REANA reproducible analyses

Tibor Šimko

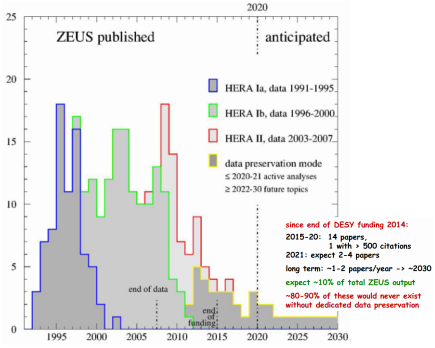
@tiborsimko

openlab summer lecture, July 28th 2022

Contents

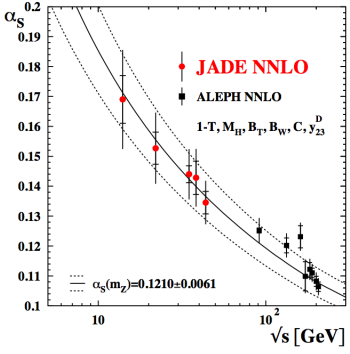
1. REANA presentation

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.

Long-term value of knowledge?

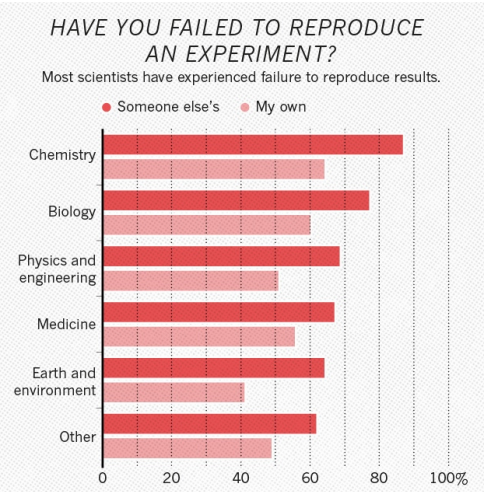
CMS collaboration

First LHCb paper `arXiv.1008.3105`

Experimental physics done by large groups of thousands of physicists.

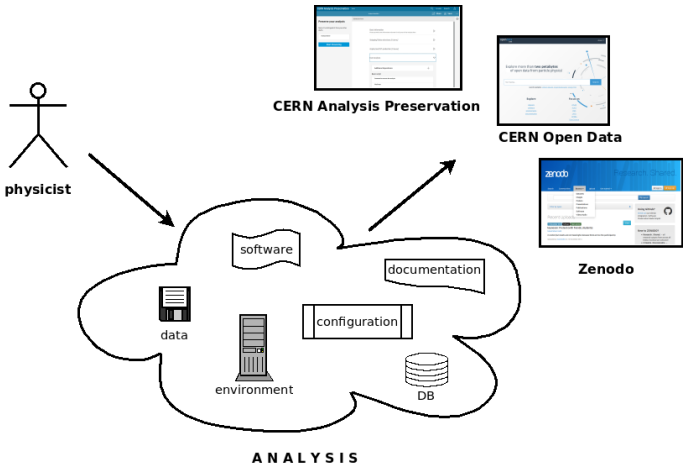
High turnover of young researchers. Half of LHCb authors remain after ten years.

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 knowlege: 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

Analysis	Data: Same	Data: Different
Same	Reproducible	Replicable
Different	Robust	Generalisable

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

The PRIMAD model

Label	Data: Parameters	Data: Raw Data	Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
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?

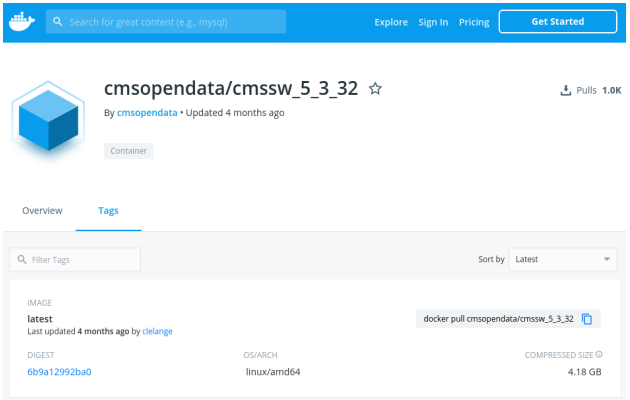
- shell commands
- notebooks and workflows

I. Data and II. Code

Data preserved in digital repositories

... as is the code

III. Environment



Containerised CMS software framework

```
❯ 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
```

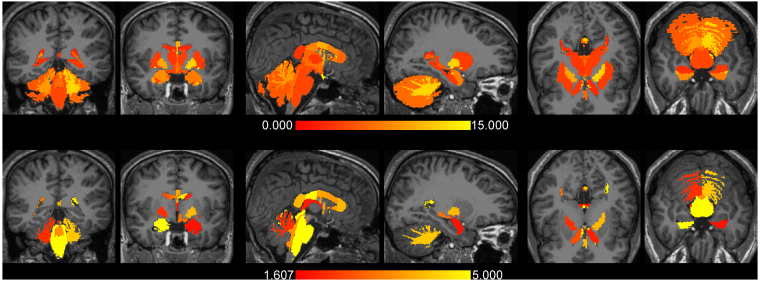
Condition DB snapshot living on CernVM File System

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

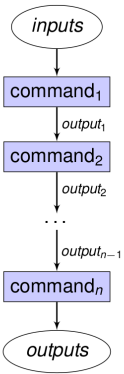
Published: June 1, 2012 • DOI: 10.1371/journal.pone.0038234



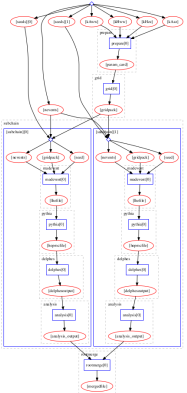
Software changes (Freesurfer 4.3.1, 4.5.0, 5.0.0): 8.8±6.6% (volume) and 2.8±1.3% (thickness)

Operating system changes (macOS 10.5, 10.6): about factor two smaller

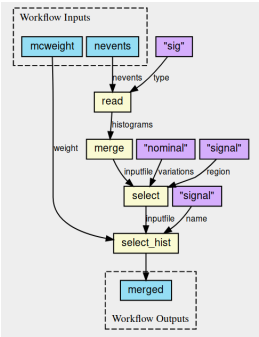
IV. Computational workflows



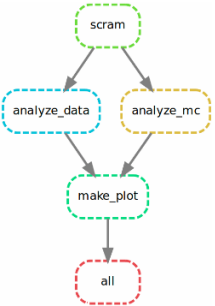
Serial



Yadage



CWL



Snakemake

reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

Scalable

Support for remote compute clouds.

Reusable

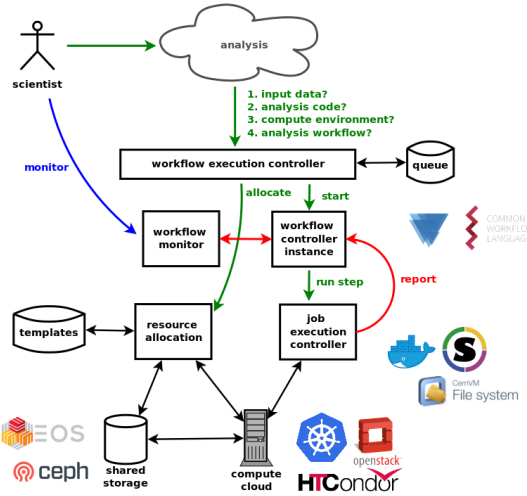
Containerise once, reuse elsewhere. Cloud-native.

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

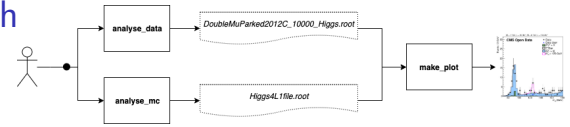
```yaml
version: 0.6.0
inputs:
  files:
    - code/gendata.C
    - code/fitdata.C
  parameters:
    events: 20000
    data: results/data.root
    plot: results/plot.png
workflow:
  type: serial
  specification:
    steps:
      - name: gendata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
        - mkdir -p results && root -b -q 'code/gendata.C(${events},"${data}")'
      - name: fitdata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
        - root -b -q 'code/fitdata.C("${data}","${plot}")'
outputs:
  files:
    - results/plot.png
```

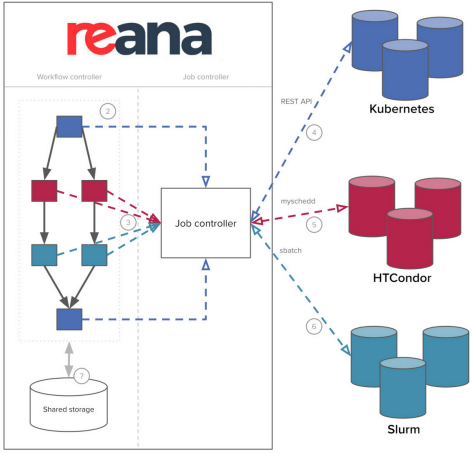
Structure data analysis by means of declarative workflows

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

An advantage of declarative approach

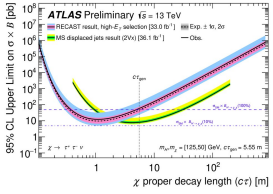
```
steps:
  analyse_data:
    run: analyse_data.cwl
    hints:
      reana:
        compute_backend: slurmcern
    out: [DoubleMuParked2012C_10000_Higgs.root]
  analyse_mc:
    run: analyse_mc.cwl
    hints:
      reana:
        compute_backend: htcondorcern
    out: [Higgs4L1file.root]
  make_plot:
    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_userlvl3.pdf]
```





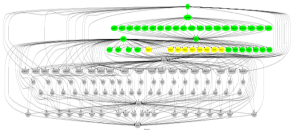
A three-step CWL hybrid workflow

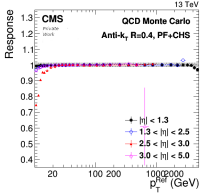
Data analysis and data production examples



ATLAS `https://cdsweb.cern.ch/record/2714064`

Data analysis example: ATLAS displaced jet search reinterpretation





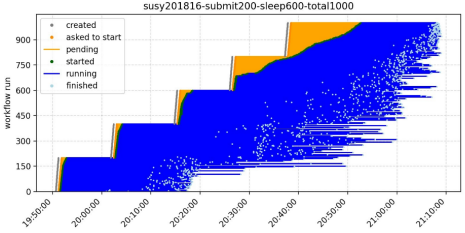
CMS `https://github.com/alintulu/reana-demo-JetMETAnalysis`

Data production example: CMS jet energy resolution and corrections

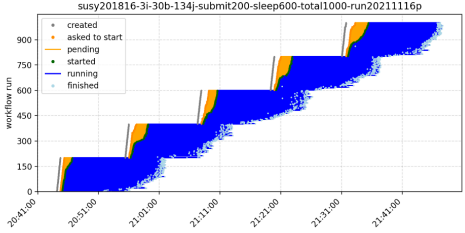
Scalability: running 100k ATLAS pMSSM workflows

ATL-SUSY-2018-16 analysis:

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



Old cluster (448 cores)



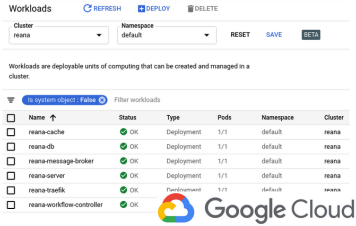
New cluster (1072 cores)

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

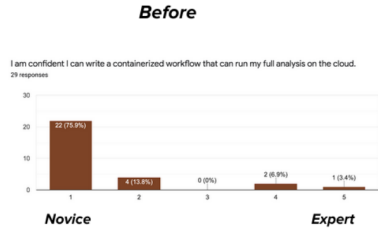


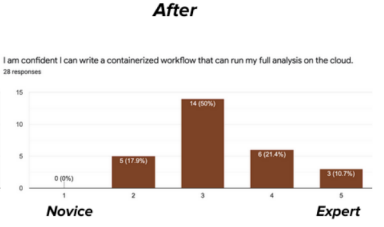
REANA on Google Cloud

REANA on US supercomputers

Sociology challenges: adopting containerised workflow paradigm

ATLAS/CMS analysis preservation workshop





“Preproducible” analyses

Nature **557** (2018) 613

WORLD VIEW · 24 MAY 2018

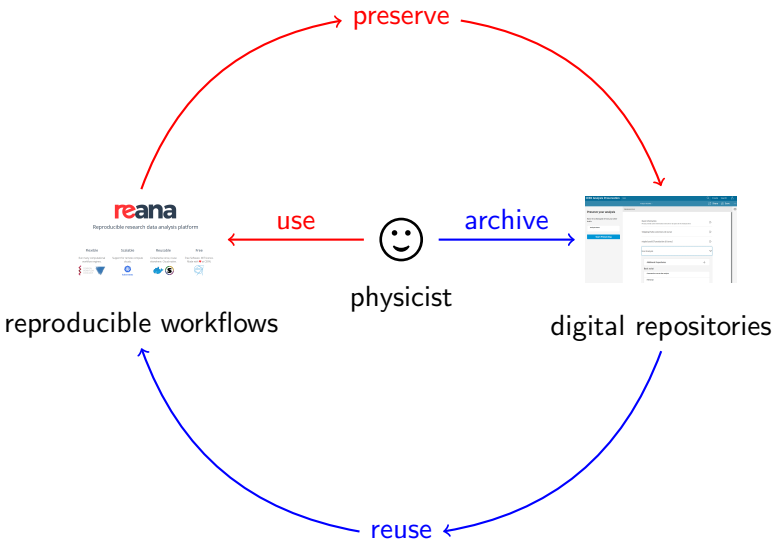
Before reproducibility must come preproducibility

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

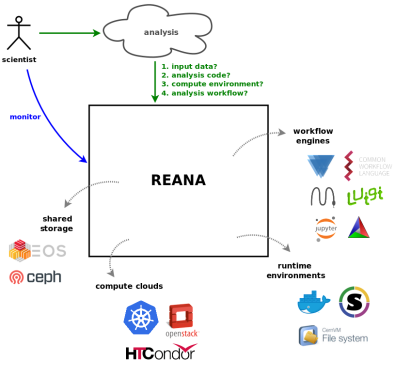
Driving preproducibility via Continuous Integration with source code management systems

Reproducibility ⇌ Preservation



Conclusions

- driving reuse through preproducibility
- 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