# Progress of analysis preservation in HEP community

Mingrui Zhao
China institute of atomic energy

2018.06.28
CEPC physics&software mini-workshop

Based on:
arXiv:1806.08787
https://indico.cern.ch/event/720455/

- A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.

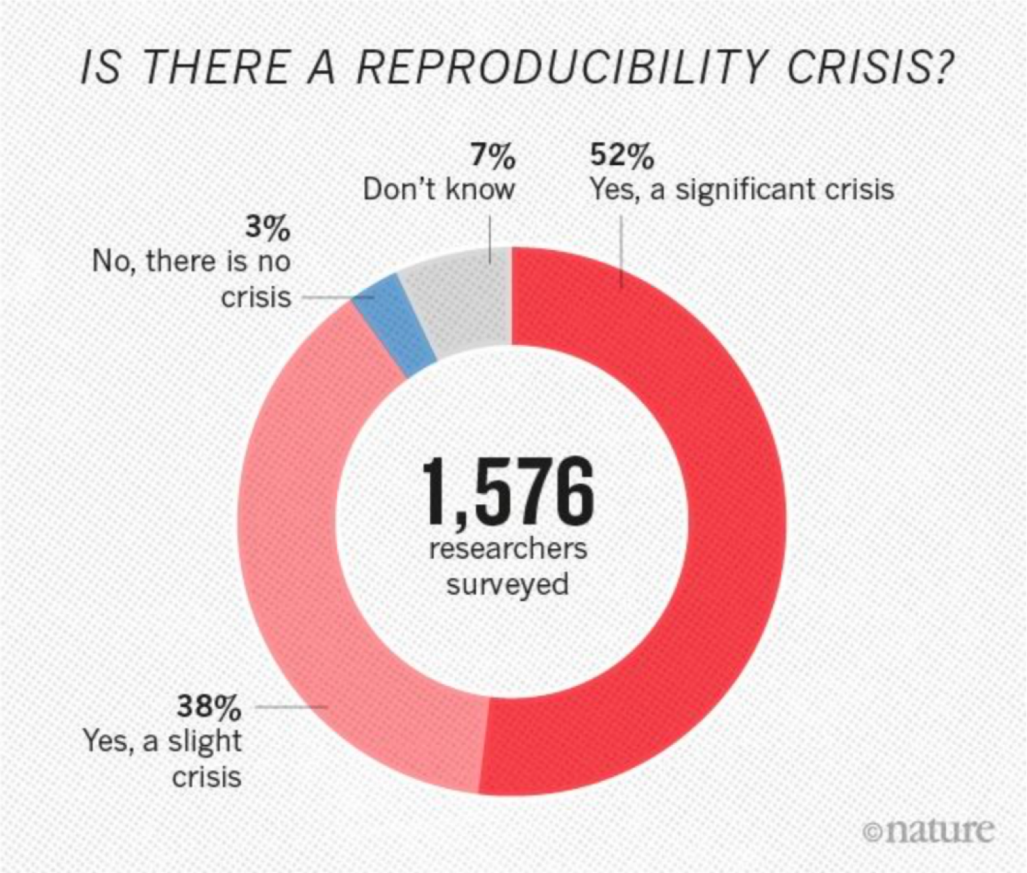
- How to believe your own experiment?
- Preserve your analysis so that your colleagues can check it.

# Problem and situation

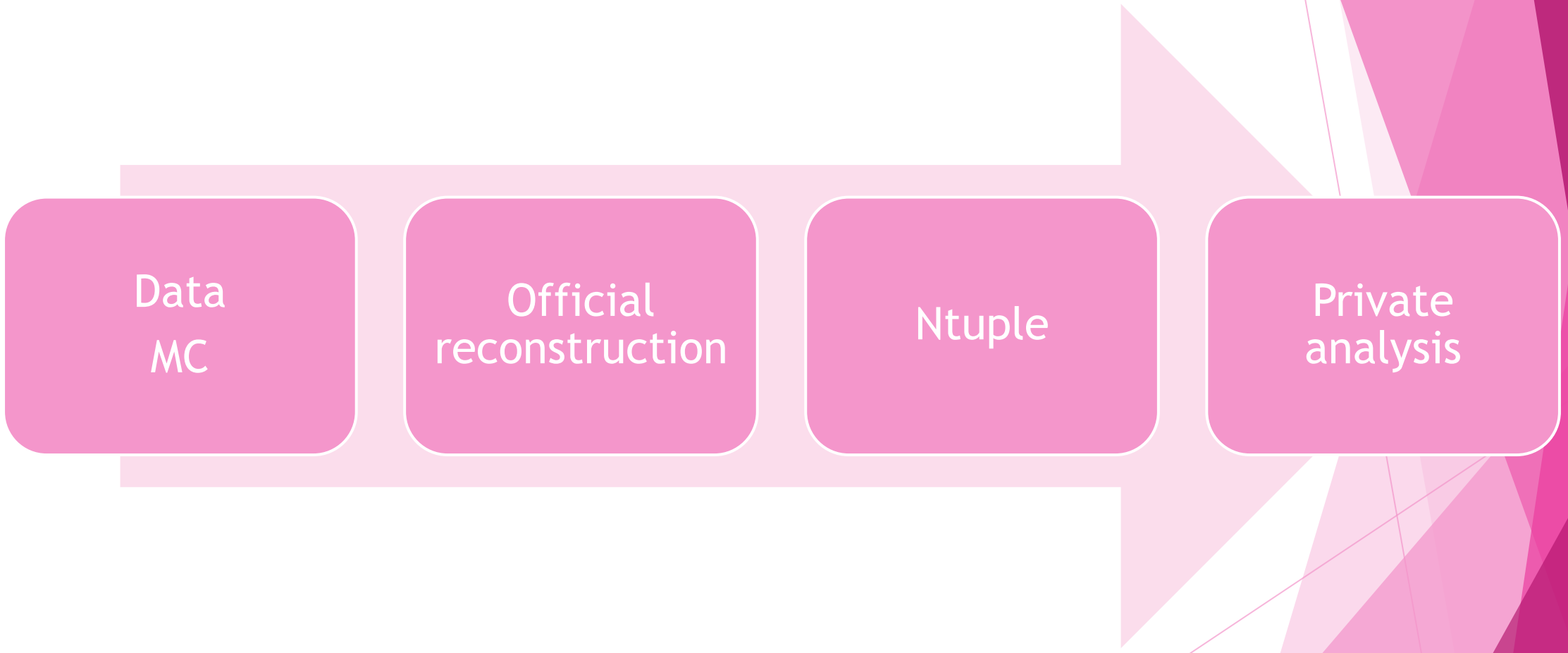
- Requirements from funding agencies.
- If a phD student leave the group, can the work be picked up by others?
- Discover “new physics” from MC?
- Sharing data with your colleague?

- A topic since (at least) SPS era.
- In LHC era, raw data preserved.
- Very few analyses can be reproduced.

*Nature* **533** (2016) 452

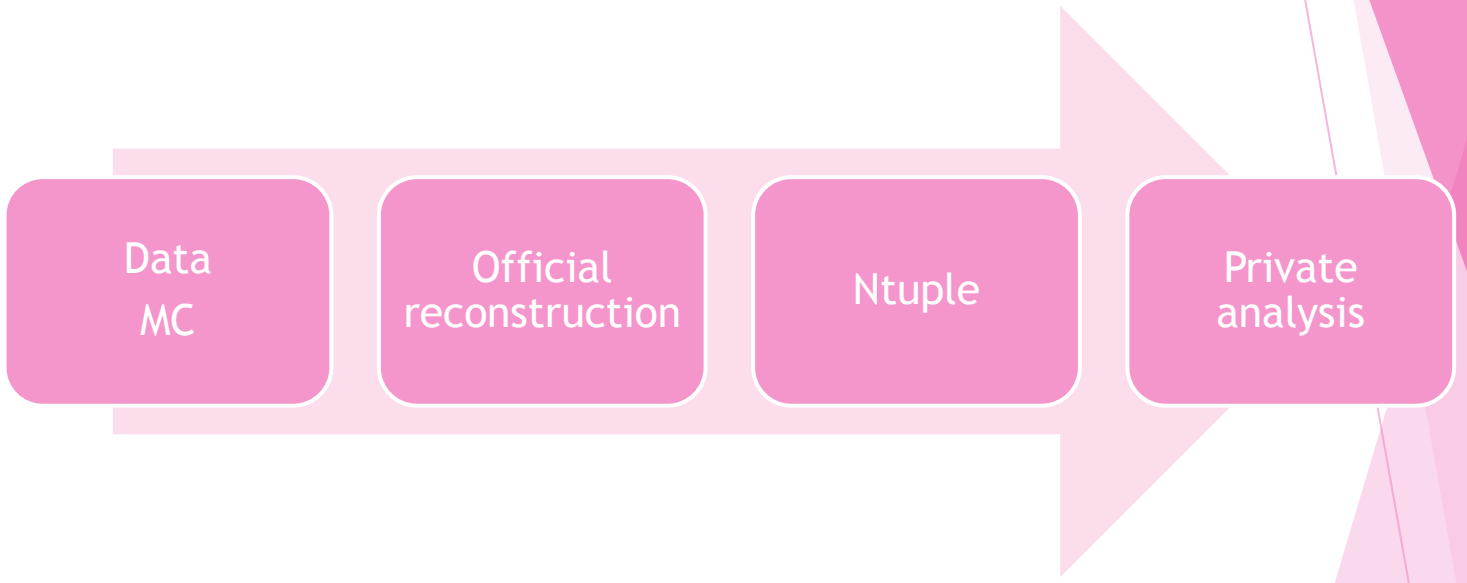


# Physics analysis



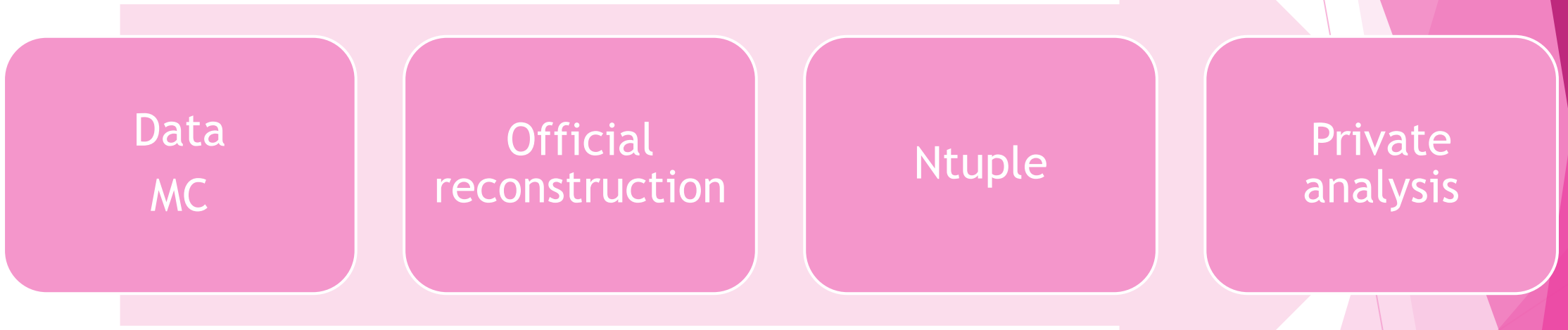
# How to preservation an analysis in principle

- Environment.
- Codes.
- ~~Database.~~
- Data.
- Analysis steps.



# Physics analysis

- Quite large number of steps
- Frequent change of scripts
- Various environments
- Data frequently moved



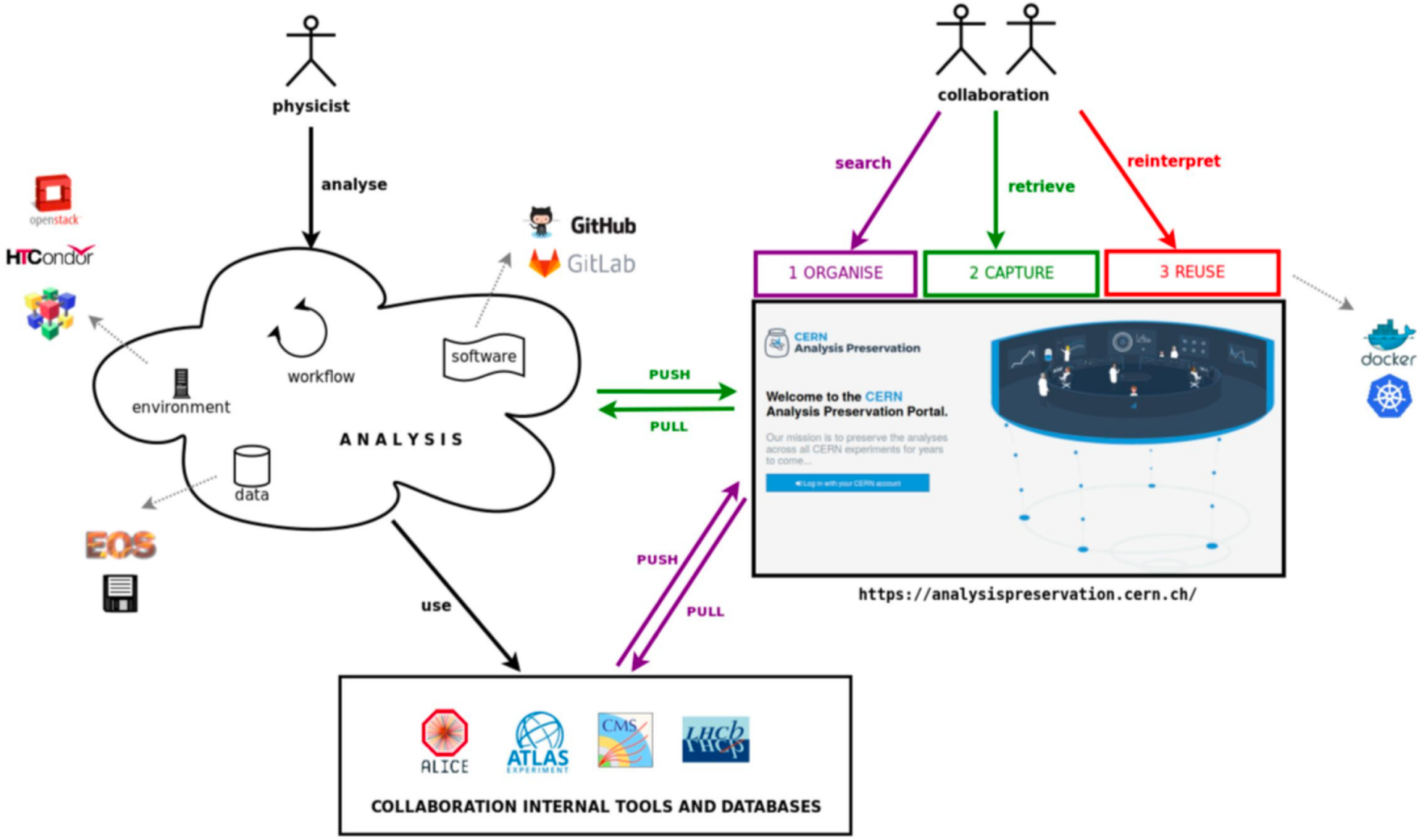

- Few fixed steps
- Fixed environment
- Official location to save data

- Connections with Ntuple

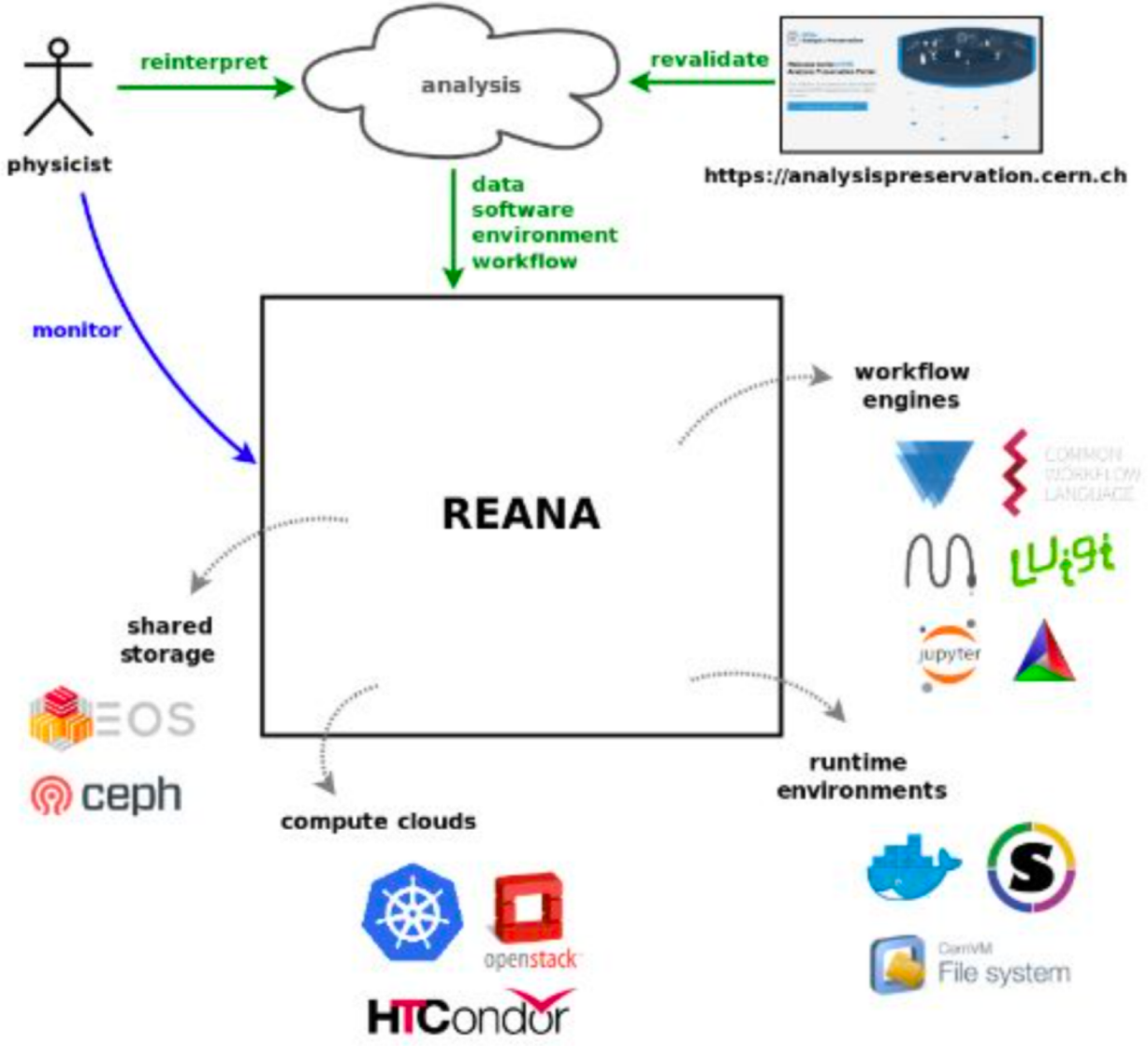
# Techniques requirements

- Environment: container(docker, singularity).
- Data-code binding.
- Re-analyzing.
- Continuous workflow execution.
- Central host.

# CERN
# Analysis Preservation



# reana


physicist
reinterpret
analysis
revalidate
https://analysispreservation.cern.ch
data
software
environment
workflow
monitor
REANA
workflow
engines
shared
storage
EOS
ceph
compute clouds
openstack
HTCondor
runtime
environments
CernVM
File system

# Chern

arXiv:1806.08787

- An architecture for organizing analysis.
- A toolkit provided.

```
[AnalysisSingleMuon] [without_silicon/Efficiency/task_PT_Theta]
 >>> ls -a
README:
task_PT_Theta
>>>> Subobjects:
[0] (task)                arbor_v1 (done)
[1] (task)                arbor_v4 (running)
[2] (task)             clupatra_v1 (done)
[3] (task)             clupatra_v4 (submitted)
```

# Workflow in Chern

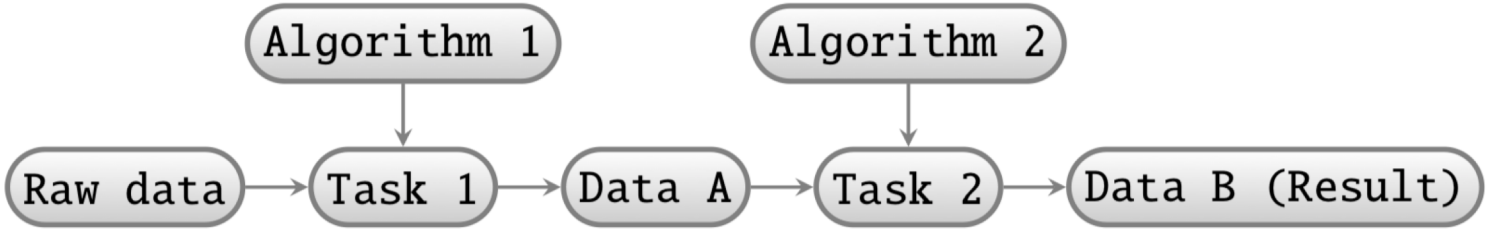



Figure 2: Example of workflow with separated algorithm and task.

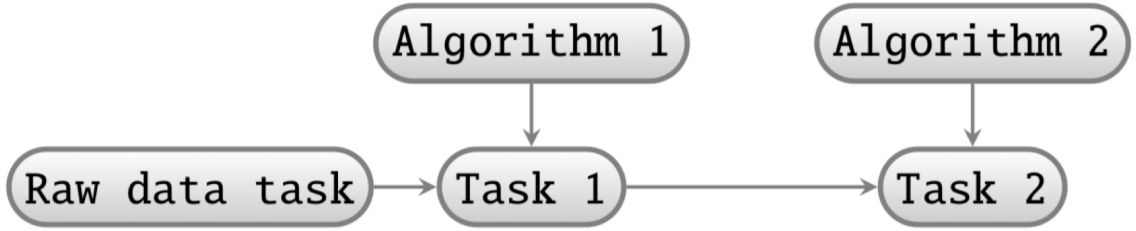
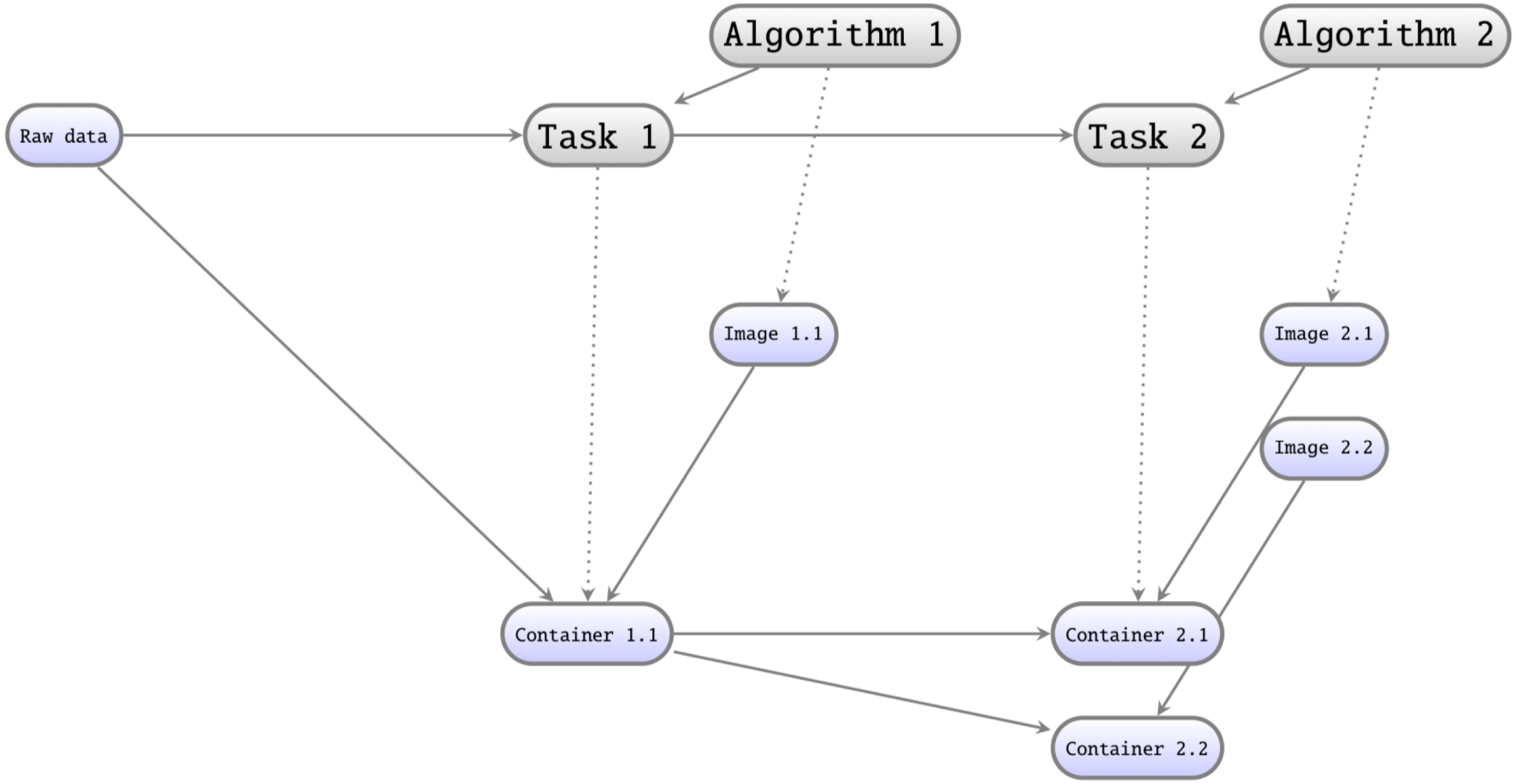



Figure 3: Example of workflow with combination of task and data.

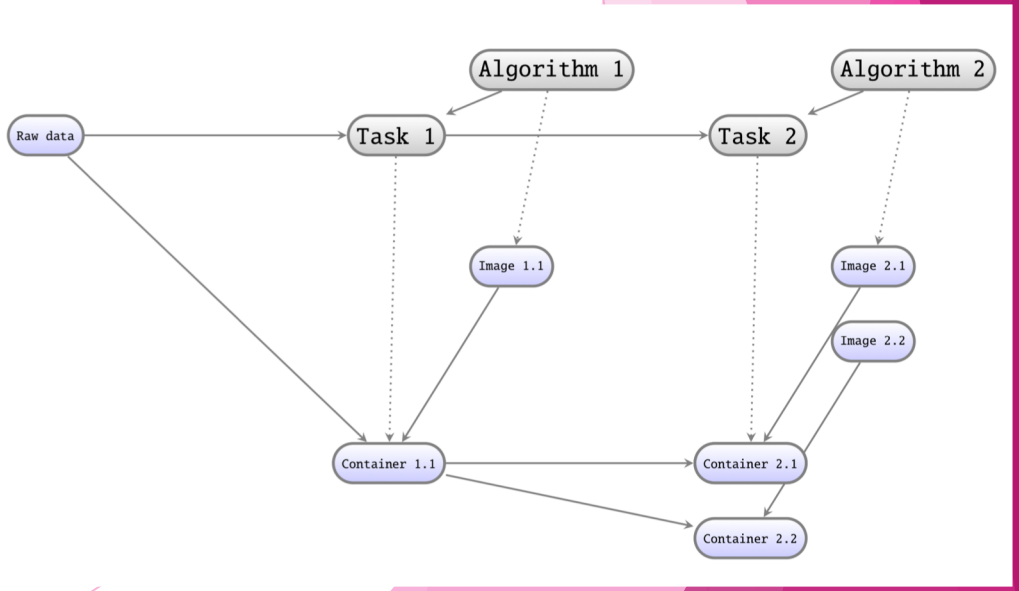
# Impression


Algorithm 1
Algorithm 2
Raw data
Task 1
Task 2
Image 1.1
Image 2.1
Image 2.2
Container 1.1
Container 2.1
Container 2.2

# Chern con’d

- Preserving the analysis at the time of analyzing.
- Possible cross-analysis link.

- Lack of strong backend.
- Lack of host.



# ALICE

## ALICE Analysis on REANA

- Get LEGO train configuration files from CAP
- Get data files from openData
- Run LEGO trains on the public data
- Get plotting macros from CAP
- Create final plots from the LEGO train result

## Wishes for Analysis Preservation

- Method to link analysis to each other
  - Same analysis in pp, pPb, PbPb
  - Data/MC
- Automatic upload of files from ALICE servers
  - Always the same files
  - Location is indicated in the CAP entry
- Automatic transfer to REANA to rerun analysis
  - Easy configuration of data files from opendata

Markus's talk on June.20.2018 CAP meeting

L Heinrich on June.20.2018 CAP meeting

# ATLAS

**Analysis Preservation: — ATLAS Sources**

```
analysis: {
  metadata: {
      ...
  },
  data: {
      ...
  }
  implementation: {
      ...
  }
}
```

Cataloguing, Discovery, Audit, Reference, Very long-term archive — ~Glance

data preservation, storing digital assets related to analysis — ~EOS / ATLS DDM

reuse of archived analyses — ~GitLab

**Analysis Glance**

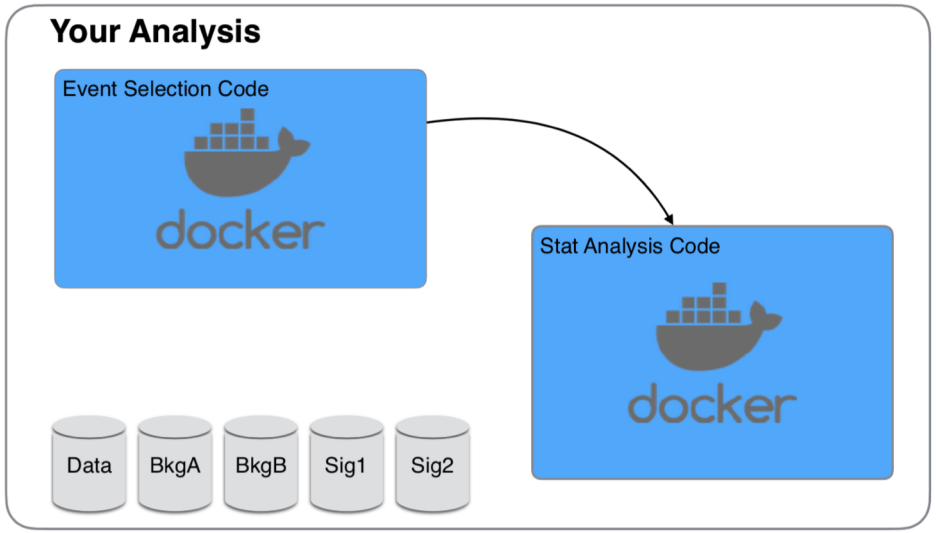
- Expression of Interest
- Analysis Definition: coordinator, team aim
- Analysis Phase:
  - Metadata
  - Signatures studied
  - Methods used
  - Links to Source code, data
  - Meetings
- EdBoard request
- Pre-Approval
- Group Approval

# ATLAS con'd

- Strong encouragement to get all code into CERN GitLab.
- Capture into self-consistent runtimes.
- Preserving the Common Stack — official ATLAS base images.
- Continuous Integration: build/test/preserve analysis code
- Work ongoing to integrate Containers into GRID infrastructure, collaboration between Google / ATLAS to investigate modern technologies in ATLAS Distributed Computing

L Heinrich on June.20.2018 CAP meeting

- make it **easy** for analyzers to preserve their Analysis Code
  - this should be automatic requiring ~no work from analyzers
  - **continuous preservation** during analysis development
- make it **easy** for analyzers to **test** the preserved Anaysis Code
  - teams should not depend on experts to run their code, verify that the preserved version works
  - should work in their usual environment (LXPLUS, home T3)
- make it **easy** for analyzers to **use** their preserved analysis code during their normal analysis activity
  - if it is more tightly integrated into their analysis workflow, **the difference between running an analysis and running a *preserved* analysis vanishes**.. original results are already produced using the latest preserved version



Lara's talk on June.20.2018 CAP meeting

# CMS

**Right now:**

Trying to make the CAP **versatile enough** to cover, to some extent, the three levels

- Searchable analysis
- Basic preservation for code/ntuples
- Some "basic" placeholders trying to include as much as possible all the steps/inputs used during the analysis workflow: control regions, efficiencies, scale factors…

**Also from the visual point of view:**

- Default version containing the **two first levels** → more user friendly
- Possibility to go to an extended version including placeholders for many more details → **first test** for a total analysis reuse

Probably we will have to go from a light approach (final plots, final ntuples) to a more sophisticated one (whole workflow)

Not all analysis will be interested in the reproducibility part → Maybe for really important analysis with a lot of visibility outside? (i.e Higgs)

# LHCb

Sebastian's talk on June.20.2018 CAP meeting

## Adoption roadmap

**New analysis:**

1. Start analysis repository in gitlab WG group
   - Add shared tools as submodules
2. Put all input data on eos WG space
3. Setup automated pipeline
   - Add analysis steps to pipeline as they are developed
4. Document usage of pipeline
5. Prepare docker container
   - Optional: setup CI

**Mature analysis:**

1. Create/fork master repository in gitlab WG group
   - Add shared tools or subprojects as submodules
2. Put all input data on eos WG space
3. Wrap the analysis in scripts for automation
4. Refactor hardcoded configuration/input/output
5. Document analysis workflow
6. Put scripted workflow into pipeline
7. Prepare docker container
   - Optional: setup CI

# CEPC official

```
Reading lcio
Loading LCIO ROOT dictionaries ...
test.slcio
WARNING!!!: The lcio file is private. Please use only privately
 and DO NOT distribute !!!
The release version: 0.1.0-rc8
The configuration  : d3cc5201c6e1e038125b35fe780f7e70
The md5 of this file  : fafcbcf141a3012f24624d8ac23c42e0
The md5 of input file : 7ddeec8ce6e535eeac3cae7d65a123f8
```

- Official release
- Official docker image
  - no version control.
- Yuki
  - A wrap of Marlin.
  - Building connections between code/environment and data.
  - Usage: yuki produce [OPTIONS] INPUT_FILE OUTPUT_FILE RELEASE CONFIG
  - No ROOT file support.

# Suggestions for analyzer

- Nope, because the analysis preservation systems are not ready.
- As a analyzer, you can not do anything.

- But some preparation?
  - LHC users can contact the experts in the collaboration.
  - Analyzer for CEPC, try the docker based software http://cepcsoft.ihep.ac.cn/guides/scratch/docs/docker/
  - Use gitlab as much as possible.
  - Never use hard core path.
  - Use environment variable as less as possible.

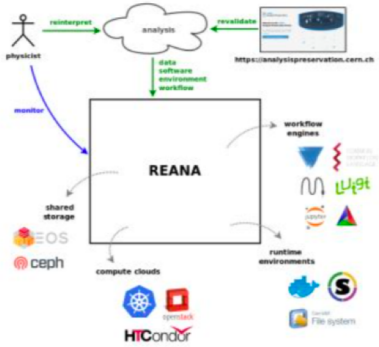
- Young people can learn everything.

# Contacts

**CERN Analysis Preservation** & **reana**


## CERN Analysis Preservation

http://analysispreservation.cern.ch
http://github.com/cernanalysispreservation
analysis-preservation-support@cern.ch




## REANA

http://www.reanahub.io
http://github.com/reanahub
reanahub
info@reanahub.io

**CERN IT** H. Hirvonsalo, D. Kousidis, D. Rodriguez, T. Šimko · **CERN SIS** S. Dallmeier-Tiessen, S. Feger, P. Fokianos, A. Lavasa, I. Tsanaktsidis, A. Trisovic, A. Trzcinska · **ALICE** D. Berzano, M. Gheata, C. Grigoras, Y. Schutz, M. Zimmermann · **ATLAS** J. Berlingen, K. Cranmer, L. Heinrich, L. Henkelmann, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, A. Pearce, S. Roiser · **DASPOS** M. Hildreth · **DPHEP** J. Shiers

# Thanks