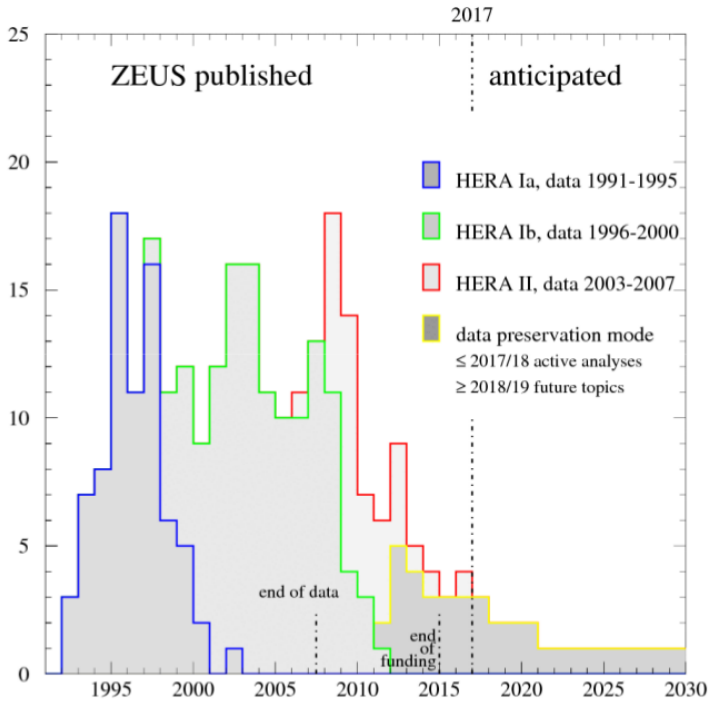
REANA: reproducible research data analysis platform

Tibor Šimko

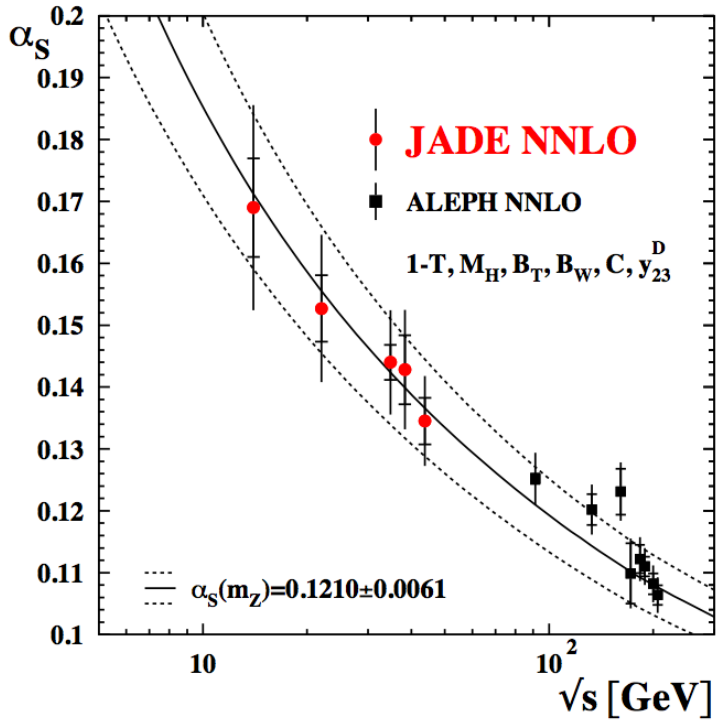
@tiborsimko

EP-SFT Group Meeting · CERN · 4 June 2018

Long-term value of data!



Achim Geiser `https://indico.cern.ch/event/588219`



DPHEP `https://arxiv.org/abs/1205.4667`

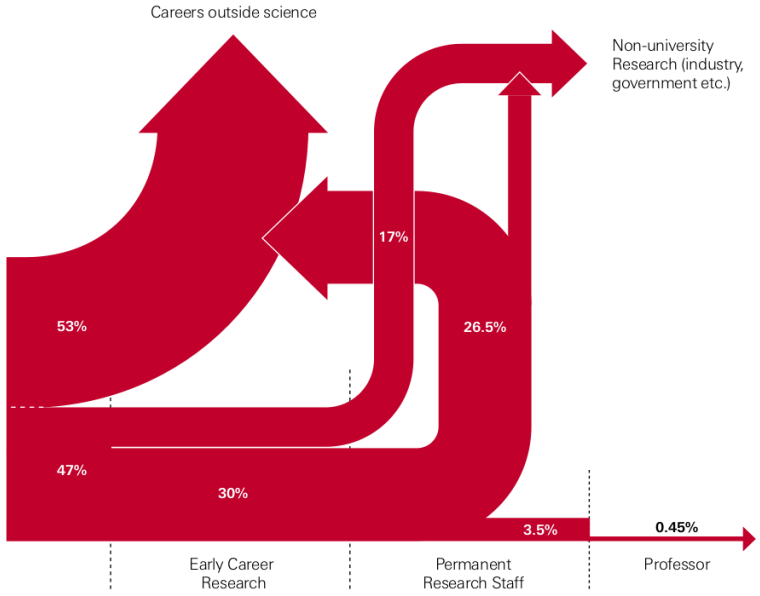
Collaborations publish papers even ∼15 years after data taking ends.

JADE data (1979–1986) still unique even ∼35 years later.

Long-term value of knowledge?

CMS collaboration

Experimental physics done by groups of ∼3000 physicists.

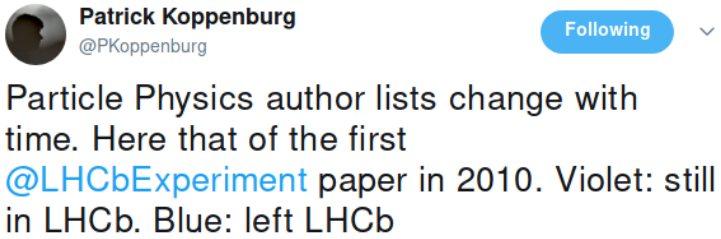


Career after PhD

THE ROYAL SOCIETY

High turnover of young researchers.

LHCb members of 2010 in 2017



LHCb Collaboration: R. Aaij, C. Abellan Beteta, B. Adeva, M. Adinolfi, C. Adrover, A. Affolder, M. Agari, Z. Ajaltouni, J. Albrecht, F. Alessio, M. Alexander, M. Alfonsi, P. Alvarez Cartelle, A.A. Alves Jr, S. Amato, Y. Amhis, J. Amoraal, J. Anderson, R. Antunes Nobrega, R. Appleby, O. Aquines Gutierrez, A. Arefyev, L. Arrabito, M. Artuso, E. Aslanides, G. Auriemma, S. Bachmann, Y. Bagaturia, D.S. Bailey, V. Balagura, W. Baldini, G. Barber, C. Barham, R.J. Barlow, S. Barsuk, S. Basiladze, A. Bates, C. Bauer, Th. Bauer, A. Bay, I. Bediaga, T. Bellunato, K. Belous, I. Belyaev, M. Benayoun, G. Bencivenni, R. Bernet, R.P. Bernhard, M.-O. Bettler, M. van Beuzekom, J.H. Bibby, S. Bifani, A. Bizzeti, P.M. Bjørnstad, T. Blake, F. Blanc, C. Blanks, J. Blouw, S. Blusk, A. Bobrov, V. Bocci, B. Bochin, E. Bonaccorsi, A. Bondar, N. Bondar, W. Bonivento, S. Borghi, A. Borgia, E. Bos, T.J.V. Bowcock, C. Bozzi, T. Brambach, J. van den Brand, L. Brarda, J. Bressieux, S. Brisbane, M. Britsch, N.H. Brook, H. Brown, S. Brusa, A. Büchler-Germann, A. Bursche, J. Buytaert, S. Cadeddu, J.M. Caicedo Carvajal, O. Callot, M. Calvi, M. Calvo Gomez, A. Camboni, W. Cameron, L. Camilleri, P. Campana, A. Carbone, G. Carboni, R. Cardinale, A. Cardini, J. Carroll, L. Carson, K. Carvalho Akiba, G. Casse, M. Cattaneo, B. Chadaj, M. Charles, Ph. Charpentier, J. Cheng, N. Chiapolini, A. Chlopik, J. Christiansen, P. Ciambrone, X. Cid Vidal, P.J. Clark, P.E.L. Clarke, M. Clemencic, H.V. Cliff, J. Closier, C. Coca, V. Coco, J. Cogan, P. Collins, A. Comerma-Montells, F. Constantin, G. Conti, A. Contu, P. Cooke, M. Coombes, B. Corajod, G. Corti, G.A. Cowan, R. Currie, B. D'Almagne, C. D'Ambrosio, I. D'Antone, W. Da Silva, E. Dane', P. David, I. De Bonis, S. De Capua, M. De Cian, F. De Lorenzi, J.M. De Miranda, L. De Paula, P. De Simone, D. Decamp, G. Decreuse, H. Degaudenzi, M. Deissenroth, L. Del Buono, C.J. Densham, C. Deplano, O. Deschamps, F. Dettori, J. Dickens, H. Dijkstra, M. Dima, S. Donleavy, P. Dornan, D. Dossett, A. Dovbnya, R. Dumps, F. Dupertuis, L. Dwyer, R. Dzhelyadin, C. Eames, S. Easo, U. Egede, V. Egorychev, S. Eidelman, D. van Eijk, F. Eisele, S. Eisenhardt, L. Eklund, D.G. d'Enterria, D. Esperante Pereira, L. Estève, E. Fanchini, C. Färber, G. Fardell, C. Farinelli, S. Farry, V. Fave, G. Felici, V. Fernandez Albor, M. Ferro-Luzzi, S. Filippov, C. Fitzpatrick, W. Flegel, F. Fontanelli, C. Forti, R. Forty, C. Fournier, B. Franek, M. Frank, C. Frei, M. Frosini, J.L. Fungueirino Pazos, S. Furcas, A. Gallas Torreira, D. Galli, M. Gandelman, P. Gandini, Y. Gao, J-C. Garnier, L. Garrido, D. Gascon, C. Gaspar, A. Gaspar De Valenzuela Cue, J. Gassner, N. Gauvin, P. Gavillet, M. Gersabeck, T. Gershon, Ph. Ghez, V. Gibson, Yu. Gilitsky, V.V. Gligorov, C. Göbel, D. Golubkov, A. Golutvin, A. Gomes, G. Gong, H. Gong, H. Gordon, M. Grabalosa Gándara, V. Gracco, R. Graciani Diaz, L.A. Granado Cardoso, E. Graugés, G. Graziani, A. Grecu, S. Gregson, G. Guerrer, B. Gui, E. Gushchin, Yu. Guz, Z. Guzik, T. Gys, G. Haefeli, S.C. Haines, T. Hampson, S. Hansmann-Menzemer, R. Harji, N. Harnew, P.F. Harrison, J. He, K. Hennessy, P. Henrard, J.A. Hernando Morata, E. van Herwijnen, A. Hicheur, E. Hicks, H.J. Hilke, W. Hofmann, K. Holubyev, P. Hopchev, W. Hulsbergen, P. Hunt, T. Huse, R.S. Huston, D. Hutchcroft, F. Iacoangeli, V. Iakovenko, C. Iglesias Escudero, C. Ilgner, J. Imong, R. Jacobsson, M. Jahjah Hussein, O. Jamet, E. Jans, F. Jansen, P. Jaton, B. Jean-Marie, M. John, D. Johnson, C.R. Jones, B. Jost, F. Kapusta, T.M. Karbach, A. Kashchuk, S. Katvars, J. Keaveney, U. Kerzel, T. Ketel, A. Keune, S. Khalil, B. Khanji, Y.M. Kim, M. Knecht, S. Koblitz, A. Konoplyannikov, P. Koppenburg, M. Korolev, A. Kozlinskiy, L. Kravchuk, R. Kristic, G. Krocker, P. Krokovny, F. Kruse, K. Kruzelecki, M. Kucharczyk, I. Kudryashov, S. Kukulak, R. Kumar, T. Kvaratskheliya, V.N. La Thi, D. Lacarrere, A. Lai, R.W. Lambert, G. Lanfranchi, C. Langenbruch, T. Latham, R. Le Gac, J.-P. Lees, R. Lefèvre, A. Leflat, J. Lefrançois, F. Lehner, M. Lenzi, O. Leroy, T. Lesiak, L. Li, Y.Y. Li, L. Li Gioi, J. Libby, M. Lieng, R. Lindner, S. Lindsey, C. Linn, B. Liu, G. Liu, S. Löchner, J.H. Lopes, E. Lopez Asamar, N. Lopez-March, P. Loveridge, J. Luisier, B. M'charek, F. Machefert, I.V. Machikhiliyan, F. Maciuc, O. Maev, J. Magnin, A. Maier, S. Malde, R.M.D. Mamunur, G. Manca, G. Mancinelli, N. Mangiafave, U. Marconi, R. Märki, J. Marks, G. Martellotti, A. Martens, L. Martin, D. Martinez Santos, A. Massaferri, Z. Mathe, C. Matteuzzi, V. Matveev, E. Maurice, B. Maynard, A. Mazurov, G. McGregor, R. McNulty, C. Mclean, M. Merk, J. Merkel, M. Merkin, R. Messi, F.C.D. Metlica, S. Miglioranzi, M.-N. Minard, G. Moine, S. Monteil, D. Moran, J. Morant, J.V. Morris, J. Moscicki, R. Mountain, I. Mous, F. Muheim, R. Muresan, F. Murtas, B. Muryn, M. Musy, J. Mylroie-Smith, P. Naik, T. Nakada, R. Nandakumar, J. Nardulli, A. Nawrot, M. Nedos, M. Needham, N. Neufeld, P. Neustroev, M. Nicol, L. Nicolas, S. Nies, V. Niess, N. Nikitin, A. Noor, A. Oblakowska-Mucha, V. Obraztsov, S. Oggero, O. Okhrimenko, R. Oldeman, M. Orlandea, A. Ostankov, J. Palacios, M. Palutan, J. Panman, A. Papadelis, A. Papanestis, M. Pappagallo, C. Parkes, C.J. Parkinson, G. Passaleva, G.D. Patel, M. Patel, S.K. Paterson, G.N. Patrick, C. Patrignani, E. Pauna, C. Pauna (Chiojdeanu), C. Pavel (Nicorescu), A. Pazos Alvarez, A. Pellegrino, G. Penso, M. Pepe Altarelli, S. Perazzini, D.L. Perego, E. Perez Trigo, A. Pérez-Calero Yzquierdo, P. Perret, G. Pessina, A. Petrella, A. Petrolini, E. Picatoste Olloqui, B. Pie Valls, D. Piedigrossi, B. Pietrzyk, D. Pinci, S. Playfer, M. Plo Casasus, M. Poli-Lener, G. Polok, A. Poluektov, E. Polycarpo, D. Popov, B. Popovici, S. Poss, C. Potterat, A. Powell, S. Pozzi, T. du Pree, V. Pugatch, A. Puig Navarro, W. Qian, J.H. Rademacker, B. Rakotomiaramanana, I. Raniuk, G. Raven, S. Redford, W. Reece, A.C. dos Reis, S. Ricciardi, J. Riera, K. Rinnert, D.A. Roa Romero, P. Robbe, E. Rodrigues, F. Rodrigues, C. Rodriguez Cobo, P. Rodriguez Perez, G.J. Rogers, V. Romanovsky, E. Rondan Sanabria, M. Rosello, G. Rospabe, J. Rouvinet, L. Roy, T. Ruf, H. Ruiz, C. Rummel, V. Rusinov, G. Sabatino, J.J. Saborido Silva, N. Sagidova, P. Sail, B. Saitta, T. Sakhelashvili, C. Salzmann, A. Sambade Varela, M. Sannino, R. Santacesaria, R. Santinelli, E. Santovetti, M. Sapunov, A. Sarti, C. Satriano, A. Satta, T. Savidge, M. Savrie, D. Savrina, P. Schaack, M. Schiller, S. Schleich, M. Schmelling, B. Schmidt, O. Schneider, T. Schneider, A. Schopper, M.-H. Schune, R. Schwemmer, A. Sciubba, M. Seco, A. Semennikov, K. Senderowska, N. Serra, J. Serrano, B. Shao, M. Shapkin, I. Shapoval, P. Shatalov, Y. Shcheglov, T. Shears, L. Shekhtman, V. Shevchenko, A. Shires, S. Sigurdsson, E. Simioni, H.P. Skottowe, T. Skwarnicki, N. Smale, A. Smith, A.C. Smith, N.A. Smith, K. Sobczak, F.J.P. Soler, A. Solomin, P. Somogy, F. Soomro, B. Souza De Paula, B. Spaan, A. Sparkes, E. Spiridenkov, P. Spradlin, A. Srednicki, F. Stagni, S. Stahl, S. Steiner, O. Steinkamp, O. Stenyakin, S. Stoica, S. Stone, B. Storaci, U. Straumann, N. Styles, M. Szczekowski, P. Szczypka, T. Szumlak, S. T'Jampens, E. Tarkovskiy, E. Teodorescu, H. Terrier, F. Teubert, C. Thomas, E. Thomas, J. van Tilburg, V. Tisserand, M. Tobin, S. Topp-Joergensen, M.T. Tran, S. Traynor, U. Trunk, A. Tsaregorodtsev, N. Tuning, A. Ukleja, O. Ullaland, U. Uwer, V. Vagnoni, G. Valenti, A. Van Lysebetten, R. Vazquez Gomez, P. Vazquez Regueiro, S. Vecchi, J.J. Velthuis, M. Veltri, K. Vervink, B. Viaud, I. Videau, D. Vieira, X. Vilasis-Cardona, J. Visniakov, A. Vollhardt, D. Volyanskyy, D. Voong, A. Vorobyev, An. Vorobyev, H. Voss, K. Wacker, S. Wandernoth, J. Wang, D.R. Ward, A.D. Webber, D. Websdale, M. Whitehead, D. Wiedner, L. Wiggers, G. Wilkinson, M.P. Williams, M. Williams, F.F. Wilson, J. Wishahi, M. Witek, W. Witzeling, M.L. Woodward, S.A. Wotton, K. Wyllie, Y. Xie, F. Xing, Z. Yang, G. Ybeles Smit, R. Young, O. Yushchenko, M. Zeng, L. Zhang, Y. Zhang, A. Zhelezov, E. Zverev (collapse list)

(Submitted on 18 Aug 2010 (v1), last revised 15 Sep 2010 (this version, v2))

https://twitter.com/PKoppenburg/status/918097093616652293

CERN Analysis Preservation

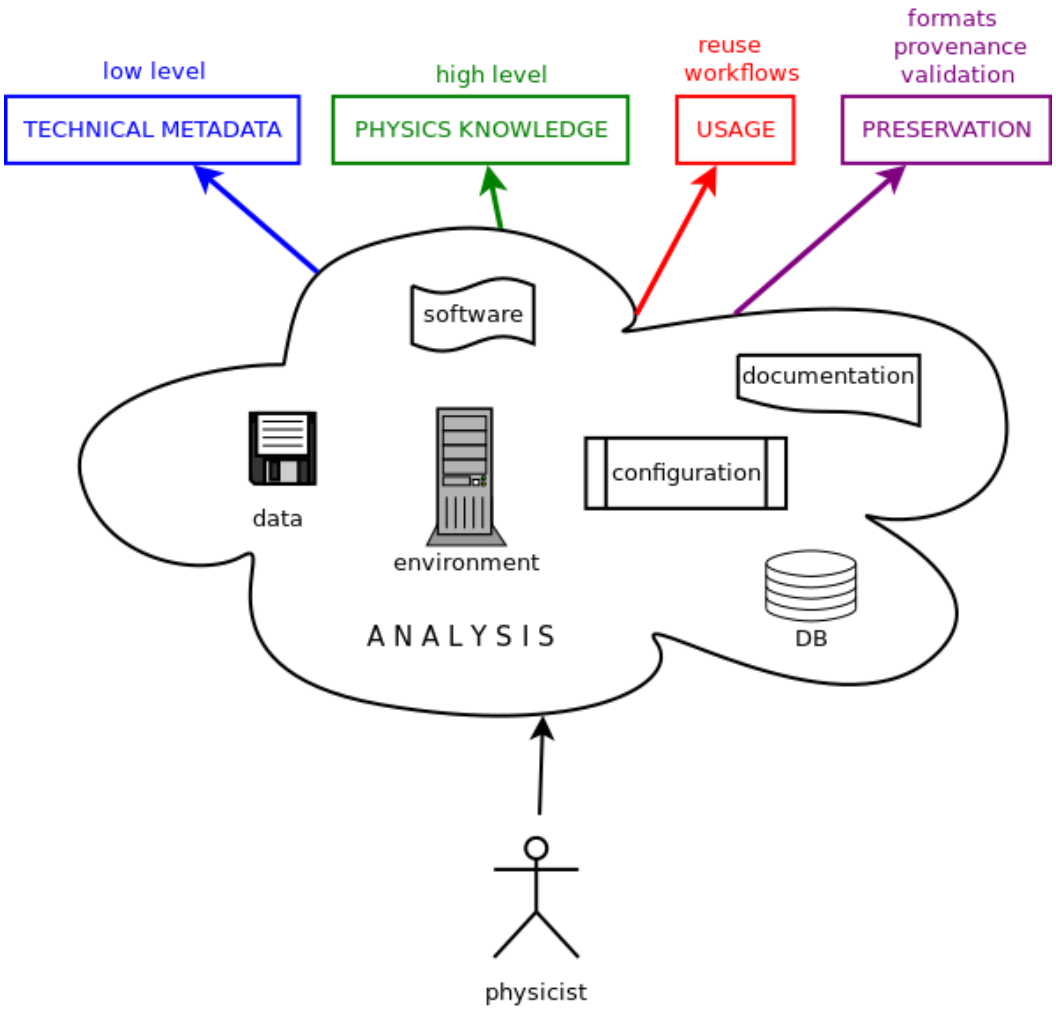
- A platform for **preserving knowledge** and **assets** of an individual physics analysis.
- Capturing the elements needed to **understand** and **reuse** an analysis even several years later:

  - ✔ data
  - ✔ software
  - ✔ environment
  - ✔ workflow
  - ✔ context
  - ✔ documentation

- Applying standard **collaboration access restrictions**

Developed by CERN IT and CERN SIS in close collaboration with LHC experiments

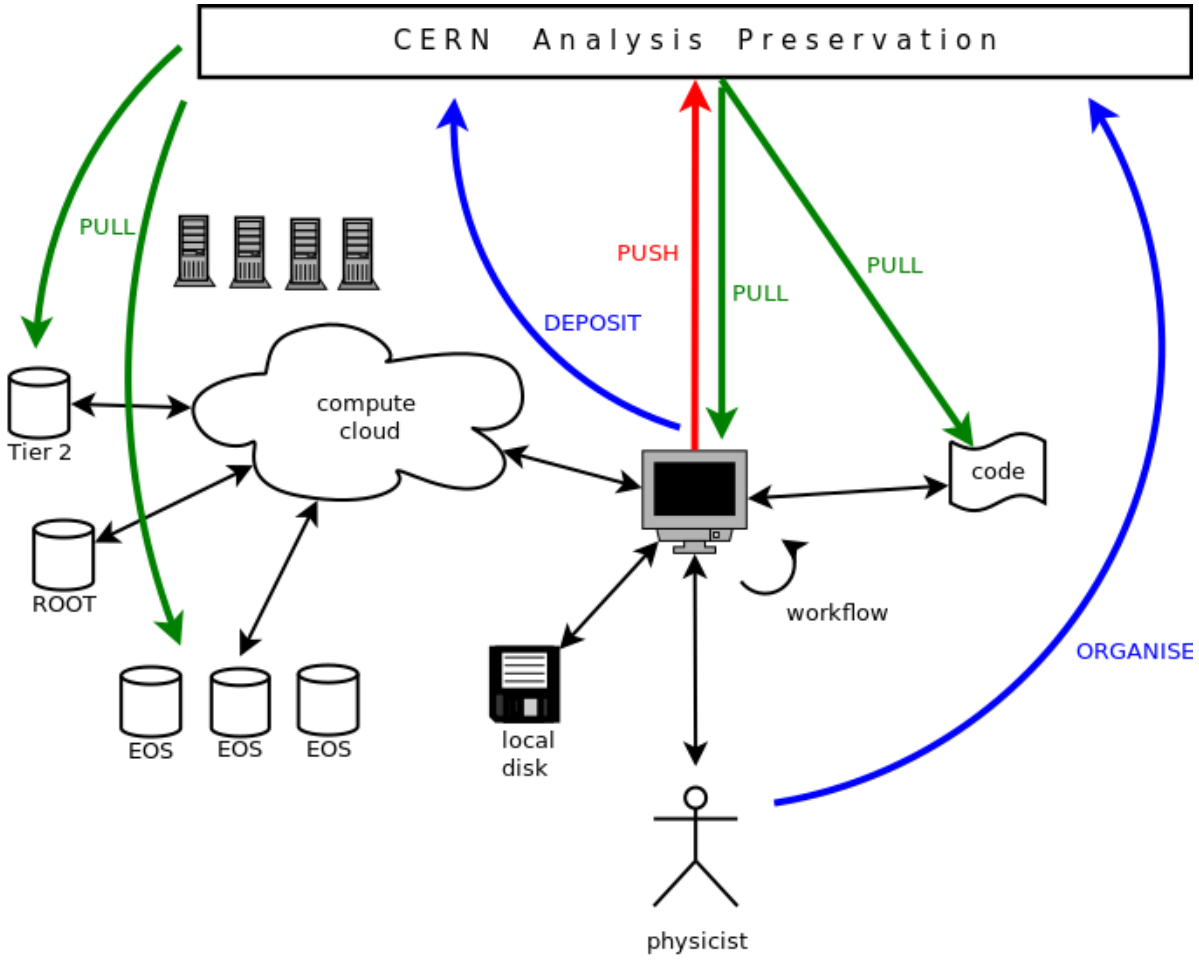
Pillar 1: Describe



- JSON Schema
- W3C DCAT
- domain-specific fields

Structuring knowledge behind research data analysis.

Pillar 2: Capture

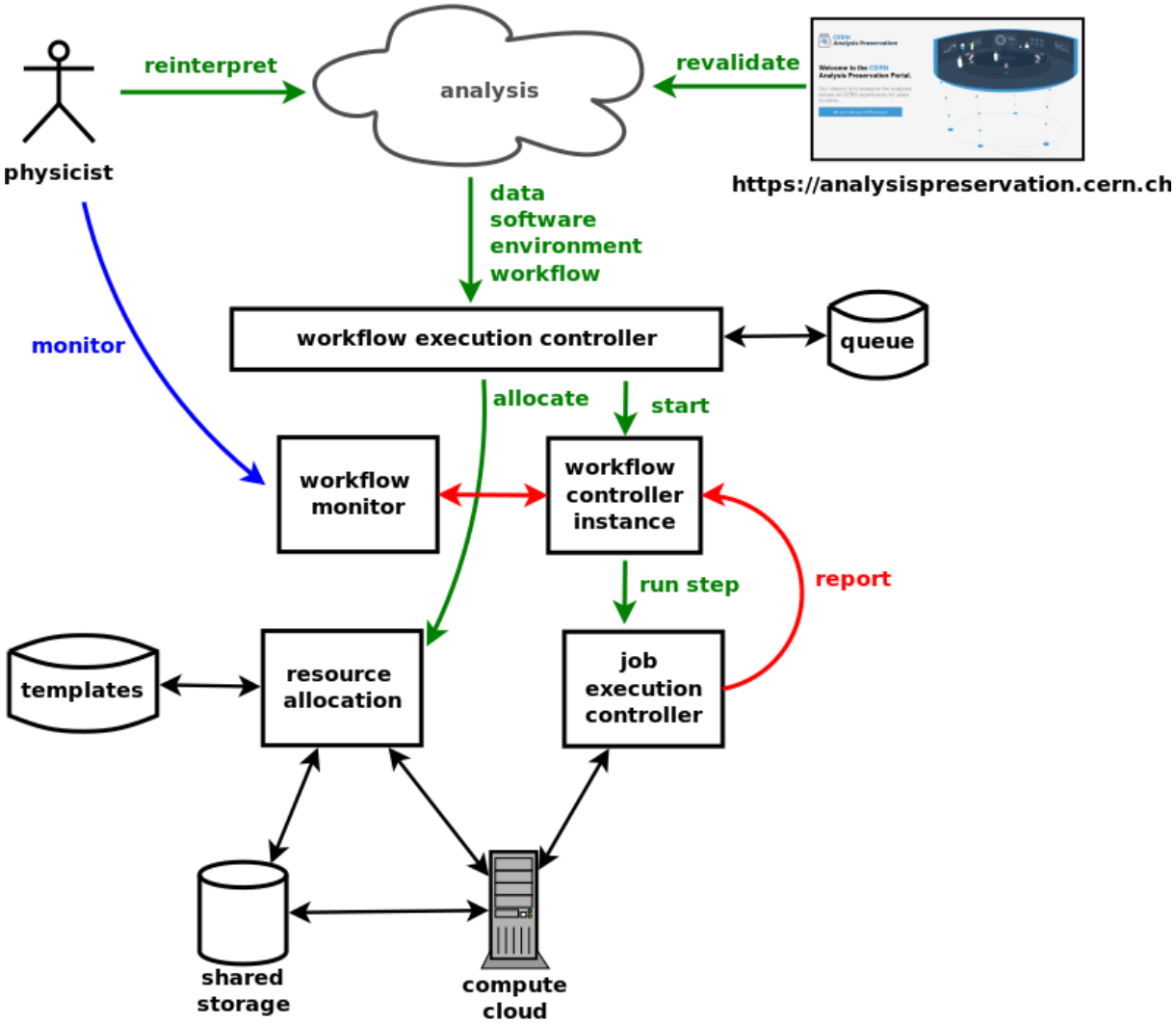


INVENIO

- datasets:
  local storage,
  cloud storage
- software:
  Git, SVN
- information:
  DBs, TWiki,
  SharePoint
- protocols:
  HTTP, XRootD

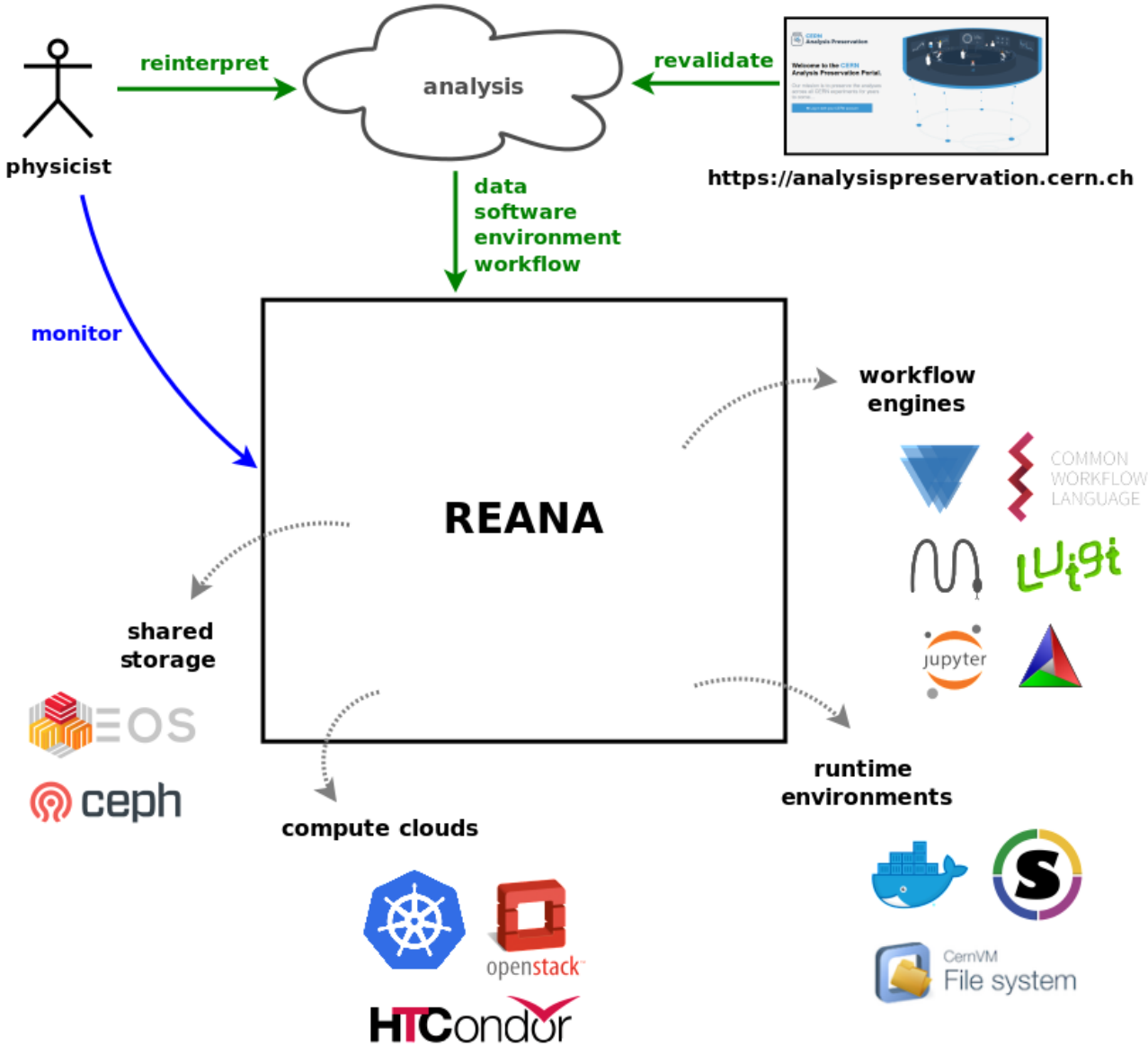
Taking consistent snapshot of analysis assets at a certain time.

Pillar 3: Reuse



Instantiating preserved analysis on the cloud.

REANA = RE usable ANA lyses



Four questions

1 Input data

What is your input data?

- input files
- input parameters

2 Analysis code

Which code analyses it?

- software frameworks
- user code

3 Compute environment

What is your environment?

- operating system
- database calls

4 Analysis workflow

Which steps did you take?

- single command
- complex workflows

Simple example

```
Region,1500,1600,1700,1750,1800,1850,1900,1950,1999,2008,2010,2012,2050,2150
World,100,100,100,100,100,100,100,100,100,100,100,100,100,100
Africa,18.8,19.7,15.5,13.4,10.9,8.8,8.1,8.8,12.8,14.5,14.8,15.2,19.8,23.7
Asia,53.1,58.4,63.9,63.5,64.9,64.1,57.4,55.6,60.8,60.4,60.4,60.3,59.1,57.1
Europe,18.3,19.1,18.3,20.6,20.8,21.9,24.7,21.7,12.2,10.9,10.7,10.5,7,5.3
Latin America and the Caribbean,8.5,1.7,1.5,2,2.5,3,4.5,6.6,8.5,8.6,8.6,8.6,9.1,9.4
Northern America,0.7,0.5,0.3,0.3,0.7,2.1,5,6.8,5.1,5,5,5,4.4,4.1
Oceania,0.7,0.5,0.4,0.3,0.2,0.2,0.4,0.5,0.5,0.5,0.5,0.5,0.5,0.5
```

1 **input**: CSV file

Regional Analysis

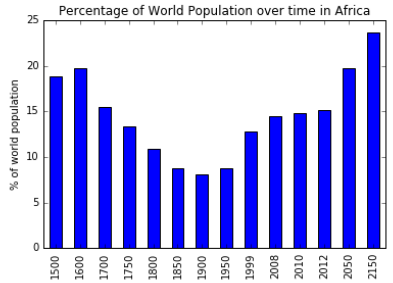
We'll start with a histogram depicting the evolution of a specific region's portion of the world population, in percentage.

```
In [6]: def histogram_by_region(region):
            local_pop=pop[['Region', str(region)]].groupby('Region').sum()

            plot=local_pop.plot(kind='bar', legend=None, title='Percentage of World Population over tim
        e in '+ str(region))

            plot.set_ylabel('% of world population')
            plot.set_xlabel('')
```

```
In [7]: histogram_by_region('Africa')
```



2 **code**: Jupyter notebook

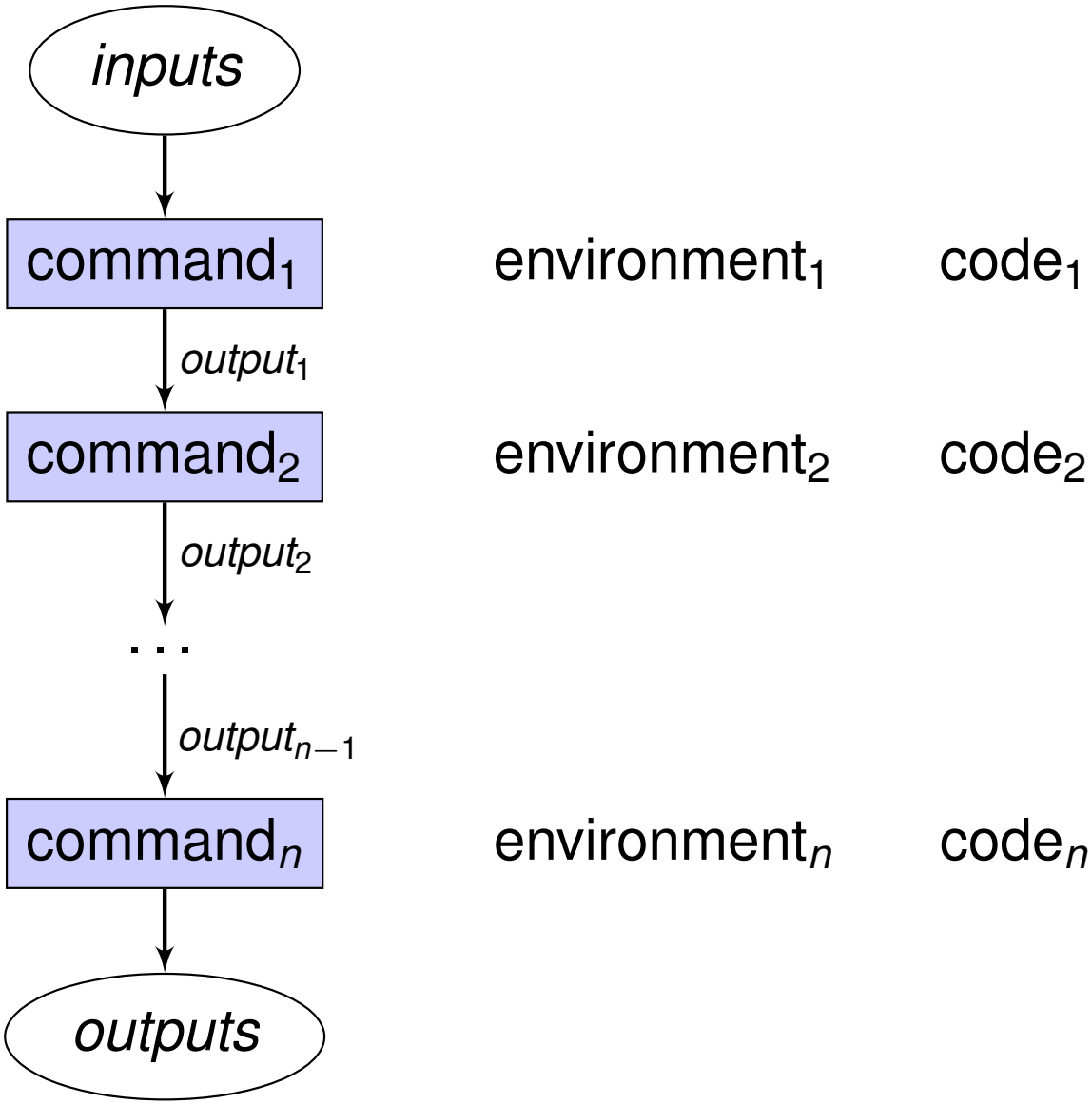
```
FROM centos:7
RUN yum install -y epel-release
RUN yum install -y \
    gcc \
    python-devel \
    python-pip
RUN pip install ipython==5.0.0 jupyter==1.0.0
ADD world_population_analysis.ipynb /code/
ADD World_historical_and_predicted_populations_in_percentage.csv /code/
WORKDIR /code
CMD ["jupyter", "nbconvert","world_population_analysis.ipynb"]
```

3 **environment**: CentOS7, IP5

4 **workflow**: `papermill ...`

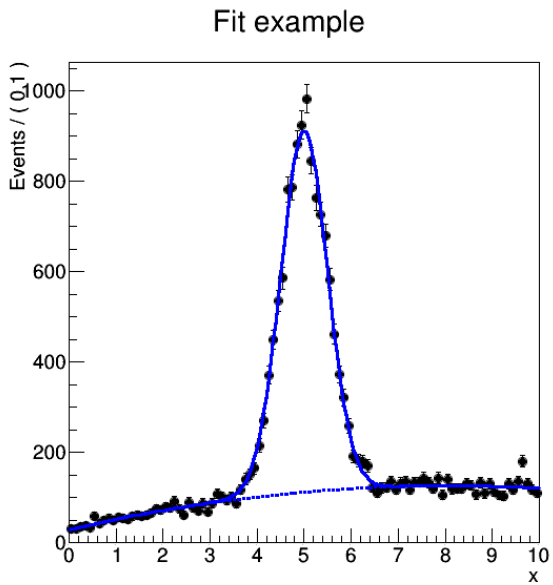
`https://github.com/reanahub/reana-demo-worldpopulation`

Simple workflows



Example: ROOT RooFit

```
workflow:
  type: serial
  steps:
    - environment: reanahub/reana-env-root6
      commands:
      - root -b -q './code/gendata.C(20000,"data.root")'
      - root -b -q './code/fitdata.C("data.root","plot.png")'
```



https://github.com/reanahub/reana-demo-root6-roofit/

User perspective: reana.yaml

- analysis structure described via `reana.yaml`

```
version: 0.2.0
code:
  files:
  - code/mycode.py
inputs:
  files:
    - inputs/mydata.csv
  parameters:
    myparameter: myvalue
environments:
  - type: docker
    image: johndoe/mypython:1.0
workflow:
  type: cwl
  file: workflow/myworkflow.cwl
outputs:
  files:
  - outputs/myplot.png
```

User perspective: CLI client

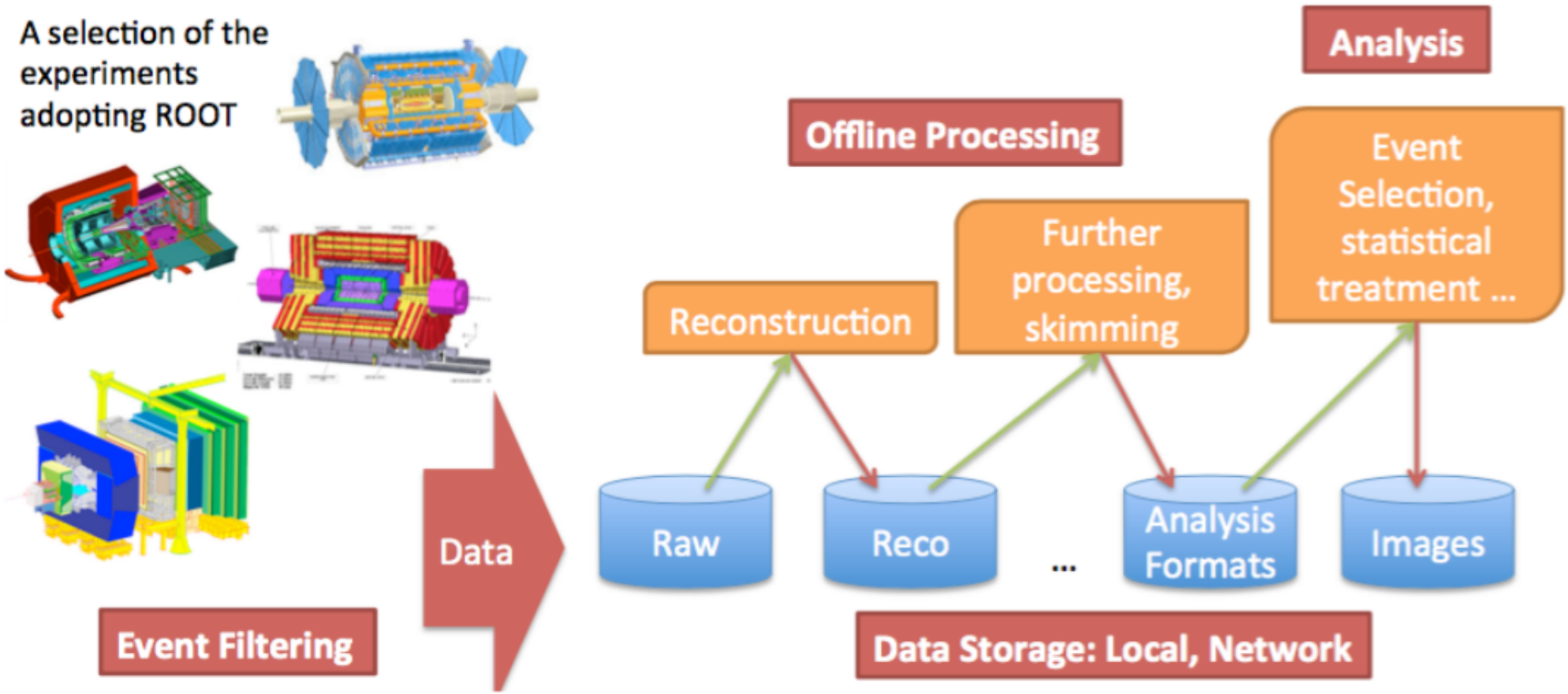
- select REANA cloud

```
$ export REANA_SERVER_URL=https://reana.cern.ch/
```

- run your workflows

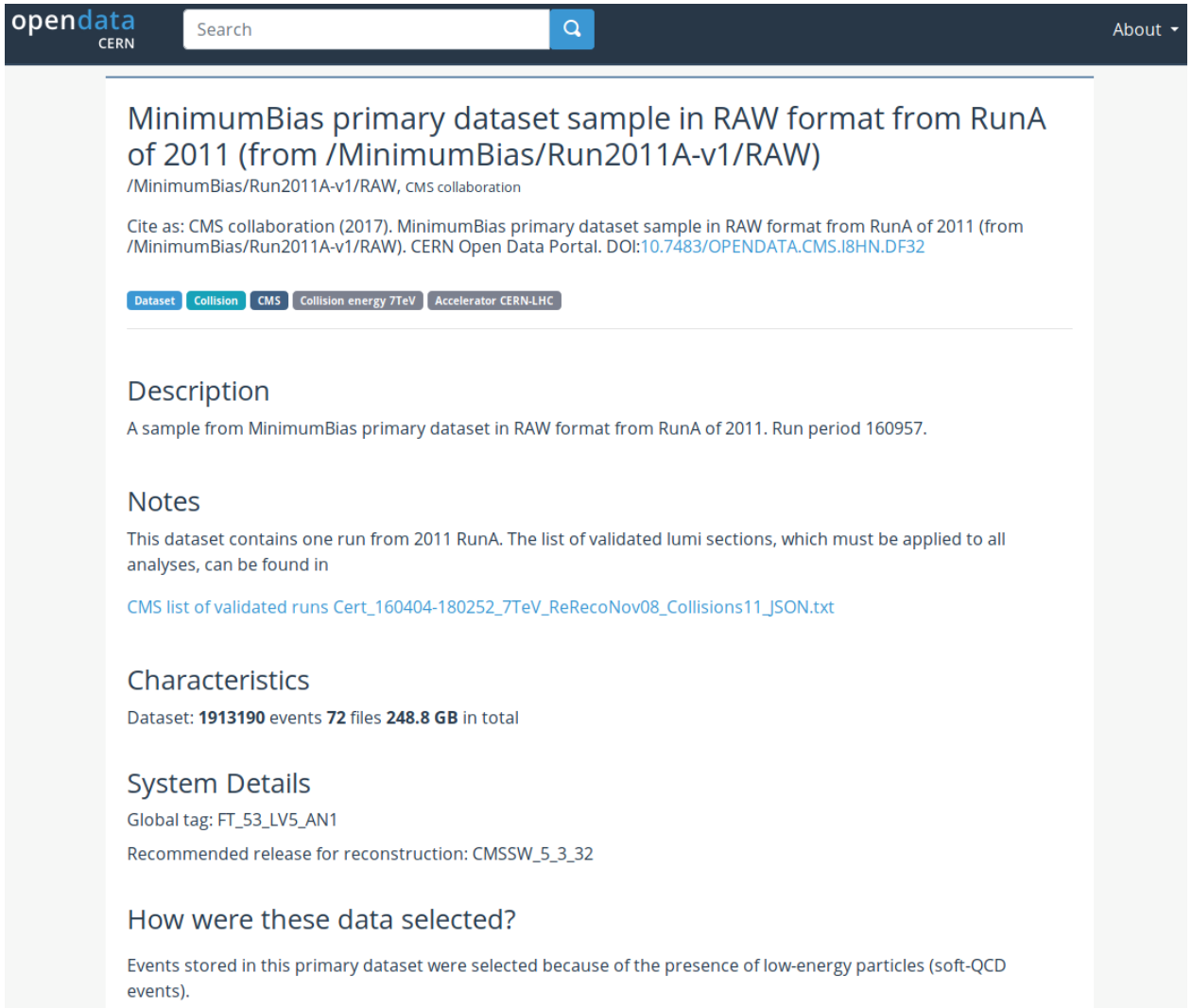
```
$ # install reana-client
$ mkvirtualenv reana-client -p /usr/bin/python2.7
$ pip install reana-client
$ reana-client ping
$ # create new workflow
$ export REANA_WORKON=$(reana-client workflow create)
$ # upload runtime code and inputs
$ reana-client code upload ./code/*
$ reana-client inputs upload ./inputs/*
$ # start workflow and check progress
$ reana-client workflow start
$ reana-client workflow status
$ # download outputs
$ reana-client outputs list
$ reana-client outputs download myplot.png
```

HEP data analysis pipelines



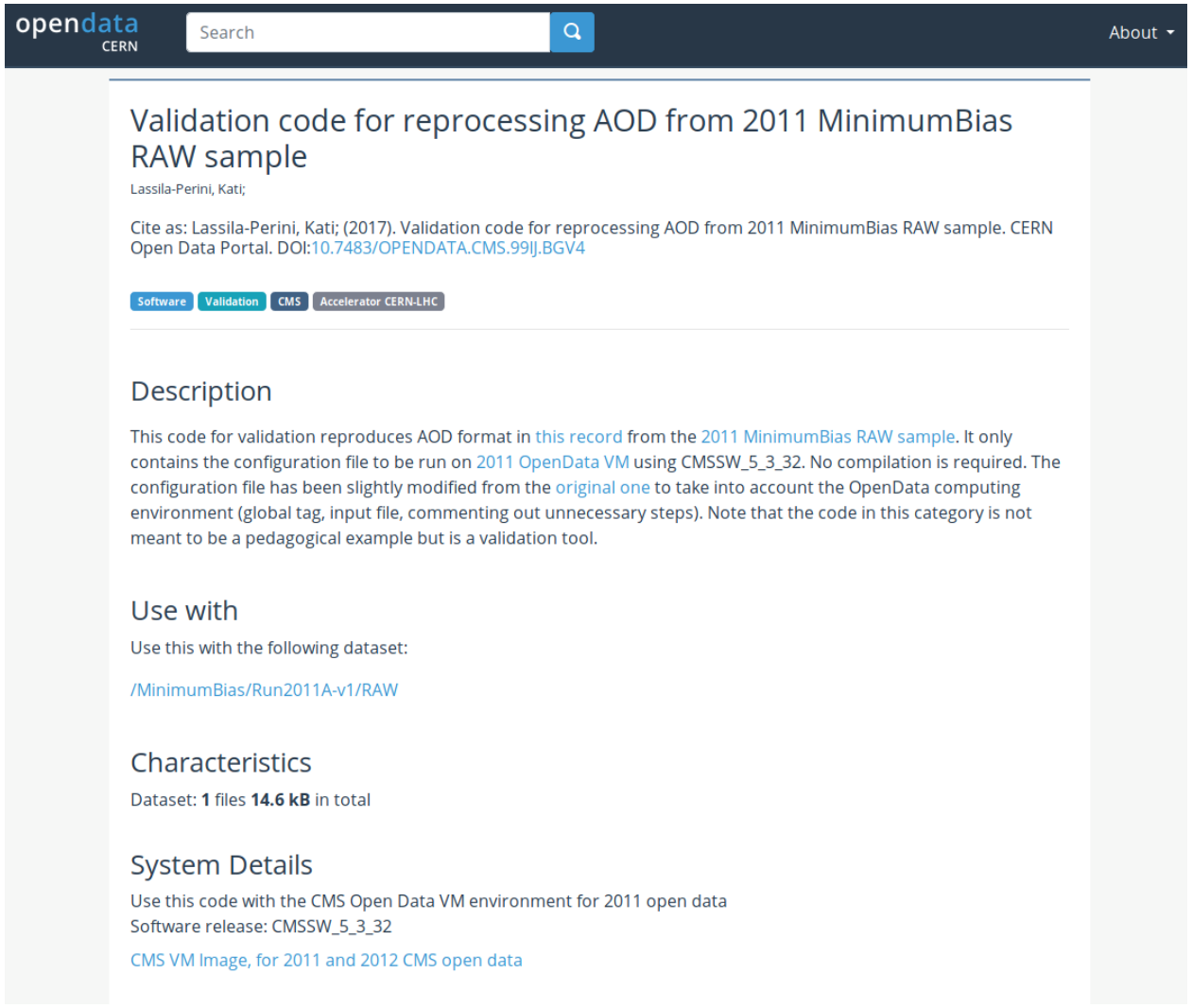
D. Krücker *et al* `https://indico.desy.de/indico/event/18343`

Example: CMS open data



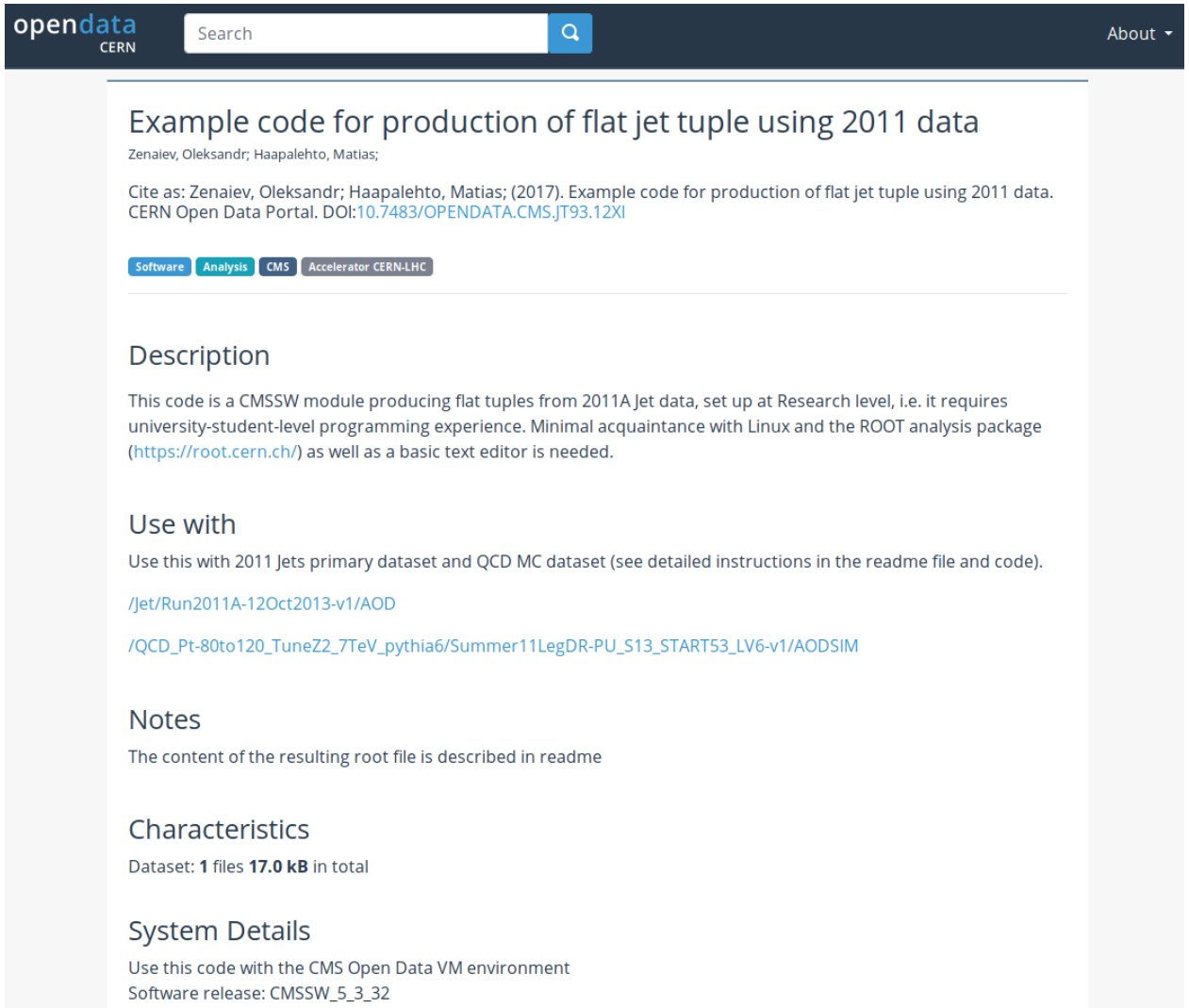
`http://opendata.cern.ch/record/35`

Example: CMS AOD production



`http://opendata.cern.ch/record/463`

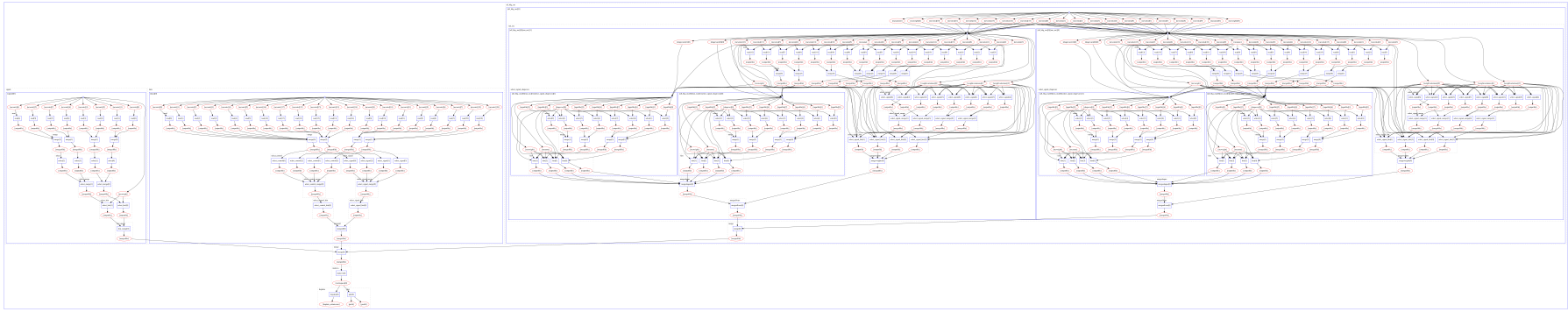
Example: CMS jet tuple production

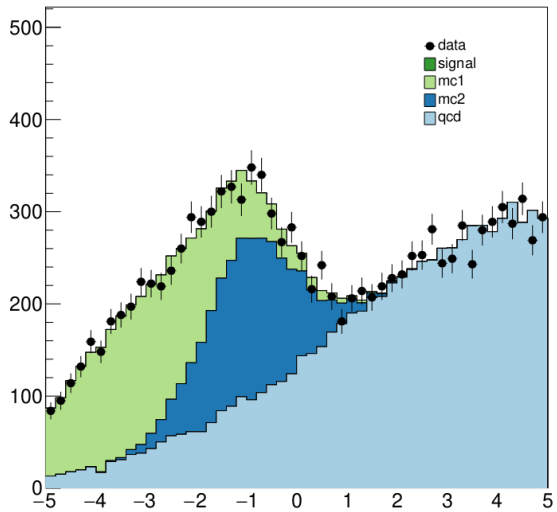


`http://opendata.cern.ch/record/5104`

Example: BSM searches

- full emulation of typical complex analysis workflows





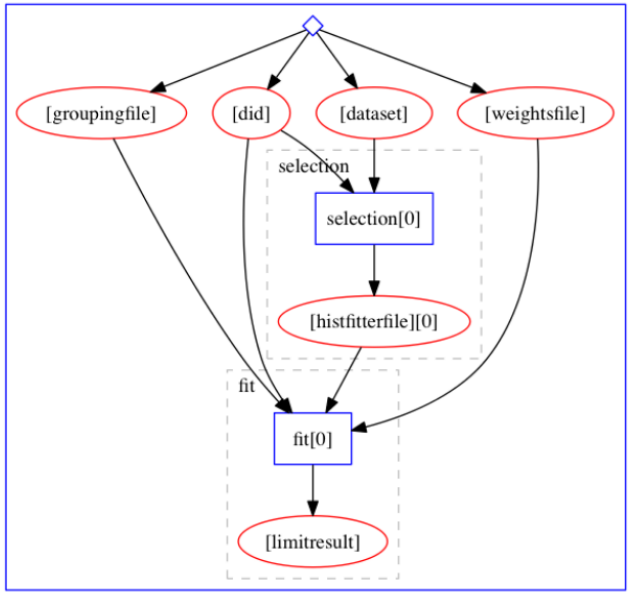
https://github.com/reanahub/reana-demo-bsm-search/

Yadage declarative workflows

- developed by 

- several ATLAS internal use case examples
  - reinterpreting MonoH(bb) exotics searches
  - adding new systematics on displaced objects
  - ...

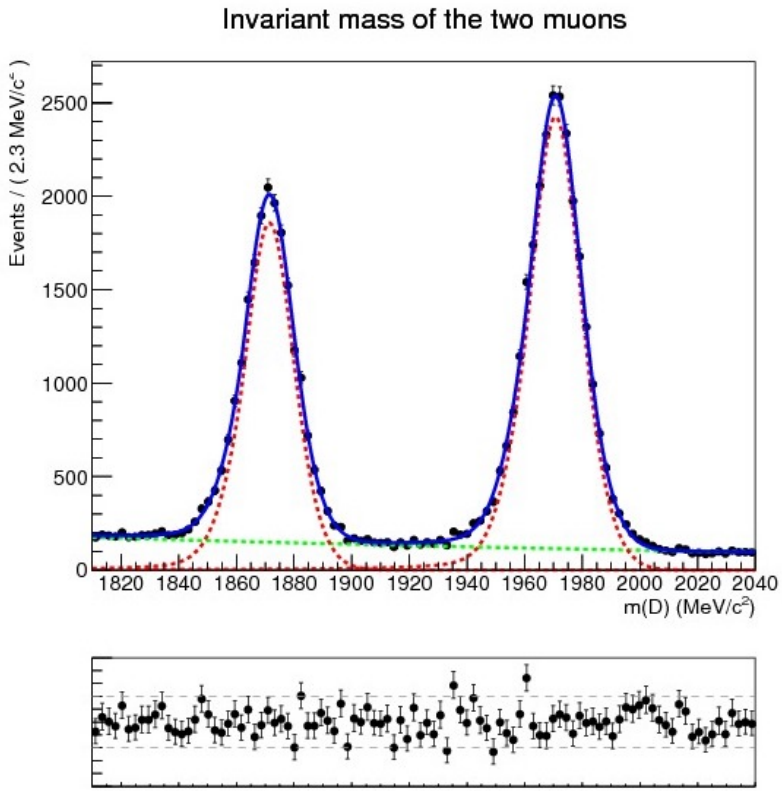
```yaml
stages:
  - name: selection
    dependencies: ['init']
    scheduler:
      scheduler_type: singlestep-stage
      parameters:
        dataset: {stages: init, output: dataset, unwrap: true}
        submitdir: '{workdir}/submitdir'
        outputprefix: '{workdir}/histfitter.root'
        did: {stages: init, output: did, unwrap: true}
      step: {$ref: 'selscript.yml#'}
  - name: fit
    dependencies: ['selection']
    scheduler:
      scheduler_type: singlestep-stage
      parameters:
        bkgtree: 'root://eosuser.cern.ch///eos/project/r/recast/Bkg_2.4.15-2-0_merged.root'
        datatree: 'root://eosuser.cern.ch///eos/project/r/recast/Data_2.4.15-2-0.root'
        outputjson: '{workdir}/fitoutput.json'
        selectionoutput: {stages: selection, output: histfitterfile, unwrap: true}
        weightsfile: {stages: init, output: weightsfile, unwrap: true}
        did: {stages: init, output: did, unwrap: true}
      step: {$ref: 'fitscript.yml#'}
```



Lukas Heinrich `http://github.com/yadage`

Example: LHCb rare charm decay

- search for $D^+_{(s)} \to \pi^+\mu^+\mu^-$ and $D^+_{(s)} \to \pi^-\mu^+\mu^+$ decays
  Phys. Lett. B 724 (2013) 203-212



https://github.com/reanahub/reana-demo-lhcb-d2pimumu/

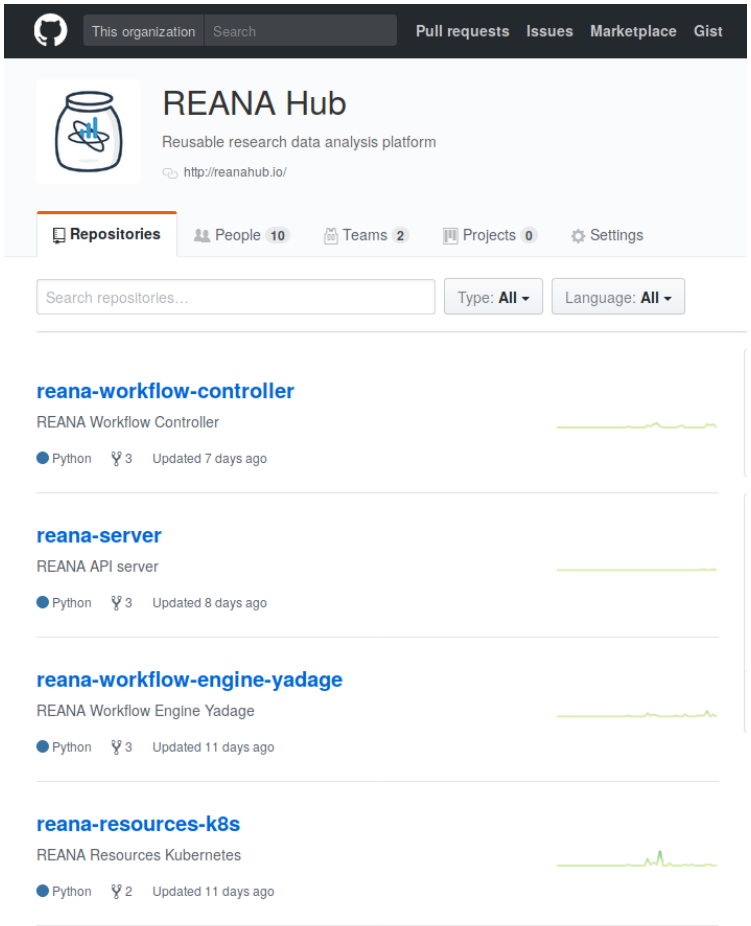
Example: ALICE LEGO train

- ALICE LEGO train validation
- large images (∼8 GB) including all dependencies (CVMFS)

```
$ root -b -q -x lego_train.C
...
I-AliPhysicsSelection::Initialize: Initializing for run 117222
I-AliPhysicsSelection::Initialize: Using Standard OADB
Physics selection cut settings:
  FMDLowCut            0.200000
  FMDHitCut            0.500000
...
===== Logical file name: /inputs/__alice__data__2011__LHC11h_2__000170387/AliESDs.root =====
...
=Analysis train= init time:       0.858033[sec]
              I/O & data mng.: 16.6985 [sec]
              task execution: 20.0355 [sec]
              total time:     CPU=24.4 [sec]  REAL=37.592[sec]
=Analysis train= Terminate time:  0.933296[sec]
Real time 0:00:40, CP time 27.390
======== lego_train.C finished with exit code: 0 ========
```

https://github.com/reanahub/reana-demo-alice-lego-train-validation/

REANA technology



REANA @ GitHub

- micro-services

python™ Flask web development, one drop at a time

- REST API

OPENAPI INITIATIVE Bravado

- services

RabbitMQ

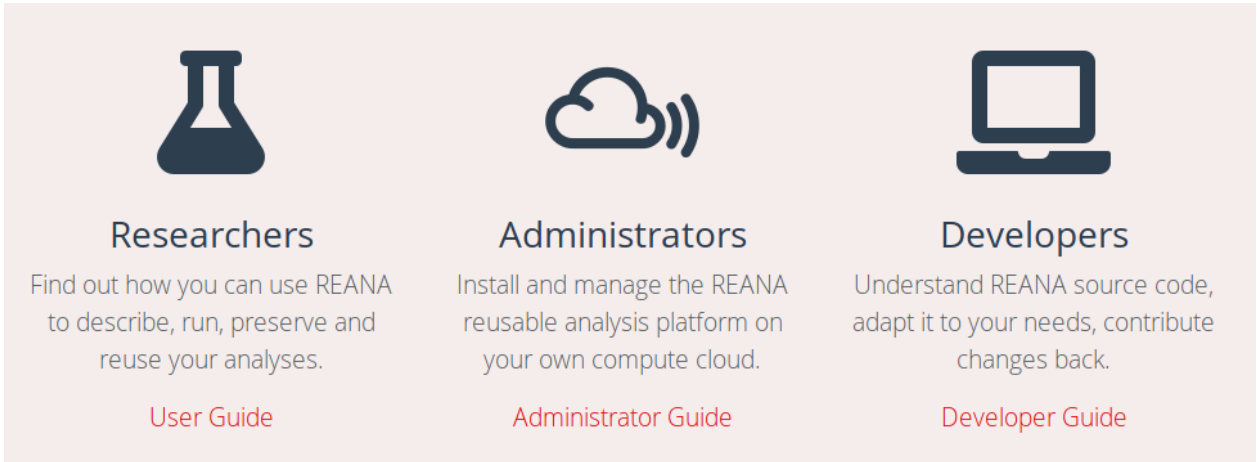
- deployments

Developer perspective

- set up REANA cluster

```
$ # install kubectl 1.9.1 and minikube 0.23.0
$ sudo dpkg -i kubectl*.deb minikube*.deb
$ minikube start --kubernetes-version="v1.6.4"
$ # install reana-cluster utility
$ mkvirtualenv reana-cluster
$ pip install reana-cluster
$ # deploy new cluster and check progress
$ reana-cluster init
$ reana-cluster status
$ # set environment variables for reana-client
$ eval $(reana-cluster env)
```

- REANA client ⟷ REANA cluster

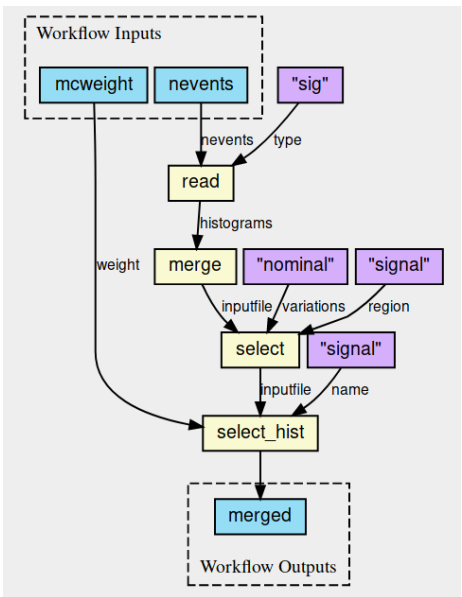


Physics meets Life sciences

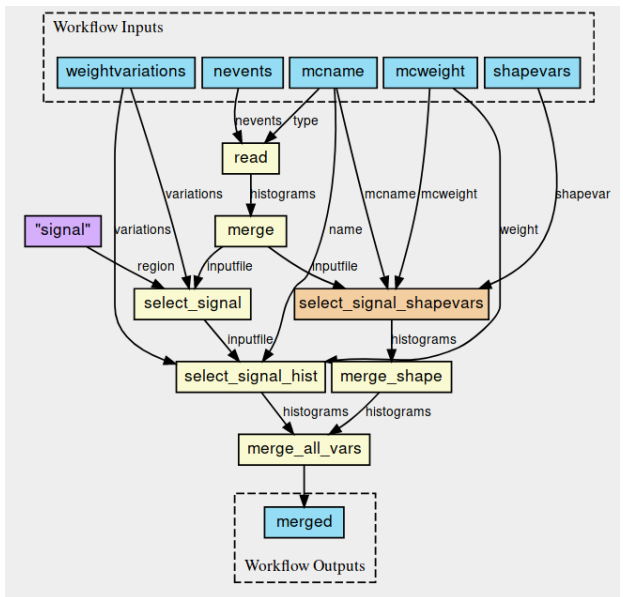
- Common Workflow Language
  - originated in bioinformatics
  - rich ecosystem: implementations, editors, tools

COMMON WORKFLOW LANGUAGE

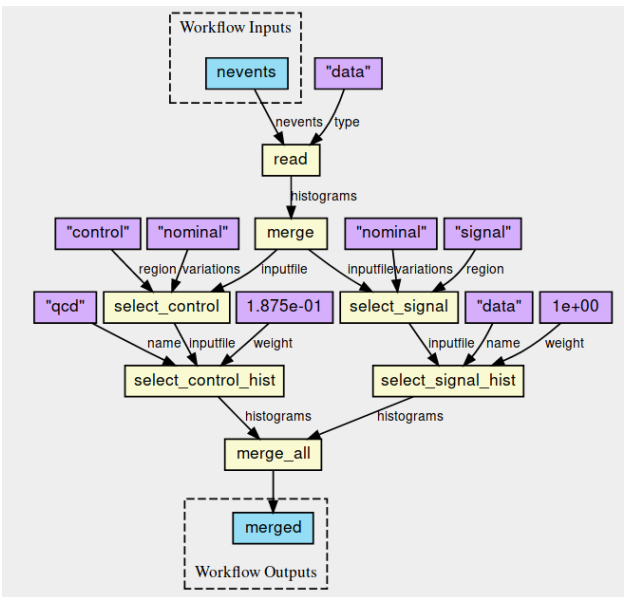
- BSM search example using CWL workflows



sig

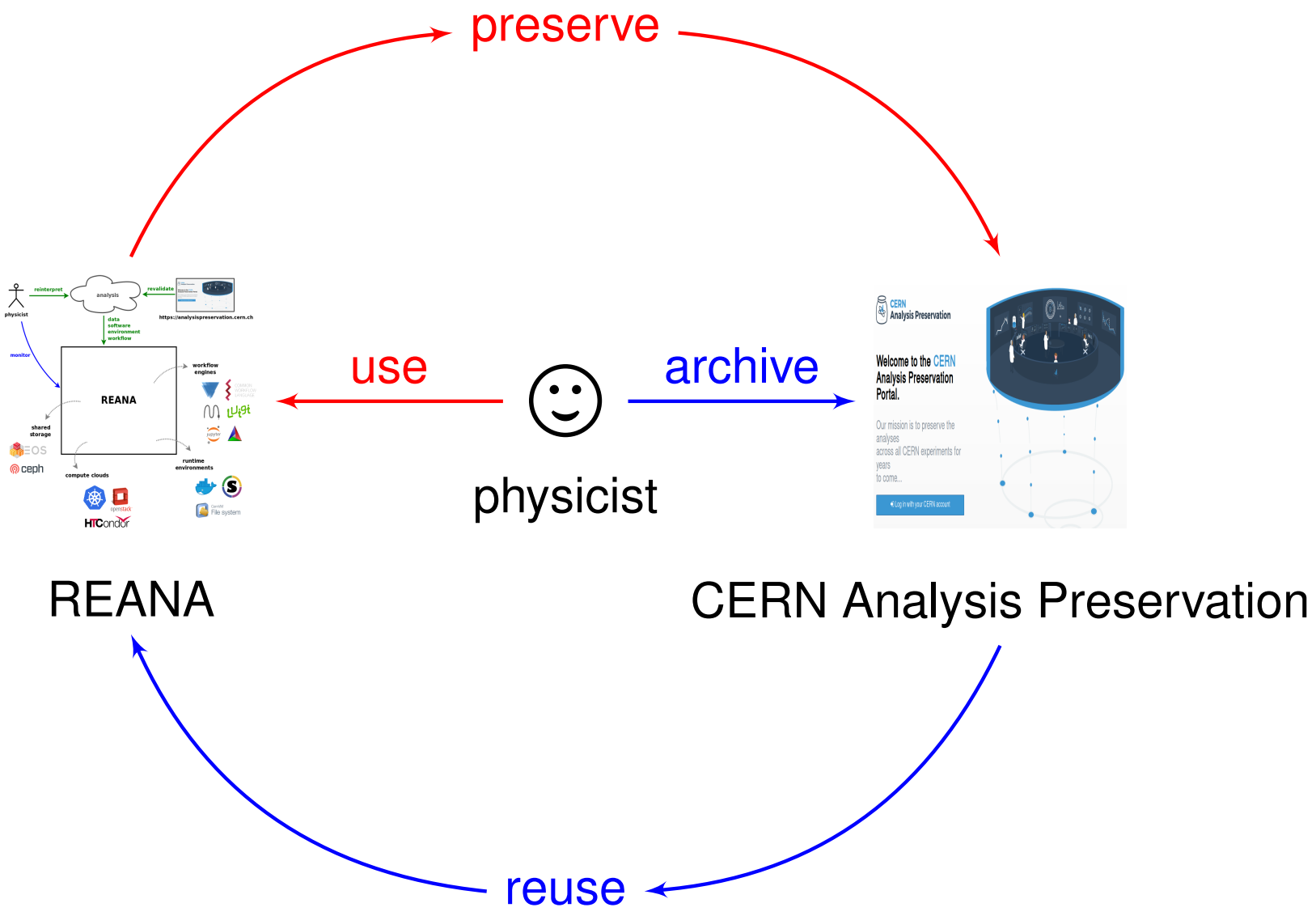


mc



data

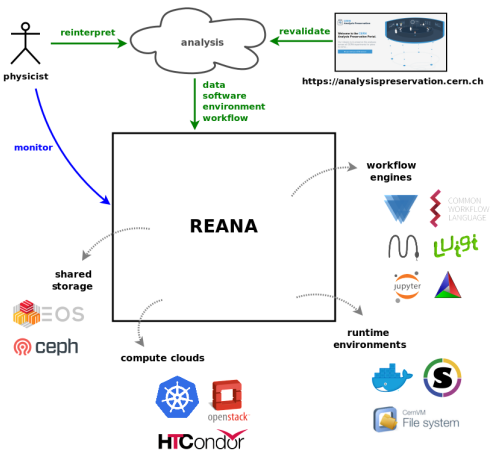
Reusability ⇌ Preservation



References

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