Preserving and Reusing High-Energy Physics Data Analyses

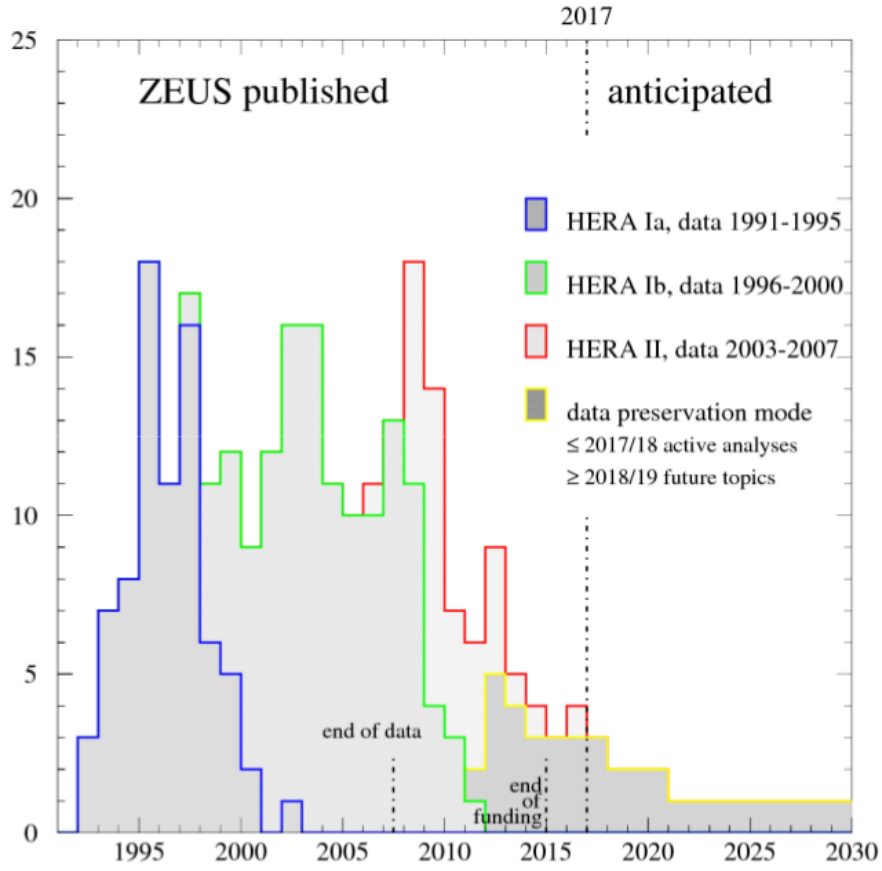
Tibor Šimko

CERN

openlab workshop · Geneva, Switzerland · 6 December 2017

Long-term value of data

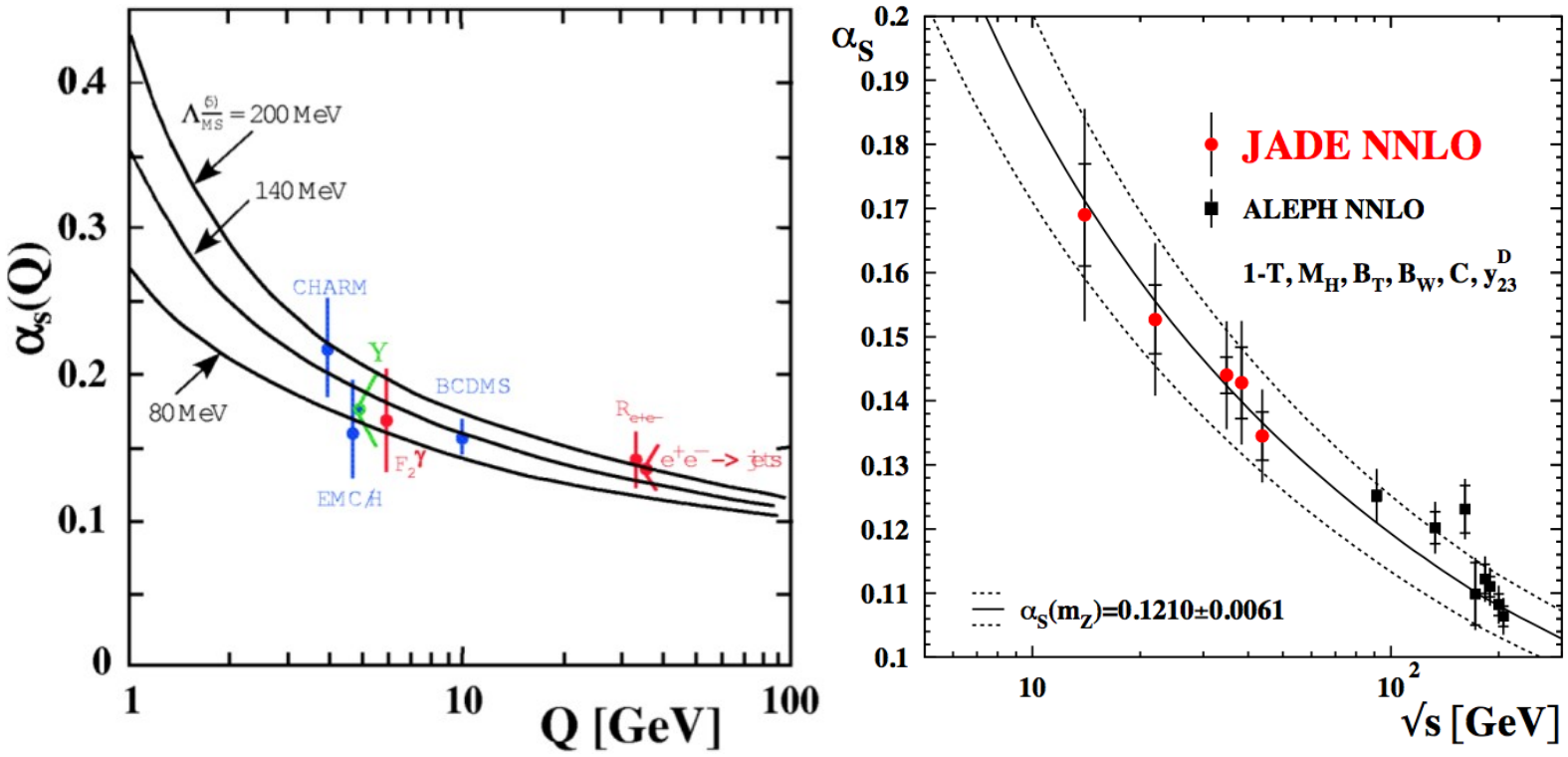
- significant number of publications after end of data taking
- example: ZEUS detector operating on HERA accelerator at DESY



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

Long-term value of data

- JADE experiment (1979–1986) on PETRA accelerator at DESY
- JADE data still cover unique e^+e^- energy range in 2017
- JADE data being re-analysed even ∼35 years later!



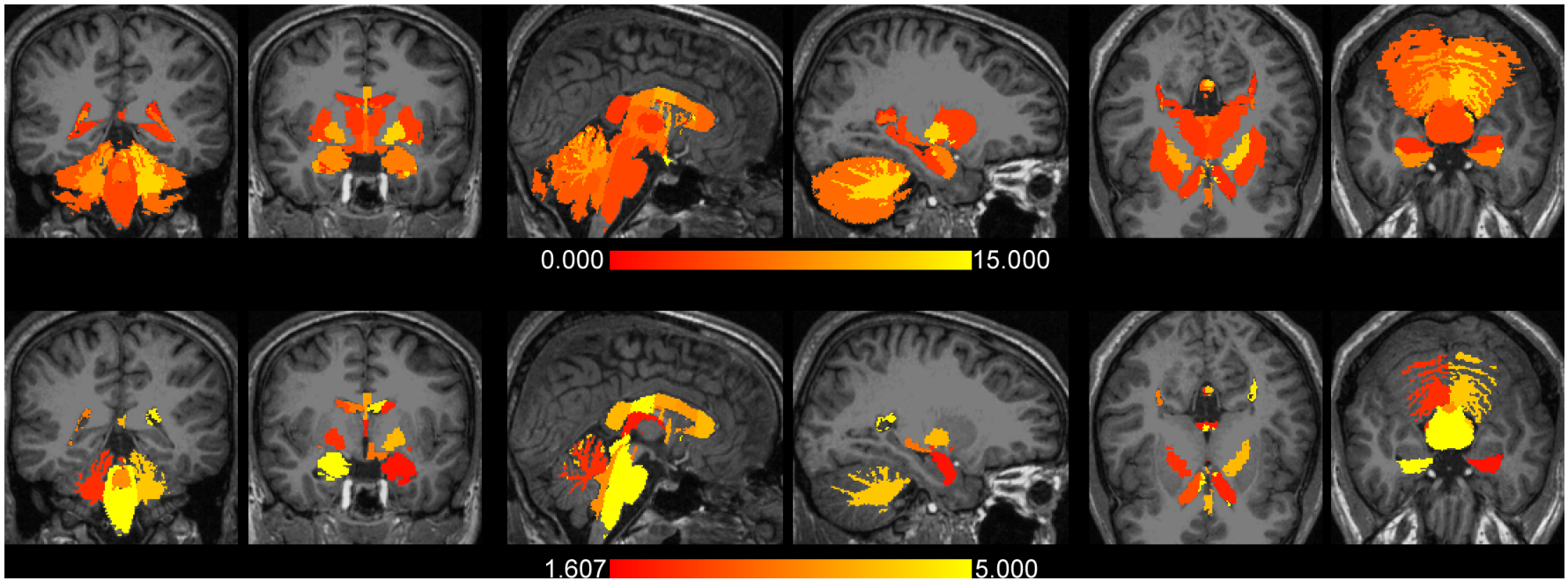
DPHEP `https://arxiv.org/abs/1205.4667`

Data alone is not enough

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

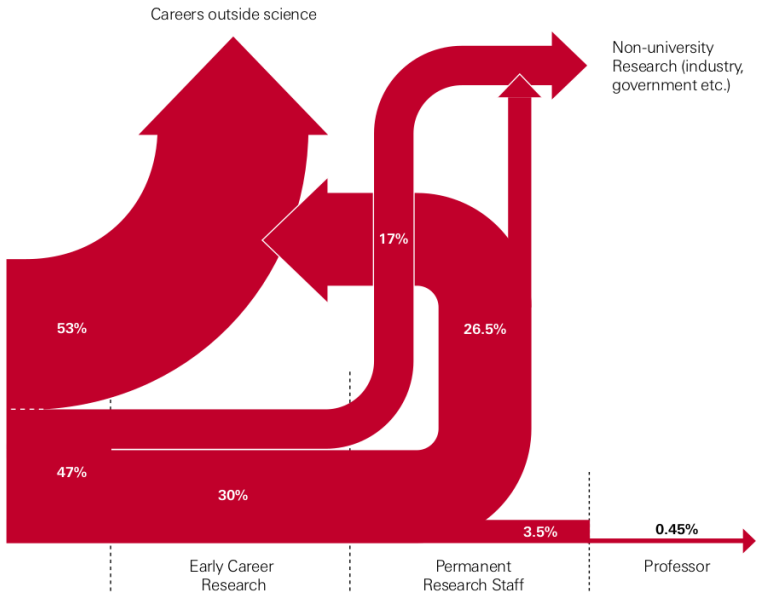


8.8±6.6% (volume) and 2.8±1.3% (cortical thickness)

Long-term value of knowledge?

CMS collaboration

Experimental physics done by groups of ∼3000 physicists.

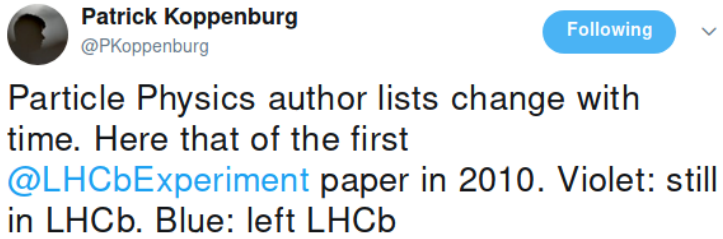


Career after PhD

THE ROYAL SOCIETY

High turnover of young researchers.

LHCb members of 2010 in 2017



Patrick Koppenburg
@PKoppenburg

Following

Particle Physics author lists change with time. Here that of the first @LHCbExperiment paper in 2010. Violet: still in LHCb. Blue: left LHCb

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

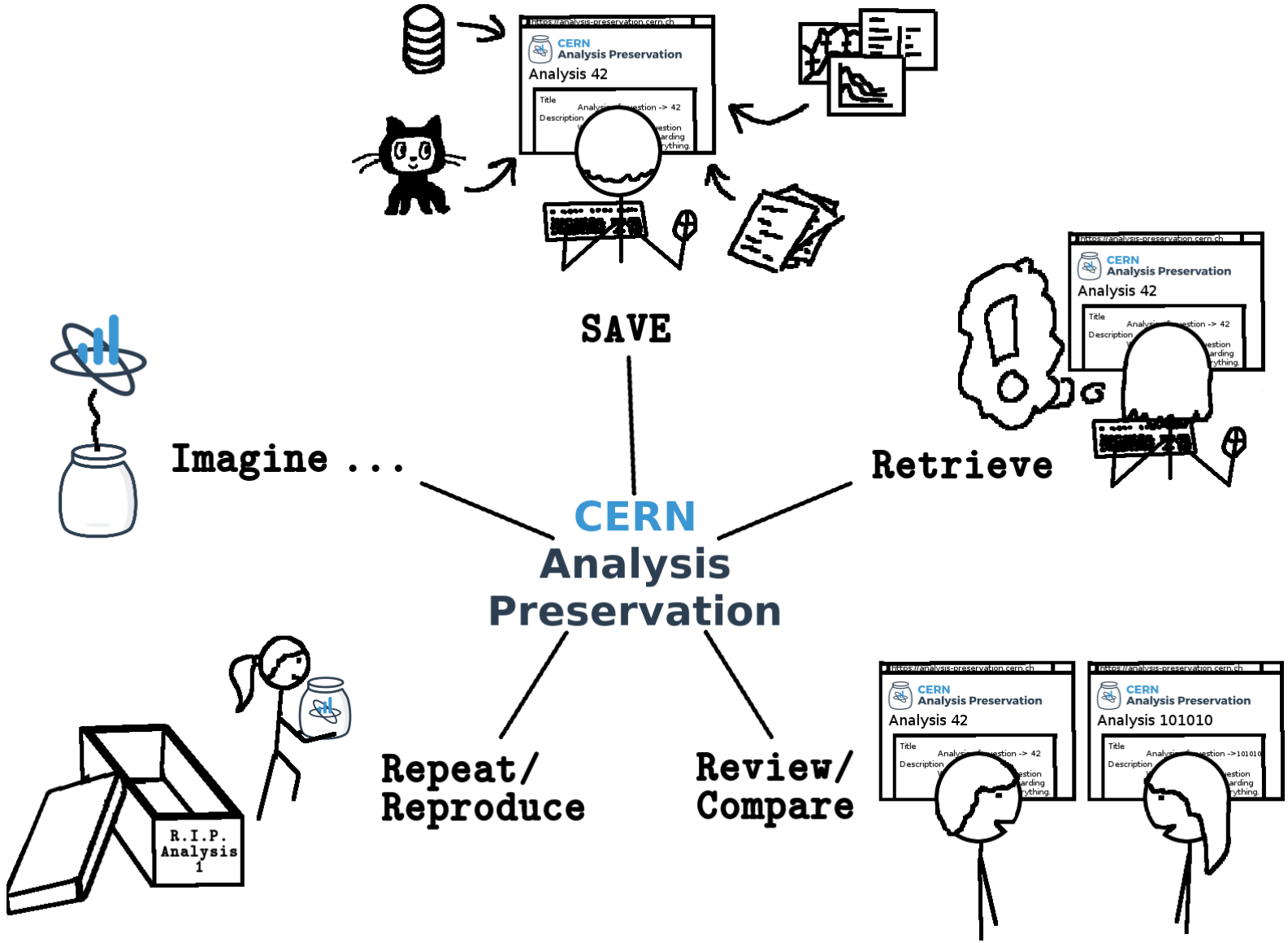
- A platform for **preserving knowledge** and **assets** of an individual physics analysis.
- Capturing the elements needed to **understand** and **reuse** an analysis even several years later:

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

- Advanced **search** for high-level physics information
- Applying standard **collaboration access restrictions**

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

User stories



LHC data analysis pyramid

~GB/analysis

~TB/analysis

~PB/year

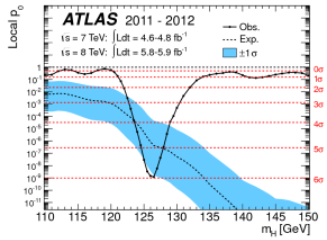
~GB/sec

1. Provision of additional documentation for the published results

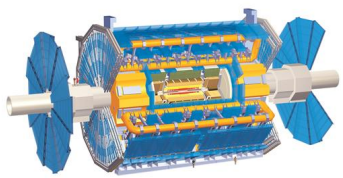
2. Simplified data formats for analysis in outreach and training exercises

3. Reconstructed data and simulations as well as the analysis level software to allow a full scientific analysis

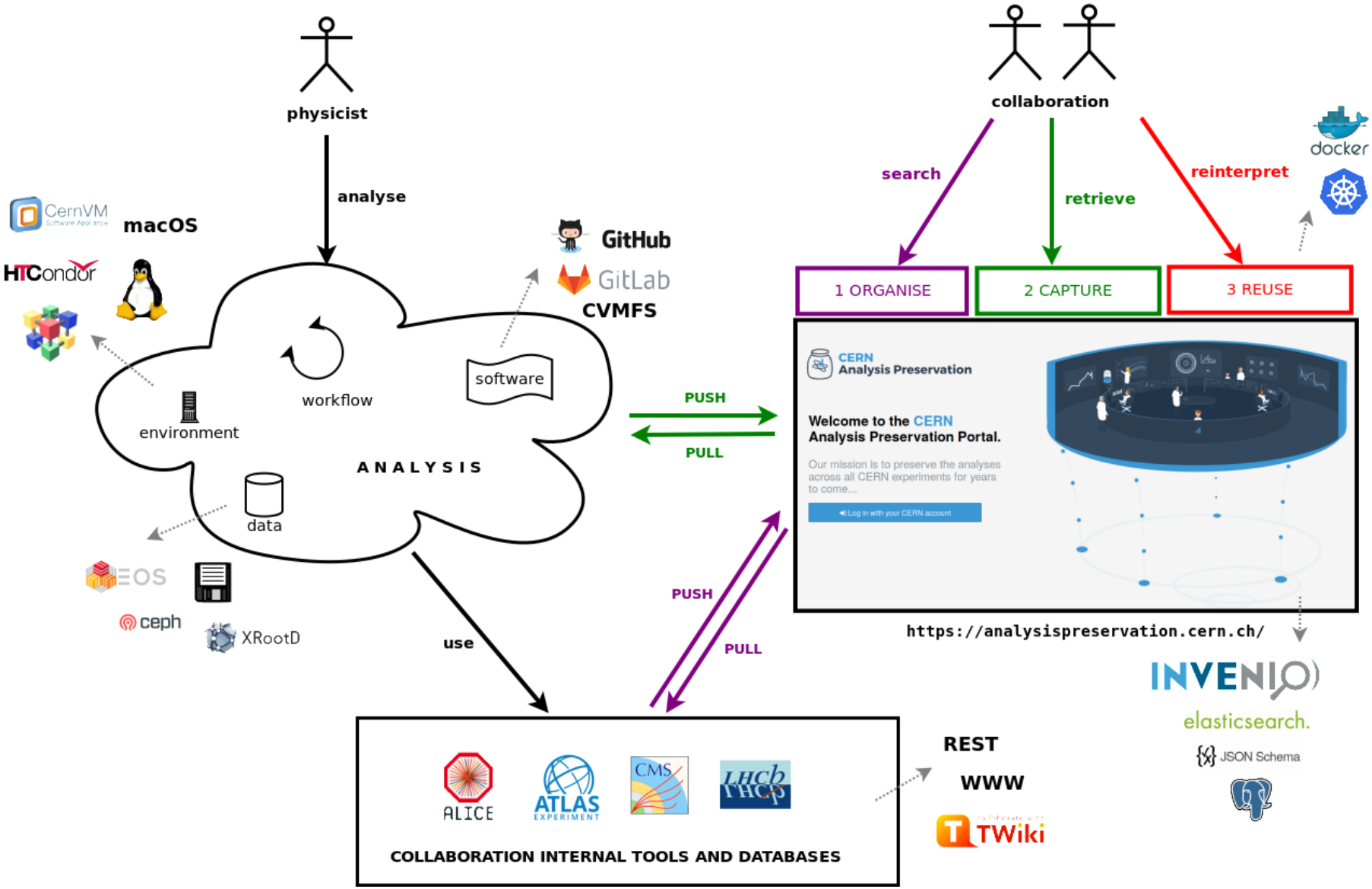
4. Basic raw level data (if not yet covered as level 3 data) and their associated software which allows access to the full potential of the experimental data



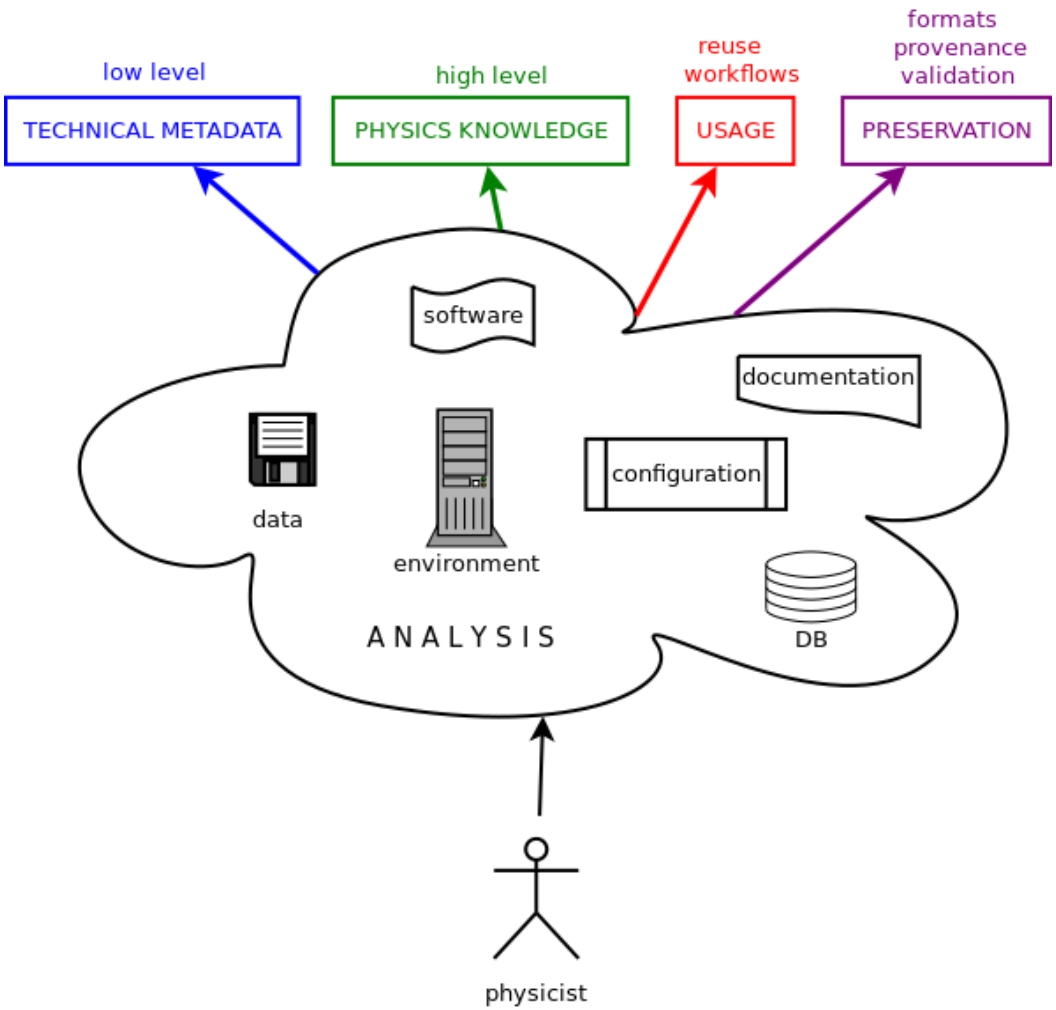
analysis



System overview



1. Describing an analysis

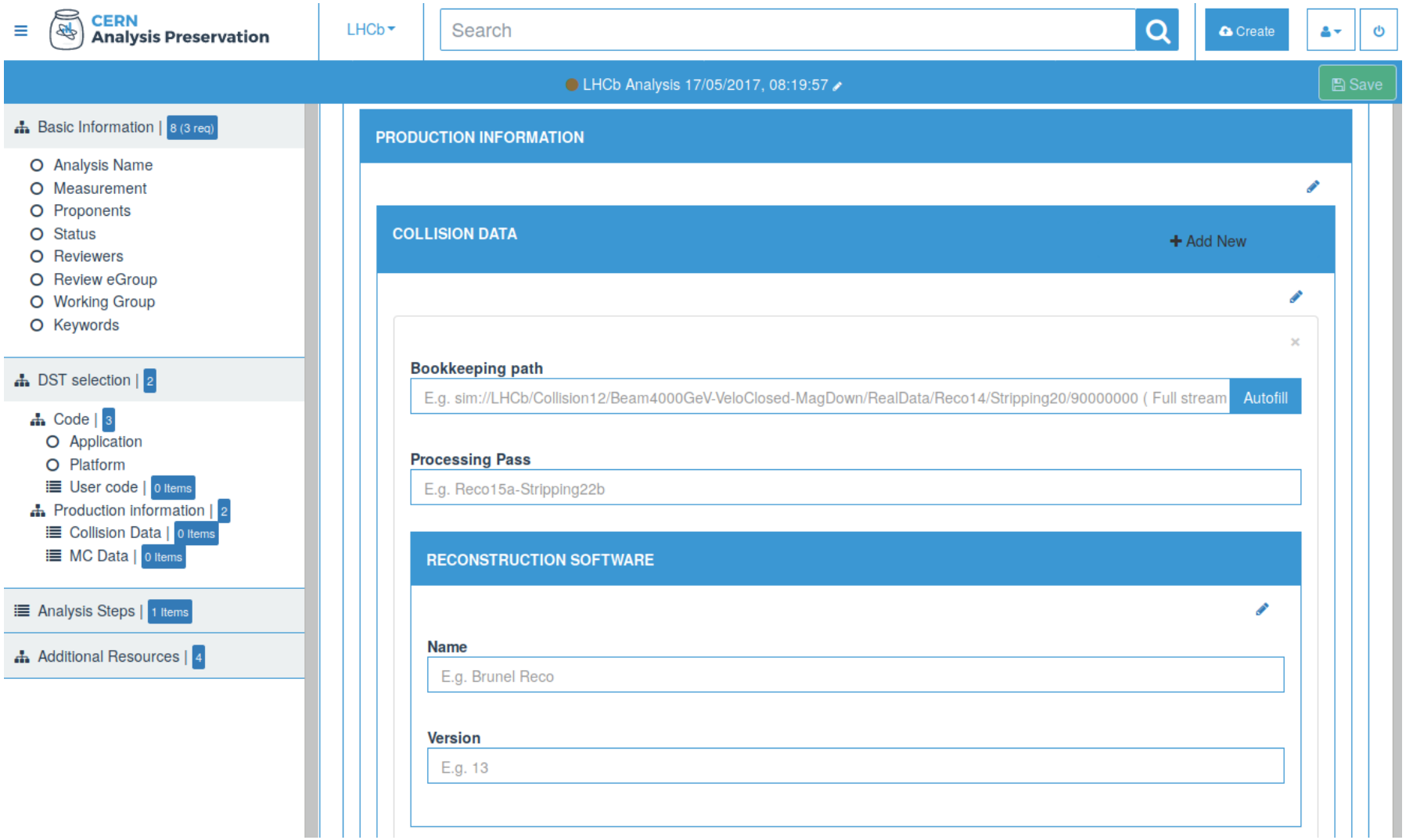


INVENIO

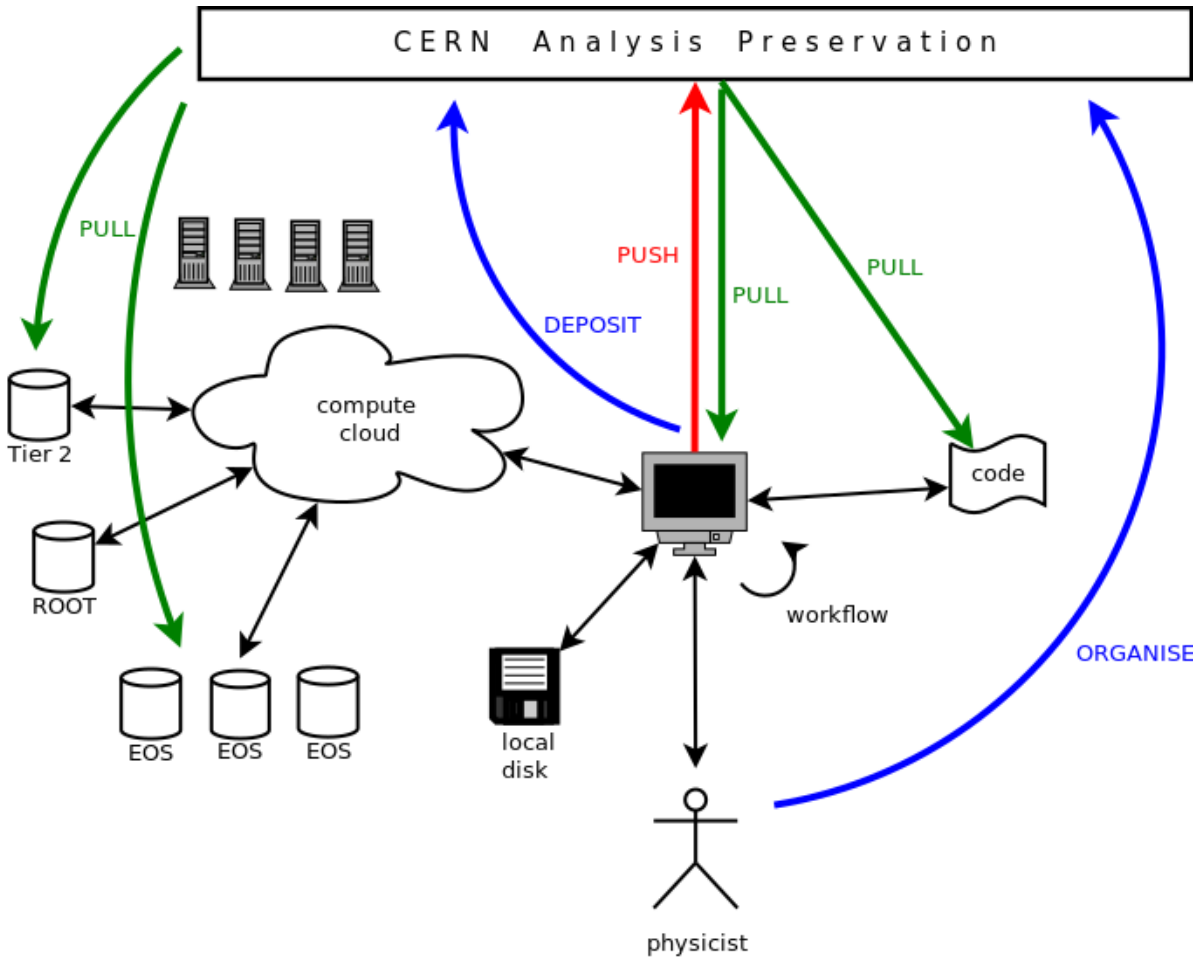
- JSON Schema
- W3C DCAT
- domain-specific fields

Structuring knowledge behind research data analysis.

Example: LHCb production info



2. Capturing an analysis

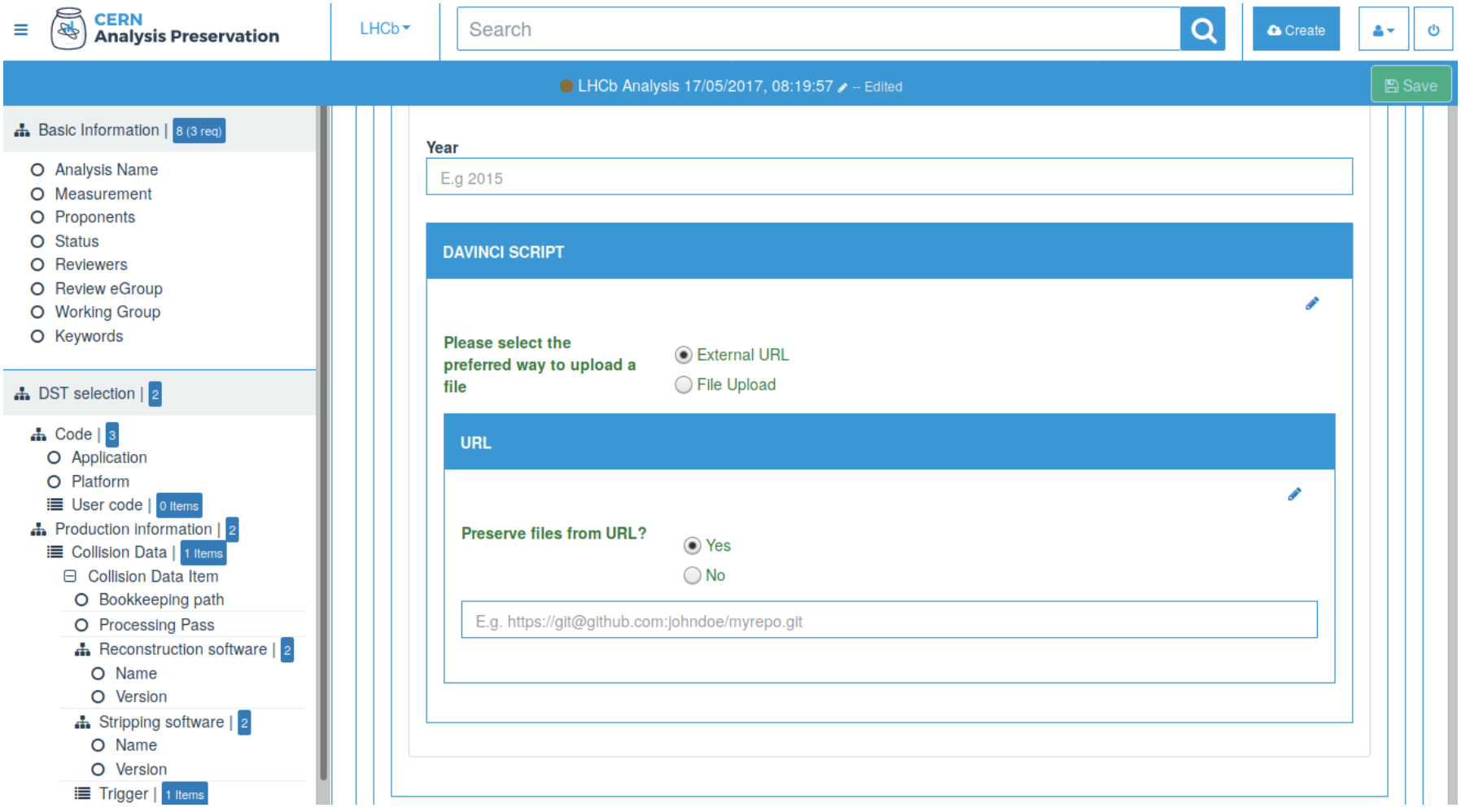


INVENIO

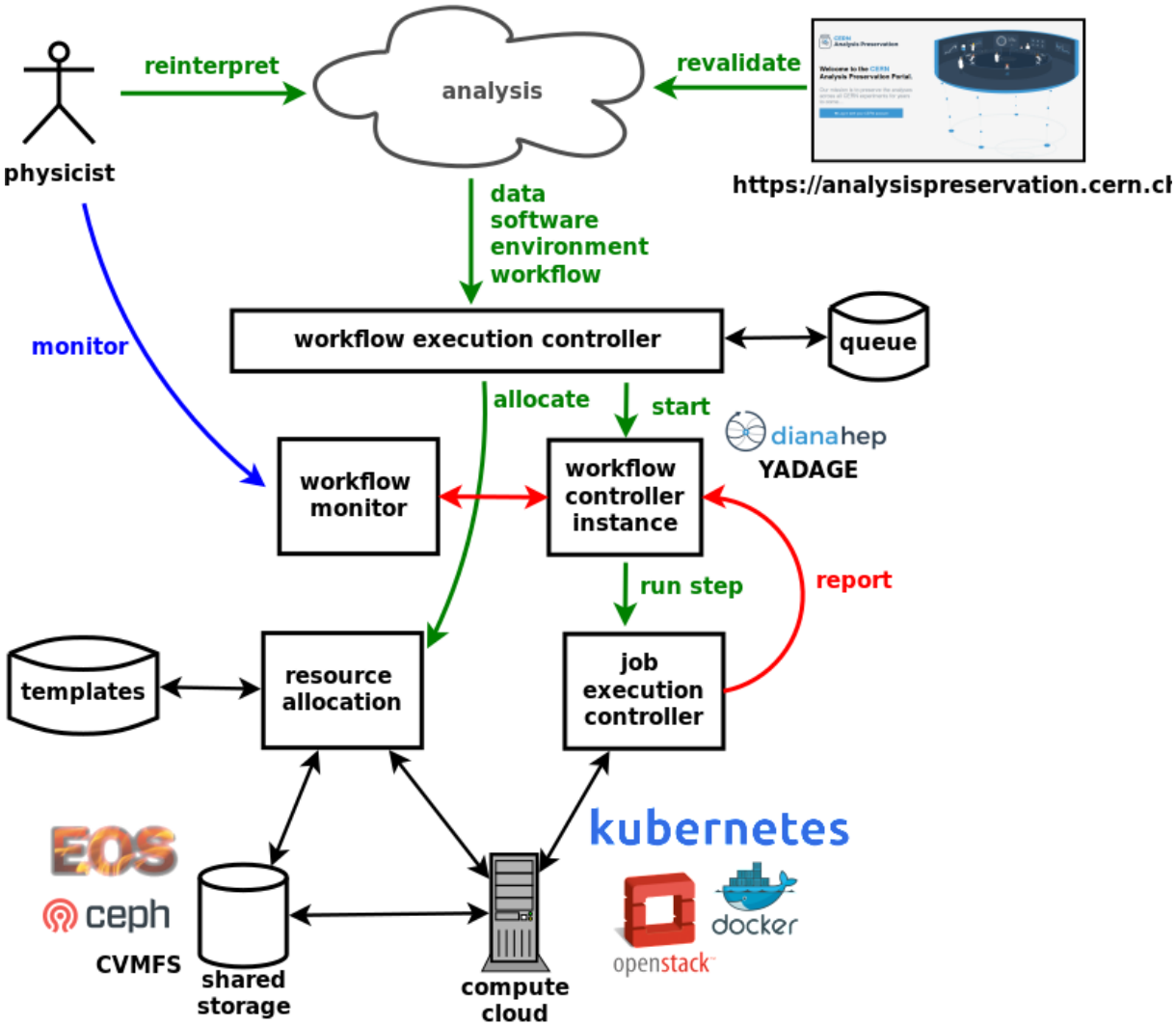
- datasets:
  local storage,
  cloud storage
- software:
  Git, SVN
- information:
  DBs, TWiki,
  SharePoint
- protocols:
  HTTP, XRootD

Taking consistent snapshot of analysis assets at a certain time.

Demo: LHCb DaVinci script



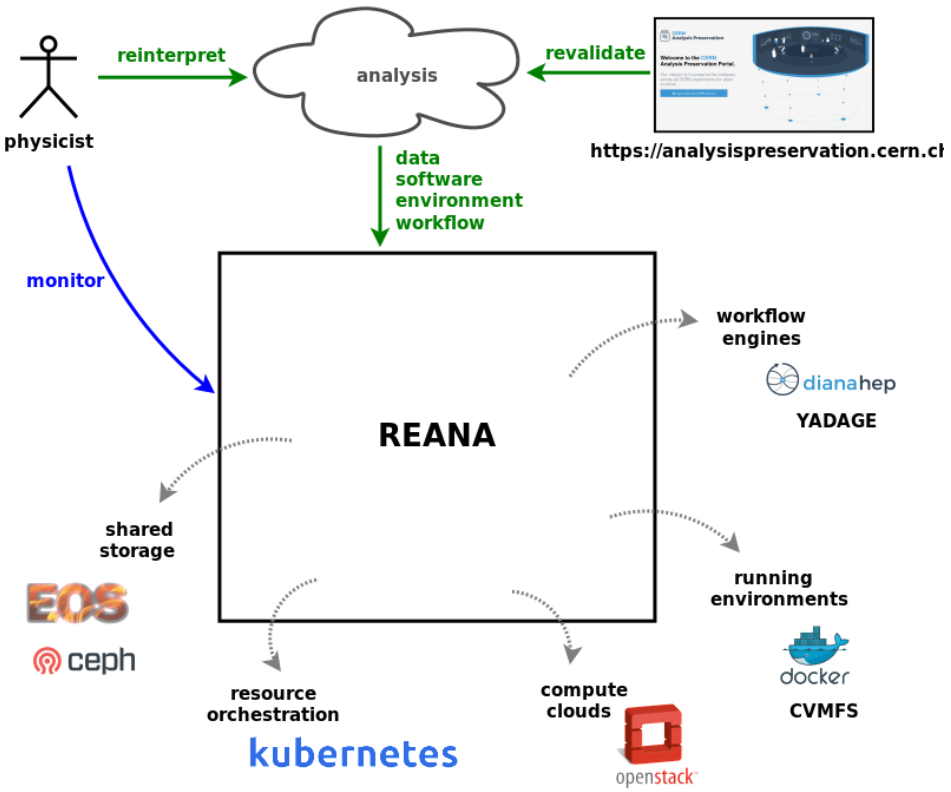
3. Reusing an analysis



Instantiating preserved analysis on the cloud.

REANA = REusable ANAlyses

- a system for **reusable analyses** execution **on the cloud**
- supporting **multiple scenarios**
  - multiple compute clouds
    → OpenStack with Kubernetes
    → HTCondor?
  - multiple runtime environments
    → Docker with CVMFS
  - multiple workflow engines
    → Yadage, CWL
  - multiple shared storage systems
    → Ceph, EOS



- close **collaboration** with

DASPOS dianahep recast

Four questions

1 Input data

What is your input data?

- input files
- input parameters

2 Analysis code

Which code analyses it?

- software frameworks
- custom 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: single command

```
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

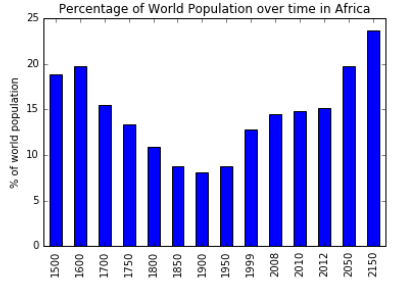
We'll start with a histogram depicting the evolution of a specific region's portion of the world population, in percentage.

```
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

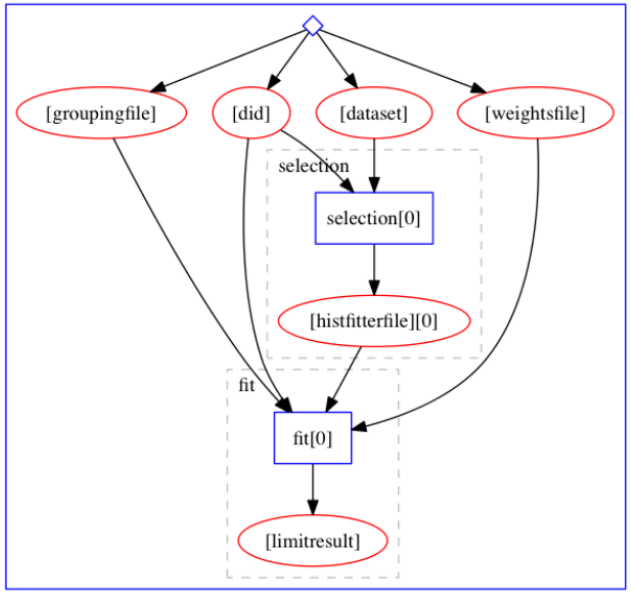
4 **workflow**: `jupyter nbconvert`

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

Complex example: workflows

- **case studies** in high-energy-physics with LHC collaborations
  - ALICE AliPhysics and post-LEGO train analysis
  - ATLAS multi-B-jets analysis, . . .
  - LHCb $\Lambda_b^0 \to \Lambda_c^+ \bar{D}^0 K^-$ analysis, data production
- diverse **community habits** in computational **workflows**
  - Yadage, Snakemake, Luigi

```
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#'}
```



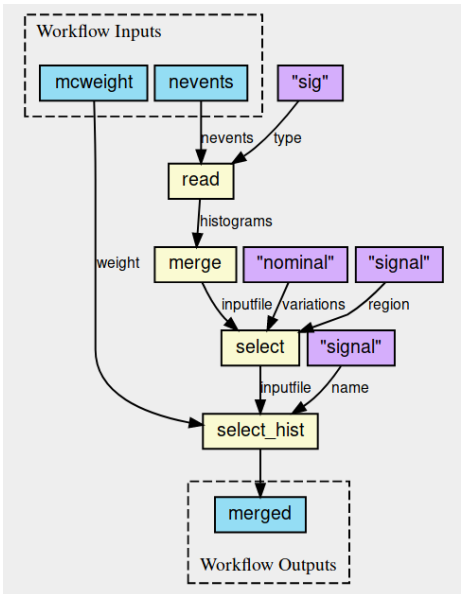
Lukas Heinrich `http://github.com/diana-hep/yadage`

Physics meets Life Sciences?

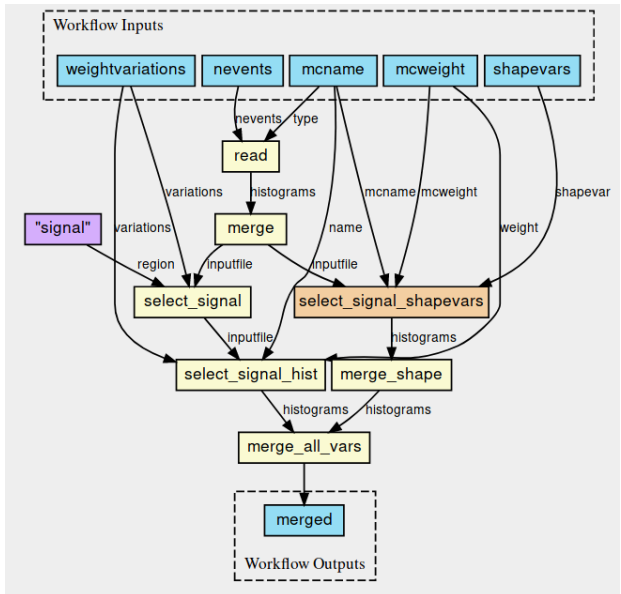
- Common Workflow Language
  - originated in bioinformatics
  - rich ecosystem: implementations, editors, tools

COMMON WORKFLOW LANGUAGE

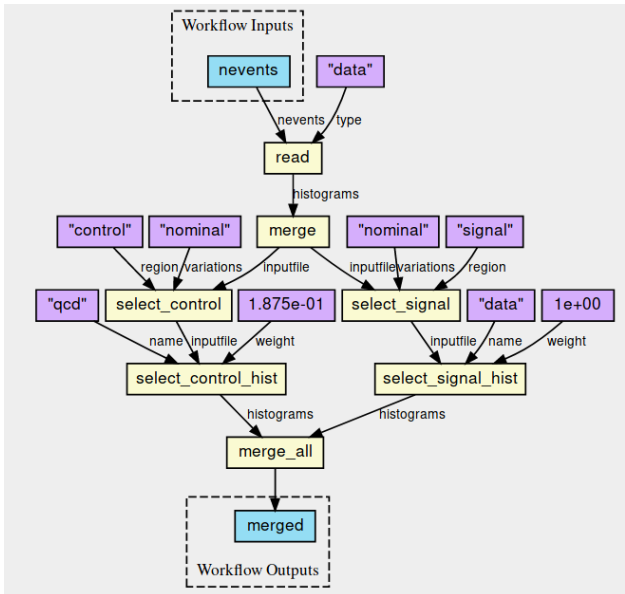
- modelling high-energy physics analyses
  - ttH analysis, ∼80 TB storage, ∼500k tasks, ∼10 days
- ATLAS full chain analysis example; CWL workflows



sig



mc



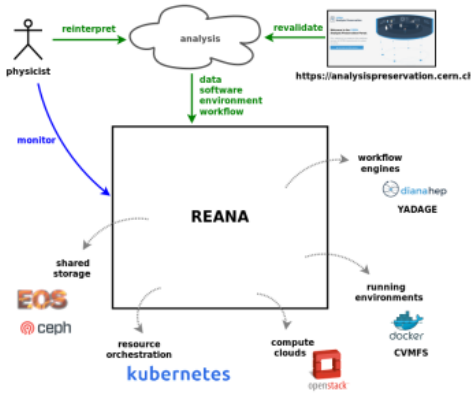
data

Reusable analyses

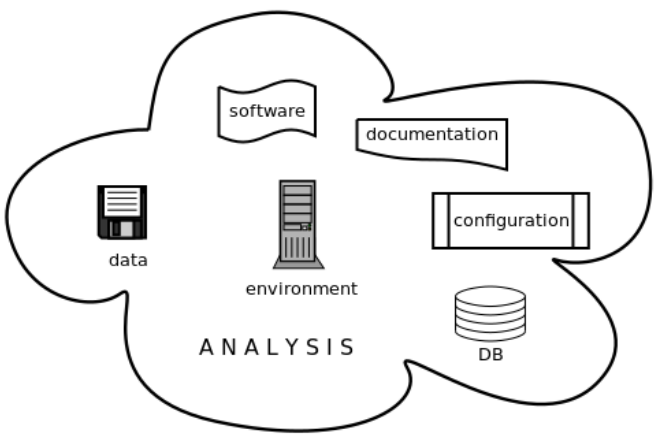
```yaml
metadata:
  author: "REANA Developers."
  title: "Reusable analysis example - ROOT6 and RooFit"
code:
  files:
    - gendata.C
    - fitdata.C
environment:
  - type: docker
    image: reanahub/reana-env-root6
workflow:
  type: yadage
  file: workflow.yml
parameters:
  events: 20000
output:
  files:
    - plot.png
```

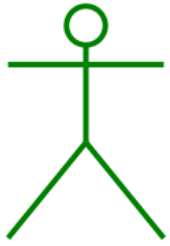
reana.yml

$ reana-client run



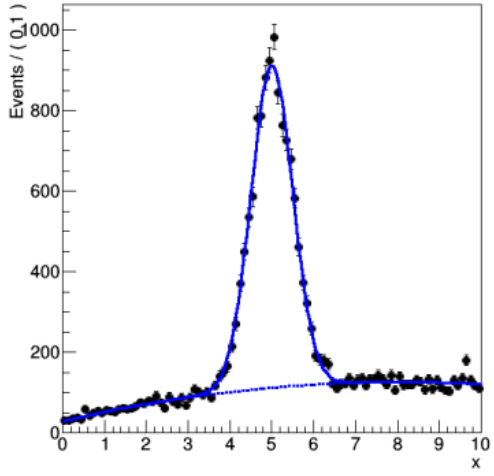
REANA cloud instance





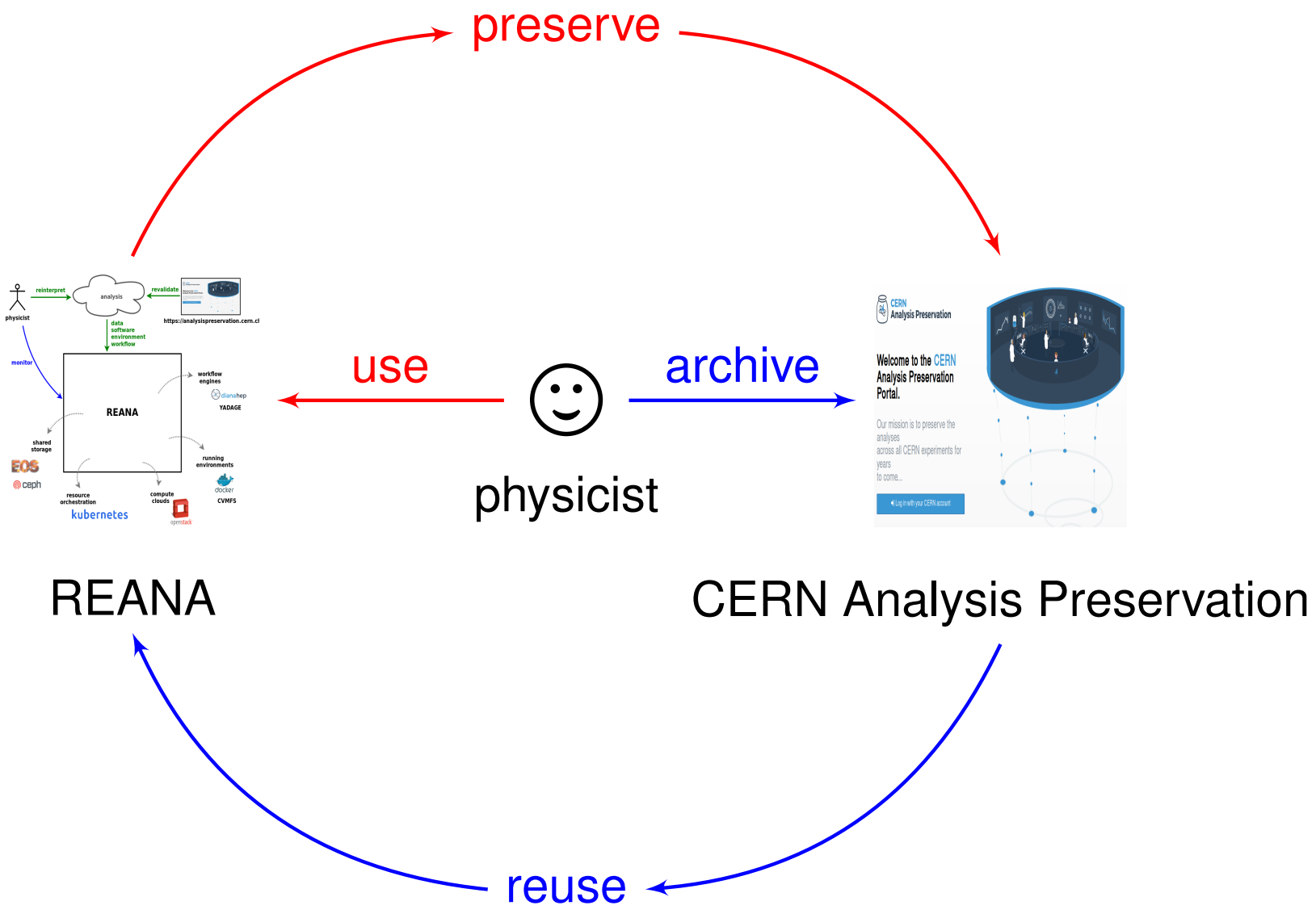
physicist

Fit example



final plot

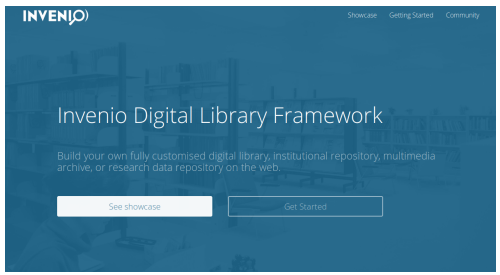
Reusability ⇌ Preservation



References

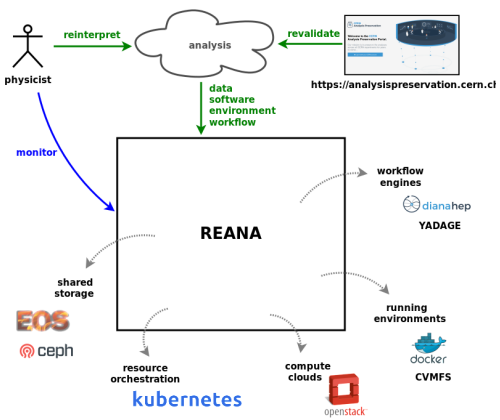
CERN Analysis Preservation

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



Invenio

- http://inveniosoftware.org
- http://github.com/inveniosoftware
- inveniosoftware
- info@inveniosoftware.org



REANA

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