

REANA reproducible analyses

Tibor Šimko

@tiborsimko

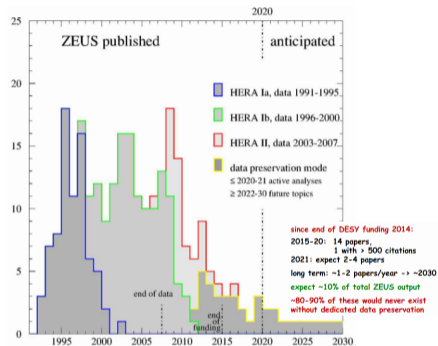
openlab summer lecture, July 28th 2022

Contents

1. REANA presentation
2. REANA demo

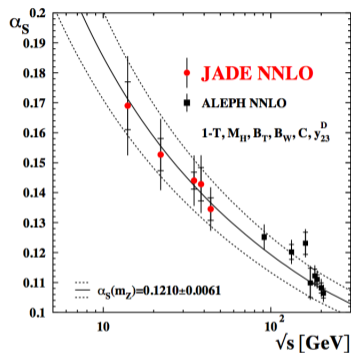
1. REANA presentation

Long-term value of data!



Achim Geiser <https://indico.cern.ch/event/1009487>

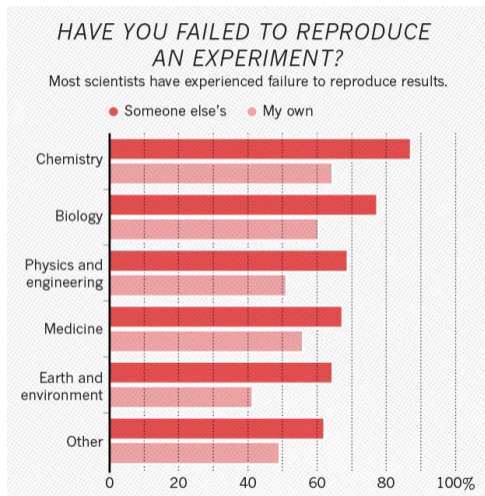
Collaborations publish papers even fifteen years after data taking ends.



DPHEP <https://arxiv.org/abs/1205.4667>

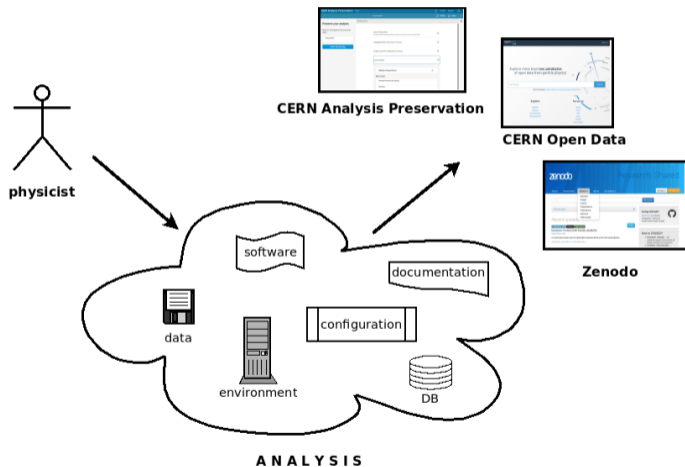
JADE data (1979–1986) still unique even forty years later.

Half of researchers cannot reproduce their own results



<https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>

Preserving analysis knowledge: data, code and more



The FAIR guiding principles for scientific data management

- ▶ **F**indable
- ▶ **A**ccesible
- ▶ **I**nteroperable
- ▶ **R**eusable

Capturing analysis assets in digital repositories to facilitate their future **reuse**

Reusability? Repeatability? Replicability? Reproducibility?

The Turing Way model

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

<https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html>

The PRIMAD model

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-	-	Determinism
Param. Sweep	x	-	-	-	-	-	-	Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-	-	Applicability across different settings
Port	-	-	x	-	-	-	-	Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-	-	Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-	-	Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	-	x	Apply code in different settings, Re-purpose
Independent x (orthogonal)							x	Sufficiency of information, independent verification

■ **Figure 1** PRIMAD Model: Categorizing the various types of reproducibility by varying the (P)latform, (R)esearch Objective, (I)mplementation, (M)ethod, (A)ctor and (D)ata, analyzing the gain they bring to computational experiments. x denotes the variable primed i.e. changed, (x) a variable that may need to be changed as a consequence, whereas - denotes no change.

https://drops.dagstuhl.de/opus/volltexte/2016/5817/pdf/dagrep_v006_i001_p108_s16041.pdf

From “reproducible” to “reusable” analyses

Four pillars of reusable computational research

I. Input data

What is your input data?

- input files
- input parameters

II. Analysis code

Which code analyses it?

- user code
- software frameworks

III. Computing environment

What is your environment?

- operating system
- database calls

IV. Computational recipes

Which steps did you take?

- shell commands
- notebooks and workflows

I. Data and II. Code

open data commons

Search

Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data

/QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12_DR53X-FU_RD1_START53_V7N-v1/AODSIM_CMS collaboration

Cite as: CMS collaboration (2017), Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data. CERIN Open Data Portal. DOI:10.7483/OPENDATA-CMS.V17B.XZNV

Download Simulated Dataset Python C++ C# R MATLAB

Description

Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in: [About CMS simulated dataset names](#).

These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.

Dataset characteristics

30125269 events, 20958 files, 9.6 TB in total.

System details

Recommended [global tag](#) for analysis: START53_V27:All
Recommended release for analysis: CMS5W_5_3_32

How were these data generated?

These data were generated in several steps (see also [CMS Monte Carlo production overview](#)):

Step SIM

Release: CMS5W_5_0_0_patch2
Global Tag: START50_V13:All
Generators: pythia6

- Production script (preview)
- Generator parameters (preview) (link)

Output dataset: /QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12-START50_V13-v1/GEN-SIM

Step HLT RECO

Release: CMS5W_5_3_14
Global Tag: START53_V7N:All

- Production script (preview)
- Configuration file for HLT (link)
- Configuration file for RECO (link)

Output dataset: /QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12_DR53X-FU_RD1_START53_V7N-v1/AODSIM

zenodo

Search

22,710 views 1,159 downloads

mwaskom/seaborn: v0.10.1 (April 2020)

Michael Waskom, Olga Borusniak, Joel Deckert, Mica DeBart, Gavriel Luboskov, Paul Helder, David C Serpentine, Tom Augspurger, Harshita Halhotra, John B. Gale, Jesse Warmuth, Allan de Barros, Cameron Pyle, Stephan Hoyer, Jake VanderPlas, Sarah Whalley, Sara Kottler, Eric Dumoulin, Peter Bachmann, Marcin Mońka, Rob Moore, Corbin Swier, Aidan Miles, Thomas Brunner, Drew D'Kaas, Tal Yarkoni, Mike Lee Williams, Constanze Elnsitz, Clark Fitzgerald, Brian

This is minor release with bug fixes for issues identified since 0.10.0.

- Fixed a bug that appeared within the bootstrapping algorithm on 32-bit systems.
- Fixed a bug where `regress` would crash on singleton inputs. Now a crash is avoided and regression estimation plotting is skipped.
- Fixed a bug where `boxenplot` would ignore user specified under/over/both values when resampling a column.
- Fixed a bug where `boxenplot` would use values from masked cells when computing default colormap limits.
- Fixed a bug where `regress` would issue an error when trying to fit against one or multiple categorical sets.
- Adapted to a change in matplotlib that caused problems with single swarm plots.
- Added the `axlims` parameter to `boxenplot` to suppress plotting of outlier data points, matching the API of `boxplot`.
- Avoided seeing an error from `statmodels` when data with an IQR of 0 is passed to `regress`.
- Added the `logprob` option to the `plotting_context` backend.
- Deprecated several utility functions that are no longer used internally (`convert_dates`, `log_probs`, `get_style`, and `next_id`).

Files (2020-04)

Name	Size
mwaskom/seaborn-v0.10.1.zip	335.0 kB

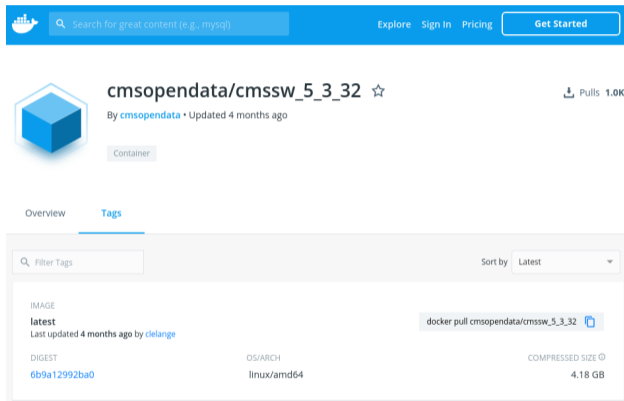
Collaborators: [fjet](#)

Show only: Literature (107) Unknown (6) Dataset (3) Software (8)

Data preserved in digital repositories

... as is the code

III. Environment



The screenshot shows the Docker Hub page for the image `cmsopendata/cmssw_5_3_32`. The page includes a search bar, navigation links for 'Explore', 'Sign In', 'Pricing', and 'Get Started'. The image details show it was created by `cmsopendata` and updated 4 months ago. It is a container image with a 'latest' tag, last updated 4 months ago by `delange`. The digest is `6b9a12992ba0`, the OS/ARCH is `linux/amd64`, and the compressed size is `4.18 GB`. A 'Tags' section is visible with a search filter and a 'Sort by Latest' dropdown.

```
> ls -l /cvmfs/cms-opendata-conddb.cern.ch/
total 1655262
drwxr-xr-x. 2 cvmfs cvmfs      24 Jan 21 2016 FT_53_LV5_AN1
drwxr-xr-x. 2 cvmfs cvmfs      24 Feb 22 2016 FT_53_LV5_AN1_RUNA
drwxr-xr-x. 2 cvmfs cvmfs     366 Jun 21 2017 FT53_V21A_AN6
drwxr-xr-x. 2 cvmfs cvmfs     365 Nov 29 2017 FT53_V21A_AN6_FULL
drwxr-xr-x. 2 cvmfs cvmfs     365 Jun 23 2017 FT53_V21A_AN6_RUNC
drwxr-xr-x. 2 cvmfs cvmfs       3 Oct 20 2017 FT_R_42_V10A
drwxr-xr-x. 2 cvmfs cvmfs      248 Nov  9 2018 START42_V17B
drwxr-xr-x. 2 cvmfs cvmfs      282 Jan 21 2016 START53_LV6A1
drwxr-xr-x. 2 cvmfs cvmfs      394 Jun 21 2017 START53_V27
drwxr-xr-x. 2 cvmfs cvmfs      296 Nov 30 2018 START53_V7N
-rw-r--r--. 1 cvmfs cvmfs 1002414080 Oct 31 2018 102X_upgrade2018_design_v9.db
-rw-r--r--. 1 cvmfs cvmfs 691593216 Oct 31 2018 80X_mcRun2_asymptotic_2016_TrancheIV_v8.db
-rw-r--r--. 1 cvmfs cvmfs  82944 Jan 21 2016 FT_53_LV5_AN1.db
-rw-r--r--. 1 cvmfs cvmfs  82944 Feb 22 2016 FT_53_LV5_AN1_RUNA.db
-rw-r--r--. 1 cvmfs cvmfs 119808 Jun 21 2017 FT53_V21A_AN6.db
-rw-r--r--. 1 cvmfs cvmfs 120832 Nov 29 2017 FT53_V21A_AN6_FULL.db
-rw-r--r--. 1 cvmfs cvmfs 120832 Jun 23 2017 FT53_V21A_AN6_RUNC.db
-rw-r--r--. 1 cvmfs cvmfs  64512 Oct 20 2017 FT_R_42_V10A.db
-rw-r--r--. 1 cvmfs cvmfs  72704 Nov  9 2018 START42_V17B.db
-rw-r--r--. 1 cvmfs cvmfs  84992 Jan 21 2016 START53_LV6A1.db
-rw-r--r--. 1 cvmfs cvmfs 130048 Jun 21 2017 START53_V27.db
-rw-r--r--. 1 cvmfs cvmfs  89088 Nov 30 2018 START53_V7N.db
```


Condition DB snapshot living on CernVM File System

Containerised CMS software framework

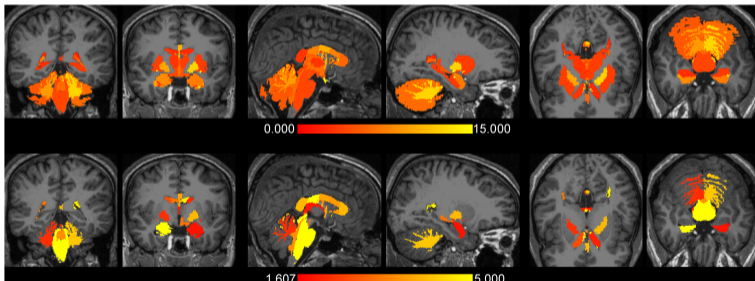


III. Environment: an example from life sciences

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

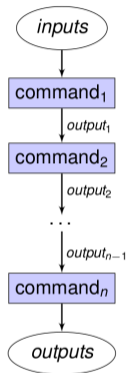
Ed H. B. M. Gronenschild , Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: June 1, 2012 • DOI: 10.1371/journal.pone.0038234



Software changes (Freesurfer 4.3.1, 4.5.0, 5.0.0): $8.8 \pm 6.6\%$ (volume) and $2.8 \pm 1.3\%$ (thickness)
Operating system changes (macOS 10.5, 10.6): about factor two smaller

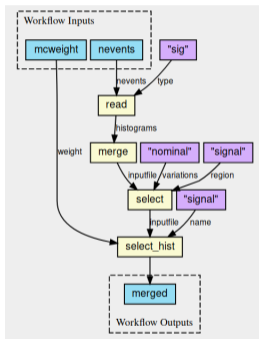
IV. Computational workflows



Serial



Yadage



CWL



Snakemake

reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



Reusable

Containerise once, reuse elsewhere. Cloud-native.



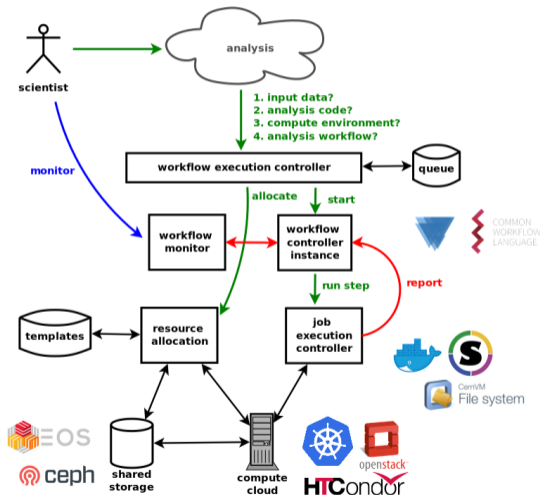
Free

Free Software. GPL licence. Made with ❤️ at CERN.



<https://www.reana.io/>

REANA architecture



Respecting diverse habits of diverse research groups

- ▶ multiple workflow systems (CWL, Serial, Snakemake, Yadage)
- ▶ multiple container technologies (Docker, Singularity)
- ▶ multiple compute backends (Kubernetes, HTCondor, Slurm)
- ▶ multiple shared storage platforms (Ceph, EOS, NFS)

REANA command-line and web interface

```
1 version: 0.6.0
2 inputs:
3   files:
4     - code/gendata.C
5     - code/fitdata.C
6   parameters:
7     events: 20000
8     data: results/data.root
9     plot: results/plot.png
10 workflow:
11   type: serial
12   specification:
13     steps:
14     - name: gendata
15       environment: 'reanahub/reana-env-root6:6.18.04'
16       commands:
17       - mkdir -p results && root -b -q 'code/gendata.C(${events},${data})'
18     - name: fitdata
19       environment: 'reanahub/reana-env-root6:6.18.04'
20       commands:
21       - root -b -q 'code/fitdata.C(${data},${plot})'
22 outputs:
23   files:
24     - results/plot.png
```

The screenshot displays the REANA web interface. At the top, there are two terminal-like windows showing the output of the `reana-client status` command. The first window shows a workflow named 'rootfit' with run number 1, created on 2021-03-08T12:47:30, and currently in a 'running' state with 0/2 tasks completed. The second window shows the same workflow after completion, with a status of 'finished' and 2/2 tasks completed. Below these, the 'reana' logo is visible. The main part of the interface shows a workflow card for 'rootfit #1', which is 'finished in 2 min 27 sec' and 'step 2/2'. It includes links for 'Job Logs', 'Step: fitdata', 'Workflow ID', 'Compute back', 'Job ID: rean', 'Docker image', and 'Status: fin'. A 'reana' logo is also present in the middle of the interface. At the bottom, there is a table listing workflow files and their modification times, and a plot titled 'Fit example' showing a peak in a histogram.

Name	Modified
code/gendata.C	2021-03-08T12:47:30
code/fitdata.C	2021-03-08T12:47:30
results/data.root	2021-03-08T12:51:00
results/plot.png	2021-03-08T12:51:10

Fit example

Structure data analysis by means of declarative workflows

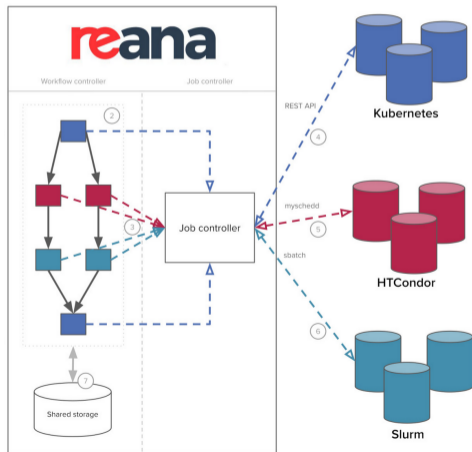
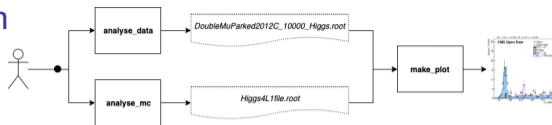
Use command-line and web interfaces to run analysis on remote compute clusters

An advantage of declarative approach

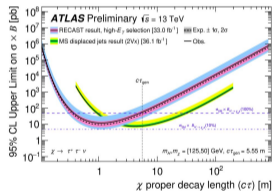
steps:

```
analyse_data:
  run: analyse_data.cwl
  hints:
    reana:
      compute_backend: slurmcern
  out: [DoubleMuParked2012C_10000_Higgs.root]
analyse_mc:
  run: analyse_mc.cwl
  hints:
    reana:
      compute_backend: htcondorcern
  out: [Higgs4L1file.root]
make_plot:
  run: make_plot.cwl
  hints:
    reana:
      compute_backend: kubernetes
  in:
    DoubleMuParked2012C_10000_Higgs: >
      analyse_data/DoubleMuParked2012C_10000_Higgs.root
    Higgs4L1file: >
      analyse_mc/Higgs4L1file.root
  out: [mass4l_combine_userlv13.pdf]
```

A three-step CWL hybrid workflow

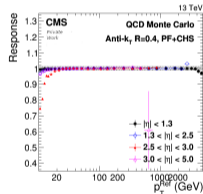
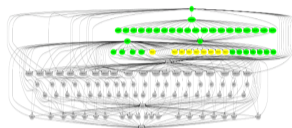


Data analysis and data production examples



ATLAS <https://cdsweb.cern.ch/record/2714064>

Data analysis example: ATLAS displaced jet search reinterpretation



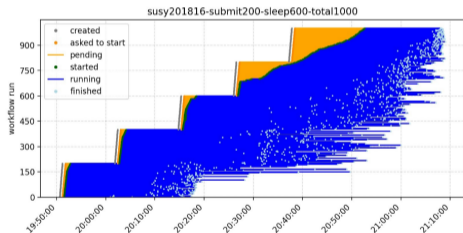
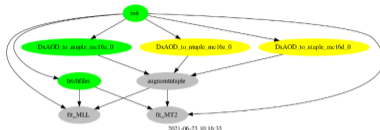
CMS <https://github.com/alintulu/reana-demo-JetMETAnalysis>

Data production example: CMS jet energy resolution and corrections

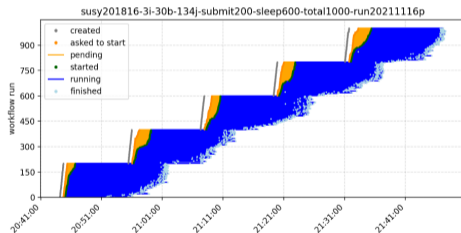
Scalability: running 100k ATLAS pMSSM workflows

ATL-SUSY-2018-16 analysis:

- ▶ NoSys: O(10 minutes); "test" payload
- ▶ AllSys: O(10 hours); "real" payload



Old cluster (448 cores)



New cluster (1072 cores)

Submitting 200 NoSys workflows every 20 minutes

REANA installations



Release	Created
0.8.0-alpha.1	2020-11-25T14:44:01.074Z
0.7.2	2021-02-04T17:45:36.640Z
0.7.1	2020-11-11T09:40:55.784Z
0.7.0	2020-10-21T09:05:18.049Z
0.7.0-alpha.2	2020-10-05T14:25:00.271Z
0.7.0-alpha.1	2020-08-14T16:28:38.333Z

Helm makes it easy to install REANA at scale

Workloads

Cluster: reana Namespace: default

Workloads are deployable units of computing that can be created and managed in a cluster.

Name	Status	Type	Pods	Namespace	Cluster
reana-cache	OK	Deployment	1/1	default	reana
reana-db	OK	Deployment	1/1	default	reana
reana-message-broker	OK	Deployment	1/1	default	reana
reana-server	OK	Deployment	1/1	default	reana
reana-traefik	OK	Deployment	1/1	default	reana
reana-workflow-controller	OK				



REANA on Google Cloud



REANA on US supercomputers

Sociology challenges: adopting containerised workflow paradigm

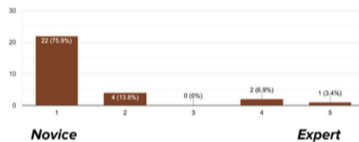


ATLAS/CMS analysis preservation workshop



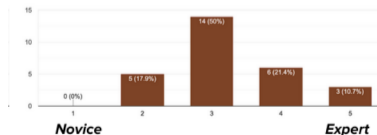
Before

I am confident I can write a containerized workflow that can run my full analysis on the cloud.
29 responses



After

I am confident I can write a containerized workflow that can run my full analysis on the cloud.
28 responses



“Preproducible” analyses

Nature 557 (2018) 613

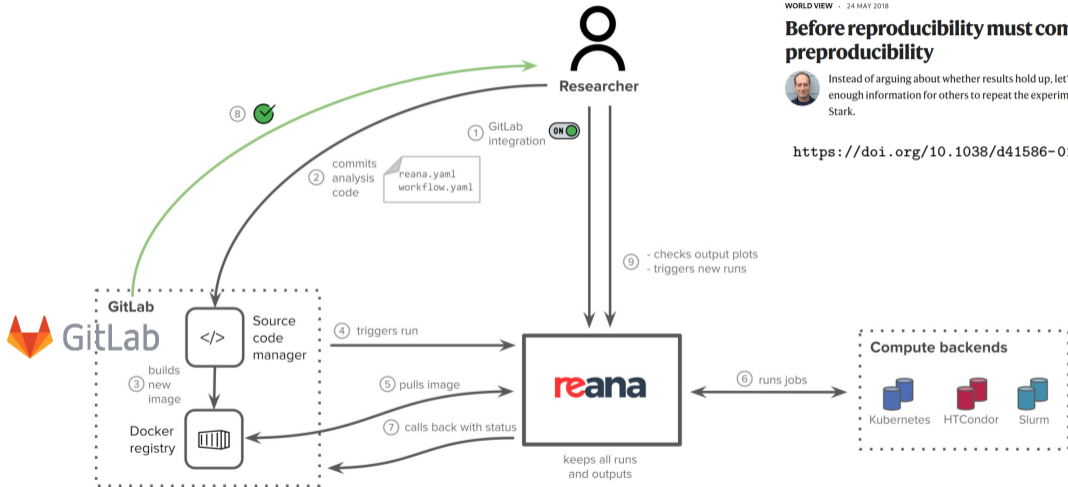
WORLD VIEW · 24 MAY 2018

Before reproducibility must come preproducibility



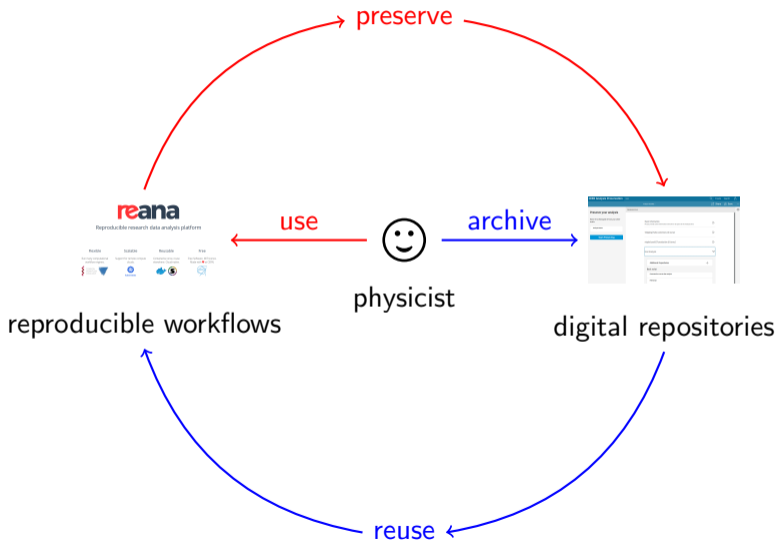
Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says Philip Stark.

<https://doi.org/10.1038/d41586-018-05256-0>



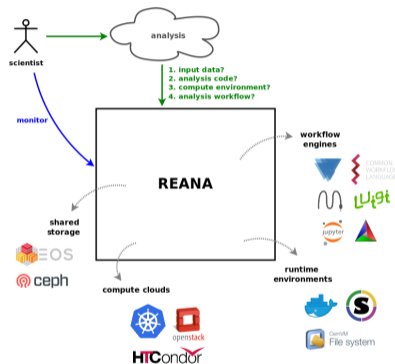
Driving preproducibility via Continuous Integration with source code management systems


Reproducibility \Leftrightarrow Preservation





Conclusions

- ▶ driving reuse through reproducibility
- ▶ data + code + environment + workflow
→ reproducible analyses
- ▶ technology challenges: large containers, complex computational workflows
- ▶ sociology challenges: declarative programming, paradigm shifting, publish-or-perish culture
- ▶ synergies with computational reproducibility needs in astronomy, life sciences



 <https://www.reanahub.io>

 <https://twitter.com/reanahub>

 <https://github.com/reanahub>

2. REANA demo