An Introduction to Research Software Sustainability and Open Science

Matt Forshaw (matthew.forshaw@ncl.ac.uk)

Introductions

What are people expecting from the session?

Do you have any particular questions or areas of interest?

Discussion (2 mins)

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

Discussion

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

–Martyn Rittman, Editor at MDPI

Science

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

Attempt to replicate major social scientific findings of past decade fails

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

Hannah Devlin
Science correspondent
@hannahdev

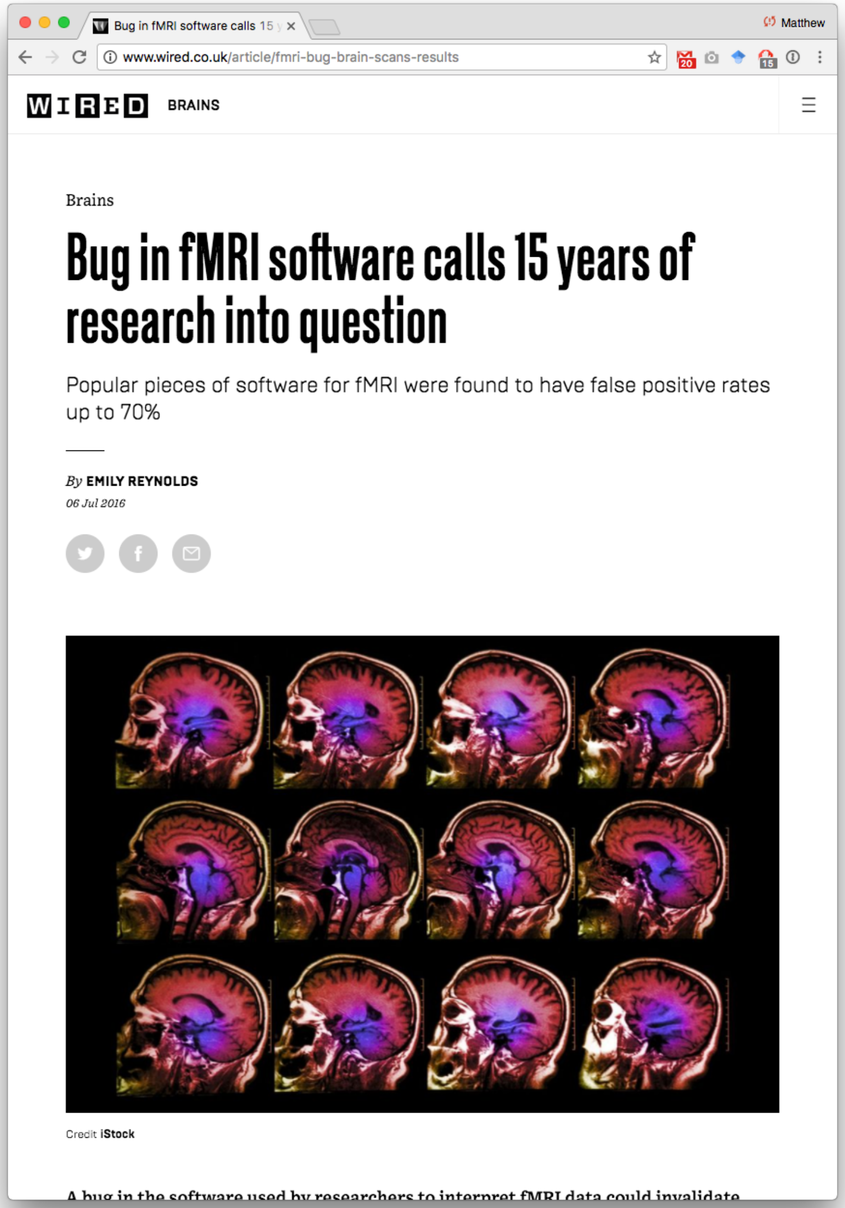
Mon 27 Aug 2018
16.00 BST

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

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

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

The research, which follows similar efforts in psychology and biomedical



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

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

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

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

By “open science” we mean any initiative that aims at lowering or erasing the technical, social, and cultural barriers that prevent scientists from sharing knowledge with one another and with individuals outside of the academic community, but also the barriers that prevent anyone from producing knowledge. More precisely, it promotes open access (i.e., no charge beyond the cost of an internet connection) to articles describing research results, data that support these results, computer code, algorithms, and any other (digital) products of scientific research. In addition to being open, these products should follow the FAIR guidelines (i.e. be findable, accessible, interoperable, and reusable) with exceptions for contractual, privacy and security reasons, potential commercial aspects and proprietary claims.”

By “open science” we mean any initiative that aims at lowering or erasing the technical, social, and cultural barriers that prevent scientists from **sharing knowledge** with one another and with individuals outside of the academic community, but also the barriers that prevent anyone from producing knowledge. More precisely, it promotes **open access** (i.e., no charge beyond the cost of an internet connection) to **articles** describing research results, **data** that support these results, computer **code**, **algorithms**, and any other (digital) products of scientific research. In addition to being open, these products should follow the **FAIR guidelines** (i.e. be findable, accessible, interoperable, and reusable) **with exceptions** for contractual, privacy and security reasons, potential commercial aspects and proprietary claims.”

EPFL

Open Science

Visibility

larger impact

Scrutiny

better quality

Reuse

higher efficiency

Public access

fair opportunity

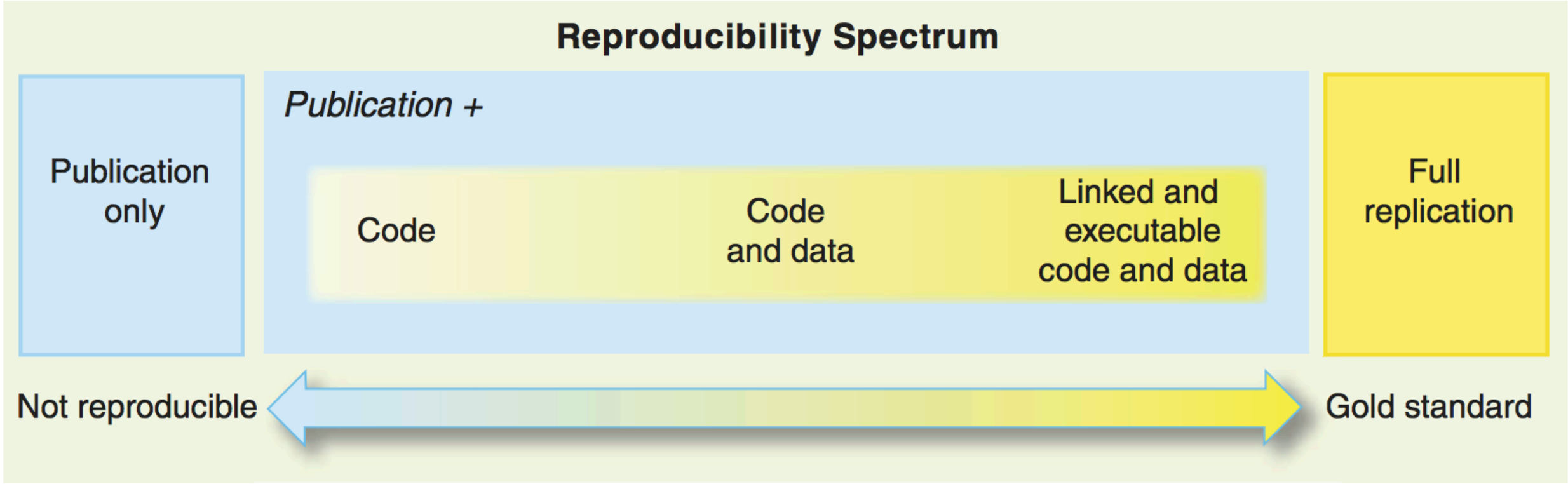
Replication

- **Replication** is the most important element of verifying and validating scientific findings
- Strengthen scientific evidence by verifying it in studies with independent…
  - Investigators
  - Data
  - Analytical methods
  - Laboratories
  - Instruments
- Replication is particularly important in studies which impact broad policy decisions or regulatory decisions
- Replication is the **gold standard** but sometimes isn't achievable.

Slides adapted from Peng *et al*

Limitations

- It is becoming more and more challenging to do replication.
- Some studies cannot be replicated, for a number of reasons
  - No time
  - No money
  - Unique
- **Reproducible Research:** Make analytic data and code available so others can reproduce your findings
  - Other authors can interrogate your approach, can re-run your computation and (hopefully) come to the same conclusion.
  - Computing power is greatly increased, allowing for more sophisticated analyses.



Peng, Roger D. "Reproducible research in computational science." *Science* 334.6060 (2011): 1226-1227.

What do we need for reproducibility

- **Analytic data** are available
  - Not the same as raw data; perhaps some subset…
  - Full experimental data may be infeasibly large to share.
- **Analytic code** are available
- **Documentation** of code and data
- Standard means of **distribution**
  - Must be easily accessible (We'll see more on Wednesday)

Who are the stakeholders?

Authors

- Produce research, and want to make their research reproducible
- Want tools for Reproducible Research to make their lives easier (at least not much harder)

Readers

- Want to reproduce, and perhaps expand upon, interesting findings
- Want tools for Reproducible Research to make their lives easier

Three steps is all you need: fast, accurate, automatic scaling decisions for distributed streaming dataflows

Vasiliki Kalavri[†], *John Liagouris*[‡], *Moritz Hoffmann*[‡],
Desislava Dimitrova[‡], *Matthew Forshaw*[‡*], *Timothy Roscoe*[†]
[†]*Systems Group, Department of Computer Science, ETH Zurich, firstname.lastname@inf.ethz.ch*
[‡]*Newcastle University, firstname.lastname@newcastle.ac.uk*

Abstract

Streaming computations are by nature long-running, and their workloads can change in unpredictable ways. This in turn means that maintaining performance may require dynamic scaling of allocated computational resources.

Some modern large-scale stream processors allow dynamic scaling but typically leave the difficult task of deciding *how much* to scale to the user. The process is cumbersome, slow and often inefficient. Where automatic scaling is supported, policies rely on coarse-grained metrics like observed throughput, backpressure, and CPU utilization. As a result, they tend to show incorrect provisioning, oscillations, and long convergence times.

We present DS2, an automatic scaling controller for such systems which combines a general performance model of streaming dataflows with lightweight instrumentation to estimate the *true* processing and output rates of individual dataflow operators.

We apply DS2 on Apache Flink and Timely Dataflow and demonstrate its accuracy and fast convergence. When compared to Dhalion, the state-of-the-art technique in Heron, DS2 converges to the optimal, backpressure-free configuration in a single step instead of six.

1 Introduction

We present DS2, a low-latency, robust controller for *dynamic scaling* of streaming analytics applications, which can vary the resources available to a computation so as to handle variable workloads quickly and efficiently.

Static provisioning is a poor fit for continuous, long-running streaming applications: it forces users to choose a single point on the spectrum between allocating resources for worst-case, peak load (which is inefficient) and suffering degraded performance during load spikes. Fixing resources *a priori* almost inevitably leads to a system which is over- or under-provisioned for much of its execution.

[*]Work done while visiting the Systems Group at ETH Zürich

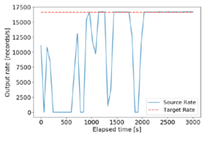
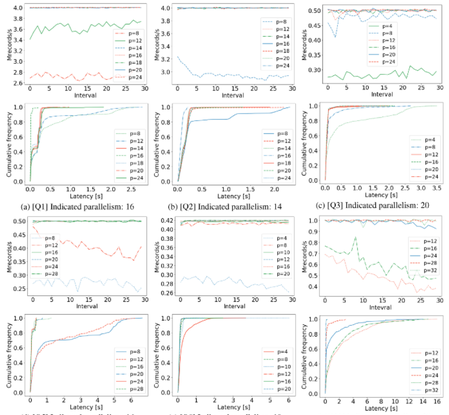


Figure 1: Effect of Dhalion's scaling decisions on the source rate when trying to match the target throughput of an under-provisioned word count dataflow.

The solution is to dynamically scale the system in response to load, an idea used extensively in cloud environments [30, 31]. This requires both a *mechanism* for scaling the computation, and a *scaling controller* which decides when and how to scale. This paper focuses on the latter; DS2 is designed to be mechanism-agnostic.

A scaling controller makes two kinds of decisions. First, it detects symptoms of over- or under-provisioning (e.g. backpressure) and decides *whether* to make a change. Detection is often straightforward and addressed by conventional monitoring tools. Second, the controller must identify the *causes* of symptoms (e.g. a bottlenecked or idle operator) and *propose* a scaling action.

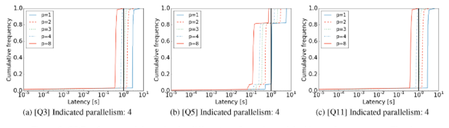
The second decision is challenging, involving performance analysis and prediction. Streaming systems supporting a form of automatic dynamic scaling (e.g. Google Cloud Dataflow [26, 5], Heron [27, 13], Pravega [11], Spark Streaming [45], and IBM System S [15]) and research prototypes (e.g. Seep [12] and StreamCloud [17]) focus on the first decision and either ignore or provide speculative, often ad-hoc solutions for the second.



(a) [Q1] Indicated parallelism: 16 (b) [Q2] Indicated parallelism: 14 (c) [Q3] Indicated parallelism: 20

(d) [Q5] Indicated parallelism: 16 (e) [Q8] Indicated parallelism: 10 (f) [Q11] Indicated parallelism: 28

Figure 8: Observed source output rates and per-record latency CDFs for different configurations of the Nexmark operators on Apache Flink.

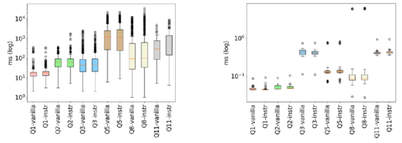


(a) [Q3] Indicated parallelism: 4 (b) [Q5] Indicated parallelism: 4 (c) [Q11] Indicated parallelism: 4

Figure 9: CDFs of per-epoch latencies for different configurations of the Nexmark operators on Timely.

Initial configuration	Q1	Q2	Q3	Q5	Q8	Q11
8	12→**16**	11→13→**14**	16→**20**	14→15→**16**	**10**	12→22→**28**
12	**16**	**14**	18→**20**	**16**	**10**	22→**28**
16	**16**	12→**14**	**20**	**16**	8→**10**	26→**28**
20	**16**	13→**14**	**20**	14→**16**	8→**10**	**28**
24	**16**	**14**	**20**	14→**16**	8→**10**	**28**
28	**16**	**14**	**20**	13→**16**	8→**10**	**28**

Table 4: DS2 convergence steps for Nexmark queries on Flink. Values are the level of parallelism of the main operator of each query. Leftmost column shows initial parallelism (from 8 to 28 instances); subsequent columns show optimal level of parallelism as estimated by DS2 in each step. Final decisions converged to by DS2 are highlighted.



(a) Flink instrumentation overhead (b) Timely instrumentation overhead

Figure 10: Policy instrumentation overhead for the Nexmark queries of Table 3 with instrumentation disabled (`vanilla`) and enabled (`instr`) for both Flink (10a) and Timely (10b).

cause backpressure.

Timely does not have a backpressure mechanism so data sources are never delayed and the observed source rates are always equal to the initial fixed rate (instead, queues grow when the system cannot keep up). We therefore simply show CDFs of per-epoch latencies with different configurations for Timely. Figure 9 shows these for **Q3**, **Q5**, and **Q11**; results are similar for other queries. Each epoch in the CDFs corresponds to 1*s* of data, which must be processed in less than 1*s*. The optimal parallelism indicated by DS2 is $p = 4$ in all queries, regardless of the starting configuration. For **Q3** (left) and **Q11** (right), $p = 4$ is clearly the configuration that can keep up with the 1*s* target (vertical line in the plots) using minimum required resources. For **Q5**, 18% of the epochs are above the target by up to 0.5*s*. Here, the larger percentage of epochs that cannot keep up is because of the window operator, which stashes data and then forwards it at certain time points. This manifests as load spikes, which require additional resources for the system to keep up. Longer decision intervals smooth out the spikes but tend to affect policy decisions towards higher optimal configurations, which is why DS2 indicated $p = 4$ (cf. § 4.2).

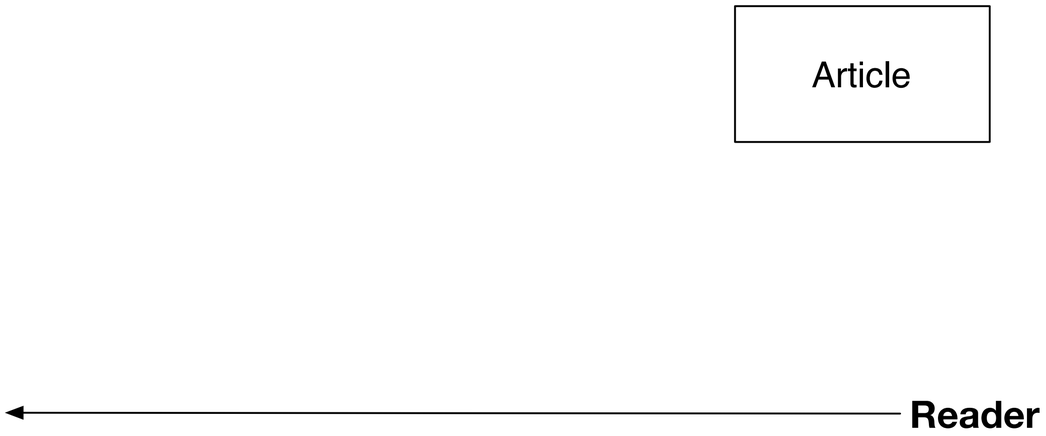
In summary, DS2 identified optimal configurations in all experiments and never overshot (provisioned more resources than needed), thereby exhibiting the remaining two SASO properties: *accuracy* and *no overshoot*.

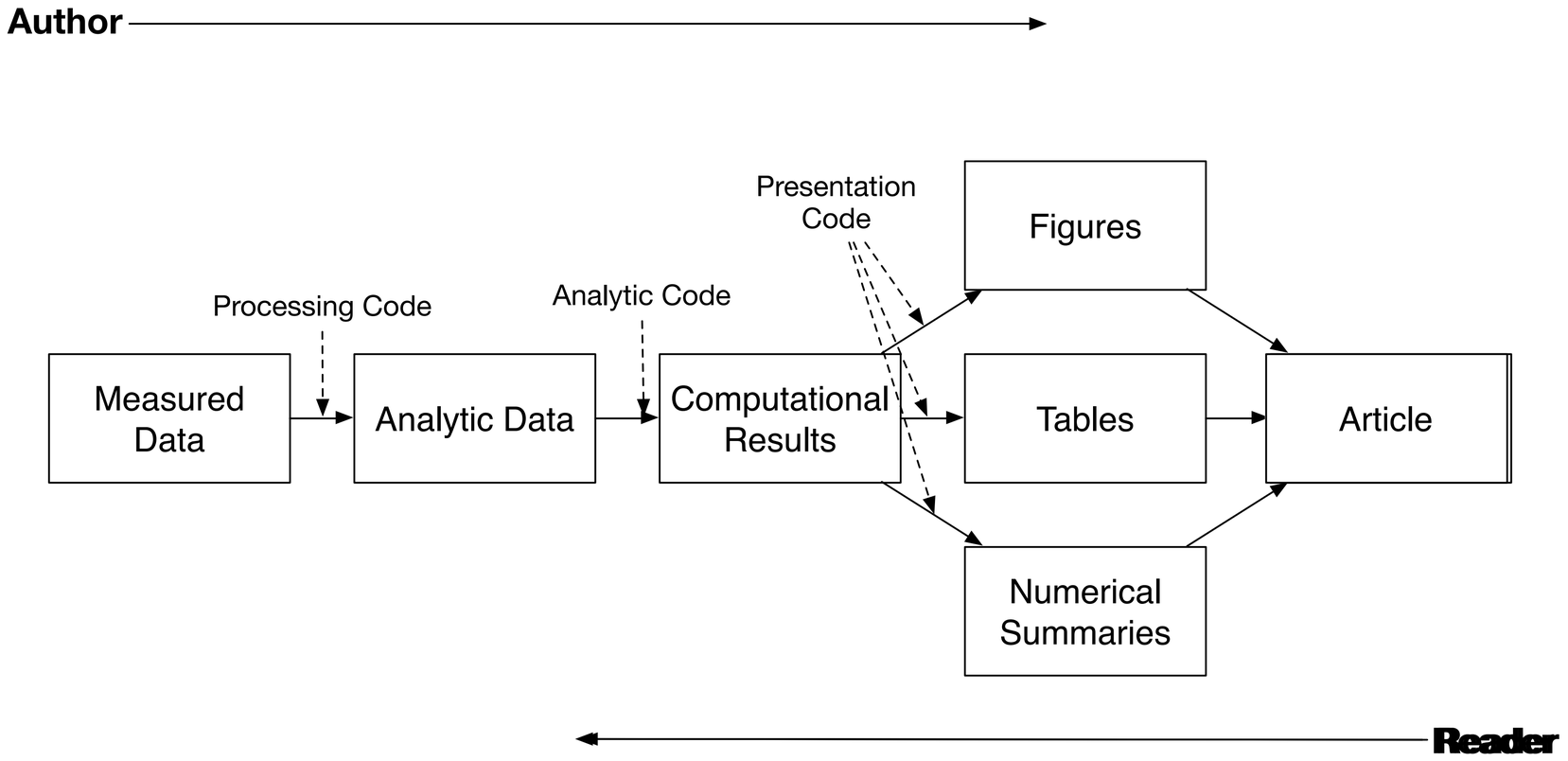
5.6 Instrumentation overhead

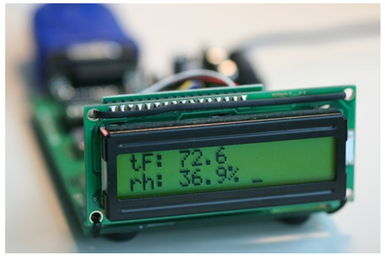
Finally, we evaluate instrumentation overhead. We run the Nexmark queries for 5 minutes with source rates from Table 3 and a 10*s* decision interval — the smallest we use in this paper, which results in the most frequently aggregated logs and has the highest potential overhead on the system performance.

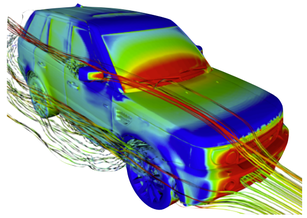
We measure per-record latency in Flink using its built-in metric and per-epoch latency in Timely using 1*s* event-time epochs. Figure 10 shows boxplots for both systems. Individual columns show latency with logging completely off (vanilla) and instrumentation activated (instr). Overheads are small: at most 13% on Flink (40*ms* absolute difference) and at most 20% on Timely (5*ms* absolute difference) across all queries. Performance penalties are an acceptable trade-off for a good scaling policy, and could be further reduced with a larger decision interval and pre-

Article


Article
Reader


Author
Presentation Code
Figures
Processing Code
Analytic Code
Measured Data
Analytic Data
Computational Results
Tables
Article
Numerical Summaries
Reader





Author

Processing Code

Analytic Code

Presentation Code

Measured Data

Analytic Data

Computational Results

Figures

Tables

Numerical Summaries

Article

Reader

Challenges

- Authors must undertake considerable effort to put data/ results on the web (may not have resources like a web server).
- Readers must download data/results individually and piece together which data goes with which code sections, etc.
- Readers may not have the same resources as authors
- Few tools to help authors/readers (although the toolbox is growing).

The current state of play…

- Authors
  - Just throw stuff up on the web
  - (Infamous) Journal supplementary materials
  - There are some central databases for various fields (e.g. biology)
- Readers
  - Just download the data and (try to) figure it out
  - Piece together the software and run it

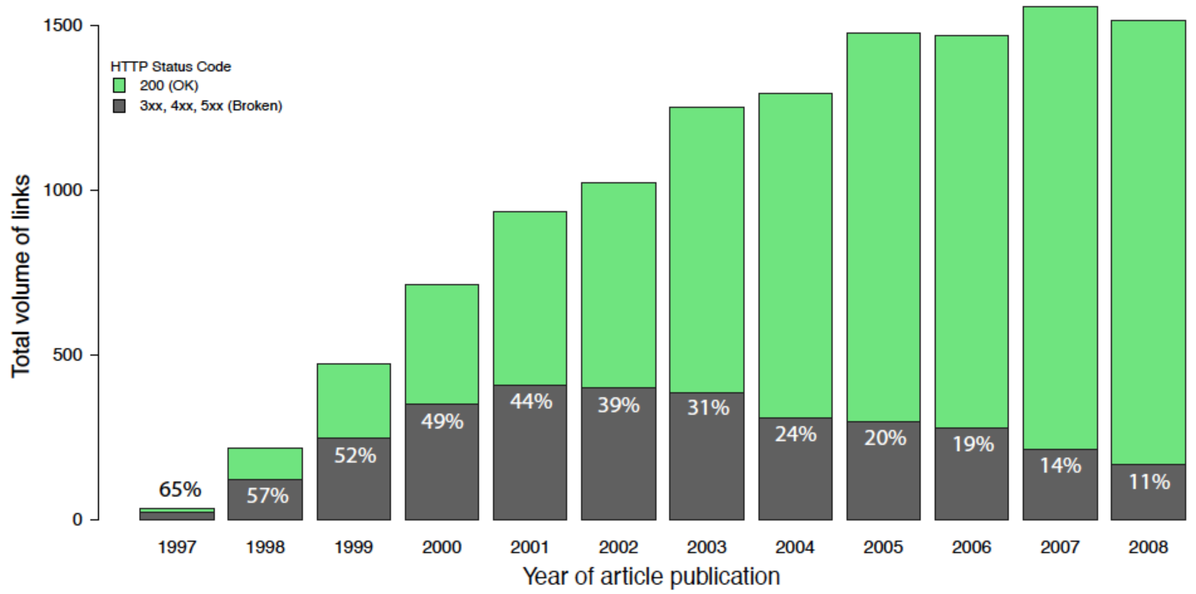


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

Pepe A, Goodman A, Muench A, Crosas M, Erdmann C (2014) How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE 9(8): e104798.

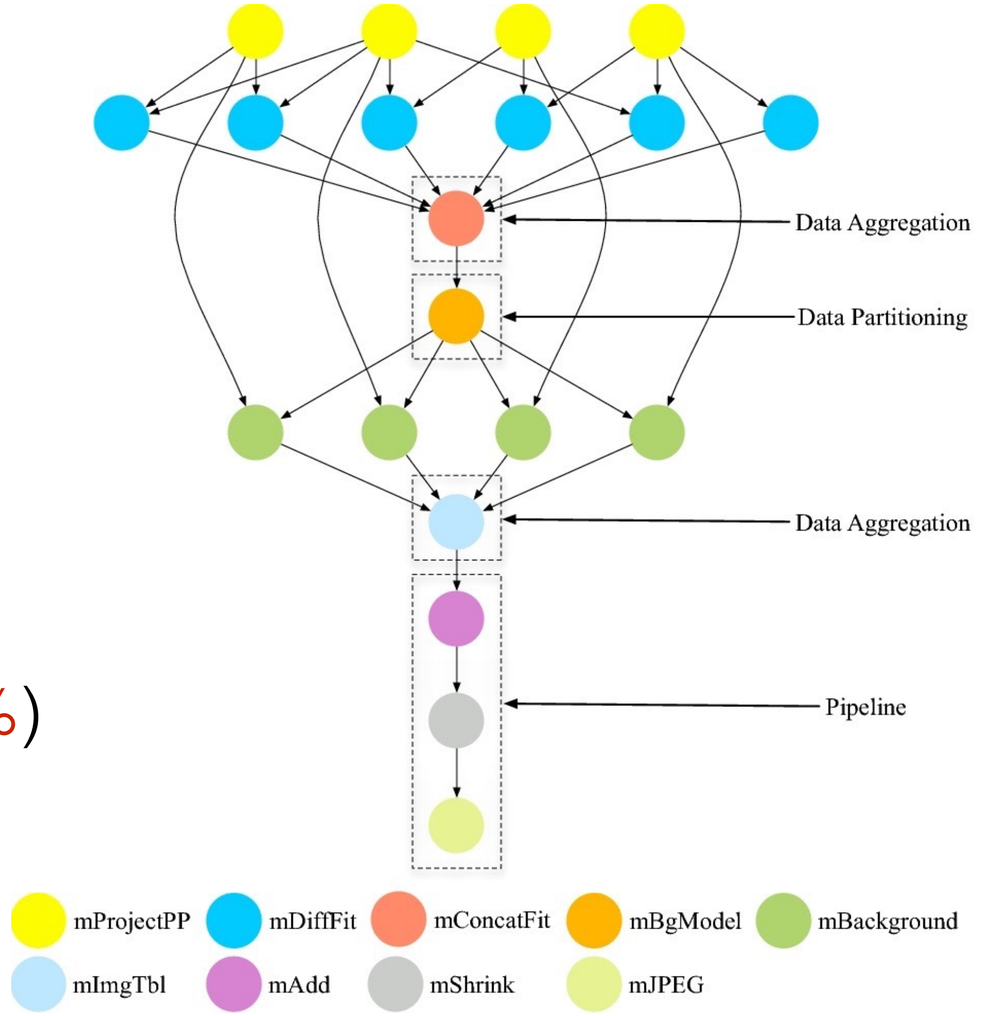
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104798

Zhao, *et al*:

18 of 92 (19.57%)

Mayer, *et al:*

341 of 1443 (23.63%)

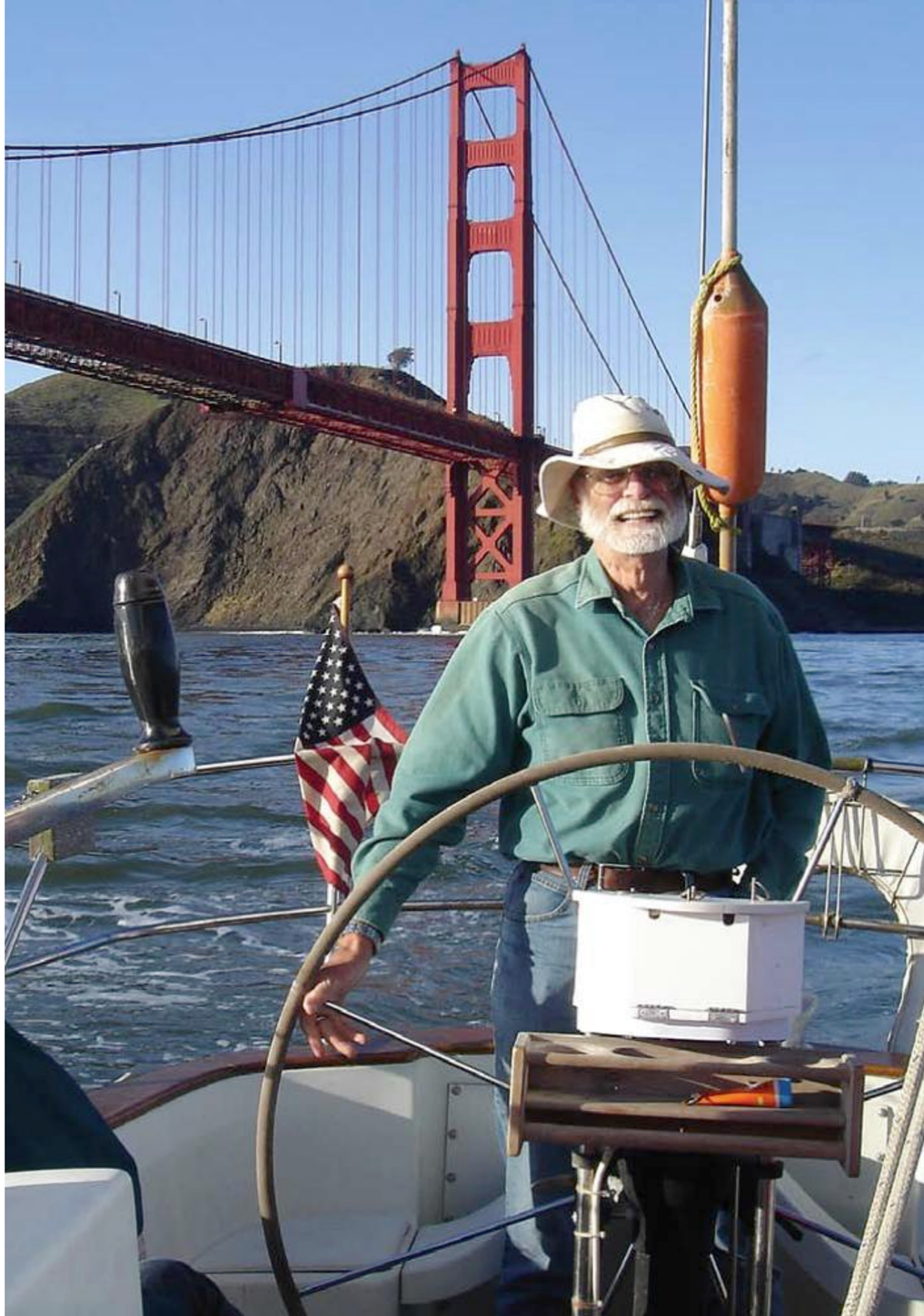


Zhao, Jun, et al. "Why workflows break—Understanding and combating decay in Taverna workflows." E-Science (e-Science), 2012 IEEE 8th International Conference on. IEEE, 2012.
http://ieeexplore.ieee.org/document/6404482/

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

“We have to do better producing tools to support the whole research cycle - from data capture and data curation to data analysis and data visualization.”

– Jim Grey (Turing Award winner)



Open Science

Visibility

larger impact

Scrutiny

better quality

Reuse

higher efficiency

Public access

fair opportunity

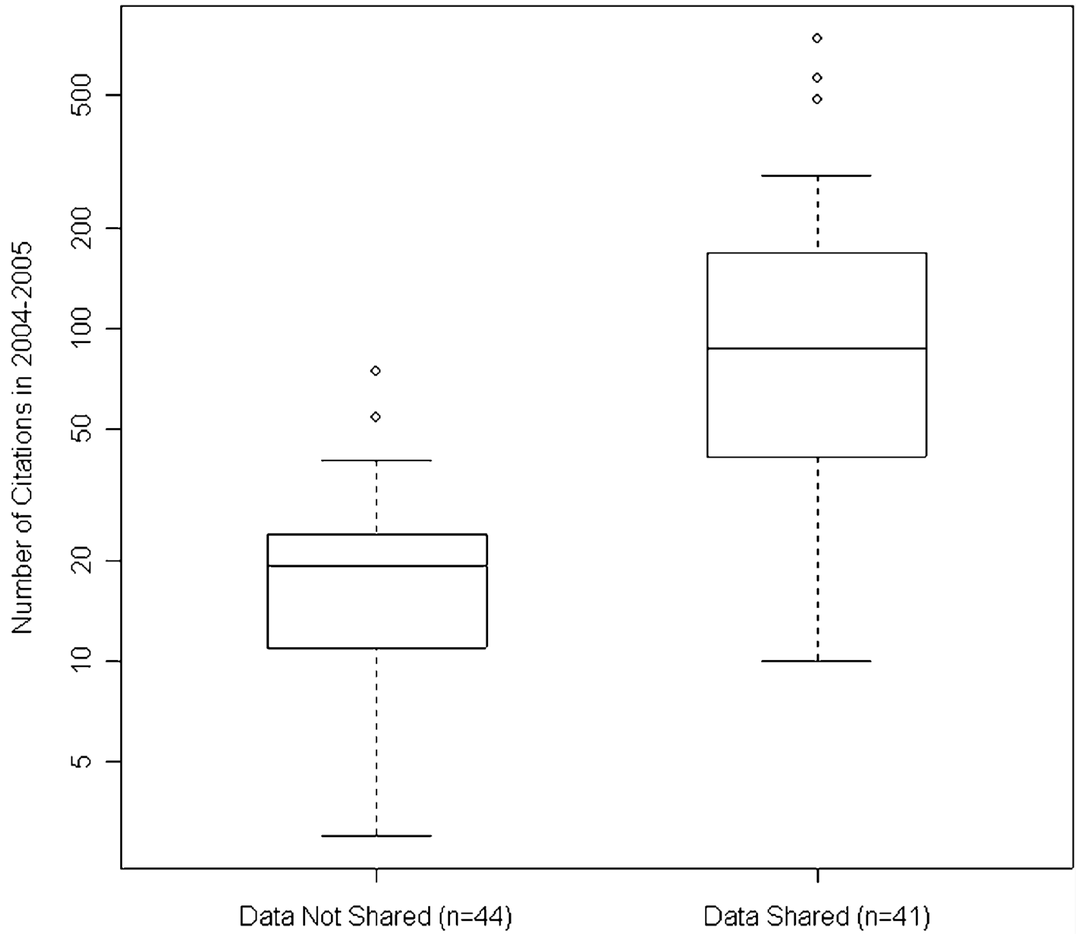
WORLD VIEW *A personal take on events*

JOHN SOARES PHOTOS

Faculty promotion must assess reproducibility

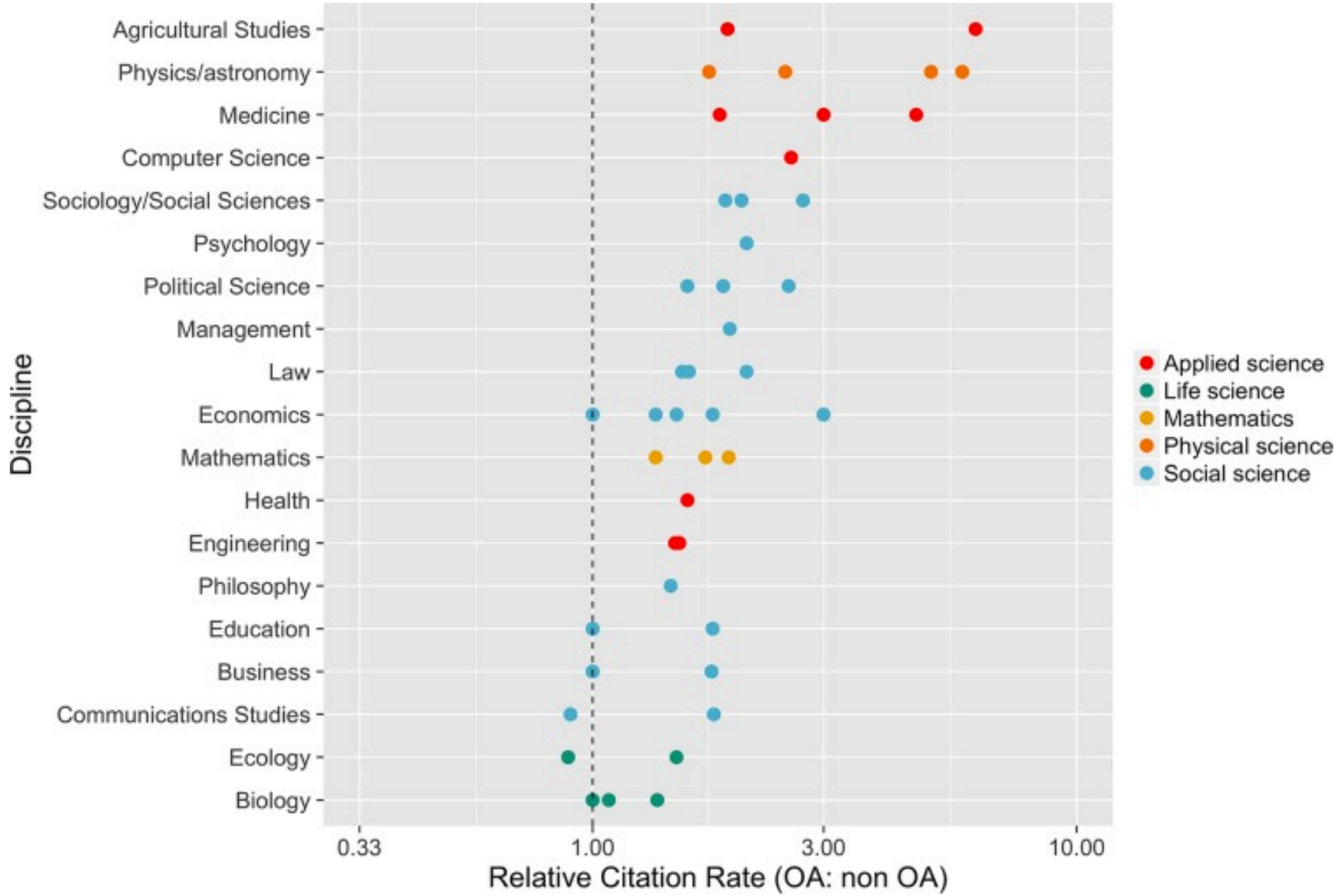
Research institutions should explicitly seek job candidates who can be frankly self-critical of their work, says **Jeffrey Flier**.

Source: Nature 2017 549, 133



Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing detailed research data is associated with increased citation rate." PloS one 2.3 (2007): e308.

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308



McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., ... & Spies, J. R. (2016). Point of view: How open science helps researchers succeed. Elife, 5, e16800.

Association for Computing Machinery

Advancing Computing as a Science & Profession

Digital Library | CACM | Queue | TechNews | Learning Center | Career Center

Join | Volunteer | myACM | Search

ABOUT ACM | MEMBERSHIP | PUBLICATIONS | SPECIAL INTEREST GROUPS | CONFERENCES | CHAPTERS | AWARDS | EDUCATION | PUBLIC POLICY | GOVERNANCE

Publications Home | About Publications | Digital Library | CACM | Journals | Magazines | Books | Conference Proceedings | ICPS | MORE

Home > Publications > Data & Software Artifacts

Software and Data Artifacts in the ACM Digital Library

ACM encourages authors to submit software and data sets with their papers. For years, ACM has provided mechanisms for authors to submit software, data sets, videos and other supplemental artifacts with their research papers. We have recently made these artifacts more discoverable through search and made them more prominent on abstract pages and Tables of Contents.

ACM's Reproducibility Task Force has been working with SIG conferences and journal EICs to understand and articulate common Best Practices in preparing and reviewing software and data artifacts, how they can be integrated with the ACM Digital Library, and how to reflect them in publication and enable their re-use. Many of ACM's technical communities are evolving standardized documentation and review processes to improve the chances for successful experiment re-runs and artifact re-use.

A number of ACM conferences and journals have already instituted formal processes and are implementing Best Practices for artifact review. ACM provides standard terms and definitions for labeling successful artifact reviews, and iconic badging for their associated articles, thereby establishing uniformity across ACM publications and any choosing to adopt its Best Practices.

Reproducibility of Results in the ACM DL

Association for Computing Machinery

Software Artifact Integrations

ACM Digital Library-Curation Platform Integrations

ACM Badges

ACM Reproducibility Task Force

DL Pilot Integrations

Reproducibility Task Force

ACM Reproducibility Task Force & Workshops

Thursday, 2019-07-11

Artifact Evaluation Track

Authors of accepted research papers are invited to submit an artifact to the ICPE Artifact Track. According to ACM's "Result and Artifact Review and Badging" policy, an "artifact" is "a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself [...] software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyze results". A formal review of such artifacts not only ensures that the "study is repeatable" by the same team, if they are available online then other researchers "can replicate the findings" as well.

In this spirit, the ICPE 2019 Artifacts Track exists to review, promote, share and catalog the research artifacts produced by any of the full papers accepted to the research track. Apart from repeatability and replicability, cataloguing these artifacts also allows reuse by other teams in reproduction or other studies. Artifacts of interest include (but are not limited to):

- Tools, libraries or frameworks, which are implementations of systems or algorithms essential for the results described in the associated paper, possibly also useful in other work.
- Data or repositories, which are essential for the results described in the associated paper, ideally also useful in other work.

The authors must ensure that at the camera ready deadline, the artifacts are generally available from a stable URL or DOI with an archival plan, such as the SPEC RG Zenodo repository (personal page is not sufficient).

If you require an exception from the conditions above, please mail the chairs before submitting.

What do you get out of it?

If your artifact is accepted, it will receive one of the following badges in the text of the paper and in the ACM Digital Library:

spec

Tweets von @ICPEcc

ACM/ SPEC ICPE 2019 hat retweetet

Cor-Paul Bezemer @corpaul

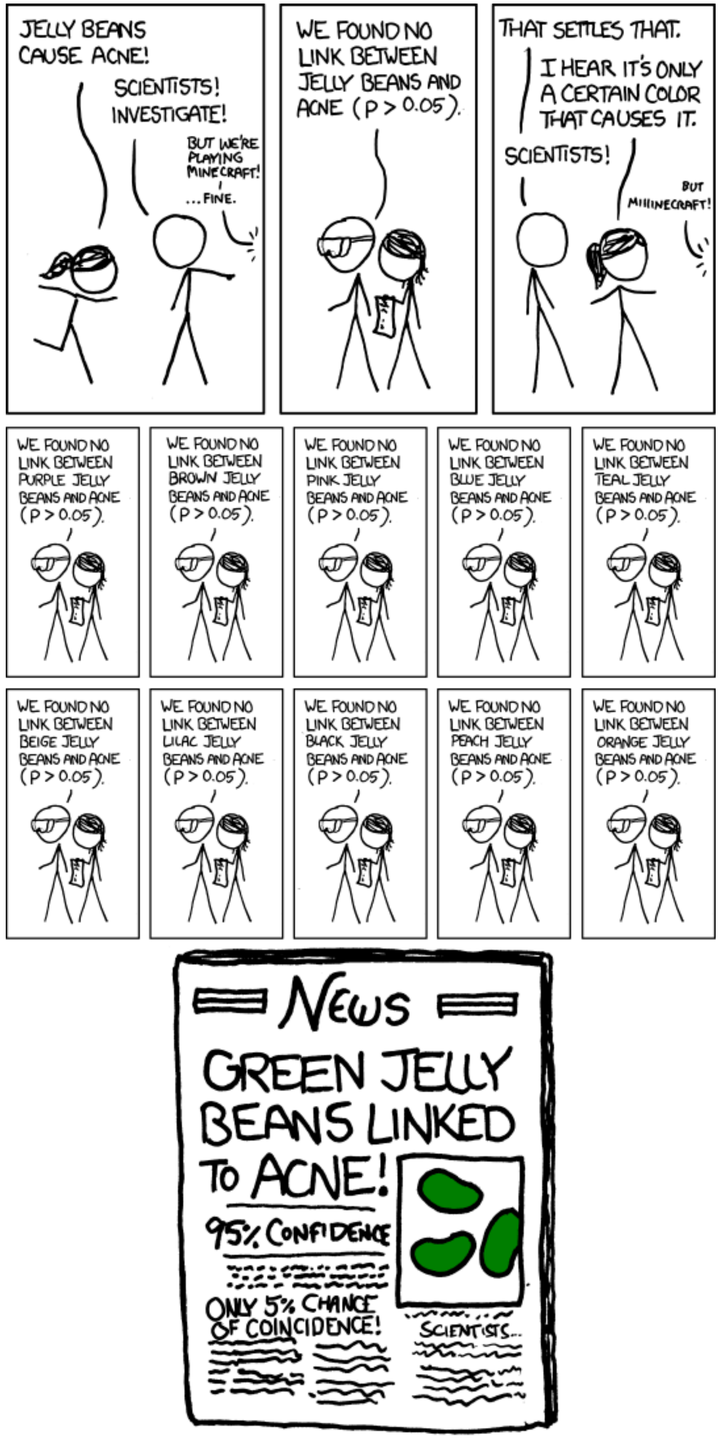
The @ICPEconf 2020 (in Edmonton!) call for workshops has been published. Are you interested in organizing a workshop on a performance engineering-related topic? We're looking forward to receiving your proposal! icpe2020.spec.org/workshops/

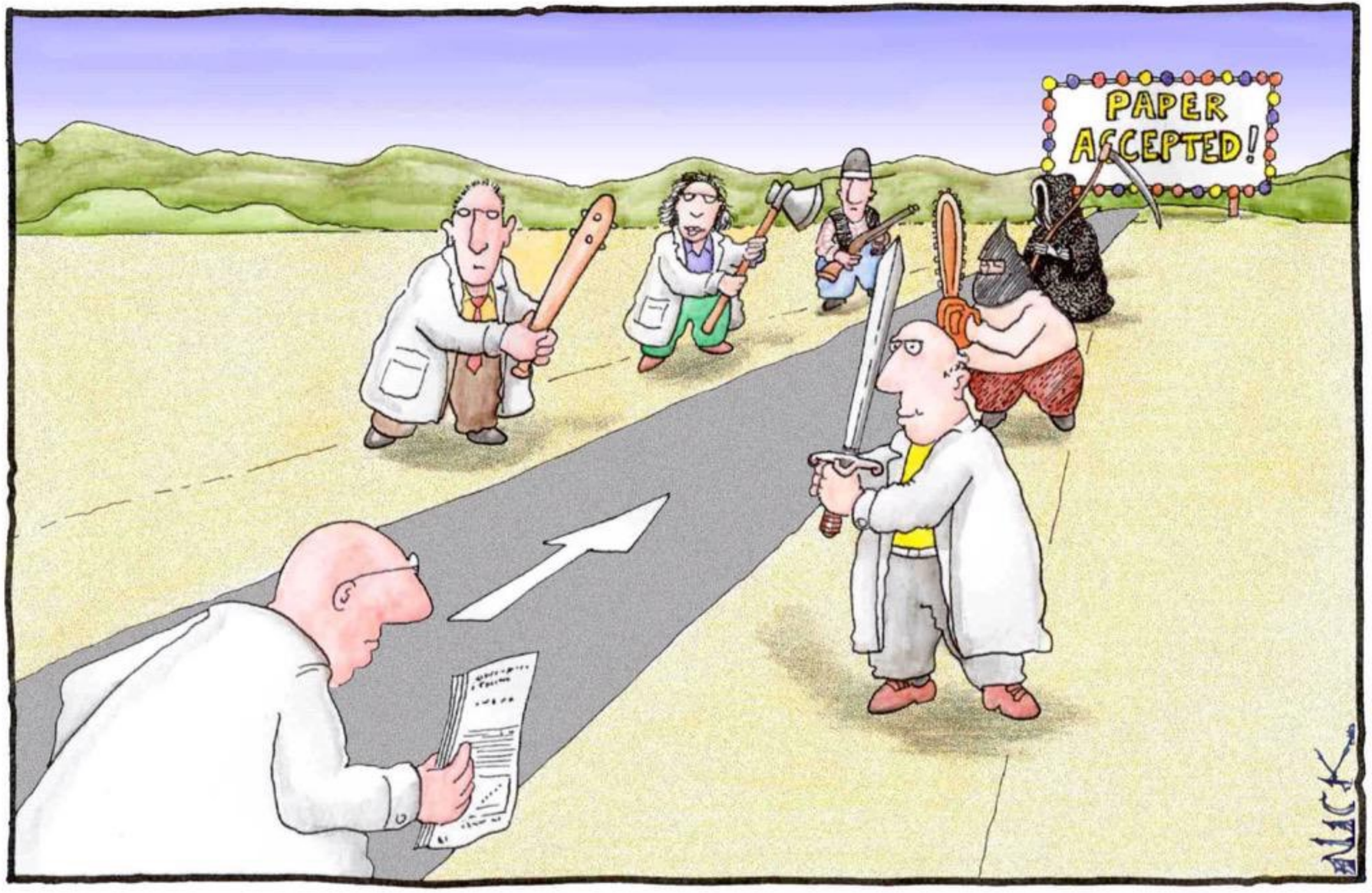
14. Juni 2019

ACM/ SPEC ICPE 2019 hat retweetet

Alexandru Iosup @Alosup

Team @LargeResearch helps out with the @ICPEconf 2020 edition. As Program Co-chair, my job will be to make sure quality and fairness continue to be the prime concerns of our reviewers, but also that

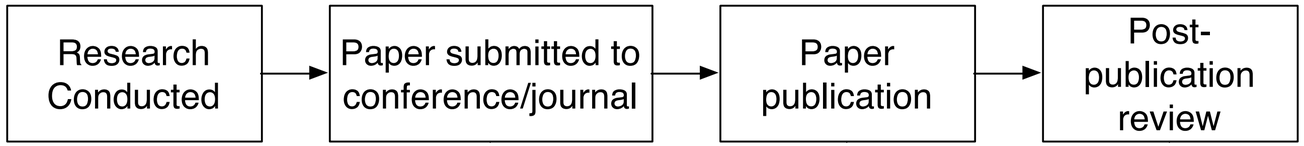

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

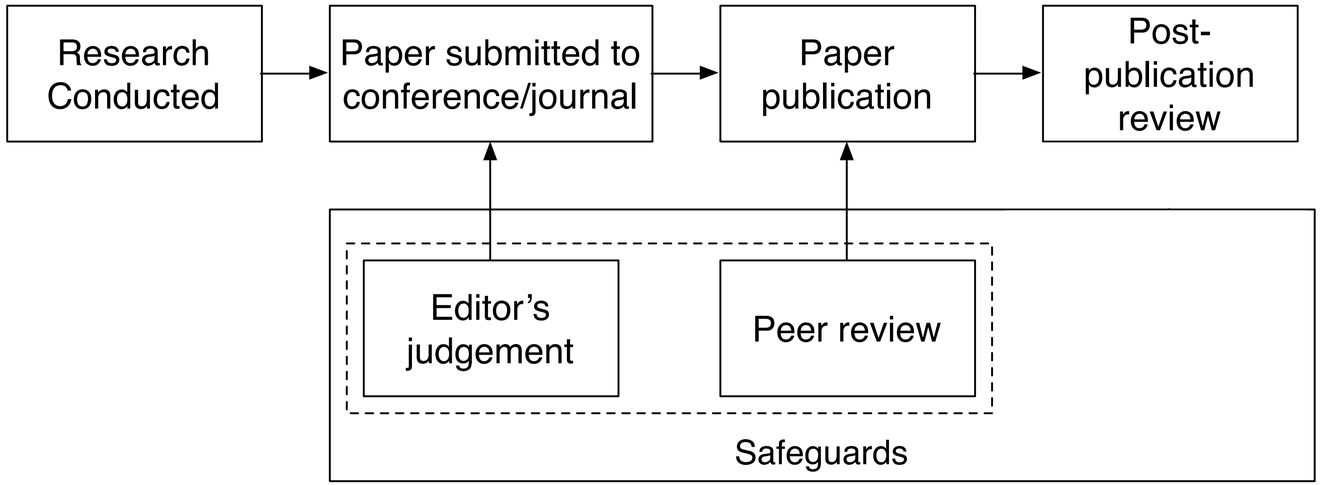


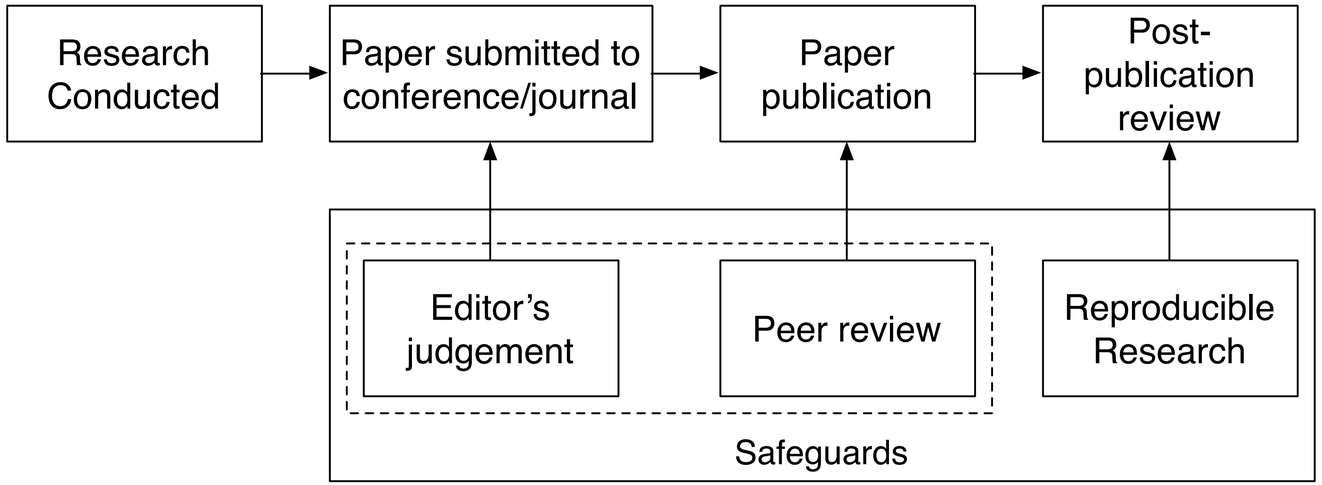
Most scientists regarded the new streamlined peer-review process as ‘quite an improvement.’

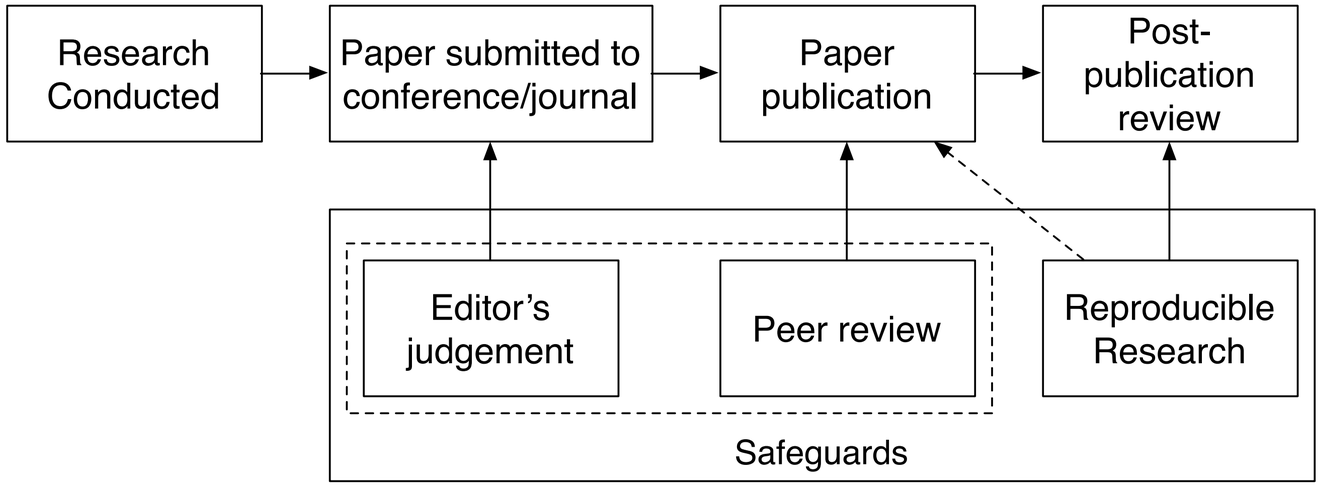
Problems with Reproducibility

- The premise of reproducible research is that with data and code available, people can check each other and the whole system is self-correcting
  - …. in the long run,
  - … where the long run is too long
- Addresses the most “downstream” aspect of the research process: *post-publication*
- Assumes everyone plays by the same rules and wants to achieve the same goals (i.e. scientific discovery)


Research Conducted
Paper submitted to conference/journal
Paper publication
Post-publication review

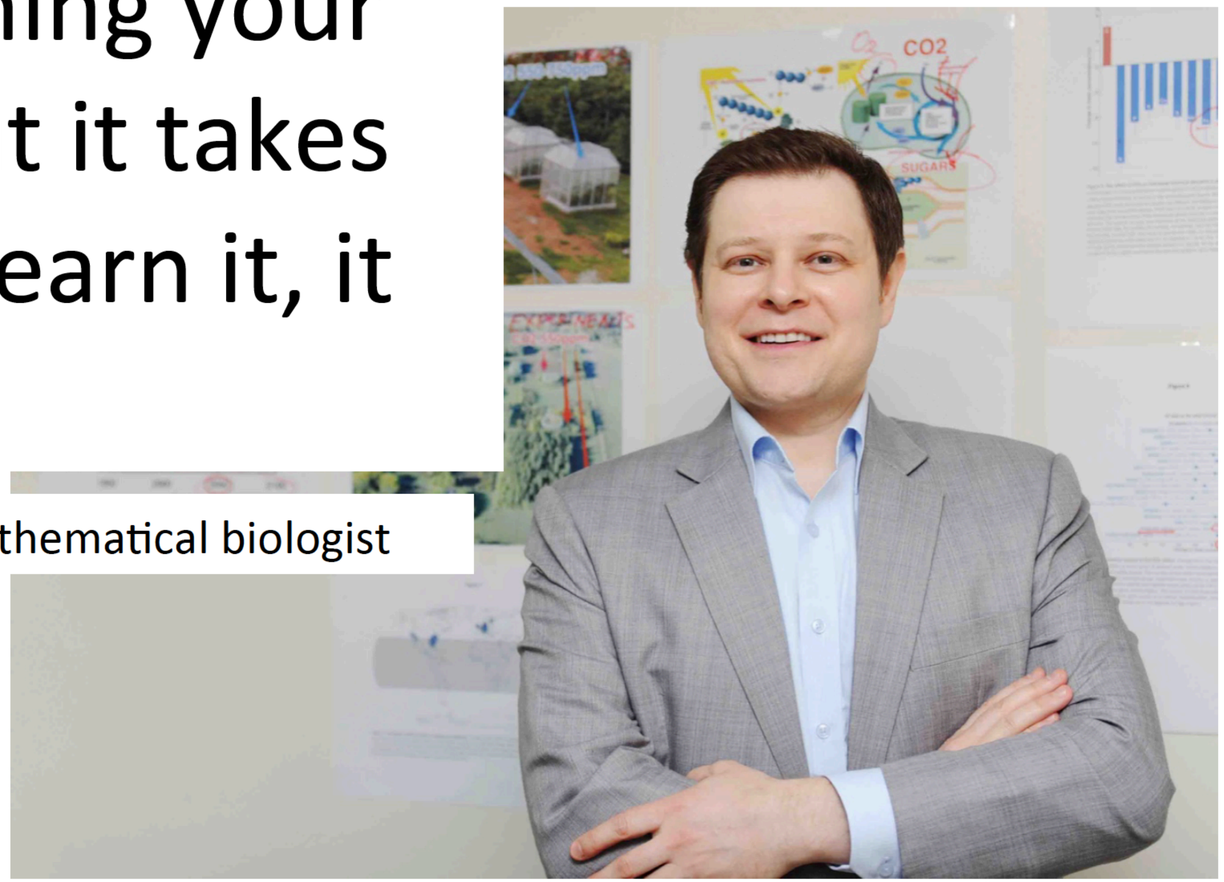

Research Conducted
Paper submitted to conference/journal
Paper publication
Post-publication review
Editor's judgement
Peer review
Safeguards


Research Conducted
Paper submitted to conference/journal
Paper publication
Post-publication review
Editor's judgement
Peer review
Reproducible Research
Safeguards


Research Conducted
Paper submitted to conference/journal
Paper publication
Post-publication review
Editor's judgement
Peer review
Reproducible Research
Safeguards

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

Irakli Loladze, mathematical biologist

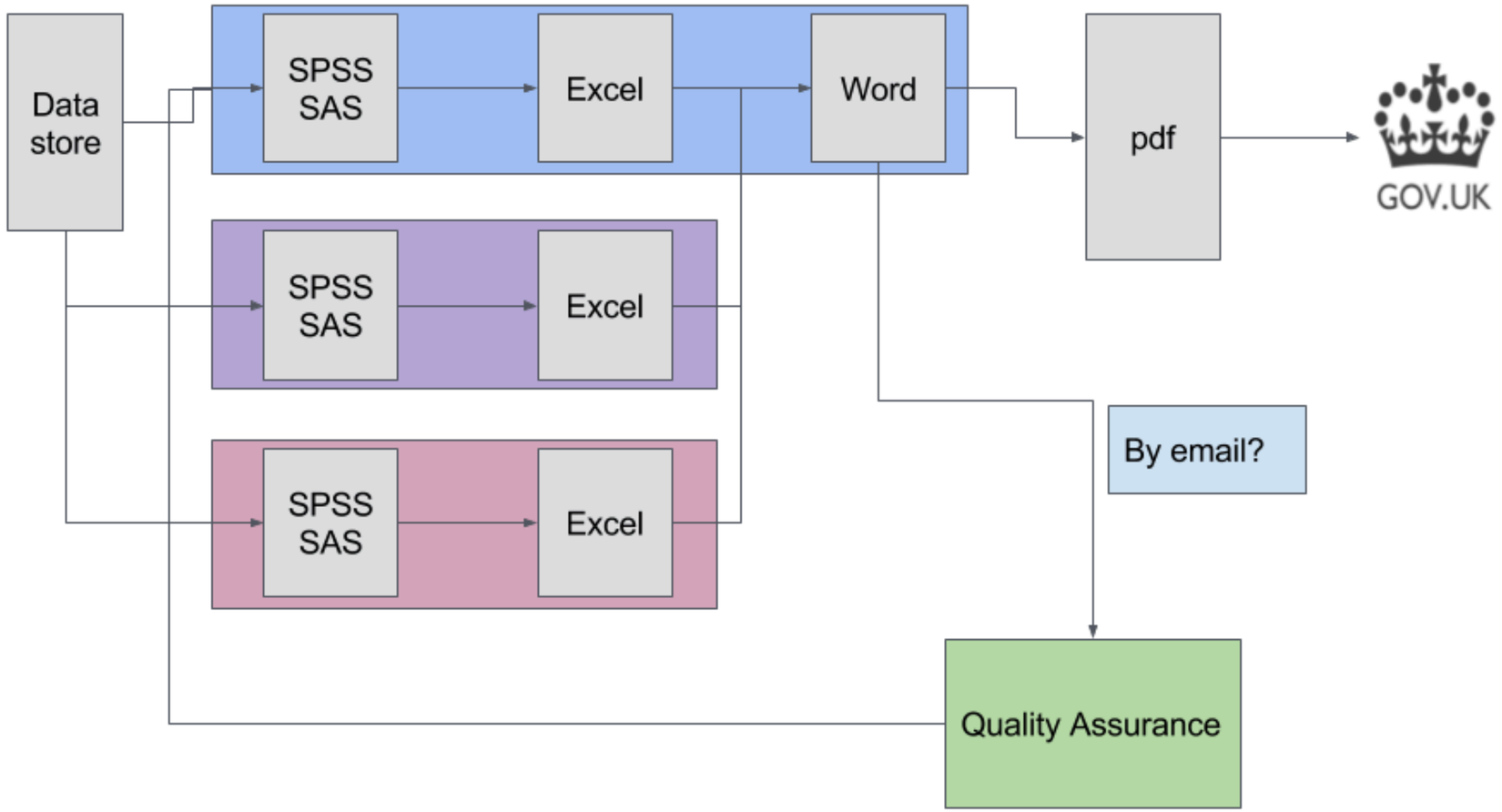


Source: Nature 2016, 533, 531–533,; Picture: http://cu.ac.kr

Return to Zero

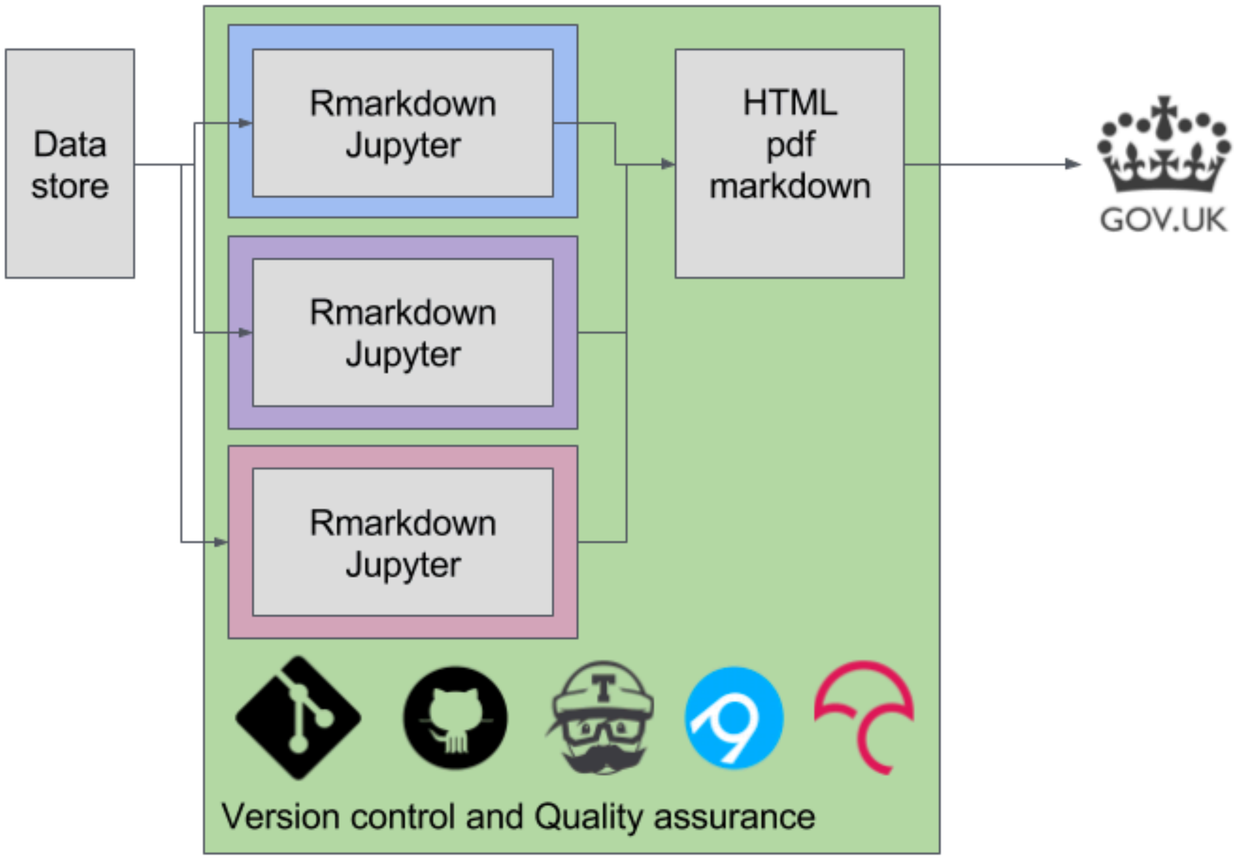
EEWeb.com

A case study of gov.uk

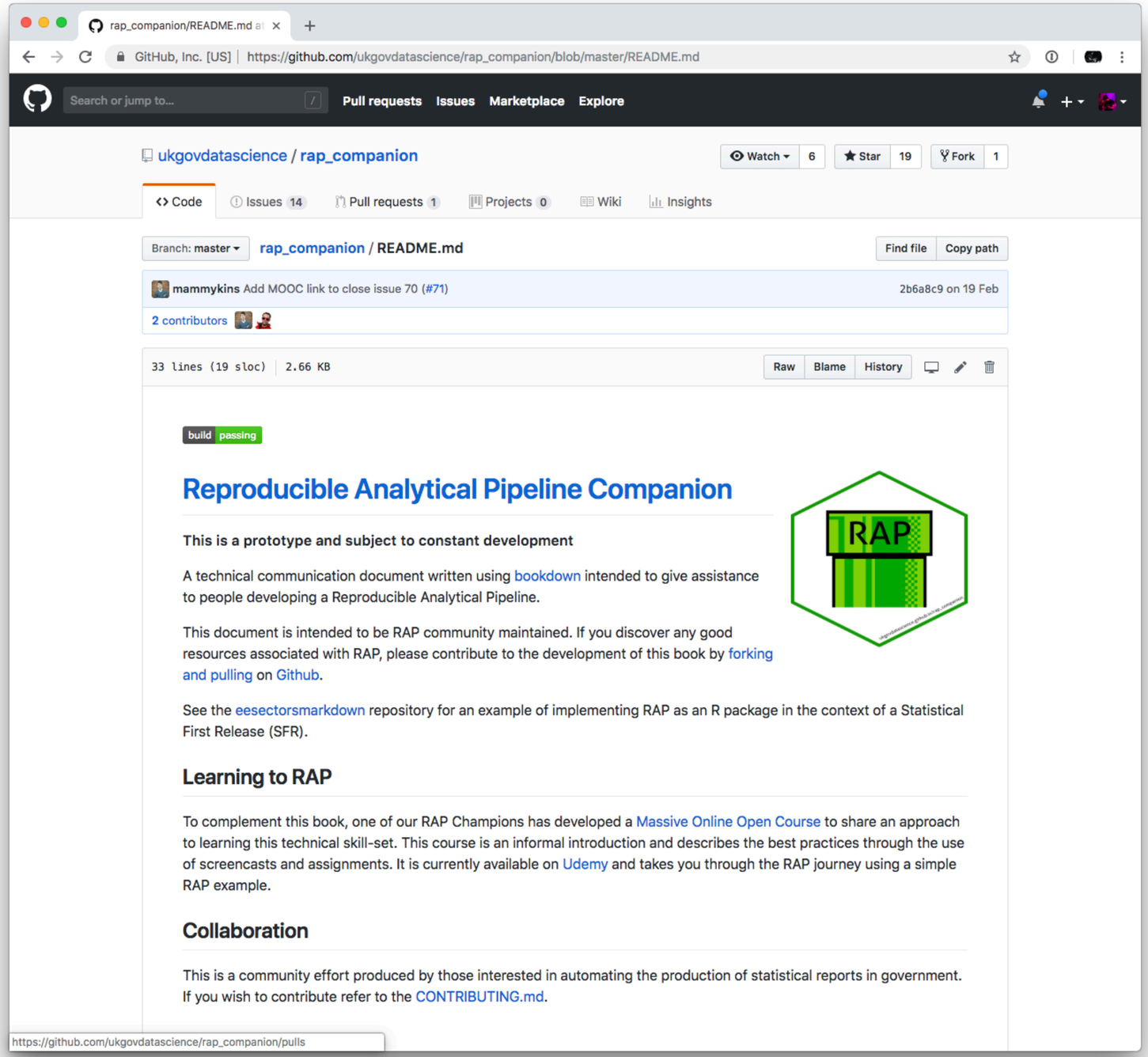


https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process

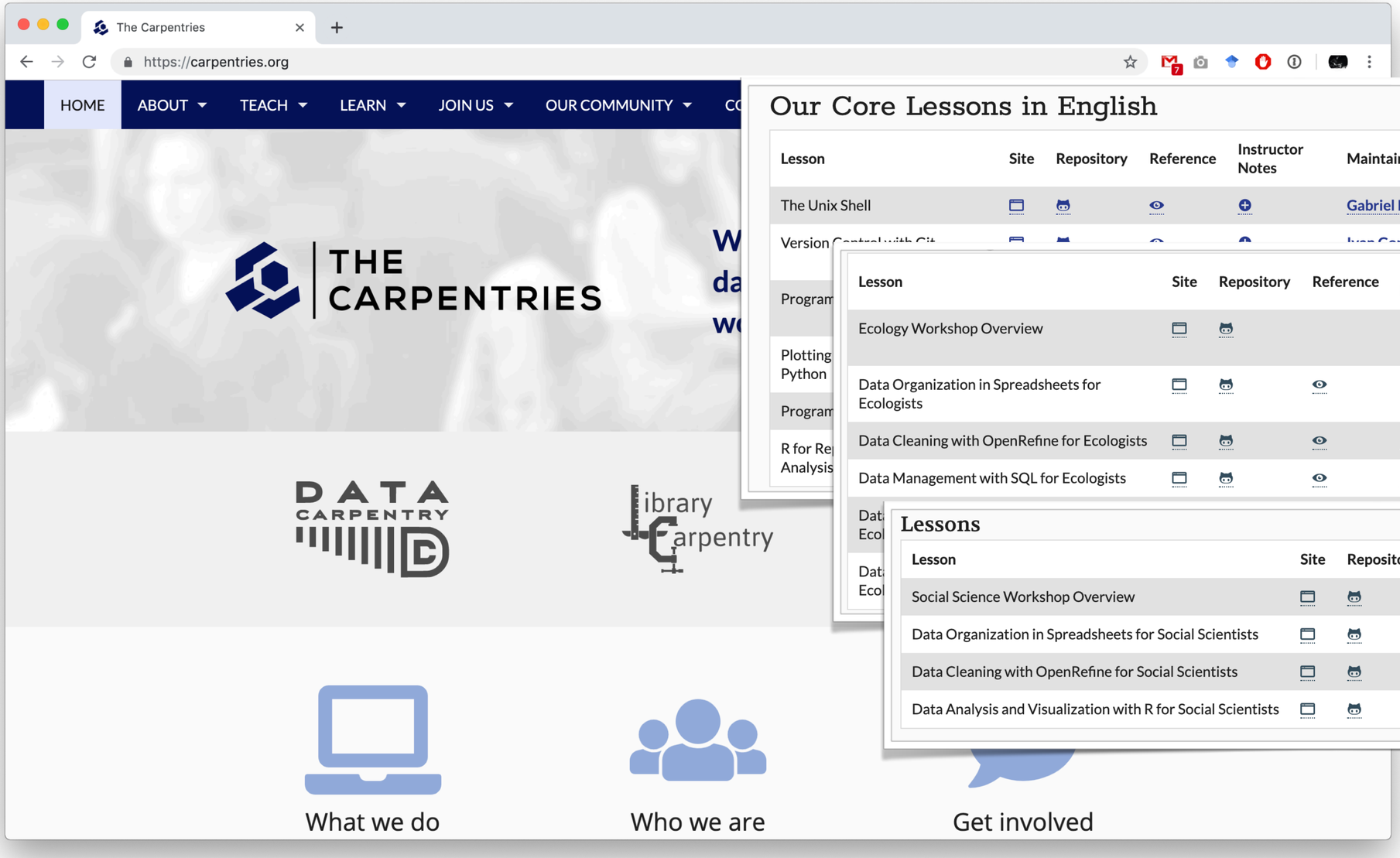
A case study of gov.uk

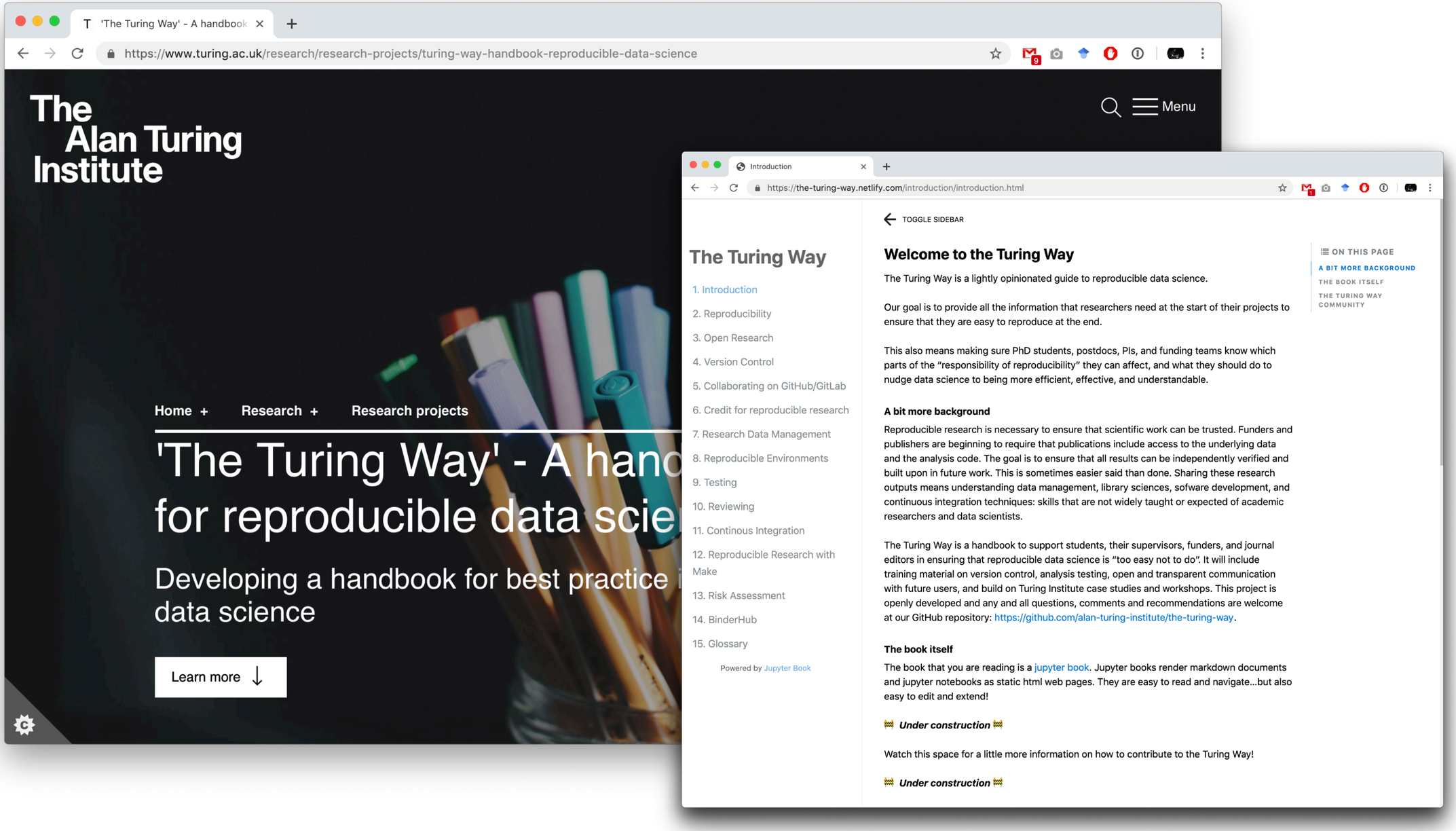


https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process



https://github.com/ukgovdatascience/rap_companion


The Carpentries
https://carpentries.org
HOME
ABOUT
TEACH
LEARN
JOIN US
OUR COMMUNITY
THE CARPENTRIES
DATA CARPENTRY
Library Carpentry
What we do
Who we are
Get involved
Our Core Lessons in English
Lesson
Site
Repository
Reference
Instructor Notes
Maintainer(s)
The Unix Shell
Gabriel Devenyi, Colin Morris, Will Pitchers, Gerard Ca
Ecology Workshop Overview
Karen Cranston, Aleksandra Pa White
Data Organization in Spreadsheets for Ecologists
Christie Bahlai, Tracy Teal, Pete
Data Cleaning with OpenRefine for Ecologists
Deborah Paul, Cam Macdonell
Data Management with SQL for Ecologists
Timothée Poisot, Rémi Rampin
Lessons
Lesson
Site
Repository
Reference
Instructor Notes
Maint
Social Science Workshop Overview
Angel
Data Organization in Spreadsheets for Social Scientists
Chris
Data Cleaning with OpenRefine for Social Scientists
Geoff
Data Analysis and Visualization with R for Social Scientists
Juan


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

How might we support researchers at Newcastle?
What barriers do you see in your roles?

Discussion

Open Science is also an opportunity:

- To make the entire research process **transparent** and **accessible**
- To increase **cooperation** and **efficiency**, and therefore impact
- To change in the way scholarly knowledge is **produced, evaluated and disseminated**
- To enable trust and **facilitate knowledge transfer** from academia to society

A fantastic research catalyst!