceded extensive analytical studies of relativistic two-body dynamics [12,13]. These advances, together with numerical relativity breakthroughs in the past decade [14–16], have enabled modeling of binary black hole mergers and accurate predictions of their gravitational waveforms. While numerous black hole candidates have now been identified through electromagnetic observations [17–19], black hole mergers have not previously been observed.

The discovery of the binary pulsar system PSR B1913+16 by Hulse and Taylor [20] and subsequent observations of its energy loss by Taylor and Weisberg [21] demonstrated the existence of gravitational waves. This discovery, along with emerging astrophysical understanding [22], led to the recognition that direct observations of the amplitude and phase of gravitational waves would enable studies of additional relativistic systems and provide new tests of general relativity, especially in the dynamic strong-field regime.

Experiments to detect gravitational waves began with Weber and his resonant mass detectors in the 1960s [23], followed by an international network of cryogenic resonant detectors [24]. Interferometric detectors were first suggested in the early 1960s [25] and the 1970s [26]. A study of the noise and performance of such detectors [27], and further concepts to improve them [28], led to proposals for long-baseline broadband laser interferometers with the potential for significantly increased sensitivity [29–32]. By the early 2000s, a set of initial detectors was completed, including TAMA 300 in Japan, GEO 600 in Germany, the Laser Interferometer Gravitational-Wave Observatory (LIGO) in the United States, and Virgo in Italy. Combinations of these detectors made joint observations from 2002 through 2011, setting upper limits on a variety of gravitational-wave sources while evolving into a global network. In 2015, Advanced LIGO became the first of a significantly more sensitive network of advanced detectors to begin observations [33–36].

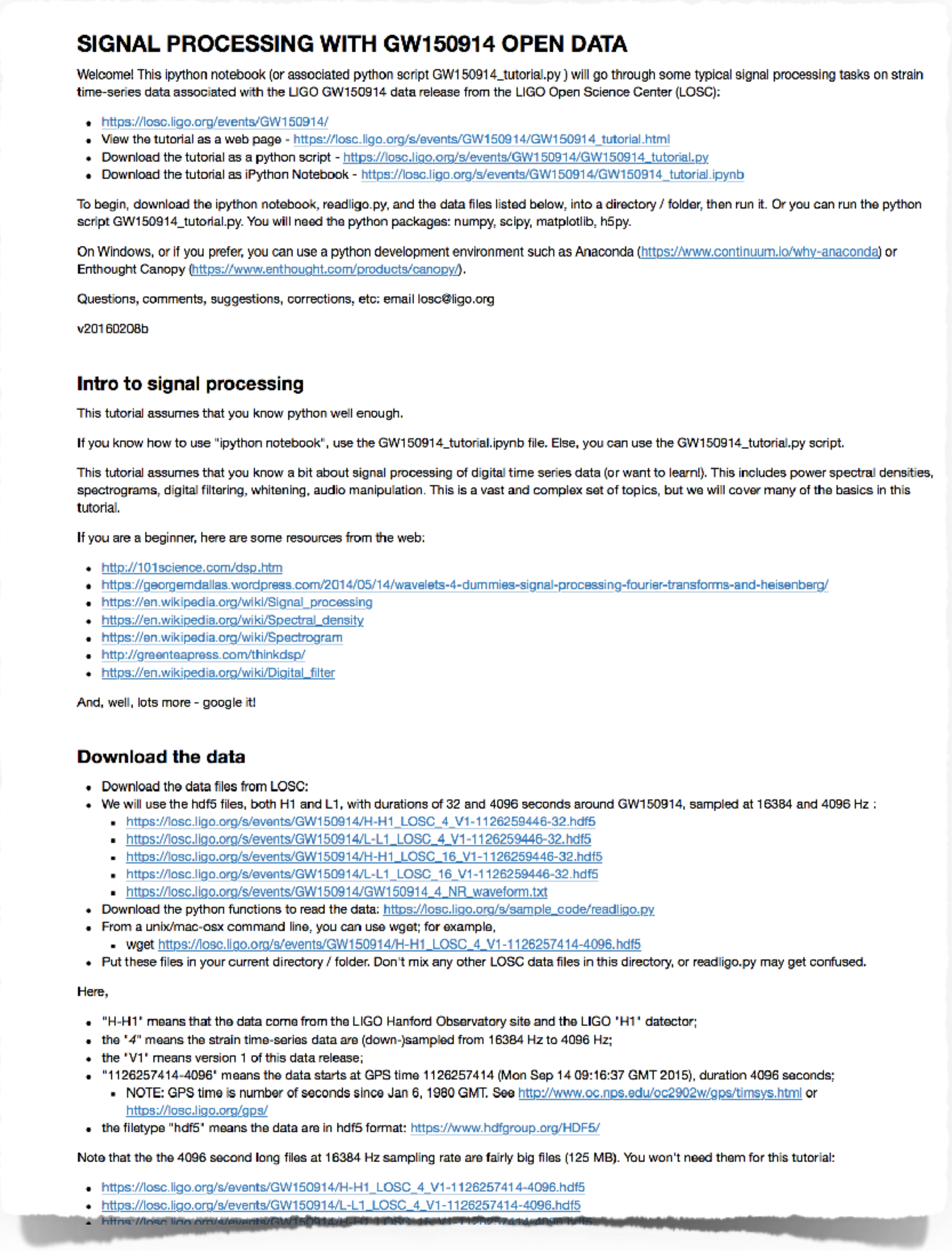
A century after the fundamental predictions of Einstein and Schwarzschild, we report the first direct detection of gravitational waves and the first direct observation of a binary black hole system merging to form a single black hole. Our observations provide unique access to the

[*] Full author list given at the end of the article.

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

0031-9007/16/116(6)/061102(16) 061102-1 Published by the American Physical Society

Original Article



SIGNAL PROCESSING WITH GW150914 OPEN DATA

Welcome! This ipython notebook (or associated python script GW150914_tutorial.py) will go through some typical signal processing tasks on strain time-series data associated with the LIGO GW150914 data release from the LIGO Open Science Center (LOSC):

- https://losc.ligo.org/events/GW150914/
- View the tutorial as a web page - https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html
- Download the tutorial as a python script - https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.py
- Download the tutorial as iPython Notebook - https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.ipynb

To begin, download the ipython notebook, readligo.py, and the data files listed below, into a directory / folder, then run it. Or you can run the python script GW150914_tutorial.py. You will need the python packages: numpy, scipy, matplotlib, h5py.

On Windows, or if you prefer, you can use a python development environment such as Anaconda (https://www.continuum.io/why-anaconda) or Enthought Canopy (https://www.enthought.com/products/canopy/).

Questions, comments, suggestions, corrections, etc: email losc@ligo.org

v20160208b

Intro to signal processing

This tutorial assumes that you know python well enough.

If you know how to use "ipython notebook", use the GW150914_tutorial.ipynb file. Else, you can use the GW150914_tutorial.py script.

This tutorial assumes that you know a bit about signal processing of digital time series data (or want to learn!). This includes power spectral densities, spectrograms, digital filtering, whitening, audio manipulation. This is a vast and complex set of topics, but we will cover many of the basics in this tutorial.

If you are a beginner, here are some resources from the web:

- http://101science.com/dsp.htm
- https://georgemdallas.wordpress.com/2014/05/14/wavelets-4-dummies-signal-processing-fourier-transforms-and-heisenberg/
- https://en.wikipedia.org/wiki/Signal_processing
- https://en.wikipedia.org/wiki/Spectral_density
- https://en.wikipedia.org/wiki/Spectrogram
- http://greenteapress.com/thinkdsp/
- https://en.wikipedia.org/wiki/Digital_filter

And, well, lots more - google it!

Download the data

- Download the data files from LOSC:
- We will use the hdf5 files, both H1 and L1, with durations of 32 and 4096 seconds around GW150914, sampled at 16384 and 4096 Hz :
  - https://losc.ligo.org/s/events/GW150914/H-H1_LOSC_4_V1-1126259446-32.hdf5
  - https://losc.ligo.org/s/events/GW150914/L-L1_LOSC_4_V1-1126259446-32.hdf5
  - https://losc.ligo.org/s/events/GW150914/H-H1_LOSC_16_V1-1126259446-32.hdf5
  - https://losc.ligo.org/s/events/GW150914/L-L1_LOSC_16_V1-1126259446-32.hdf5
  - https://losc.ligo.org/s/events/GW150914/GW150914_4_NR_waveform.txt
- Download the python functions to read the data: https://losc.ligo.org/s/sample_code/readligo.py
- From a unix/mac-osx command line, you can use wget; for example,
  - wget https://losc.ligo.org/s/events/GW150914/H-H1_LOSC_4_V1-1126257414-4096.hdf5
- Put these files in your current directory / folder. Don't mix any other LOSC data files in this directory, or readligo.py may get confused.

Here,

- "H-H1" means that the data come from the LIGO Hanford Observatory site and the LIGO "H1" datector;
- the "4" means the strain time-series data are (down-)sampled from 16384 Hz to 4096 Hz;
- the "V1" means version 1 of this data release;
- "1126257414-4096" means the data starts at GPS time 1126257414 (Mon Sep 14 09:16:37 GMT 2015), duration 4096 seconds;
  - NOTE: GPS time is number of seconds since Jan 6, 1980 GMT. See http://www.oc.nps.edu/oc2902w/gps/timsys.html or https://losc.ligo.org/gps/
- the filetype "hdf5" means the data are in hdf5 format: https://www.hdfgroup.org/HDF5/

Note that the the 4096 second long files at 16384 Hz sampling rate are fairly big files (125 MB). You won't need them for this tutorial:

- https://losc.ligo.org/s/events/GW150914/H-H1_LOSC_4_V1-1126257414-4096.hdf5
- https://losc.ligo.org/s/events/GW150914/L-L1_LOSC_4_V1-1126257414-4096.hdf5

Companion Notebook

A WELL KNOWN STORY

Once upon a time, there was a post-doc…

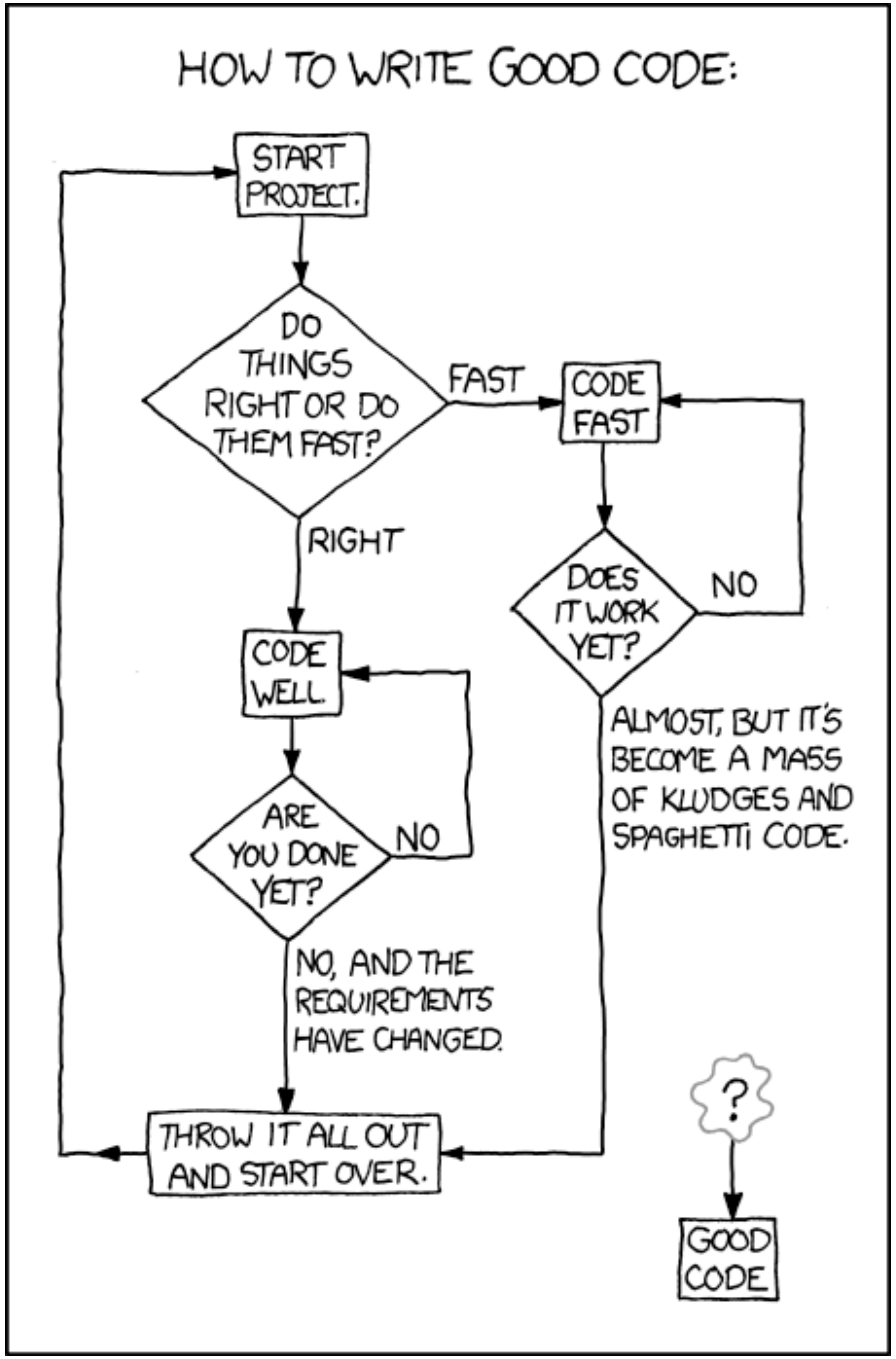
Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study. M. Guthrie, A. Leblois, A. Garenne, and T. Boraud, Journal of Neurophysiology, 109, 2013

Nice paper, good results, but…

- No public repository, no version control
- Sources were mixing actual computation and GUI code
- Model was split into a hundred files, main file 6,000 lines long
- Several configuration files, no data saved
- Model description included ambiguous information

Model was hardly reproducible.

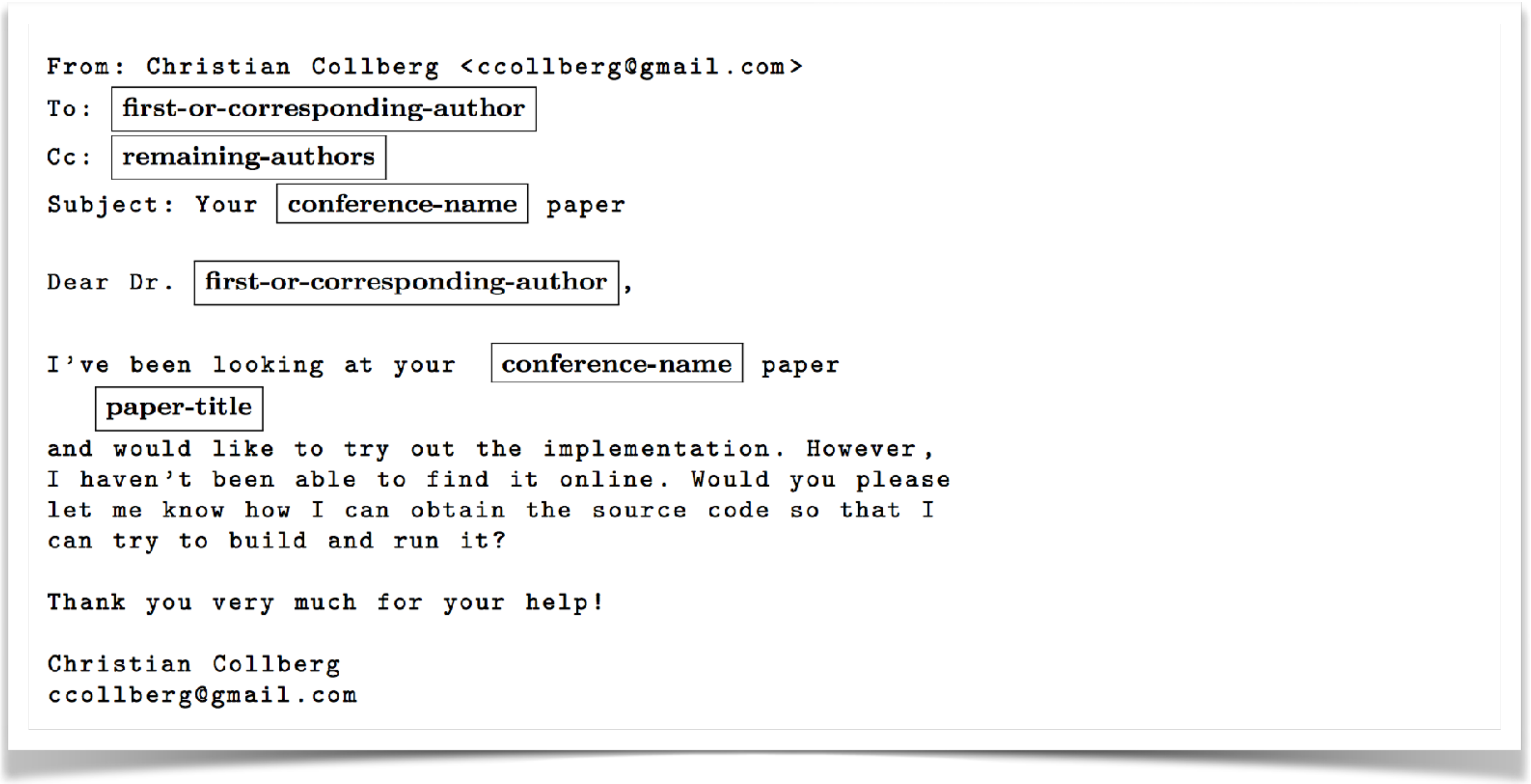
> *You can download our code from the URL supplied. Good luck downloading the only postdoc who can get it to run, though…*
> Ian Holmes


HOW TO WRITE GOOD CODE:
START PROJECT.
DO THINGS RIGHT OR DO THEM FAST?
FAST
CODE FAST
RIGHT
DOES IT WORK YET?
NO
CODE WELL
ALMOST, BUT IT'S BECOME A MASS OF KLUDGES AND SPAGHETTI CODE.
ARE YOU DONE YET?
NO
NO, AND THE REQUIREMENTS HAVE CHANGED.
?
THROW IT ALL OUT AND START OVER.
GOOD CODE

COLLBERG ET AL. 2014

Ma’am, the dog ate my program

We describe a study into the extent to which Computer Systems researchers share their code and data and the extent to which such code builds. Starting with 601 papers from ACM conferences and journals, we examine 402 papers whose results were backed by code. For 32.3% of these papers we were able to obtain the code and build it within 30 minutes; for 48.3% of the papers we managed to build the code, but it may have required extra effort; for 54.0% of the papers either we managed to build the code or the authors stated the code would build with reasonable effort.



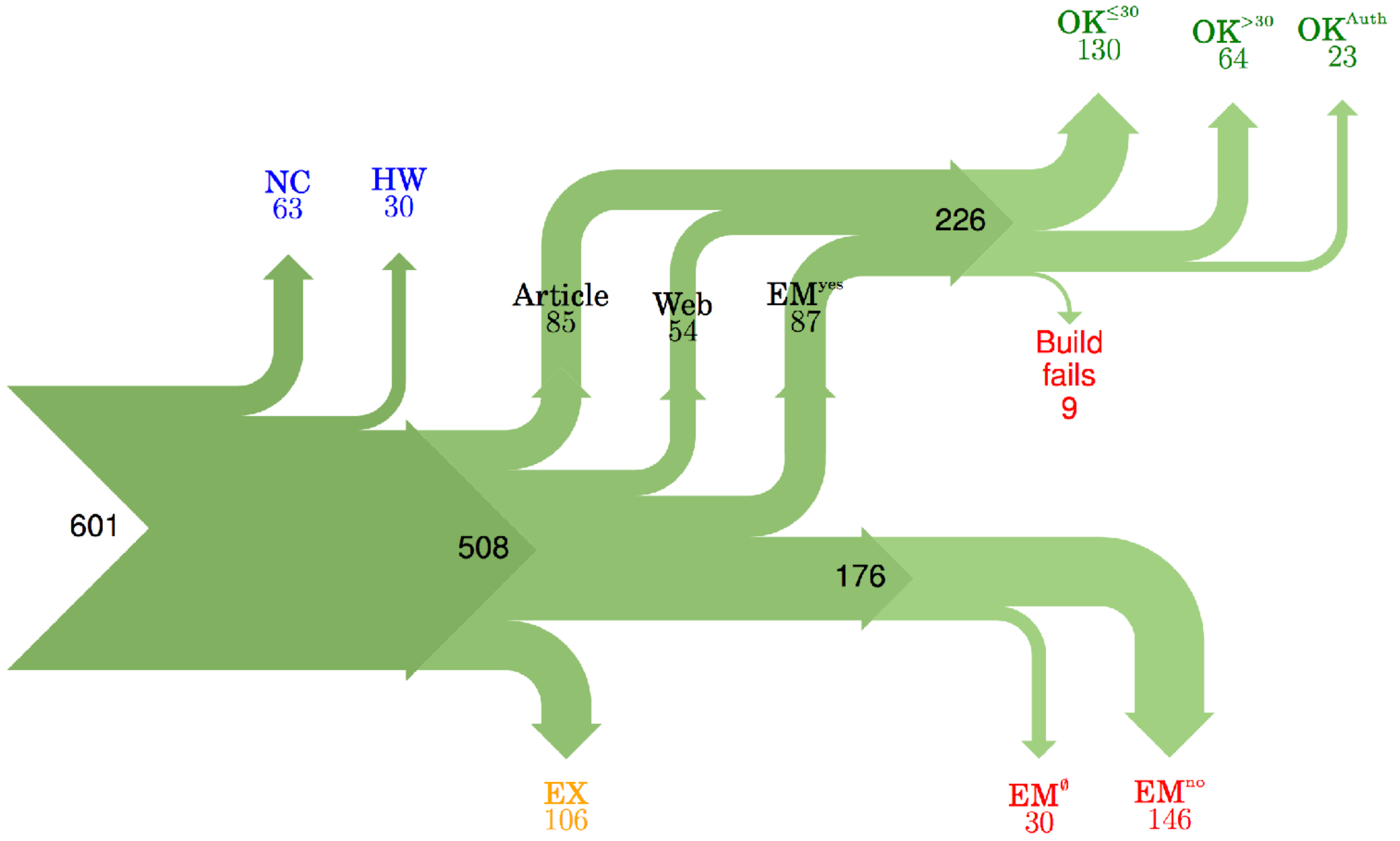
From: Christian Collberg <ccollberg@gmail.com>
To: first-or-corresponding-author
Cc: remaining-authors
Subject: Your conference-name paper

Dear Dr. first-or-corresponding-author,

I’ve been looking at your conference-name paper
paper-title
and would like to try out the implementation. However,
I haven’t been able to find it online. Would you please
let me know how I can obtain the source code so that I
can try to build and run it?

Thank you very much for your help!

Christian Collberg
ccollberg@gmail.com

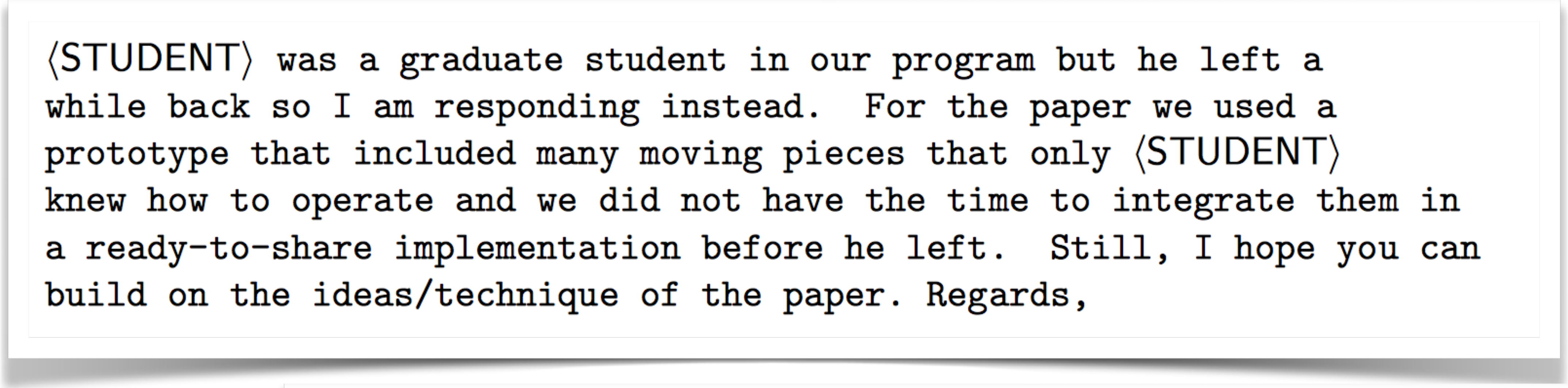


COLLBERG ET AL. 2014

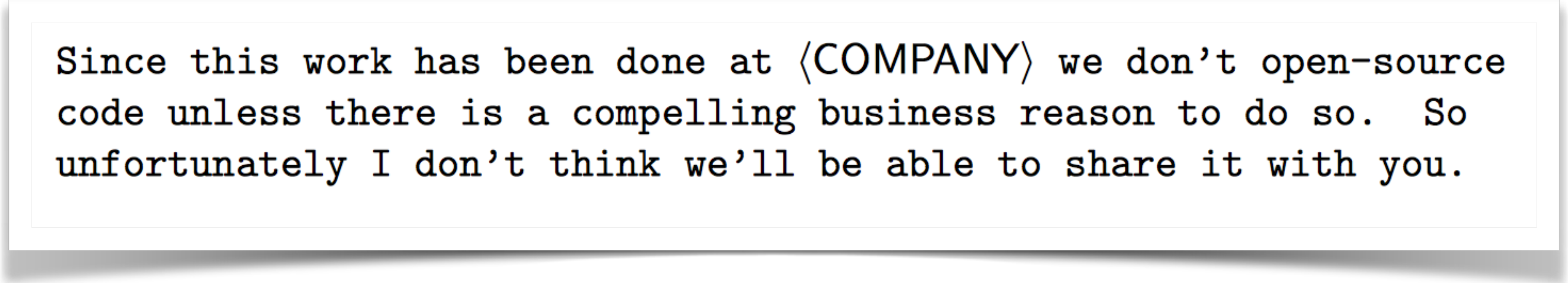
Ma’am, the dog ate my program

Reasons why code cannot be shared:

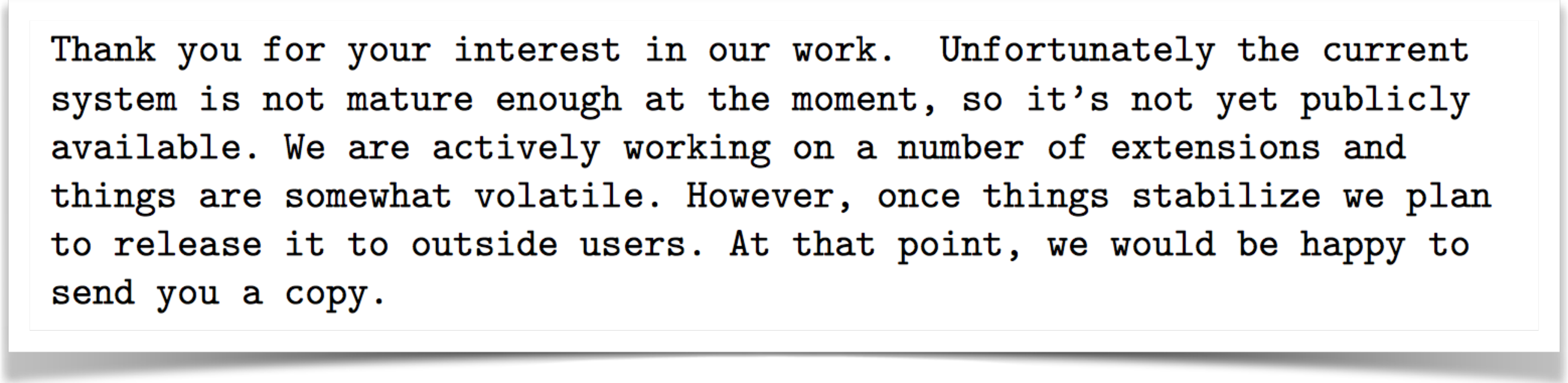
- → Versioning Problems
- → Code Will be Available Soon
- → No Intention to Release
- → Programmer Left
- → Bad Backup Practices
- → Commercial Code
- → Proprietary Academic Code
- → Industrial Lab Issues
- → Unavailable Subsystems
- → Multiple Reasons
- → Intellectual Property
- → Research vs. Sharing
- → Security and Privacy
- → Design Issues
- → Too Busy to Help



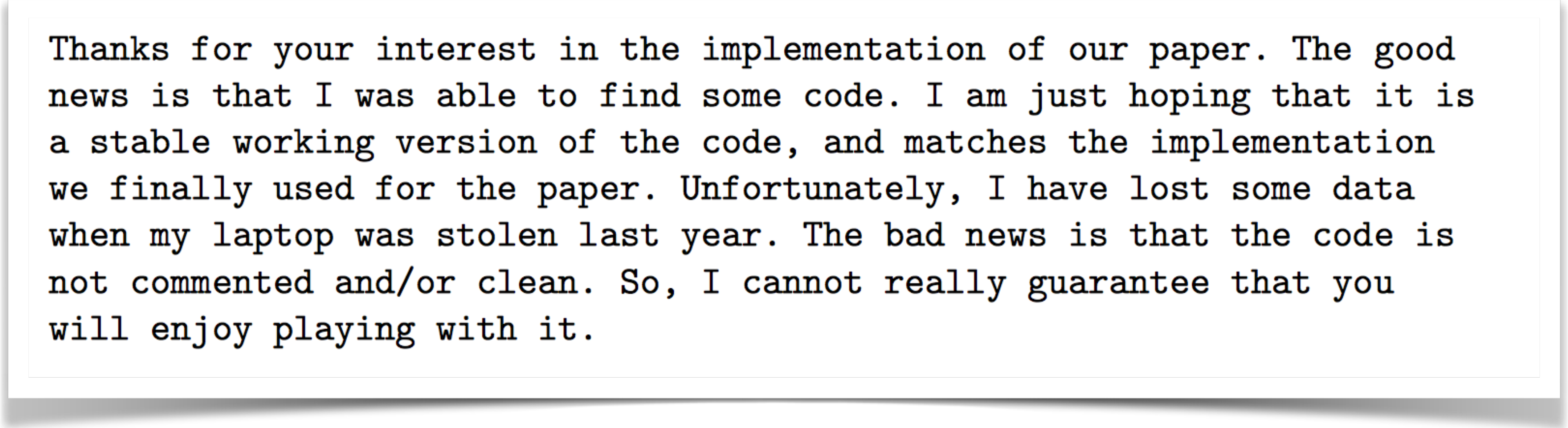
⟨STUDENT⟩ was a graduate student in our program but he left a while back so I am responding instead. For the paper we used a prototype that included many moving pieces that only ⟨STUDENT⟩ knew how to operate and we did not have the time to integrate them in a ready-to-share implementation before he left. Still, I hope you can build on the ideas/technique of the paper. Regards,



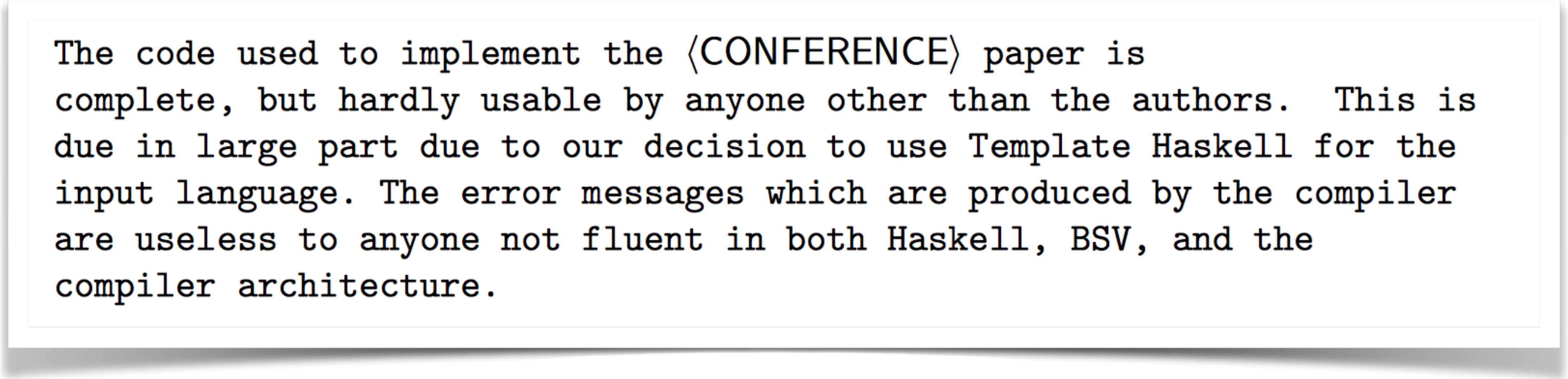
Since this work has been done at ⟨COMPANY⟩ we don’t open-source code unless there is a compelling business reason to do so. So unfortunately I don’t think we’ll be able to share it with you.



Thank you for your interest in our work. Unfortunately the current system is not mature enough at the moment, so it’s not yet publicly available. We are actively working on a number of extensions and things are somewhat volatile. However, once things stabilize we plan to release it to outside users. At that point, we would be happy to send you a copy.



Thanks for your interest in the implementation of our paper. The good news is that I was able to find some code. I am just hoping that it is a stable working version of the code, and matches the implementation we finally used for the paper. Unfortunately, I have lost some data when my laptop was stolen last year. The bad news is that the code is not commented and/or clean. So, I cannot really guarantee that you will enjoy playing with it.



The code used to implement the ⟨CONFERENCE⟩ paper is complete, but hardly usable by anyone other than the authors. This is due in large part due to our decision to use Template Haskell for the input language. The error messages which are produced by the compiler are useless to anyone not fluent in both Haskell, BSV, and the compiler architecture.

LET'S PYTHONIZE IT

A brand new implementation

Remember? *Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study.* M. Guthrie, A. Leblois, A. Garenne, and T. Boraud, Journal of Neurophysiology, 109, 2013.→ 100 files, 6,000 lines of Delphi

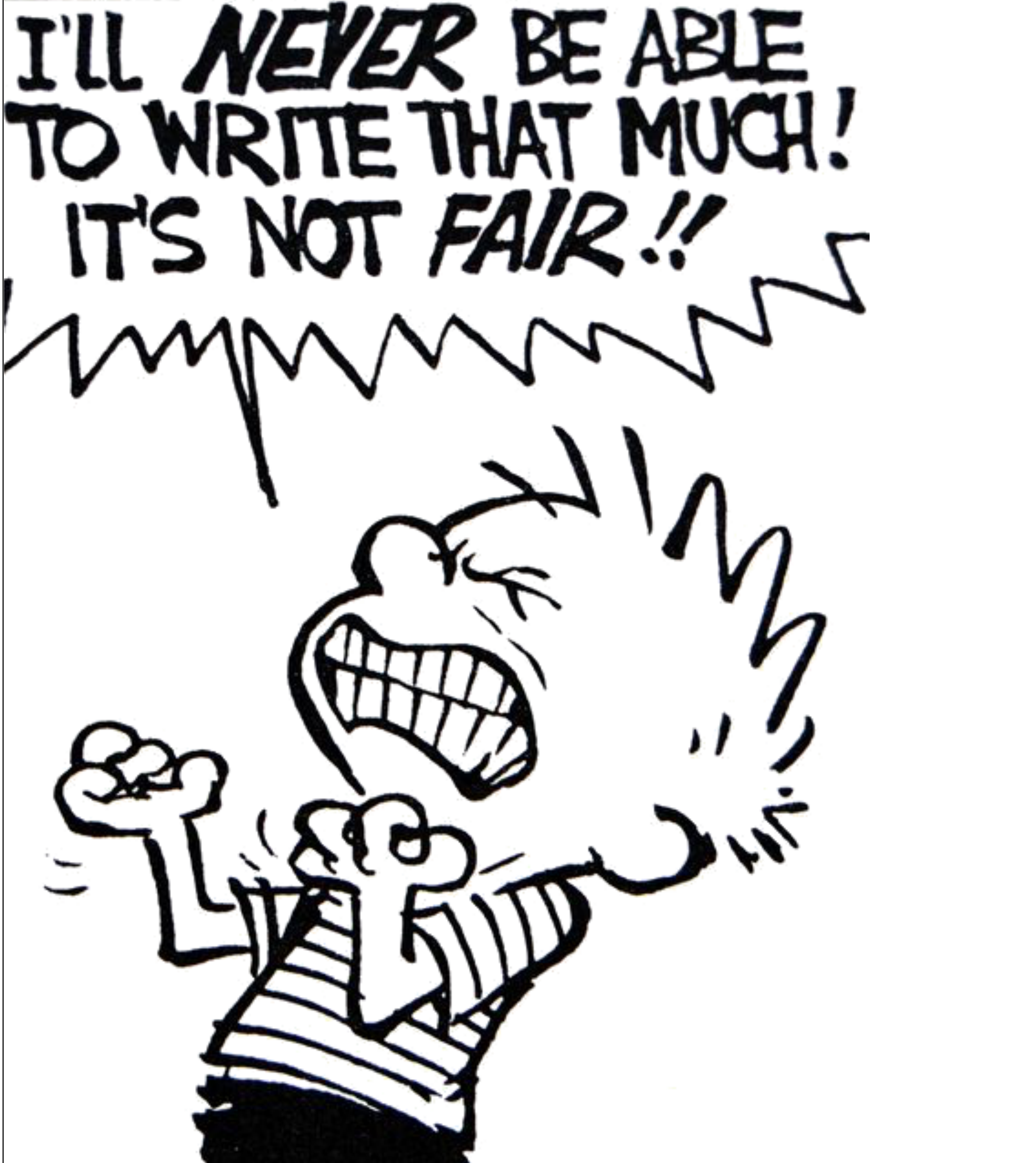
I asked my PhD student (M. Topalidou) to write a brand new implementation. Together, it took us three months of hard work to replicate the model using

- Python language and numerical libraries
- DANA library for intuitive description
- IPython notebook for interactive sessions

Source is now a single file of 200 readable notebook available on GitHub. Without this replication effort, original model would have been useless for our research.

Because of strong incentives for innovation and weak incentives for confirmation, direct replication is rarely practiced or published… Innovative findings produce rewards of publication, employment, and tenure; replicated findings produce a shrug.

Brian Nosek, The Reproducibility Project, 2012



A LONG JOURNEY INTO

Reproducible computational neuroscience

Any model in Science is doomed to be proved wrong or incomplete and replaced by a more accurate one. In the meantime, for such replacement to happen, we have first to make sure that models are actually reproducible such that they can be tested, evaluated, criticized and ultimately modified, replaced or even rejected.

This is where the shoe pinches.

If we cannot reproduce a model in the first place, we're doomed to re-invent the wheel again and again, preventing us from building an incremental computational knowledge.

> My field of research is quite different from computational neuroscience, but I recognize the problem described in this paper very well. The core issue has in my opinion been identified in the comment by Jan Moren: there is no obvious way to publish complex scientific models other than as part of simulation software.
> Konrad Hinsen, 2015

WHAT DO WE DO

NEXT ?

WE ARE NOT INTERESTED IN SCIENCE

‘cause we are $cientific publi$her$

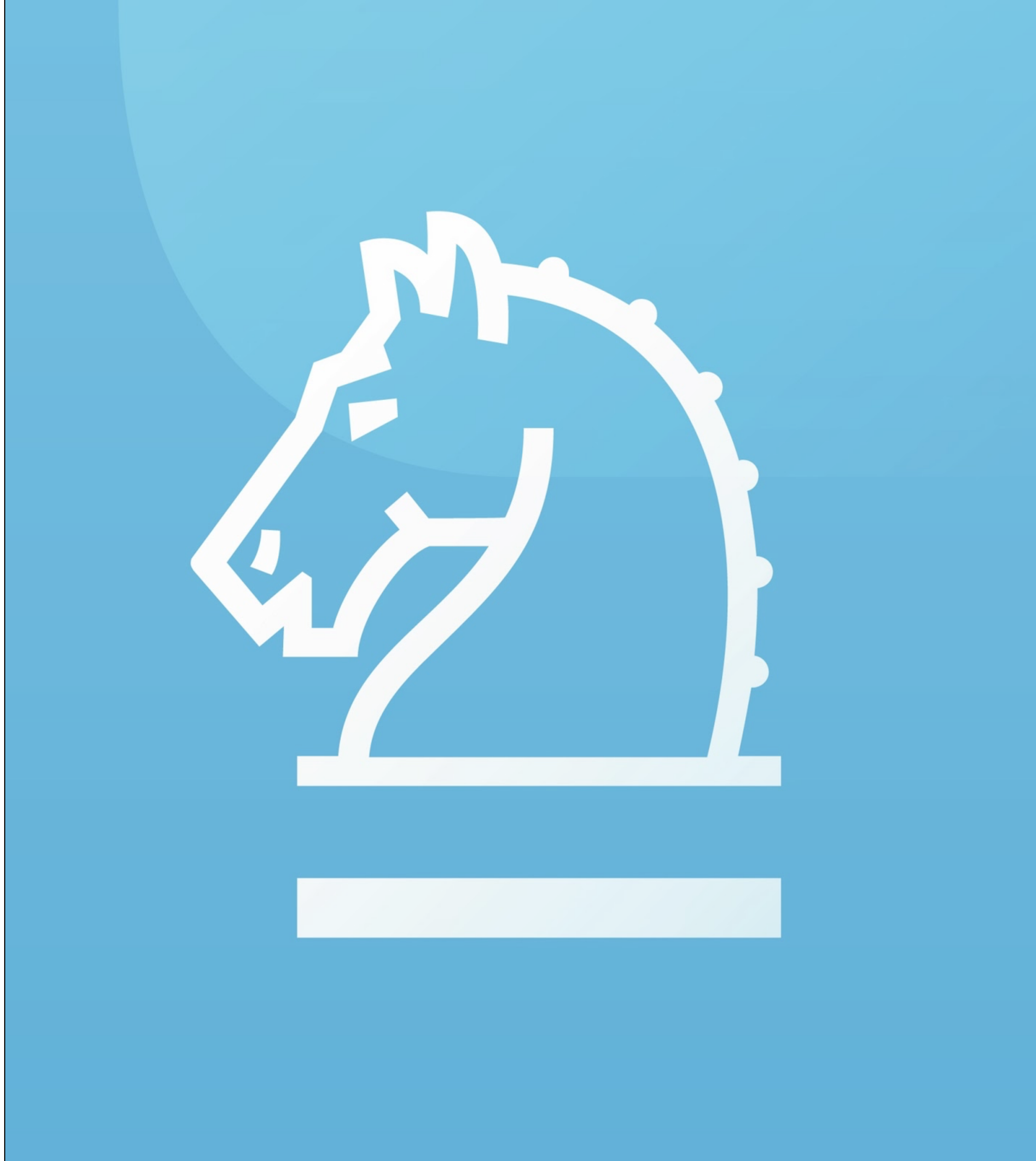
— Elsevier, can I publish my replication in your journal?
— Nope!

— Hi Springer, interested in replication?
— Failure or success?
— Success!
— Nope!

— Hello Mr Wiley, did you hear about reproducible Science?
— tut…. tut…. tut…

— Dear beloved Frontiers, can you review this?
— Ha ha ha…. No.

— Well, well, well…



INTRODUCING

The ReScience journal

ReScience is an open peer-reviewed journal that target any computational research and encourage the explicit replication of already published research promoting new and open-source implementations.

ReScience lives on github where each new implementation is made available together with explanations (article).

Each published article is archived on Zenodo and code is saved by Software Heritage

ReScience in numbers:

4 editors-in-chief
12 associate editors
110 registered reviewers
72 published articles
100% replication rate (strong bias)

Reproducible science is good. Replicated science is better.

BEFORE WE BEGIN

The R quintuplet (R^5)

Rerunnable
  Can you re-run your program ?
  One day, one week, one month, one year (just kidding) apart ?

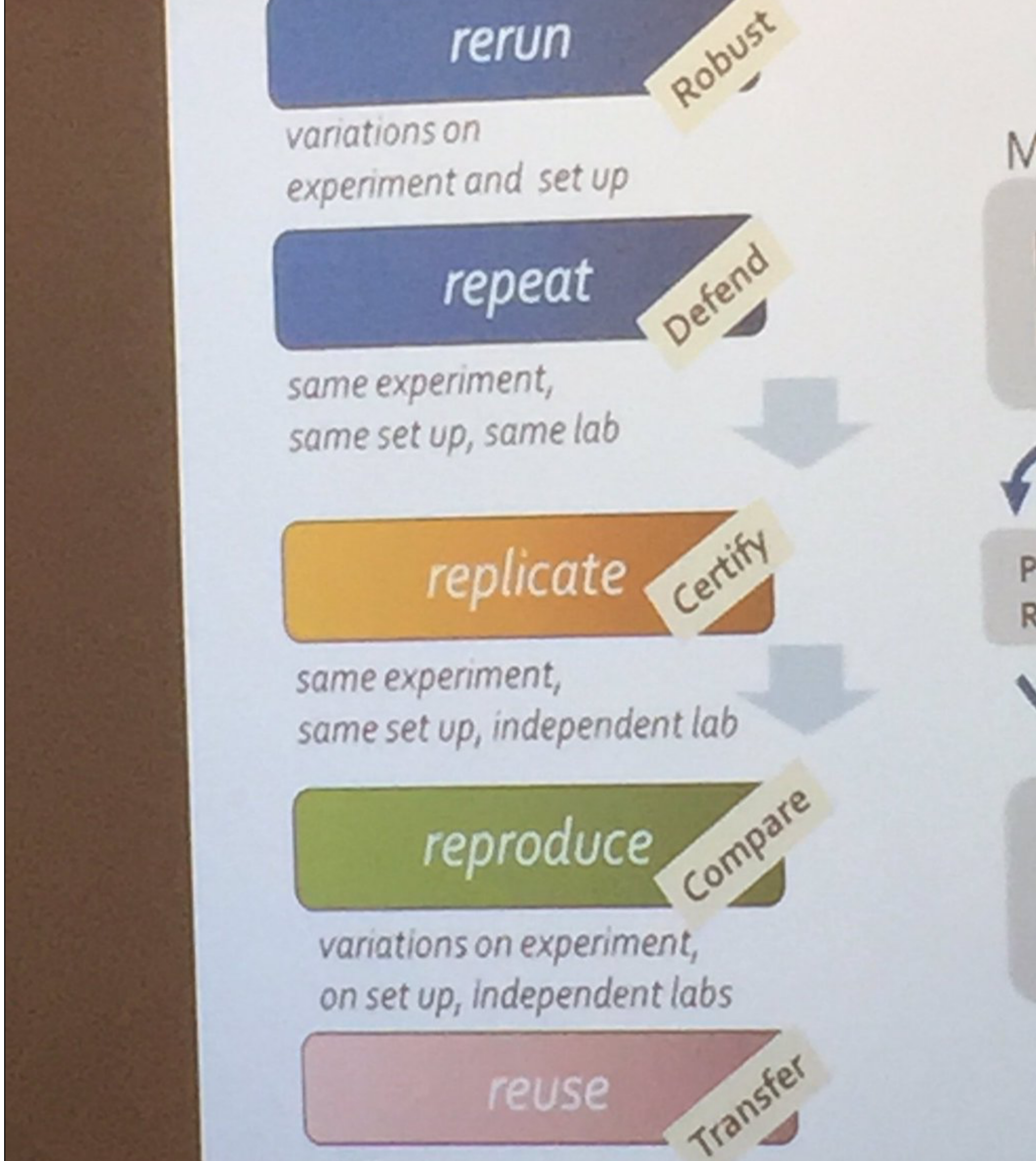
Repeatable
  Can you re-run your program and get same results ?
  Did you save everything, including random seed ?

Reproducible
  Can someone re-run your program and get same results ?
  Did you save the software stack ?

Replicable
  Can someone reimplement your model and get same results ?
  Did you describe everything ?

Reusable
  Can someone reuse your program using different data ?
  Is your software data-dependent ?

Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming
Code into Scientific Contributions - Benureau & Rougier, 2018



rerun
Robust
variations on
experiment and set up

repeat
Defend
same experiment,
same set up, same lab

replicate
Certify
same experiment,
same set up, independent lab

reproduce
Compare
variations on experiment,
on set up, independent labs

reuse
Transfer

IT’S MORE COMMON THAN YOU THINK

Replications in the wild

What is a replication?

Bob reads Alice’s paper, takes note of all model properties and then implements the model himself using a method of his choice.

Bob confirms Alice’s result by obtaining qualitatively the same results.

Alice’s model has been replicated.

Who wants to write replication?

During the course of a PhD, it is often the case that a student will try to replicate results from the literature, possibly interacting with the original authors.

Such replication generally lives inside the hard-drive of the computer’s student while it would be actually useful for the whole scientific community.

Who wants to review & publish such replication?

We do!

CHOICES HAD TO BE MADE

Why GitHub ?

GitHub offers a web-based git repository hosting service with great specific features (issue, pull request, etc).

→ Version control
→ Public repositories
→ Transparency and verifiability
→ Easy exploration of new ideas

A kind of modern lab for the computer scientist.

→ Popular among developers (Google, Microsoft, etc.)
→ Ergonomic & efficient
→ Free (as in beer)

But

→ Closed sources
→ Ran by a private company
→ Can close tomorrow

SUBMISSION PROCESS

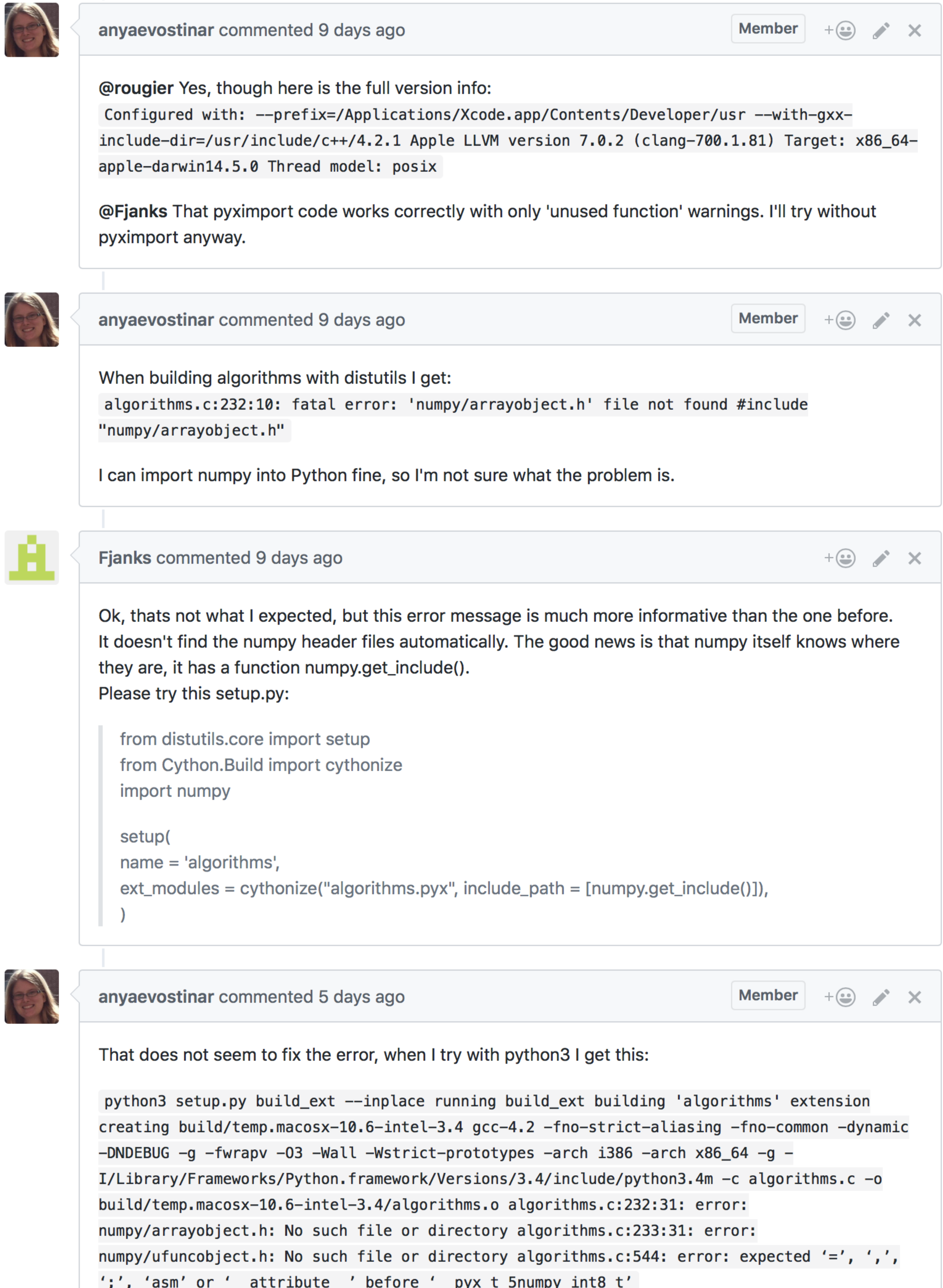
Open peer-review

Editor is publicly assigned by editor-in-chief.

Reviewers are publicly invited to review (they can decline the invitation of course)

The actual review takes place in the discussion area of the issue. Anybody can enter the discussion unless this discussion is locked.

This means anybody can give advice and/or comment because this discussion is public.



anyaevostinar commented 9 days ago

@rougier Yes, though here is the full version info:

```
Configured with: --prefix=/Applications/Xcode.app/Contents/Developer/usr --with-gxx-include-dir=/usr/include/c++/4.2.1 Apple LLVM version 7.0.2 (clang-700.1.81) Target: x86_64-apple-darwin14.5.0 Thread model: posix
```

@Fjanks That pyximport code works correctly with only 'unused function' warnings. I'll try without pyximport anyway.

anyaevostinar commented 9 days ago

When building algorithms with distutils I get:

```
algorithms.c:232:10: fatal error: 'numpy/arrayobject.h' file not found #include "numpy/arrayobject.h"
```

I can import numpy into Python fine, so I'm not sure what the problem is.

Fjanks commented 9 days ago

Ok, thats not what I expected, but this error message is much more informative than the one before.
It doesn't find the numpy header files automatically. The good news is that numpy itself knows where they are, it has a function numpy.get_include().
Please try this setup.py:

```
from distutils.core import setup
from Cython.Build import cythonize
import numpy

setup(
name = 'algorithms',
ext_modules = cythonize("algorithms.pyx", include_path = [numpy.get_include()]),
)
```

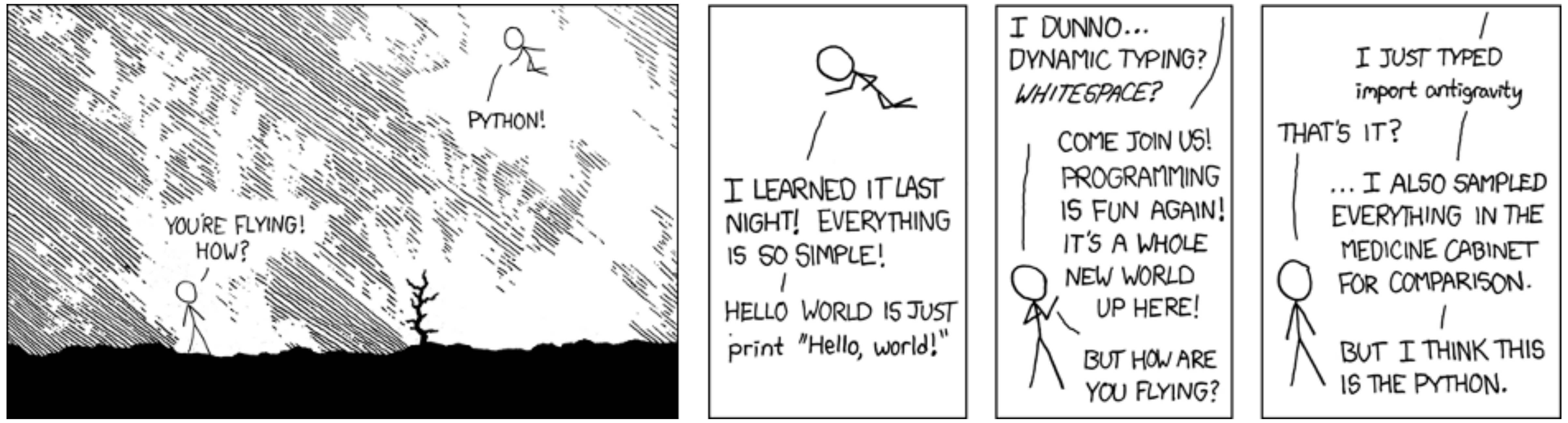
anyaevostinar commented 5 days ago

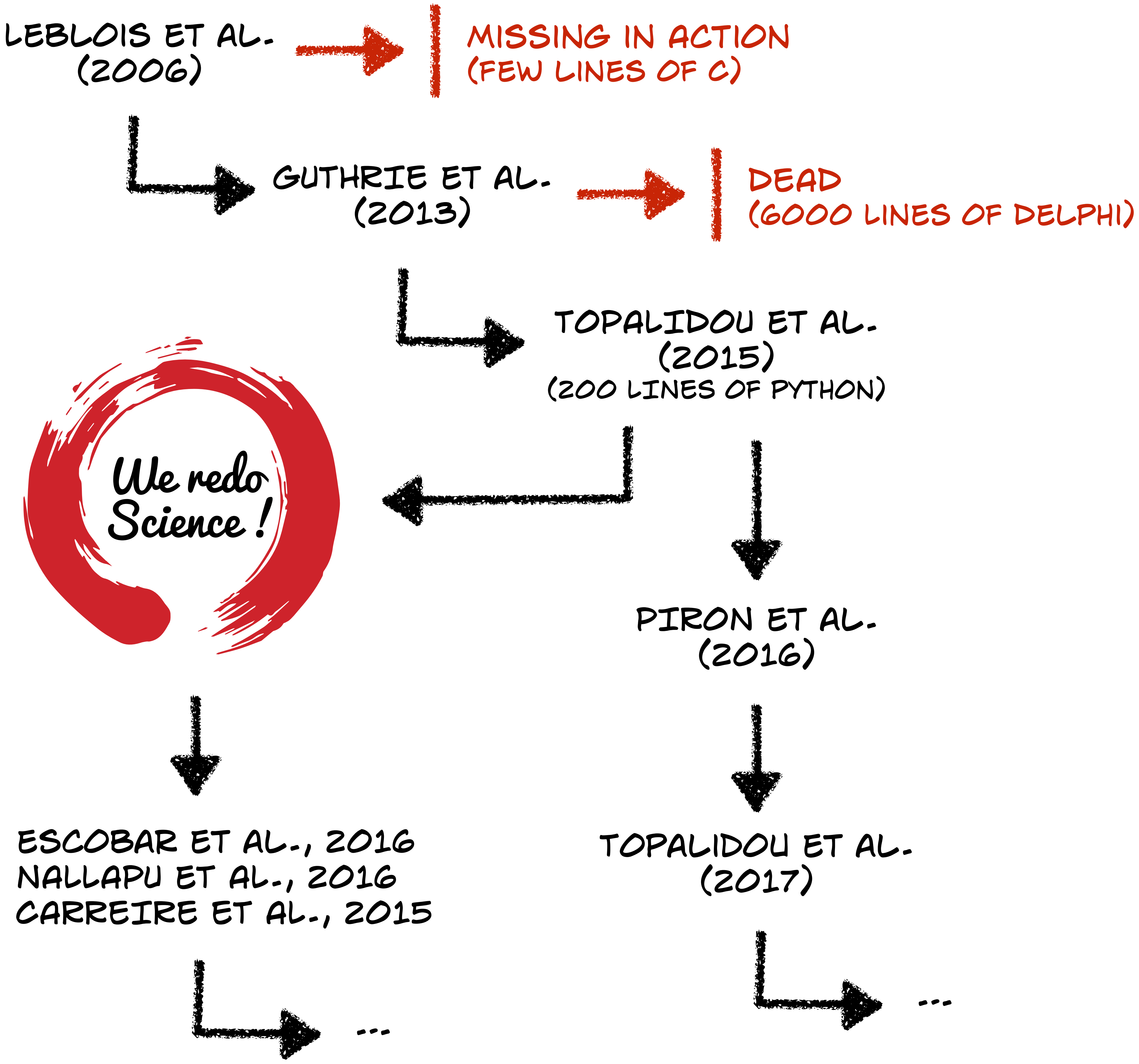
That does not seem to fix the error, when I try with python3 I get this:

```
python3 setup.py build_ext --inplace running build_ext building 'algorithms' extension creating build/temp.macosx-10.6-intel-3.4 gcc-4.2 -fno-strict-aliasing -fno-common -dynamic -DNDEBUG -g -fwrapv -O3 -Wall -Wstrict-prototypes -arch i386 -arch x86_64 -g -I/Library/Frameworks/Python.framework/Versions/3.4/include/python3.4m -c algorithms.c -o build/temp.macosx-10.6-intel-3.4/algorithms.o algorithms.c:232:31: error: numpy/arrayobject.h: No such file or directory algorithms.c:233:31: error: numpy/ufuncobject.h: No such file or directory algorithms.c:544: error: expected '=', ',', ';', 'asm' or '__attribute__' before '__pyx_t_5numpy_int8_t'
```

PYTHON THE ALMIGHTY

The Lazarus effect


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.



LEBLOIS ET AL. (2006)

MISSING IN ACTION (FEW LINES OF C)

GUTHRIE ET AL. (2013)

DEAD (6000 LINES OF DELPHI)

TOPALIDOU ET AL. (2015) (200 LINES OF PYTHON)

We redo Science!

PIRON ET AL. (2016)

ESCOBAR ET AL., 2016
NALLAPU ET AL., 2016
CARREIRE ET AL., 2015

TOPALIDOU ET AL. (2017)

...

...

RESCIENCE.GITHUB.IO/FAQ

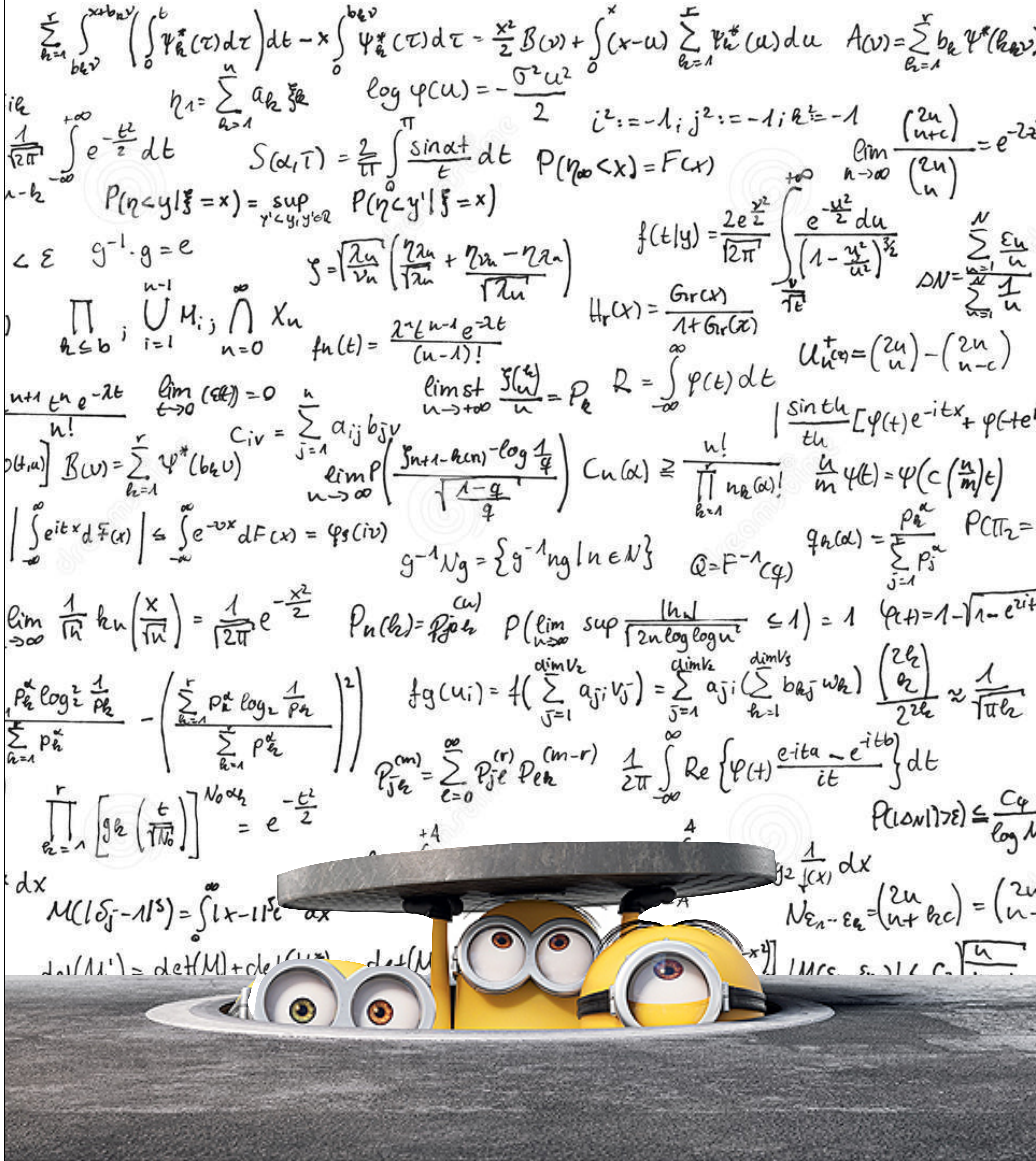
Frequently Asked Questions

What kind of research can I replicate?

Any computational research in any domain of science as long as there is an editor from the Board who has the expertise to edit your submission. The editorial board is growing to increase the scientific domains being covered. If no editor is able to edit your submission, you can also propose a guest editor (who must be willing to work with our GitHub-based editorial processes). about replication of my own work?

I'm a student, can I submit?

Yes ! Students are strongly encouraged to submit their work. Although the ReScience publishing model is a bit different from other academic journals, it can give students a first experience at peer-reviewed scholarly publishing, including meeting standards of scientific rigor and addressing reviewers' comments. Publishing in ReScience is also a way to actively contribute to open science while adding to one's publication record.



RESCIENCE.GITHUB.IO/FAQ

Frequently Asked Questions

What if I cannot replicate a result?

Some research may not be replicable. Before declaring a research result non-replicable, we require extra caution to be taken. In addition to scrutiny of your submission by reviewers and editors, we will contact the authors of the original research, and issue a challenge to the ReScience community to spot and report (using the issue tracker) errors in your implementation.

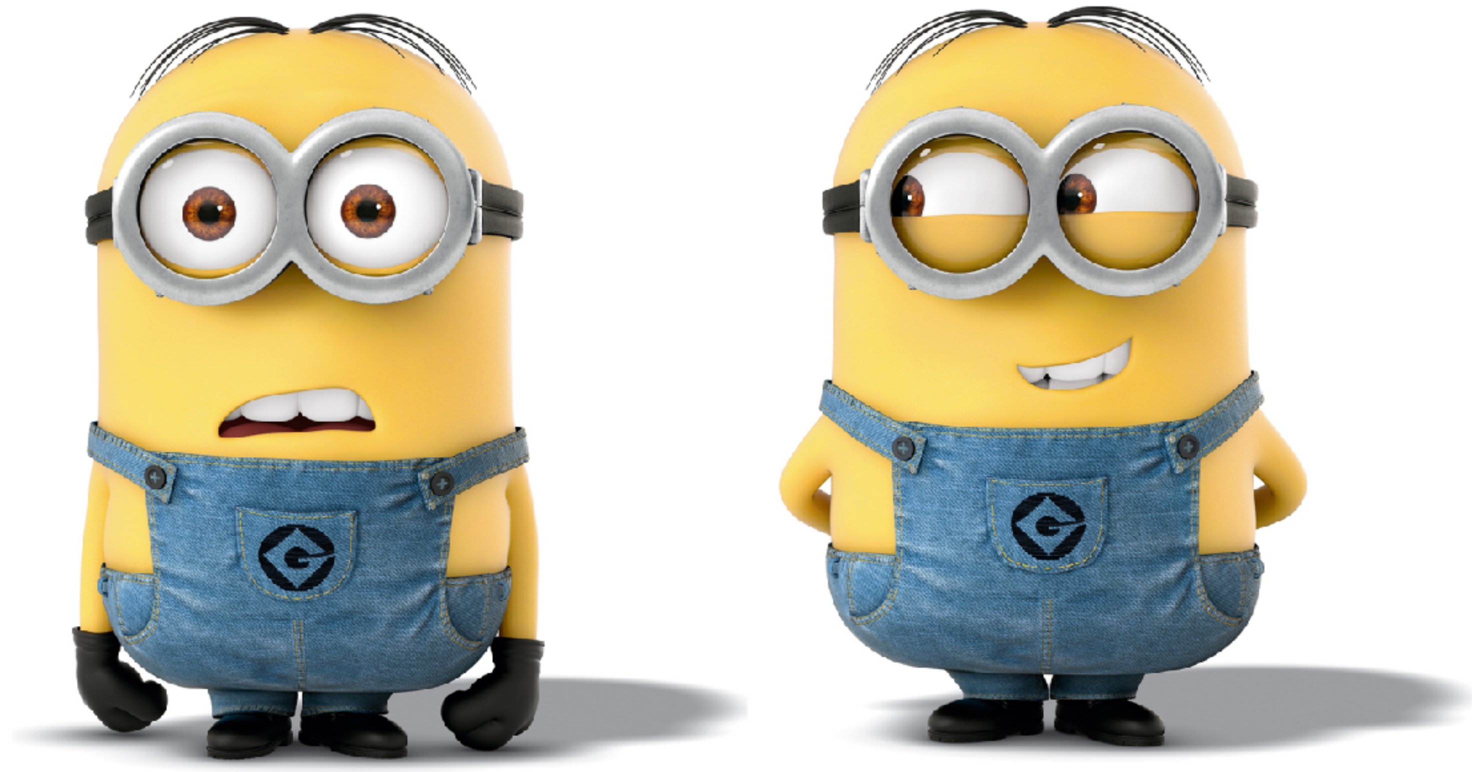
If no errors are found, your work will be accepted and the original research will be declared non-replicable.

What about replication of my own work?

No. Mistakes in the implementation of research questions and methods are often due to biases authors invariably have, consciously or not. One's biases will inevitably carry over to how one approaches a replication.

Perhaps even more importantly, we aim at the cross-fertilization of research and trying to replicate the work of one's peers might pave the way for a future collaboration, or may give rise to new ideas as a result of the replication effort.

This changed in 2020…



Ten Years Reproducibility Challenge

RESCIENCE SPECIAL ISSUE

FREE TO READ - FREE TO PUBLISH

Would you dare to run the code from your past self ?

(the one that does not answer mail)

WHAT'S NEXT?

ReScience X

FREE TO READ - FREE TO PUBLISH

SINCE JANUARY 2021