Boosting productivity of researches integrating IT services into an interactive analysis platform

D. Piparo (CERN EP-SFT) for the SWAN team

https://cern.ch/swan

Aug 31st, 2018

GridKa School

The Work Behind These Slides

- I am here on behalf of the SWAN team, spanning two CERN Departments: Experimental Physics and Information Technology
- The team: E. Tejedor, D. Castro, P. Mato, E. Bocchi, J. Moscicki, M. Lamanna, P. Kothuri, D. P.
- **I'd like to thank Diogo Castro and Enric Tejedor for contributing so much to this talk.**

Introduction


CERN
Accelerating Science

CERN Mission

https://kt.cern/success-stories/cern-virtual-machine-file-system-euclid-science-data-centres

Innovation

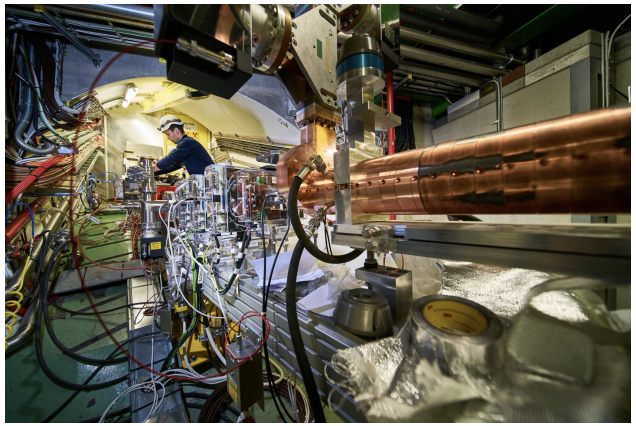
Education

CERN

Uniting People

Research

https://root.cern.ch/code-snippet/learning-root

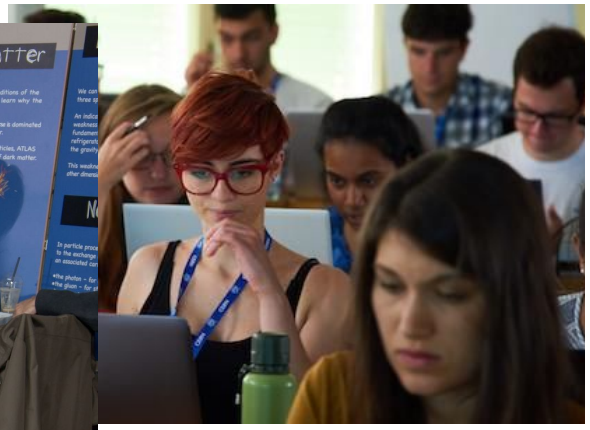


https://home.cern/about/experiments/awake

CERN Mission

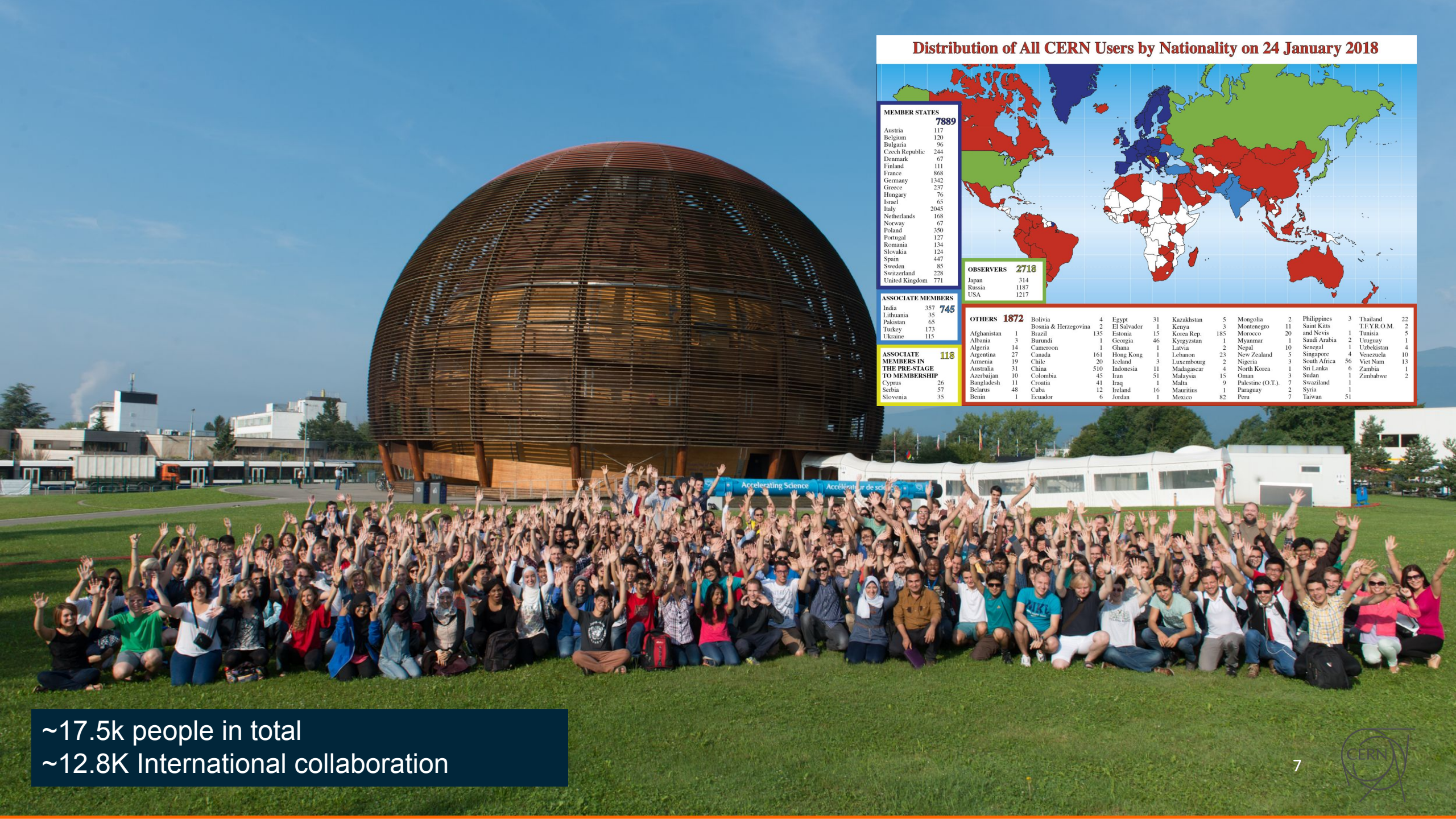
https://kt.cern/success-stories/cern-virtual-ma
e-file-system-euclid-science-data-centres

https://home.cern/about/experiments/awake



://root.cern.ch/code-snippet/learning-root

http://careers.cern/



Distribution of All CERN Users by Nationality on 24 January 2018

MEMBER STATES	7889
Austria	117
Belgium	120
Bulgaria	96
Czech Republic	244
Denmark	67
Finland	111
France	868
Germany	1342
Greece	237
Hungary	76
Israel	65
Italy	2045
Netherlands	168
Norway	67
Poland	350
Portugal	127
Romania	134
Slovakia	124
Spain	447
Sweden	85
Switzerland	228
United Kingdom	771

OBSERVERS	2718
Japan	314
Russia	1187
USA	1217

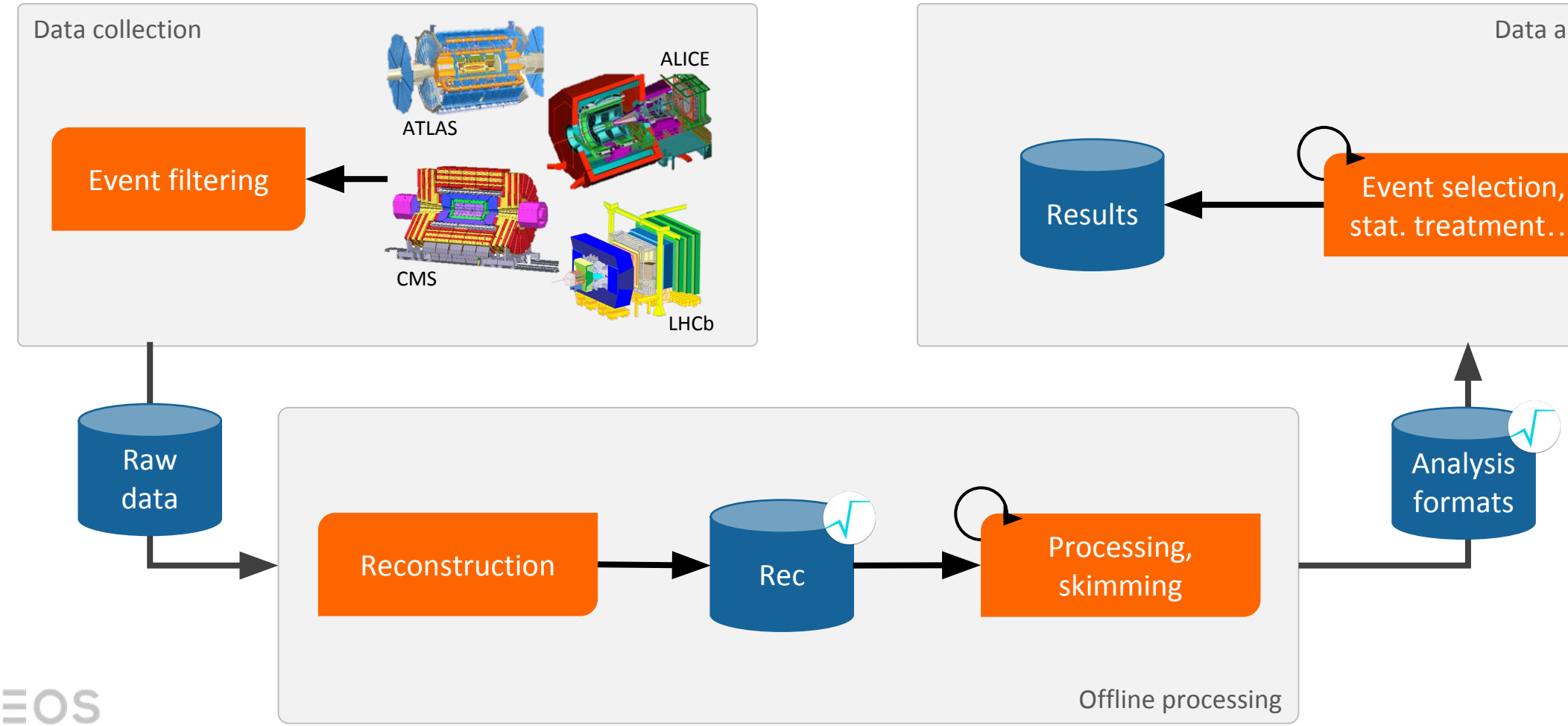
ASSOCIATE MEMBERS		745
India	357	
Lithuania	35	
Pakistan	65	
Turkey	173	
Ukraine	115	

ASSOCIATE MEMBERS IN THE PRE-STAGE TO MEMBERSHIP	118
Cyprus	26
Serbia	57
Slovenia	35

OTHERS	1872																			
Afghanistan	1	Bolivia	4	Egypt	31	Kazakhstan	5	Mongolia	2	Philippines	3	Thailand	22
Albania	3	Bosnia & Herzegovina	2	El Salvador	1	Kenya	3	Montenegro	11	Saint Kitts and Nevis	1	T.F.Y.R.O.M.	2
Algeria	14	Brazil	135	Estonia	15	Korea Rep.	185	Morocco	20			Tunisia	5
Argentina	27	Burundi	1	Georgia	46	Kyrgyzstan	1	Myanmar	1	Saudi Arabia	2	Uruguay	1
Armenia	19	Cameroon	1	Ghana	1	Latvia	2	Nepal	10	Senegal	1	Uzbekistan	4
Australia	31	Canada	161	Hong Kong	1	Lebanon	23	New Zealand	5	Singapore	4	Venezuela	10
Azerbaijan	10	Chile	20	Iceland	3	Luxembourg	2	Nigeria	3	South Africa	56	Viet Nam	13
Bangladesh	11	China	510	Indonesia	11	Madagascar	4	North Korea	1	Sri Lanka	6	Zambia	1
Belarus	48	Colombia	45	Iran	51	Malaysia	15	Oman	3	Sudan	1	Zimbabwe	2
Benin	1	Croatia	41	Iraq	1	Malta	9	Palestine (O.T.)	7	Swaziland	1		
		Cuba	12	Ireland	16	Mauritius	1	Paraguay	2	Syria	1		
		Ecuador	6	Jordan	1	Mexico	82	Peru	7	Taiwan	51		

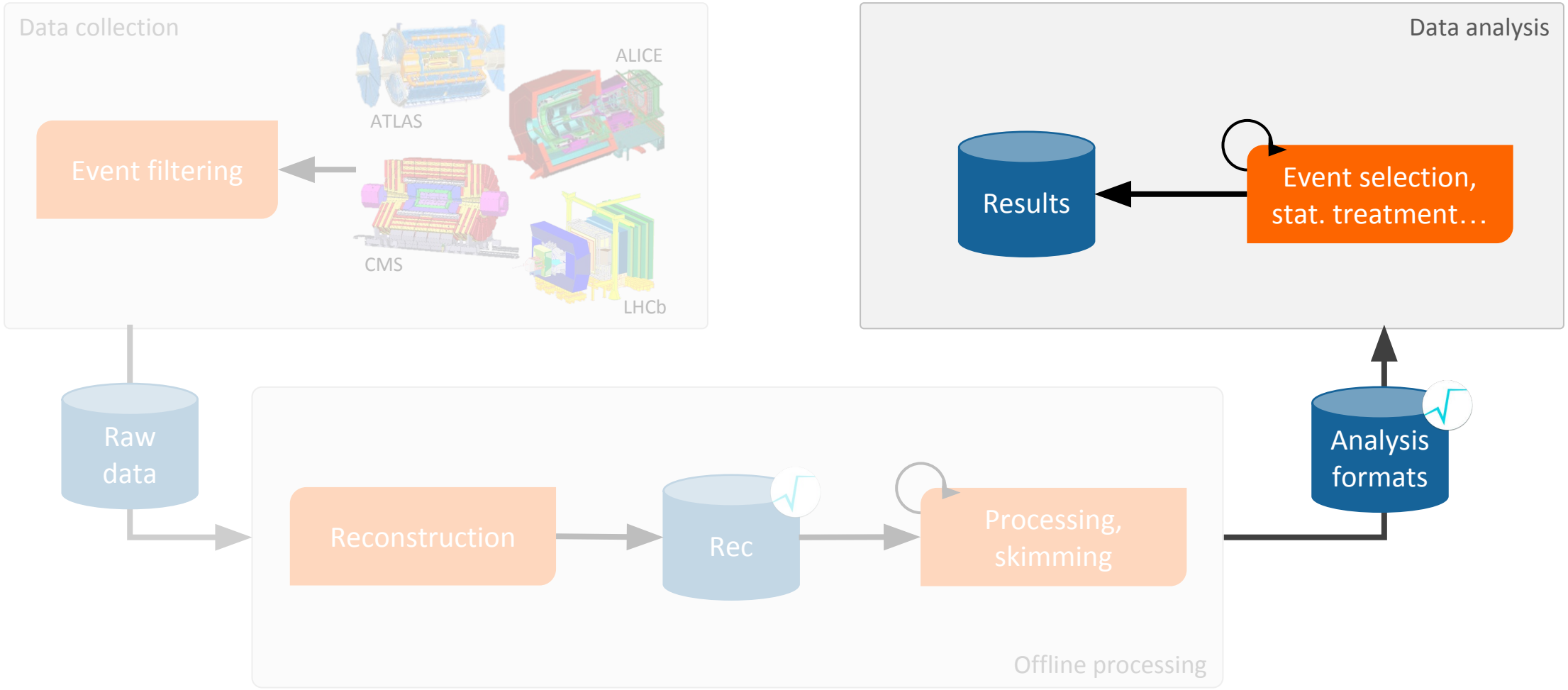
~17.5k people in total
~12.8K International collaboration

The steps towards data analysis



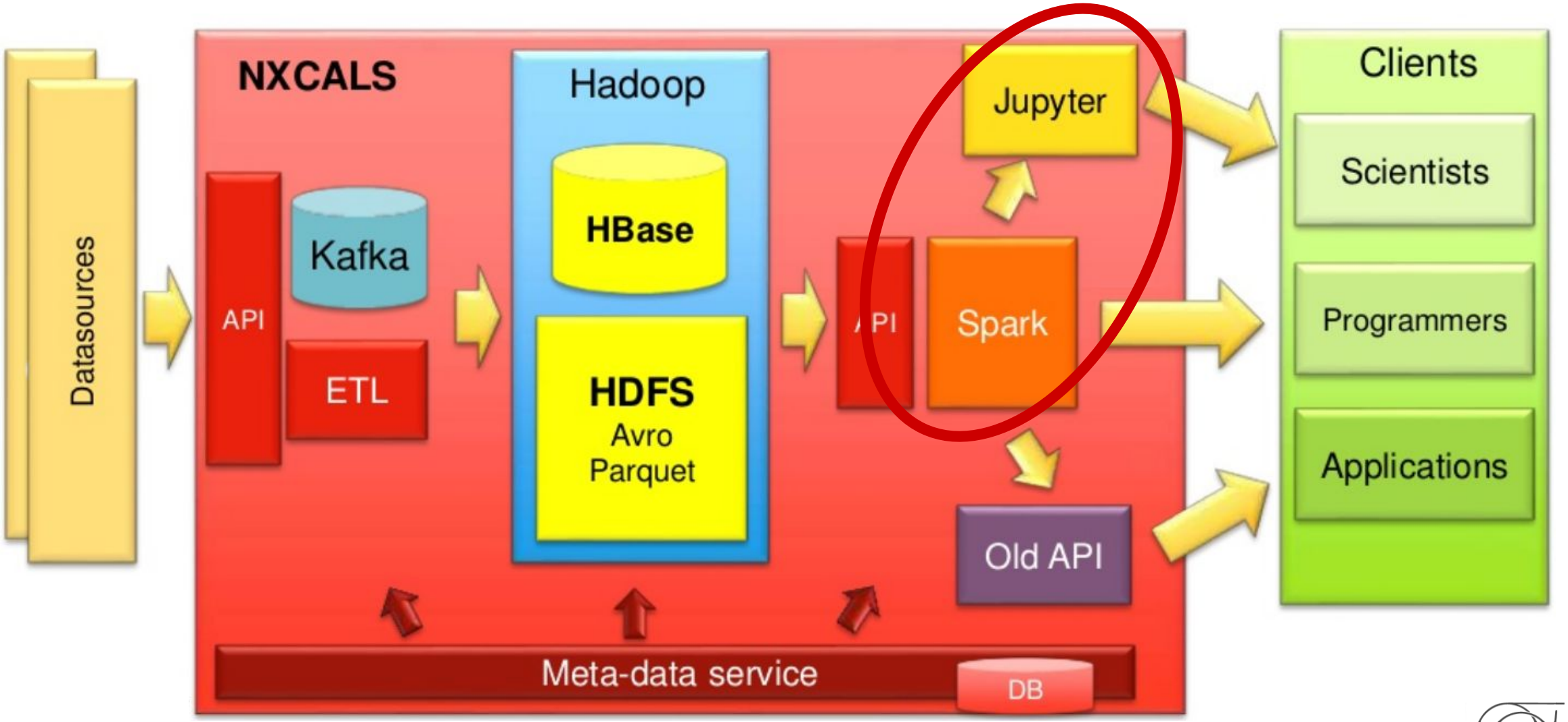
~ 300 PB

Service for Web based Analysis



Not Only HEP Analysis: LHC Controls Data

Not Only HEP Analysis: LHC Controls Data



Service for **W**eb based **AN**alysis: **SWAN**

ROOT

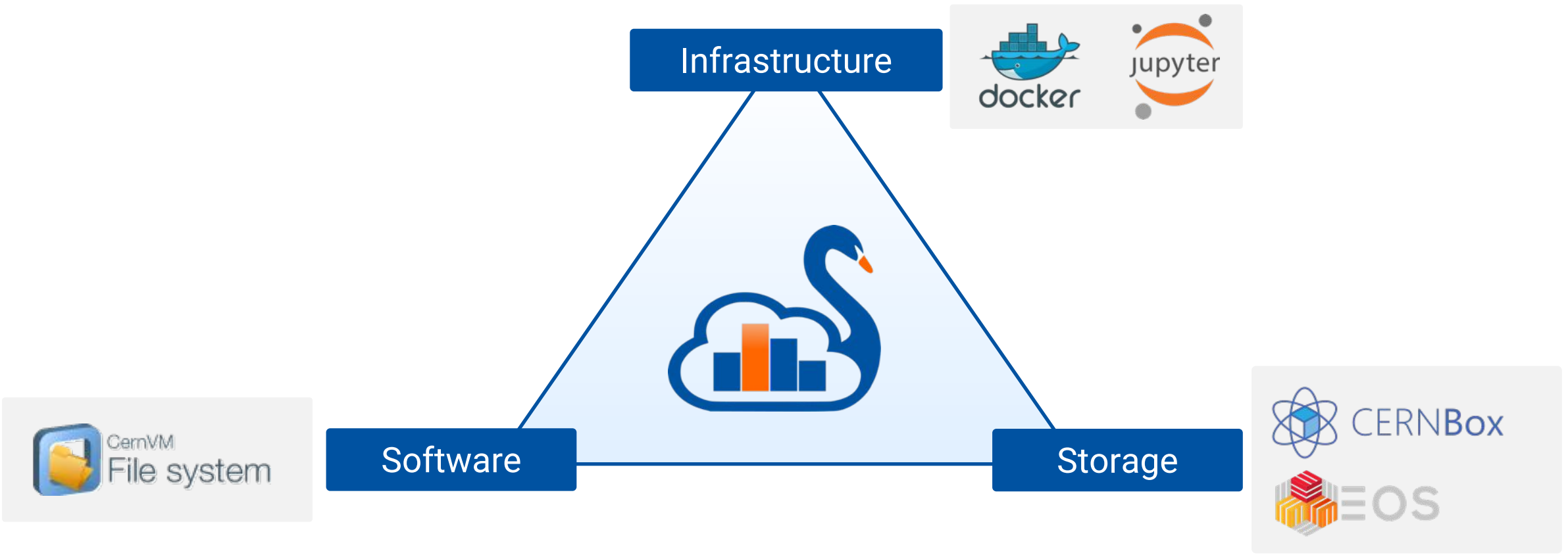
- **The** physics Data Analysis Framework
  - Created at CERN
- Used for data analysis, I/O, ML, advanced graphics…
- Integrated with Jupyter notebooks
  - **First C++ kernel based on ROOT C++**
- The origin story of SWAN
- http://root.cern

Service for Web based Analysis

- Analysis only with a web browser
  - **Blur boundaries between web and local applications**
  - **Available everywhere and at anytime**
  - Providing access also to Jupyter notebooks
- Easy to use (but powerful)
  - No local installation and configuration needed
- Create easily shareable scientific results: plots, data, code
  - Storage is crucial: mass & synchronized
- Integration with CERN resources
  - Access software, user/experiments data, mass processing power
- Synergy among analysis ecosystems : ROOT, R, Python, ...

https://en.wikipedia.org/wiki/Argentina

Integrating services



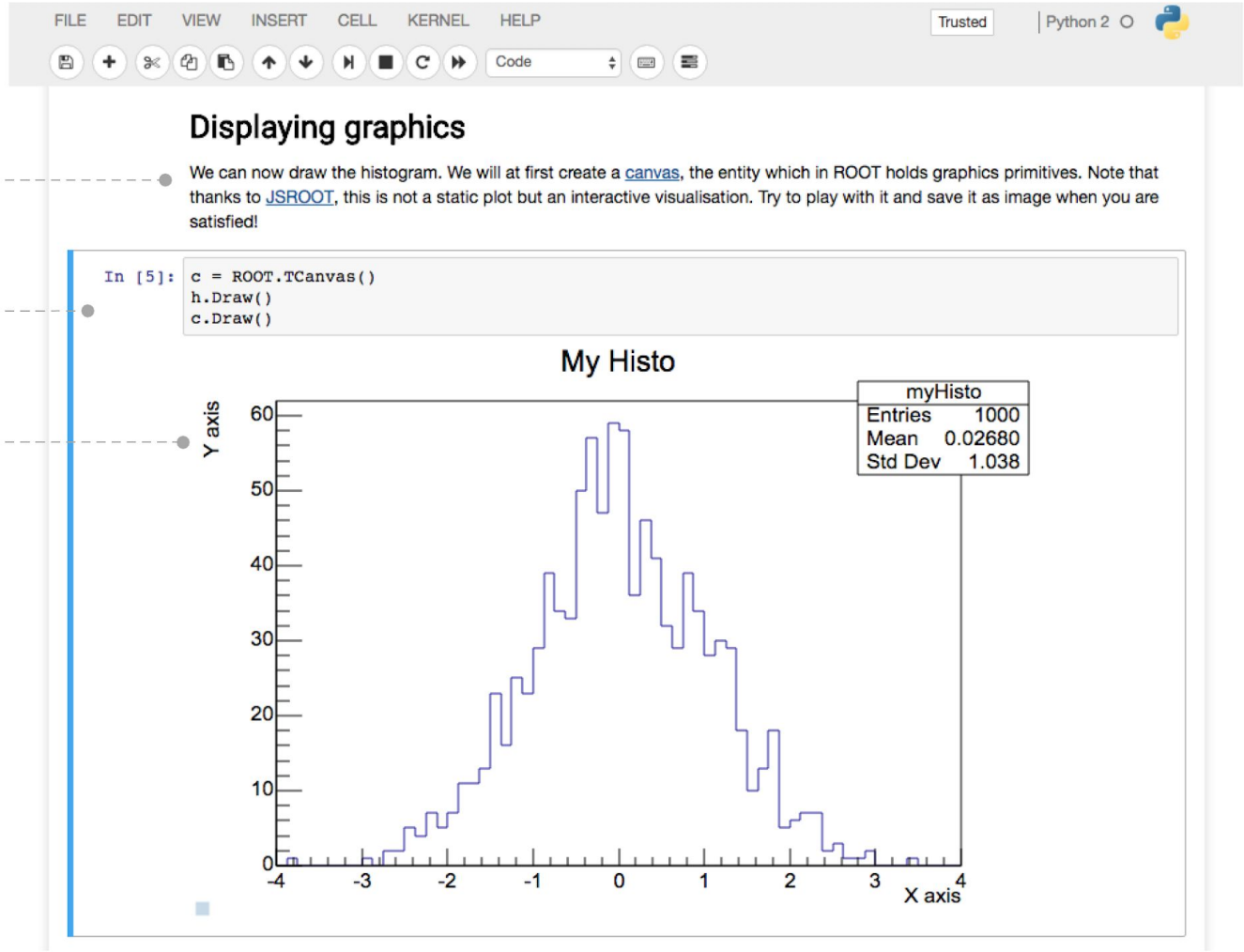
Open Source Technologies

Jupyter Notebooks

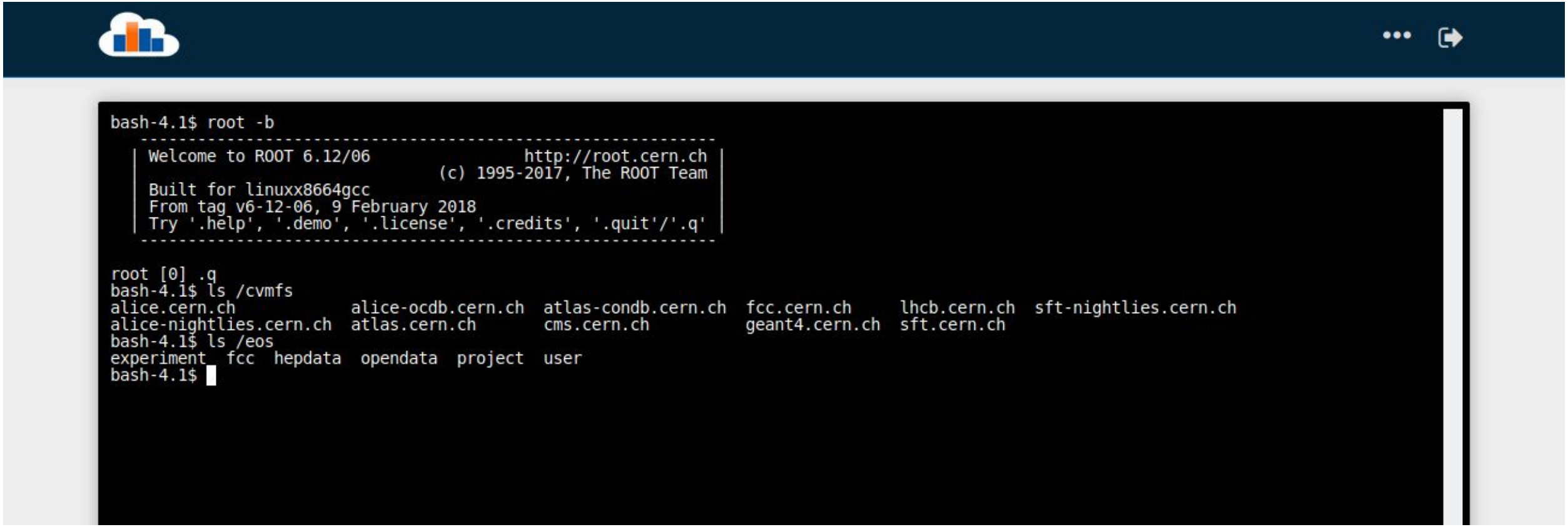
Text

Code

Graphics

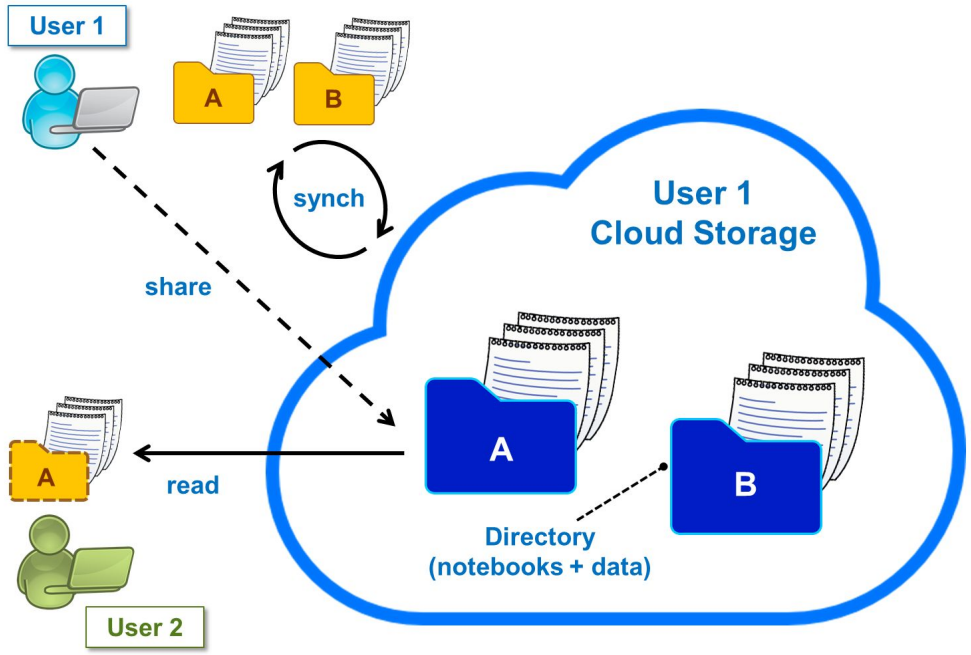


Does it Have a Terminal?



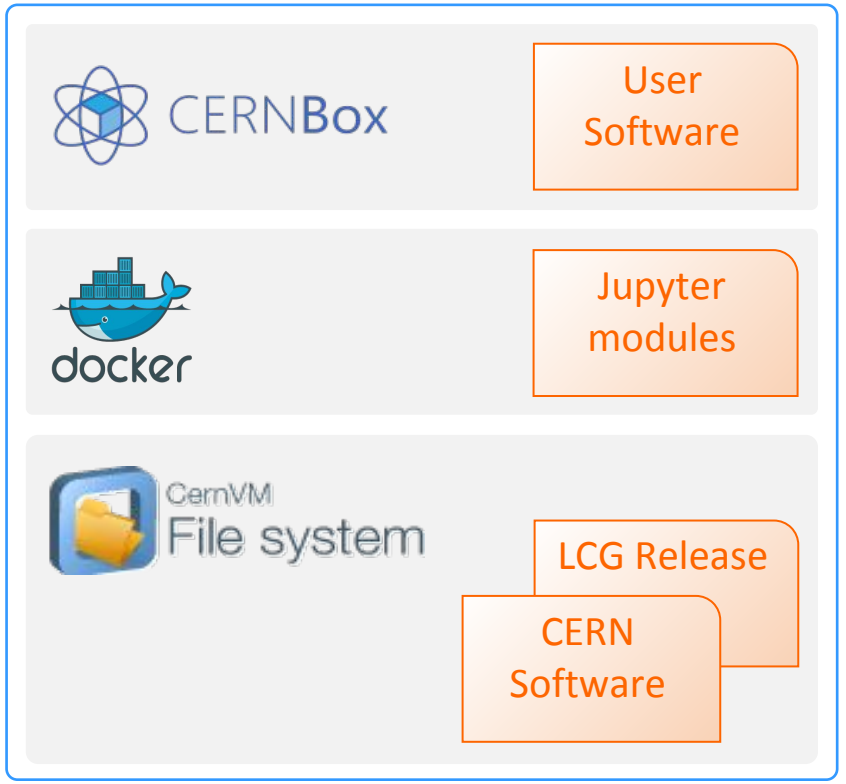
Storage

- Uses EOS disk storage system
  - All experiment data potentially available
- CERNBox is SWAN's home directory
  - Storage for your notebooks and data
- Sync&Share
  - Files synced across devices and the Cloud
  - Collaborative analysis

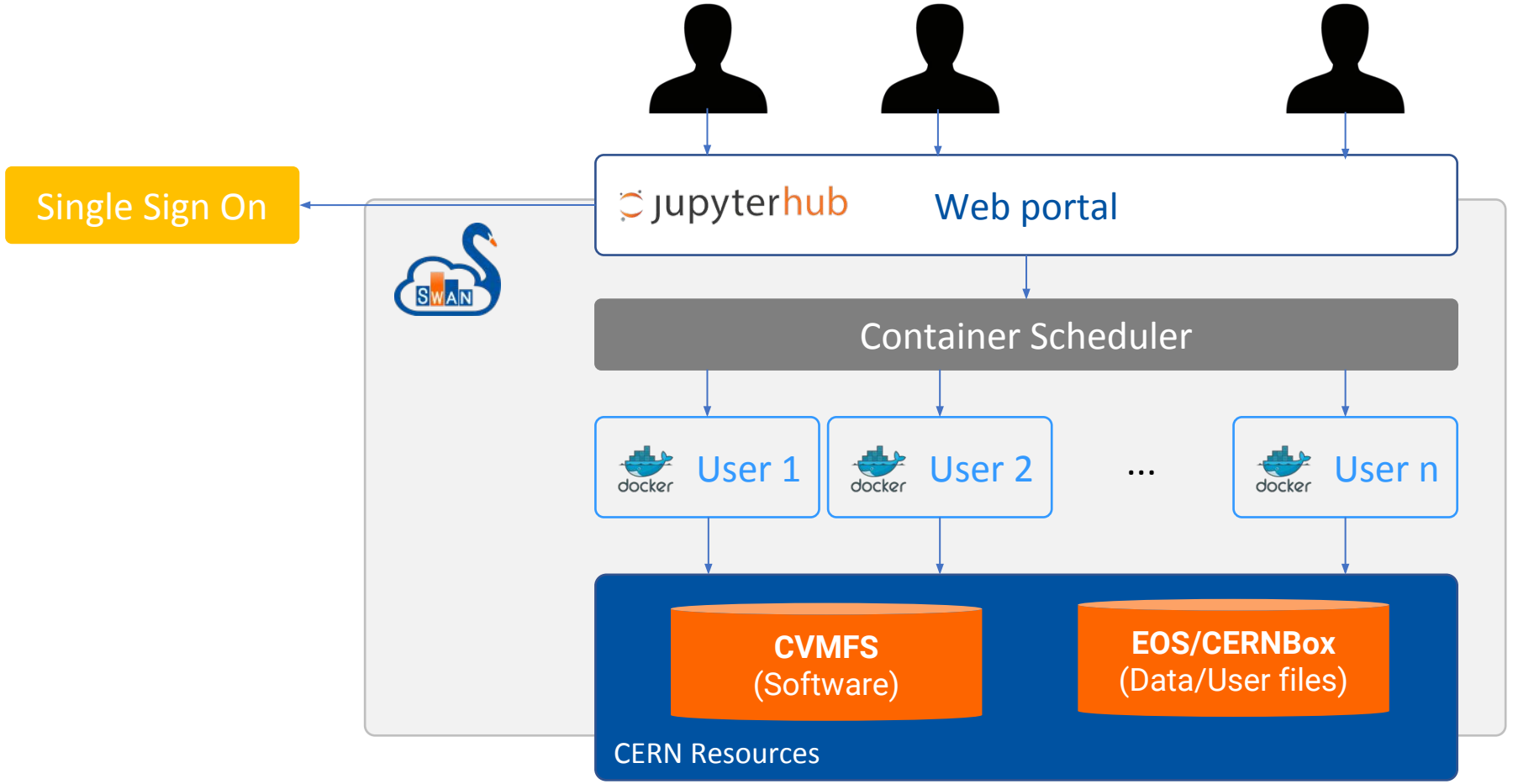


Software

- Software distributed through CVMFS
  - "LCG Releases" - pack a series of compatible packages
  - **Single Docker image, many software stacks**
  - Reduced Docker Image size
  - Lazy fetching of software
- Possibility to install libraries in user cloud storage
  - Good way to use custom/not mainstream packages
  - Configurable environment
- Multiple "kernels" ("links" between Jupyter and a language) available
  - Python 2 or 3, ROOT C++, R, Octave



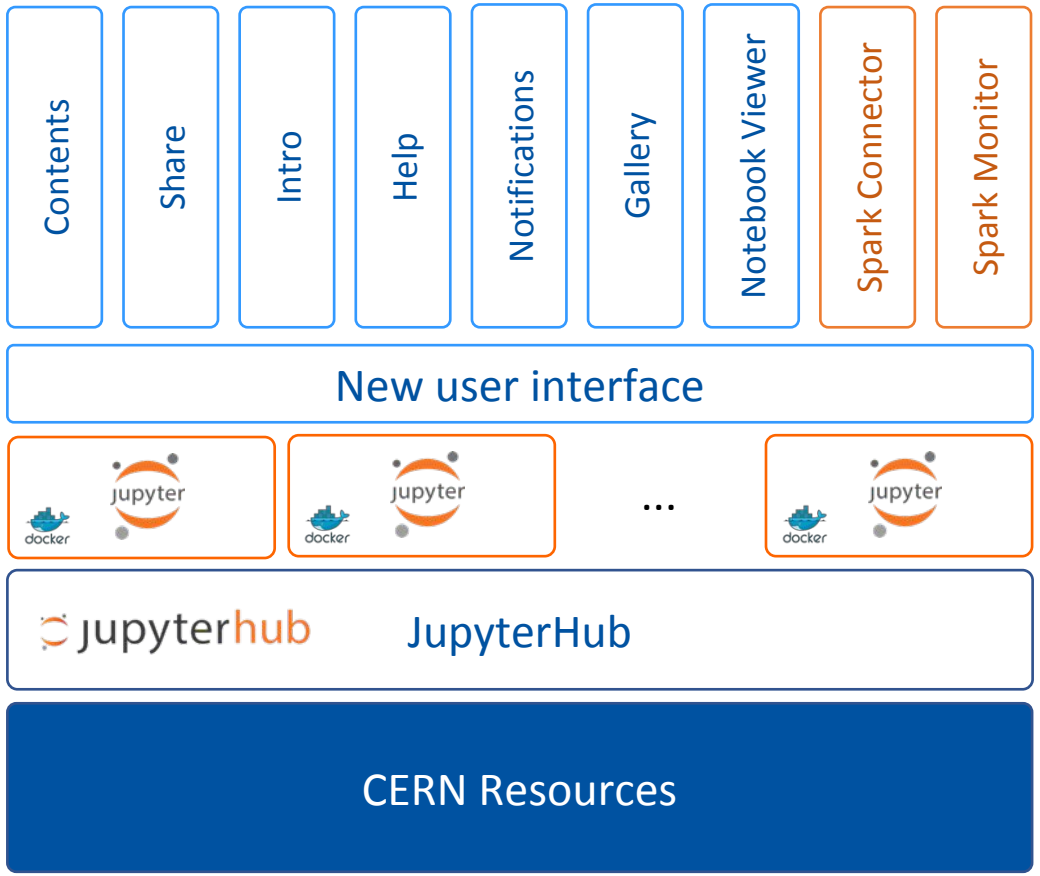
Architecture



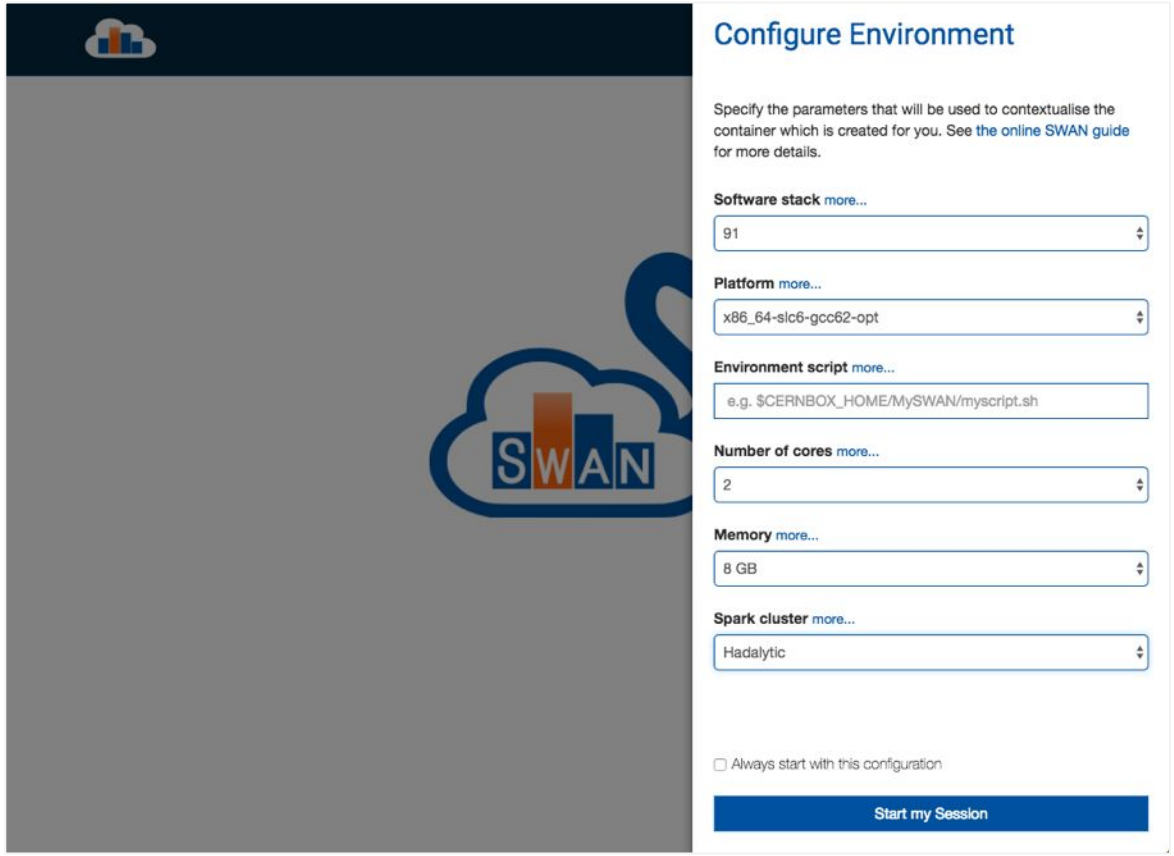
Leveraging the power of Jupyter

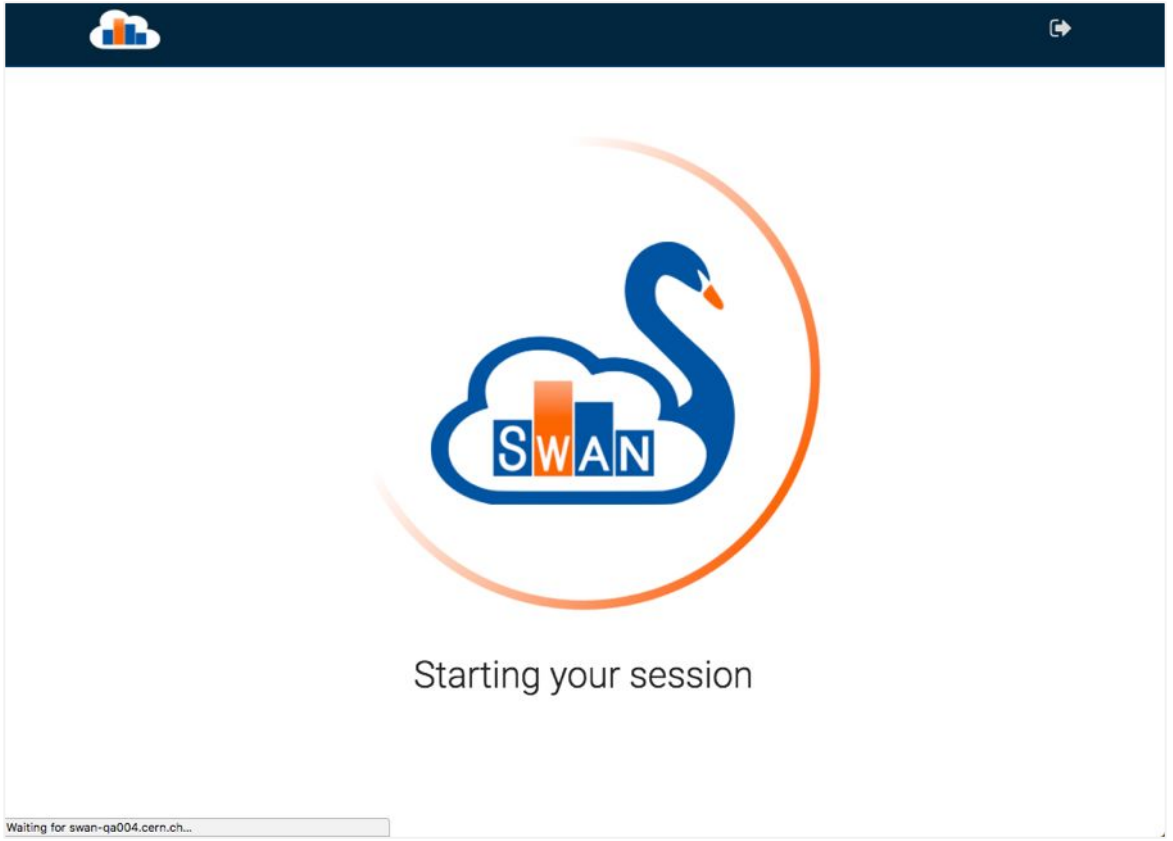
Service for Web based Analysis

Extensions, extensions everywhere

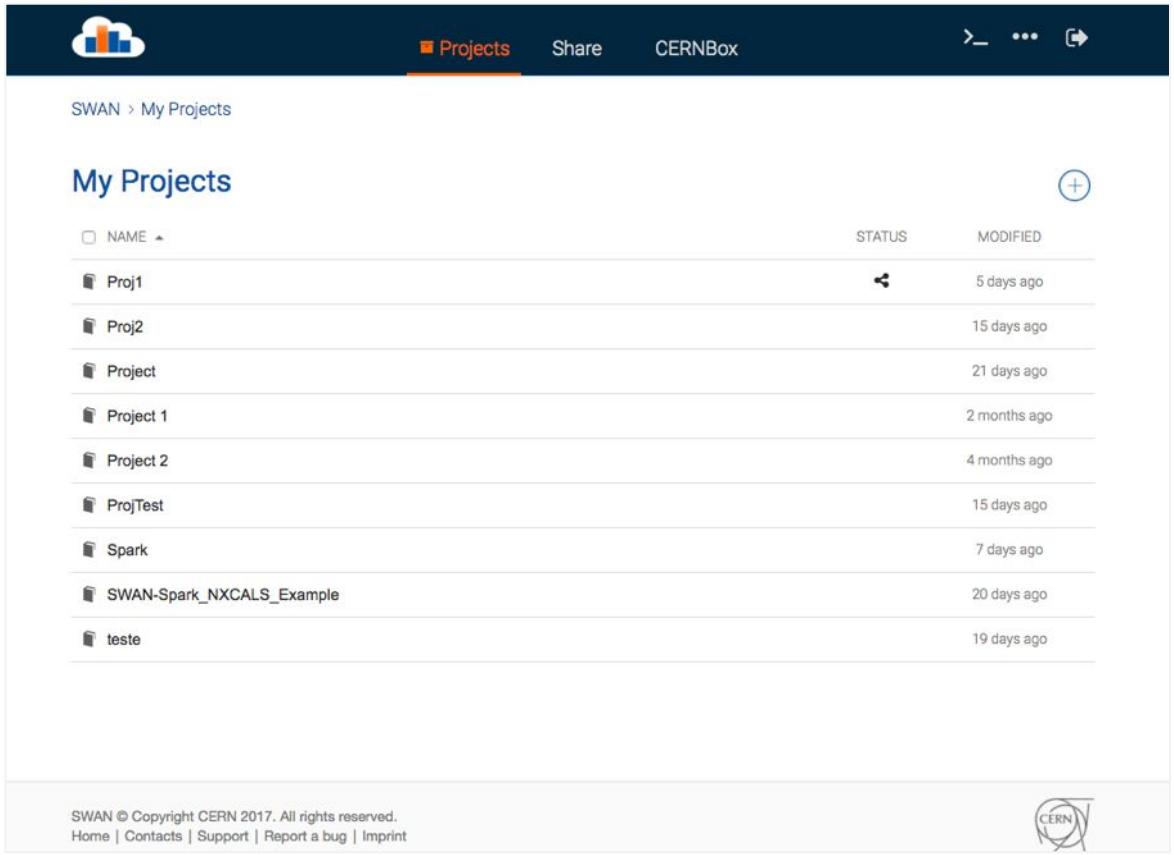


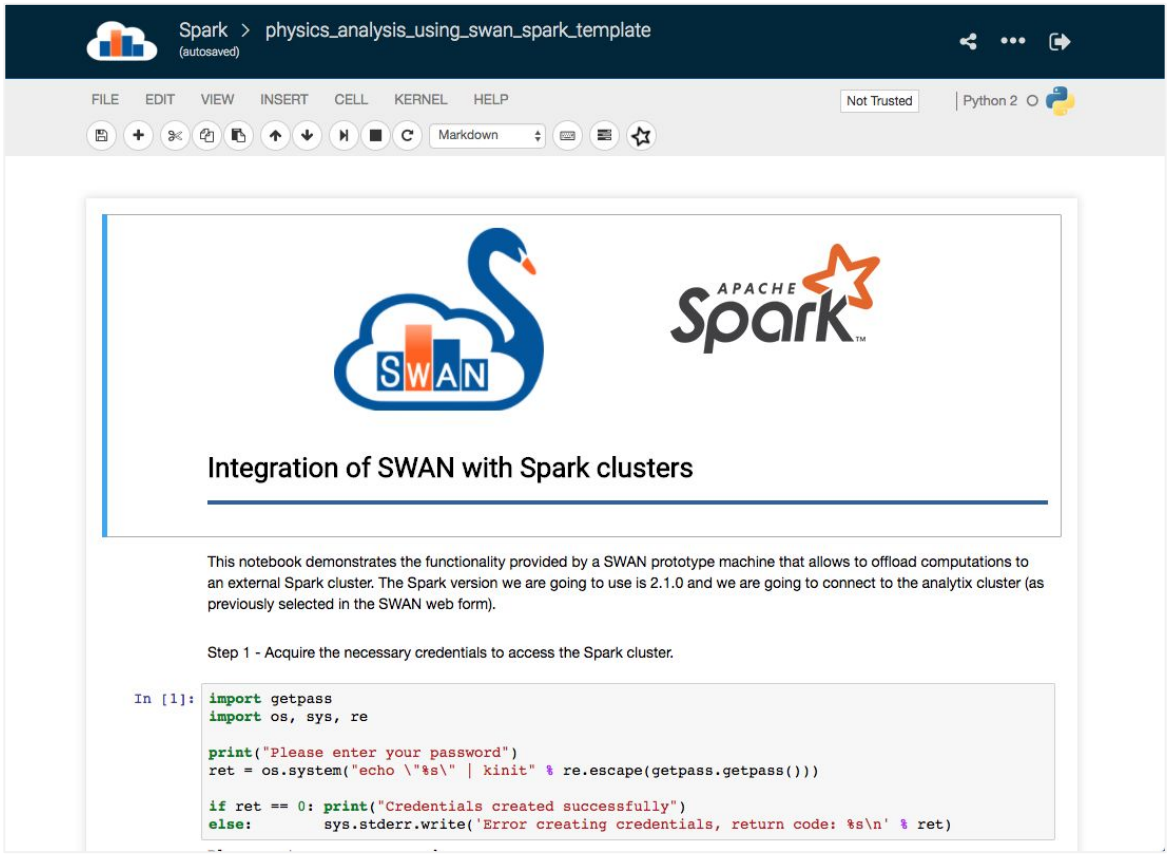
New User Interface





New User Interface



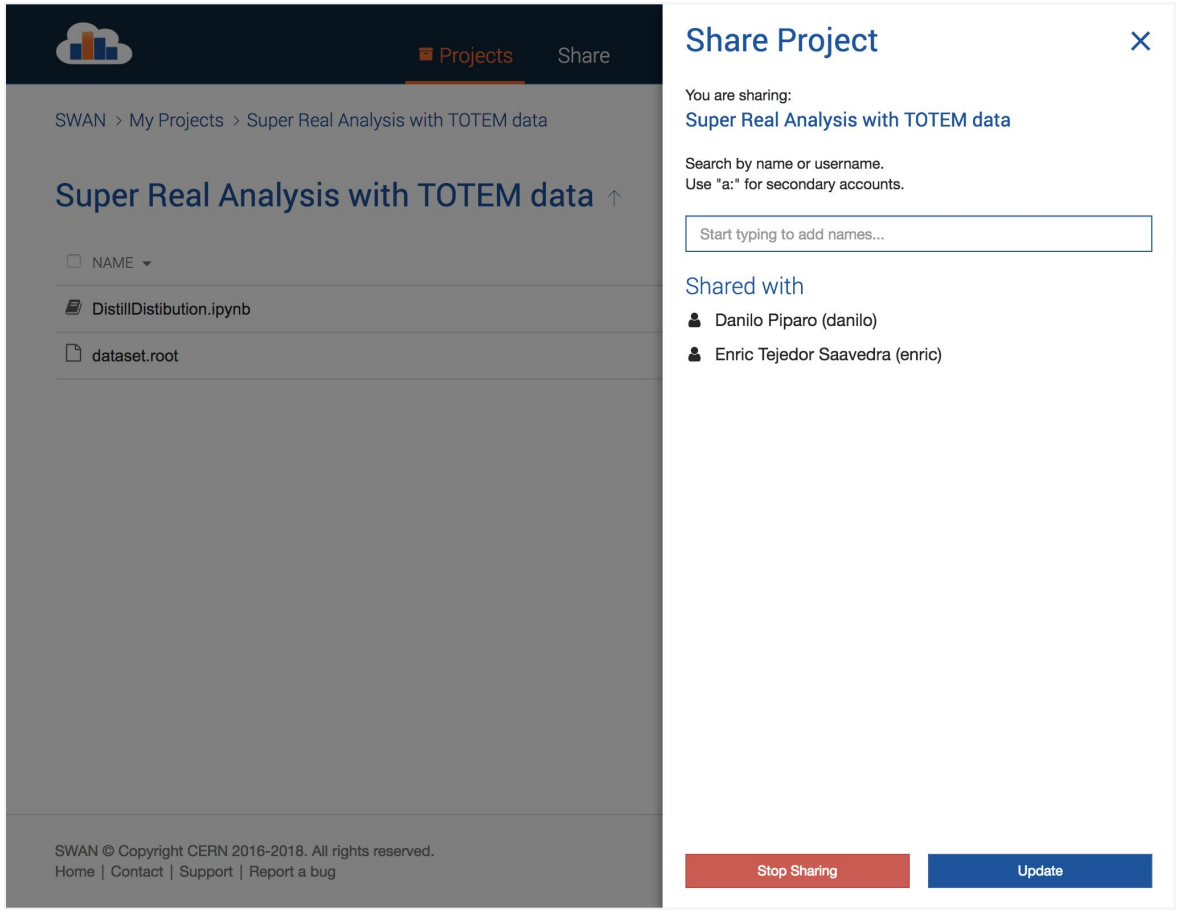


Collaborative Analysis

Service for Web based Analysis

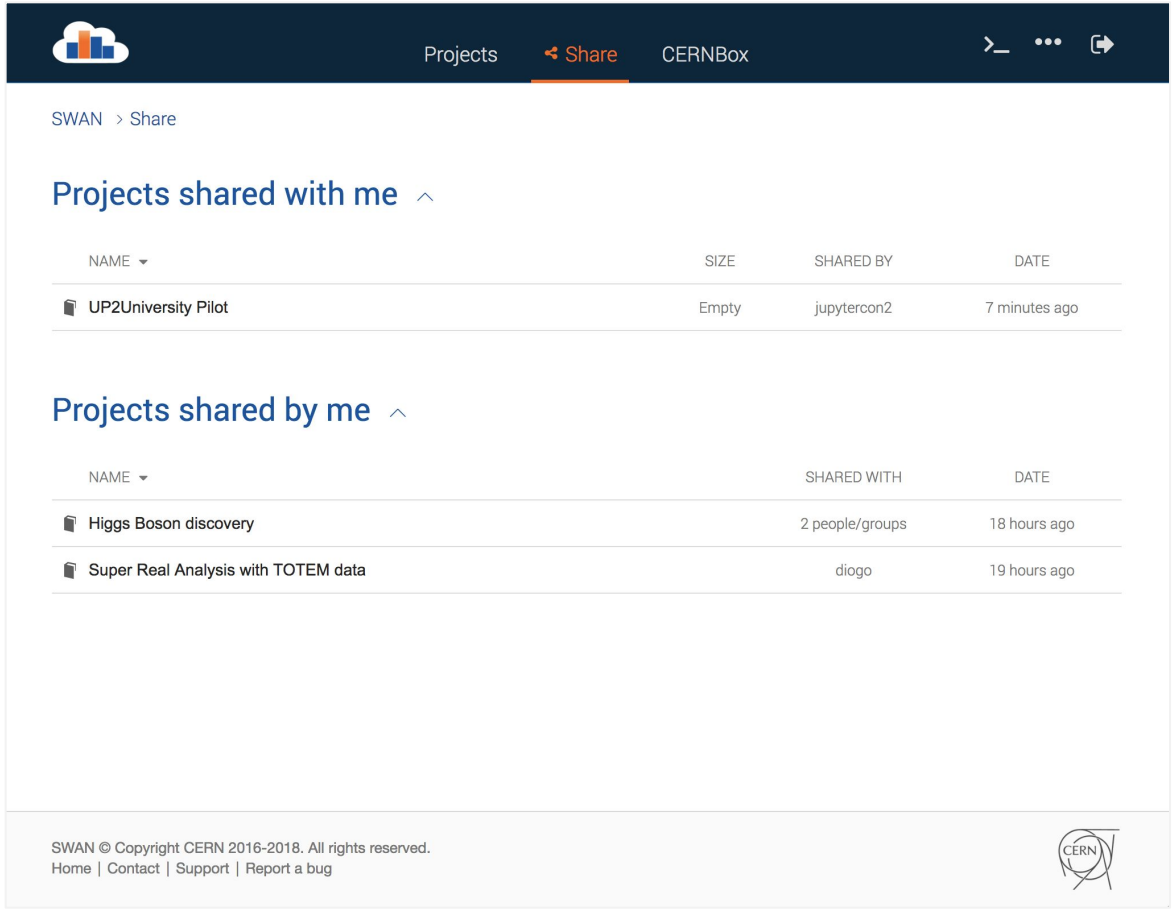
Sharing made easy

- Sharing from inside SWAN interface
  - Integration with CERNBox
- Users can share “Projects”
  - Special kind of folder that contains notebooks and other files, like input data
  - Self contained



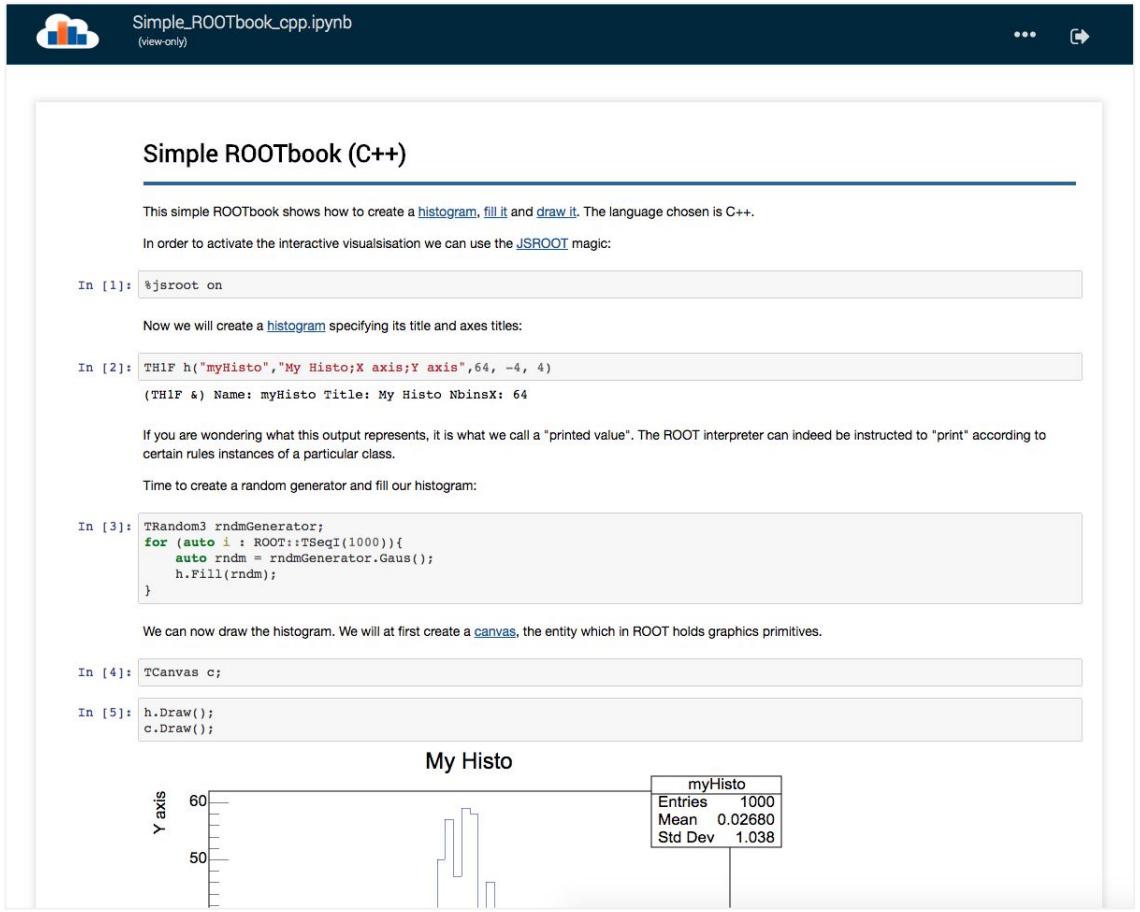
The Share tab

- Users can list which projects...
  - they have shared
  - others have shared with them



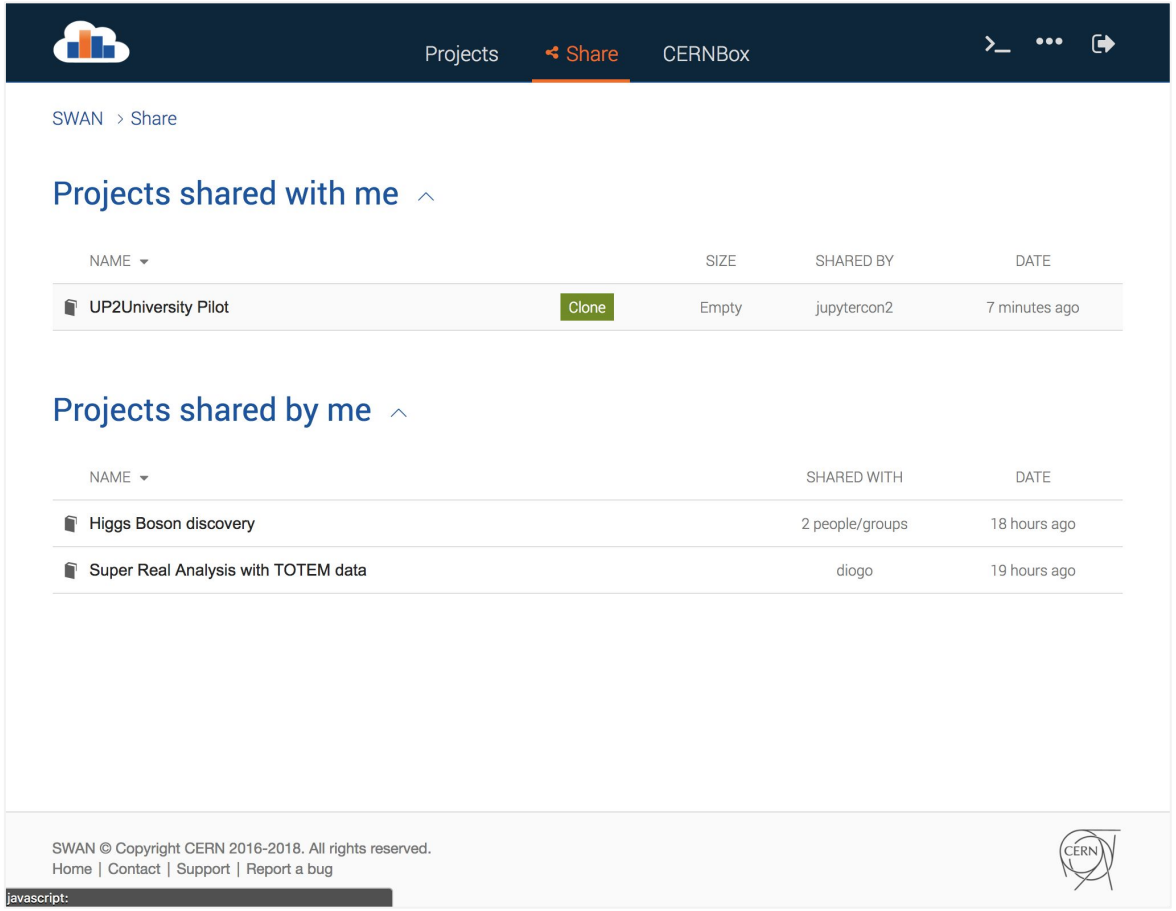
Inspecting a Project

- Users can inspect shared project contents
  - Browsing of the files
  - Static rendering of notebooks
- Useful to decide whether to accept or not the shared project



Accepting a Shared Project

- If accepted, project is cloned to the receiver's CERNBox
  - The receiver will work on his own copy
- Concurrent editing not (yet?) supported by Jupyter
  - Safer to clone



Access to Computing Resources

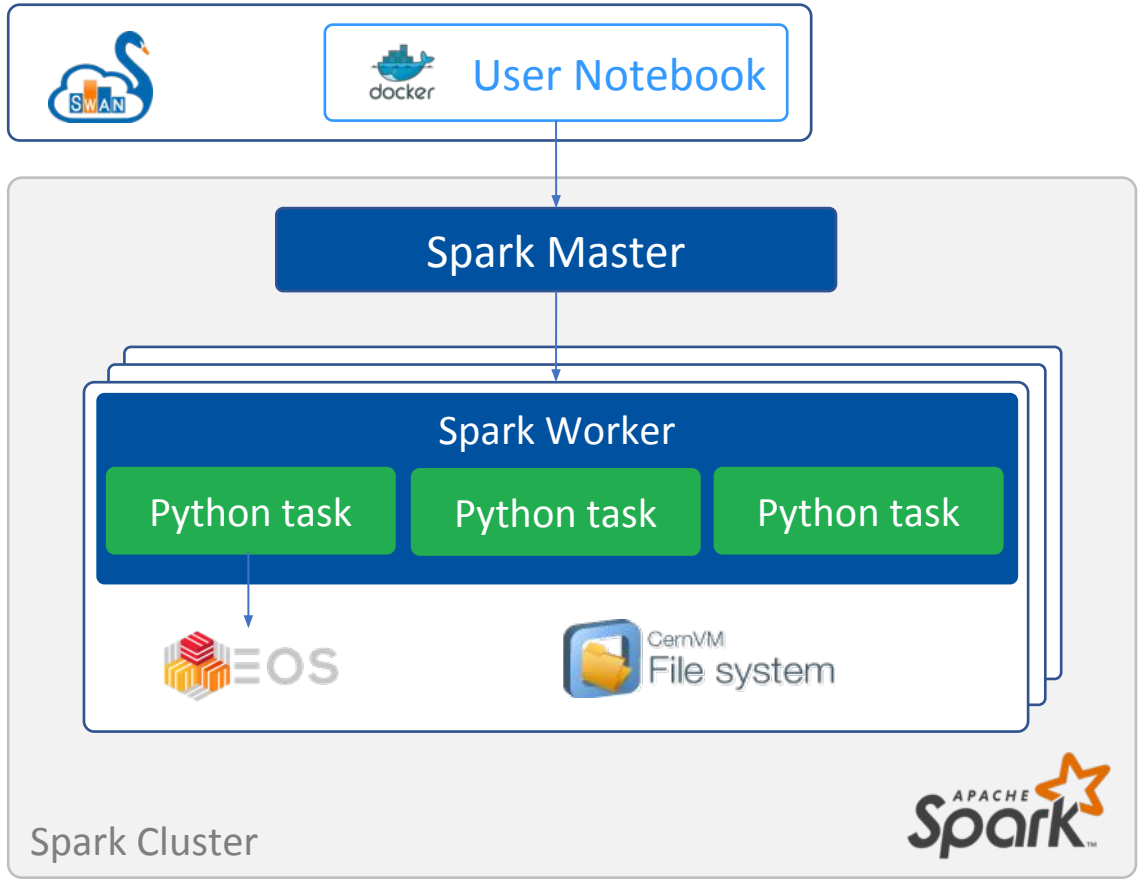
Spark integration

Motivation

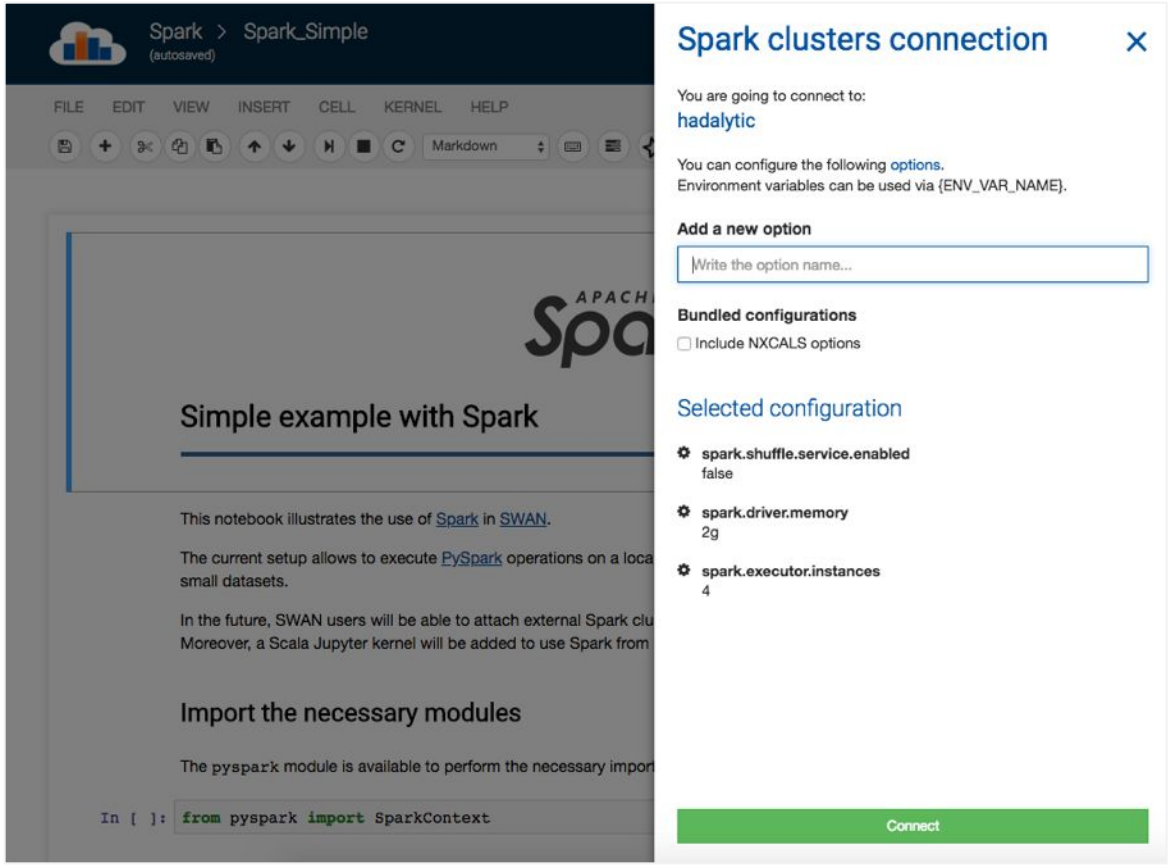
- Beams department at CERN built and operates the LHC
  - Data: monitoring LHC accelerator hardware devices
  - Analysis of logs
  - Different from physics data of the experiments
- With increase in data, they adopted Spark
  - Spark: a modern cluster management system allowing distributed and interactive data analysis
  - But it was missing a unified platform for analytics visualization

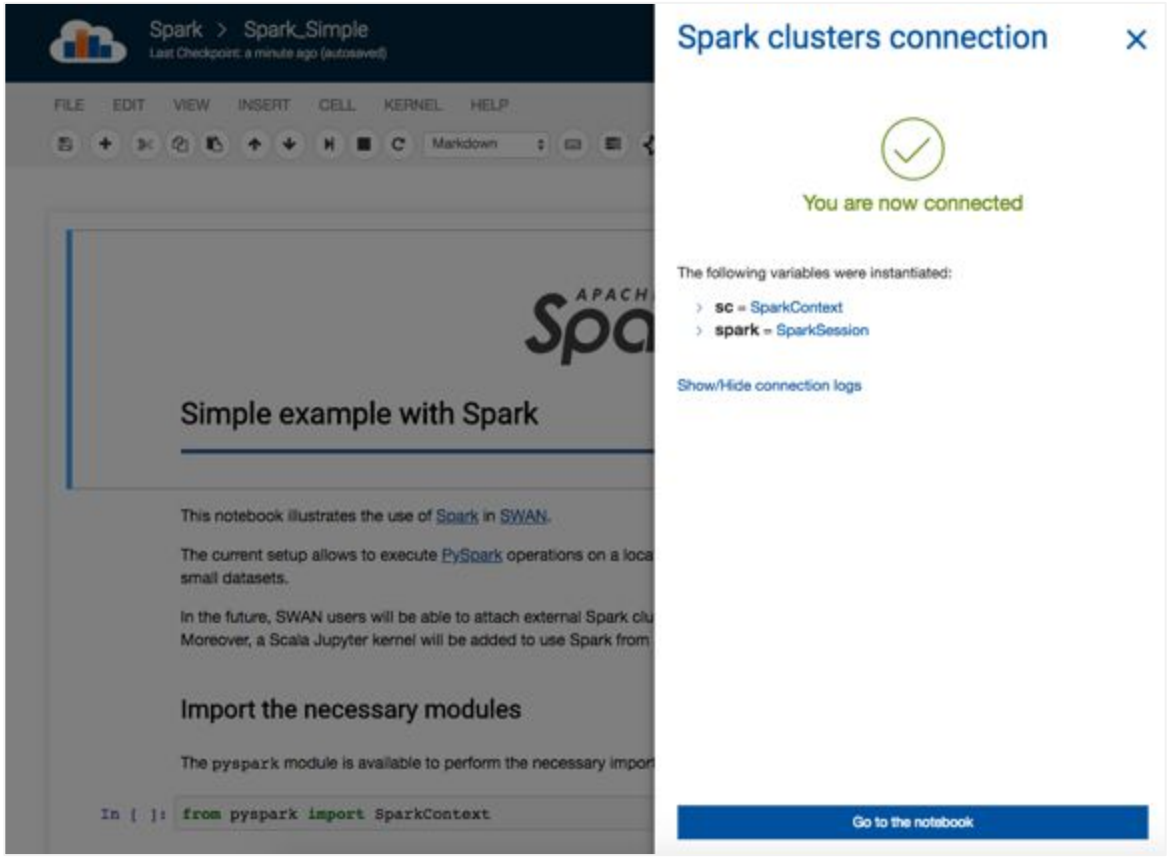
Integration with Spark

- Connection to CERN Spark Clusters
- Same environment across platforms
  - User data - EOS
  - Software - CVMFS
- Graphical Jupyter extensions developed
  - Spark Connector
  - Spark Monitor



Spark Connector





Spark Monitor

- Allows to monitor Spark jobs from within the notebook
  - Interactive display appears **automatically** in the output of the cell

```
In [ ]: def f(x):
            global a
            a+=x
        RDD9.foreach(f)
        RDD9.foreach(f)
        print(a.value)
        #Display should appear automatically
```

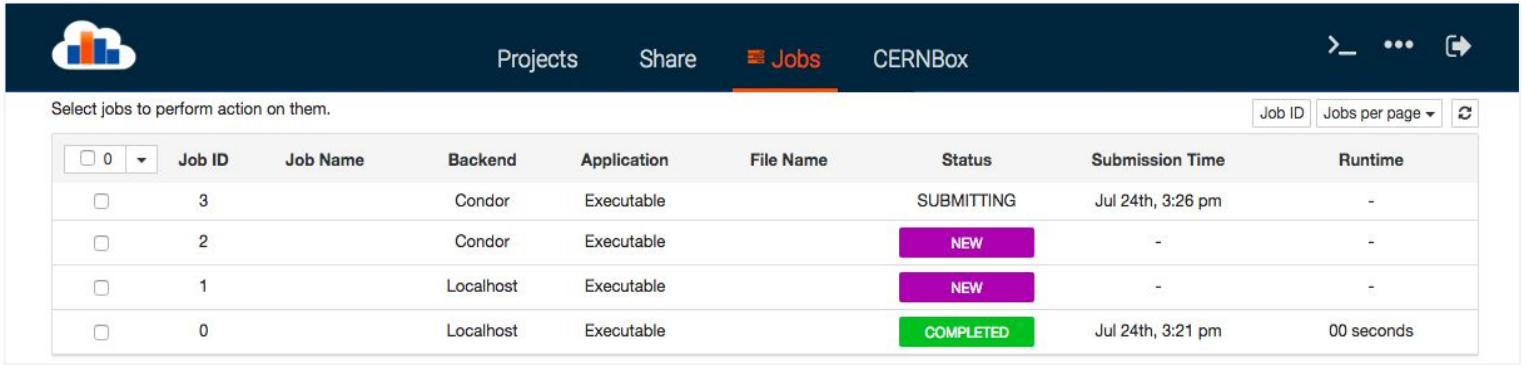
- Multiple views
  - Progress bars for jobs and stages
  - Task timeline
  - Resource utilization

Access to Computing Resources

Batch jobs

Connecting More Resources

- Ongoing effort: submit **batch jobs** from the notebook
  - Monitoring display
  - Jobs tab




```
In [6]: %%ganga
        j=ganga.Job()
        j.submit()
```

Backend: LOCALHOST Application: EXECUTABLE Splitter: None 0 SUBJOBS

Job ID	Job Name	Status	Subjobs	Submission Time	Runtime
0		COMPLETED	No Subjobs	Jul 24th, 3:21 pm	00 seconds

Education, Outreach

Notebooks for Teaching

- SWAN and notebooks are very useful for teaching
  - Teachers prepare exercises in the form of notebooks
  - Students complete and run the exercises
- Used frequently as support for:
  - CERN Summer student courses: ~150 students, data analysis with ROOT
  - CERN School of Computing exercises: ~70 students, parallelism
  - CERN ATLAS PhD student courses: ~50 students, declarative data analysis
  - Machine learning tutorials at CERN

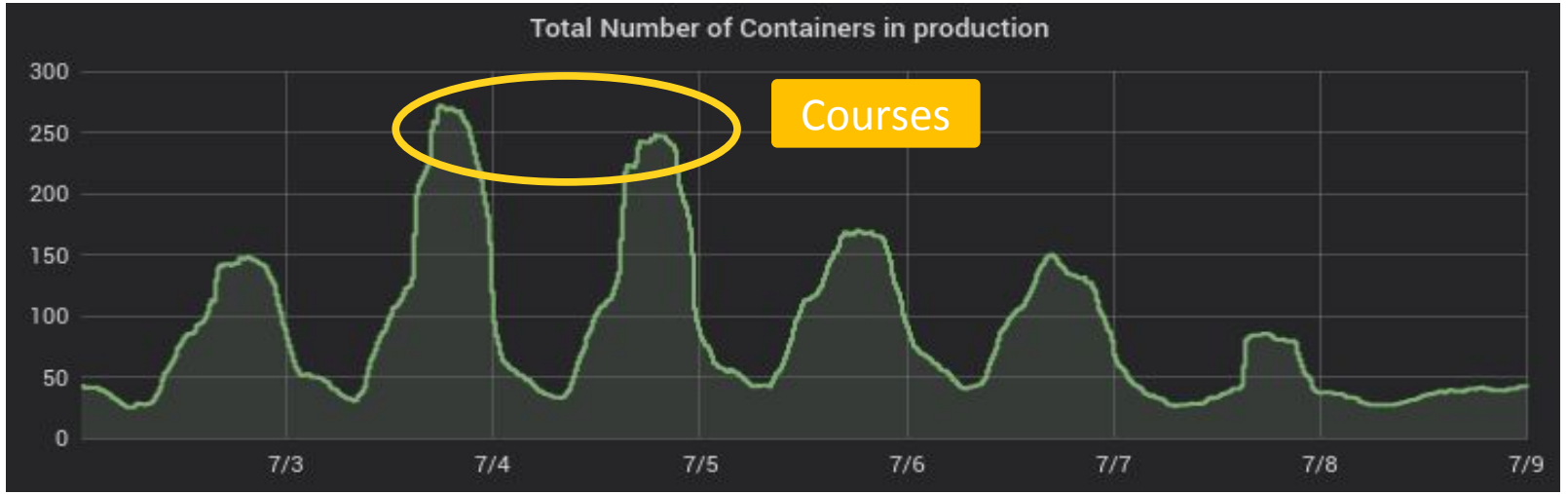
Science Box: SWAN on Premises

- UP2University European Project
  - Bridge the gap between secondary schools, higher education and the research domain
  - Partner universities (OU, UROMA, NTUA, …), pilot schools
  - http://up2university.eu
- SWAN used by students to learn physics and other sciences
  - Let them use the very same tools & services used by scientists at CERN
  - Integrating Jupyter in teaching environment
  - Complete SWAN package, with CERNBox, EOS and CVMFS

SWAN Users

Stats

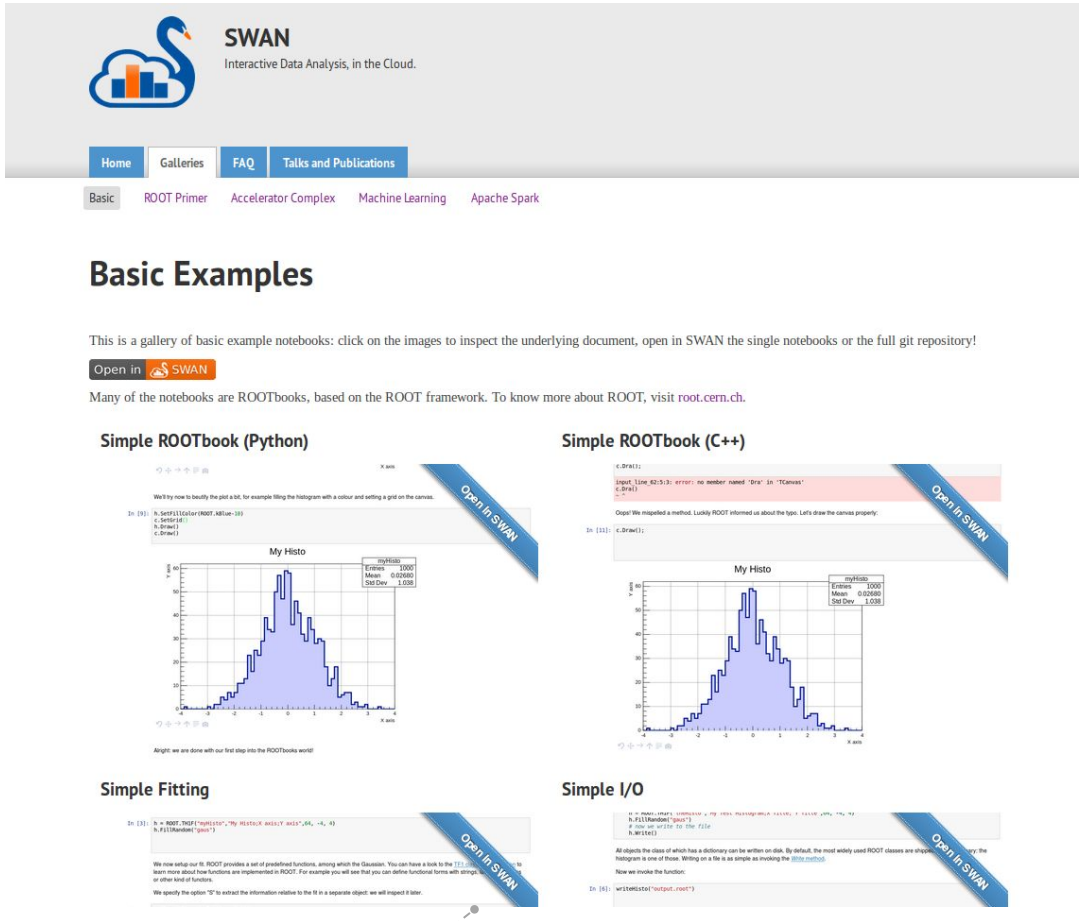
- **~200 user sessions** a day on average
  - Users doubled this year with new SWAN interface
- Spark cluster connection: 15 – 20 % of users
  - SWAN as entry point for accessing computational resources



User community

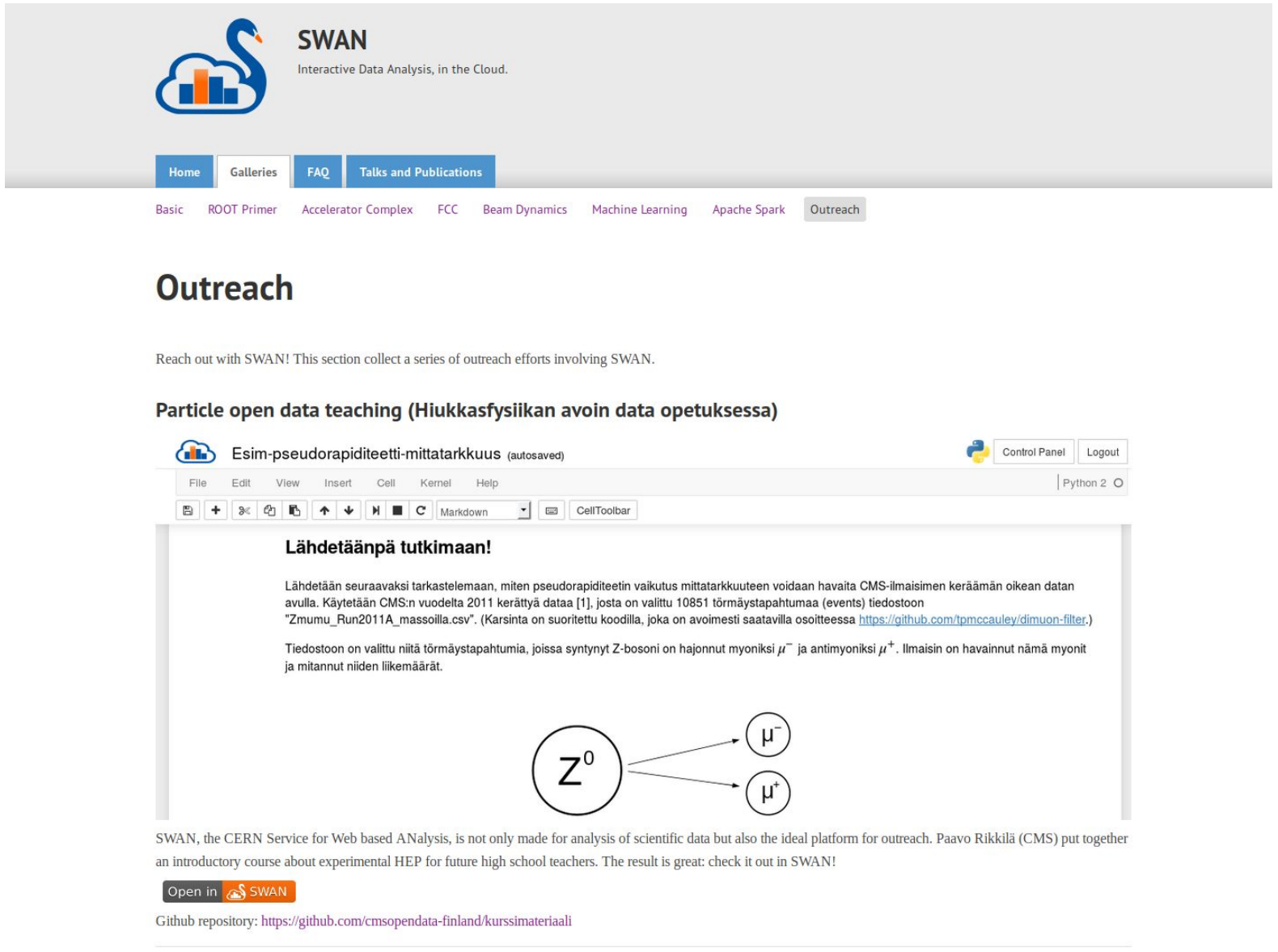
- SWAN development is guided by our user community
  - New features (libs, kernels, ...) are requested by users from their real usage needs
- Gallery of examples
  - Made in collaboration with our users
  - Almost 50 notebooks in 7 categories

Example notebooks at **cern.ch/swan**



Access with only a click

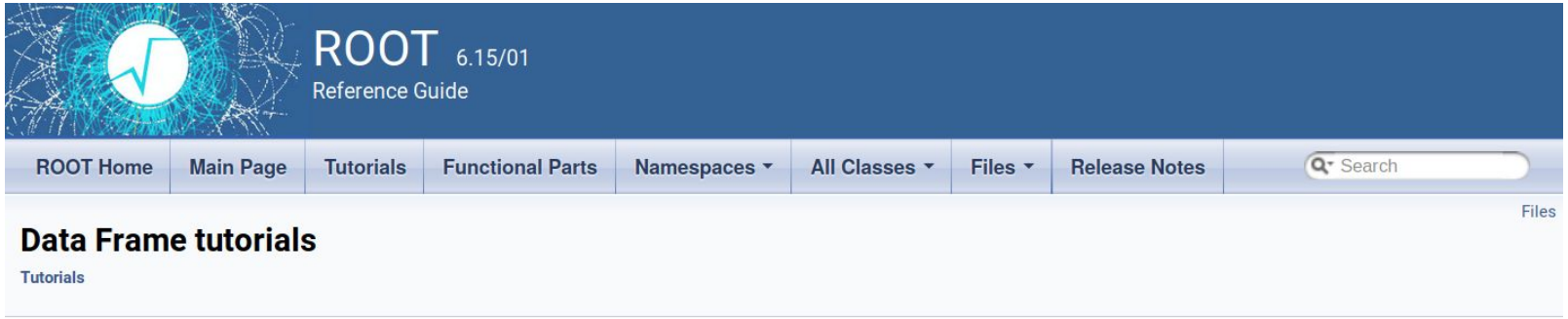
Education, Too



Example notebooks at **cern.ch/swan**

- Finnish High School program

Integrate Software Packages



Example notebooks at **root.cern**

Looking ahead

Future work/challenges

- Move to Jupyterlab
  - Porting the current extensions
  - ROOT is already migrating its graphics to JavaScript only
  - **Ultimate blur between local/cloud resources and apps**
- New architecture made to scale
  - Based on Kubernetes, leveraging CERN Container Service
- Exploitation of GPUs
  - ML found its place in HEP
  - Speed up computation of GPU-ready libraries (e.g. TensorFlow)
- Moving beyond LHC
  - New challenges from HL-LHC
  - Even small dataset will not fit inside users laptops: need R&D - Analysis Facilities?

Where to find us

Where to find us

- Contacts
  - swan-talk@cern.ch
  - http://cern.ch/swan
- Repository
  - https://github.com/swan-cern/
- Science Box: SWAN distribution **deployable on premises**
  - https://cern.ch/sciencebox
  - Includes all SWAN components: CERNBox/EOS, CVMFS, JupyterHub
  - Deployable through Kubernetes or docker-compose

CERN

Conclusions

Conclusion

- We successfully integrated CERN computing services with Jupyter
  - Increase value of existing services' portfolio (storage, batch, auth)
  - Successfully boosting productivity of scientists
  - Jupyter became a new entry-point to our computing resources and data
- Successful model for teaching activities
  - Tutorials for software, data analysis, ML, Spark…
  - European project focused on higher education
- Blending desktop and web application can help reach more users
  - Can Jupyterlab help?