CS3 MESH4 EOSC

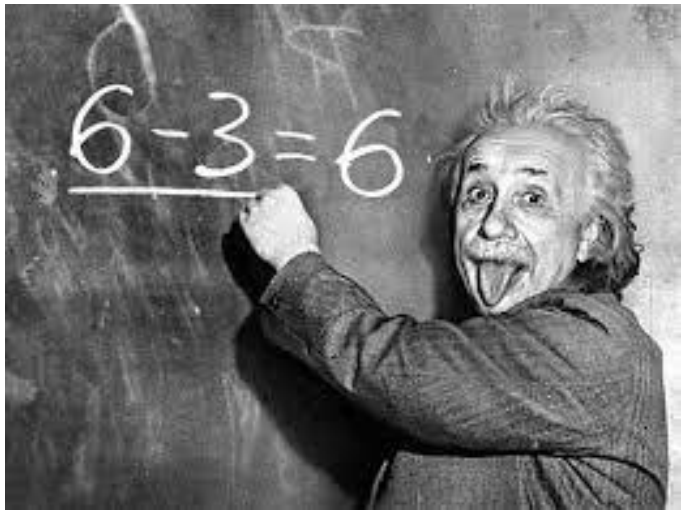
Connecting European Data

Collaborative data science: High Energy Physics and beyond

Marcin Sieprawski
Head of Big Data Lab, Software Mind

CS3MESH4EOSC has received funding from the European Union's Horizon 2020 Research and Innovation programme under **Grant Agreement No. 863353**.

Objectives/Agenda: High Energy Physics use case



- Not a High Energy Physics talk
  - only a part of the picture of data analysis in HEP (last stage of analysis)

- HEP as an excellent use case for ScienceMesh distributed Data Science environments
  - Present the tools we provide to facilitate HEP research
  - Why it is relevant in a wider perspective
    - can be transferred to other sciences and business
  - Get feedback

High Energy Physics use case – intro

- Using particle accelerators to smash particles together at high speeds in the search for new particles
  - Identifying events of interest
  - Analysing data streams from detectors
- The Large Hadron Collider (LHC)
  - analysing the myriad of particles
- These experiments are run by collaborations of scientists from institutes all over the world

Image: Simulation of a particle shower resulting from a particle collision.
© CERN

HEP: data challenges, collaborative data science

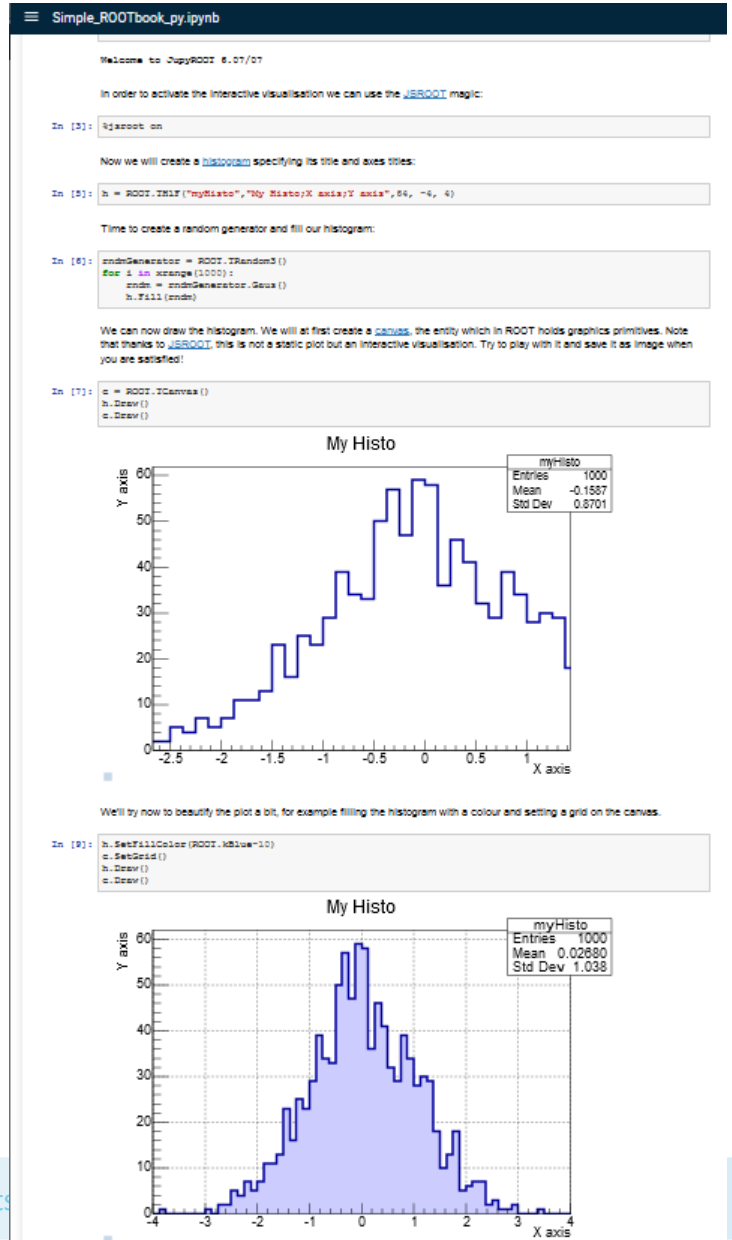
- LHC produces unprecedented volumes of data
  - Raw stream from detectors: **600TB per second**, or **50 000 PB per day**
    - The raw data per event ~1MB, 600 million events per second
  - Storing only fraction of this (total volume of **all stored data: 350PB**)
    - Filtering of events/data need to be smart and fast!
  - Large distributed infrastructure for transferring and large-scale processing
- Constant innovation in tools and methods for analytics.
  - Data streams from LHC increase with each upgrade
- Distributed teams of scientists from institutes all over the world
  - a variety of storage systems and processing tools.
- **One of the challenges: providing tools in this distributed environment for effective collaboration in Data Science**

Data Science

- In main stream since 2012: Harvard Business Review
  - „Data Scientist: The Sexiest Job of the 21st Century”
- In business: using data to increase competitive advantage
- Data analytics approaches
  - Data Warehouse, Business Intelligence, Data Mining …
- Origin of Data Science:
  - using methods „always” used in science to analyse data
  - Most of those methods were invented for physics
  - HEP – pionieers in using many approaches (ag. Machine Learning)

Jupyter Notebook

- A free, open-source, interactive web tool: a computational notebook
  - combine software code, computational output, explanatory text and multimedia resources in a single document
  - rapid uptake, an enthusiastic community of user–developers
  - Python / R / more
  - 10 milion public Jupyter Notebooks on GitHub (December 2020)
    - 2.5M in September 2018, 200k in 2015
- “de facto” platform used by data scientists to build interactive applications and to tackle big data and AI problems
- Replacing Business Intelligence tools



Jupyter Notebook: SWAN service @ CERN

UI/Core

Software

Analysis platforms

Storage

Compute

Infrastructure

1

Jupyter Notebook in HEP: SWAN galery

https://swan-gallery.web.cern.ch/

Gallery

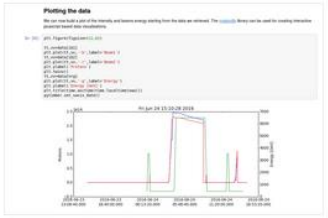
- Basic Examples
- ROOT Primer
- Accelerator Complex
- Beam Dynamics
- Machine Learning
- Apache Spark
- Outreach
- AWAKE

Accelerator Complex

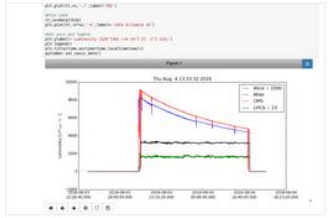
This gallery shows examples of machine studies relative to the CERN accelerators' complex.

Open in SWAN

LHC Page1

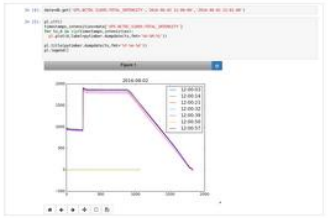


Experiments' Luminosities

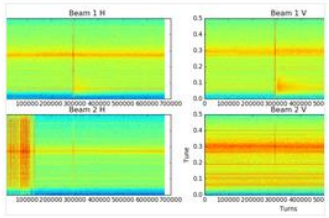


PyTimber Tutorial

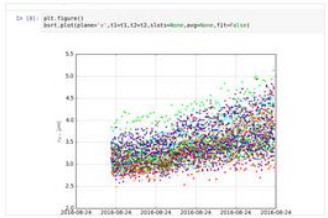
SPS Intensity



LHC BBQ Example



BSRT Example

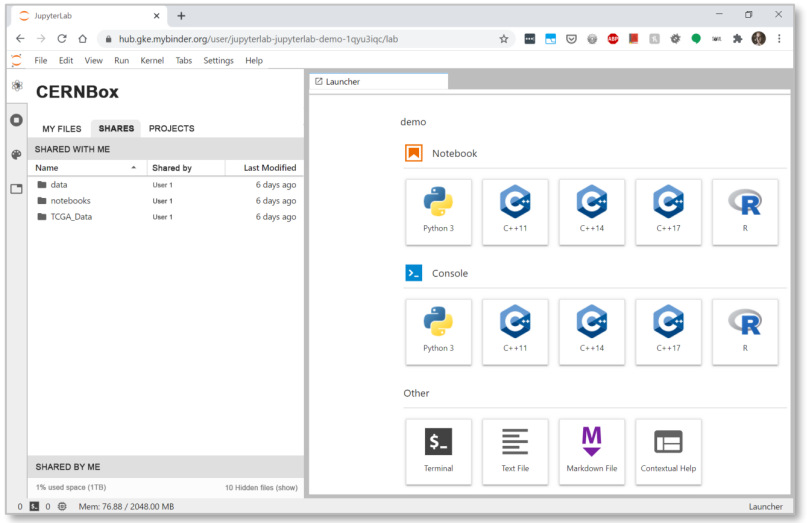


Previous
ROOT Primer

Next
Beam Dynamics

JupyterLab extension (**Cs3Api4Lab**)

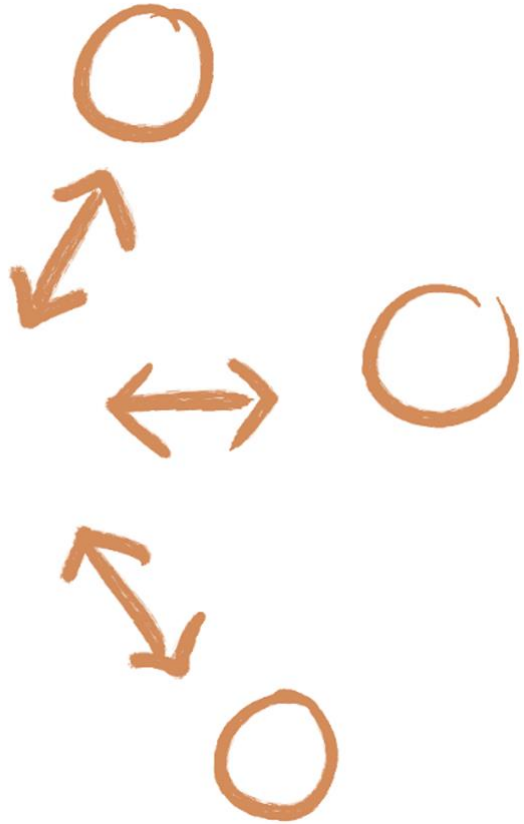
- Integration with ScienceMesh IOP (CS3 APIS)



jupyterlab

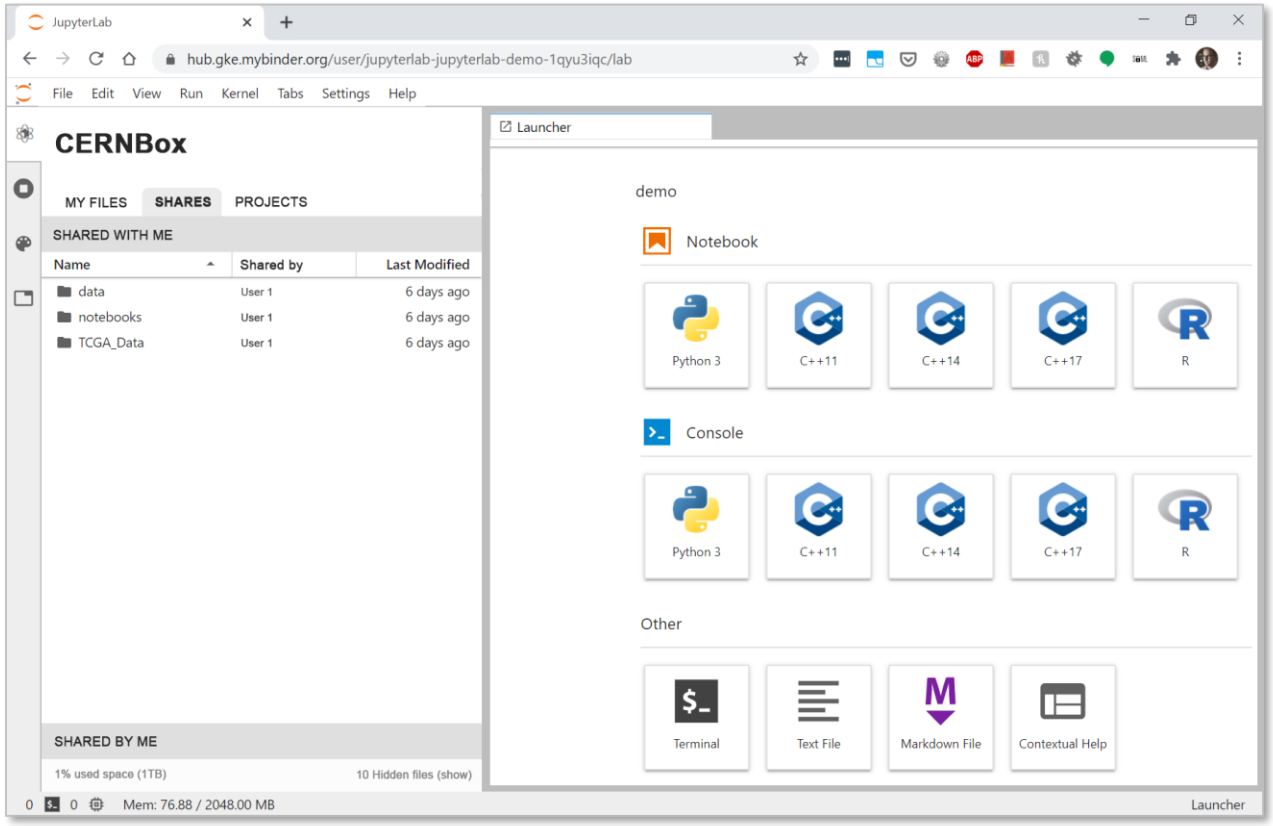
CS3 APIs

Science Mesh
Inter-operability Platform



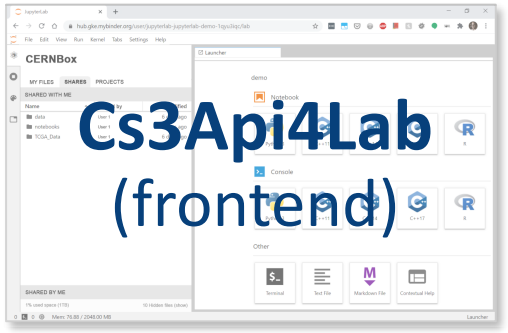
JupyterLab extension (**Cs3Api4Lab**): Frontend

- Full client in Lab
- File browser – share functionalities
  - Shared by/with tab
  - Sharing buttons
  - Entries in the context menus
  - Pop-up windows: file information and sharing status
  - Account info
- File browsing



JupyterLab extension (**Cs3Api4Lab**): Backend

- Replaces ContentsManager and Checkpoints
- REST endpoints for integration with the frontend:
  - API for content operations
  - API for checkpoints operations (todo)
  - API for share operations
- Connecting IOP: gRPC (CS3 APIs)



Cs3Api4Lab (frontend)

jupyterlab

REST

Cs3Api4Lab (backend)

jupyterlab

CS3 APIs (gRPC)

Science Mesh
Inter-operability Platform

Distributed Data Science environments

Done (since CS3 conference)

- New implementation of file browser
- User information (look-up users, share with a user by name)
- Locking mechanism for concurent updating of notebooks

Current

- JupyterLab 3
- Concurrent updating of notebooks
- Unification share APIs (shares, OCM shares)

Next

- Stabilisation/tests
  - Testing with multiple remote instances
  - OCM
- Mount the file system, to allow local access from the kernel

MVP

github.com/sciencemesh/cs3api4lab

*ready to check it out

- Accessing shared files and folders
- Sharing
- Info panel (sharing info)

Case studies:

CERN

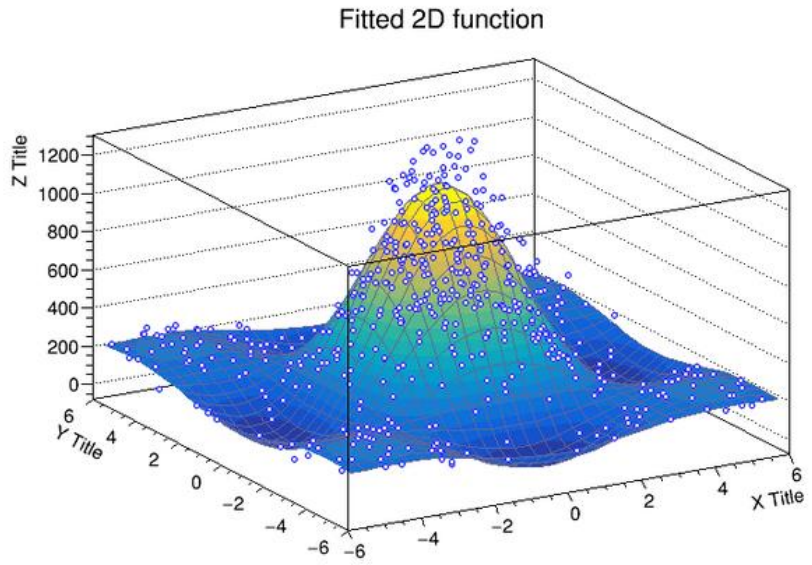
Joint Research Centre JRC

PSNC

Software Mind

ailleron

Collaborative Data Science: Beyond High Energy Physics



- All scientific disciplines nowadays are data-driven
  - Data analytics play an increasing role in all types of research
  - Distributed data science environments => all fields of study
  - A more effective collaboration between scientific institutions
- Business: develop new products in all sectors
  - Finance, IoT, SmartCities, energy and many others
- Gartner - Critical Capabilities for Data Science and Machine Learning Platforms
  - (13 March 2021)
  - **By 2023, 30% of organizations will harness the collective intelligence** of their analytics communities, outperforming competitors that rely solely on centralized analytics or self-service.
  - **By 2024, 70% of enterprises will use cloud and cloud-based AI infrastructure** to operationalize AI, thereby significantly alleviating concerns about integration and upscaling.

CS3 MESH4 EOSC

Connecting European Data

Thank you!

Discover more on...

cs3mesh4eosc.eu

company/cs3mesh4eosc

CS3org

CS3MESH4EOSC Project

https://www.youtube.com/channel/UCHKcZEkMqXjCvc3MLFjFxbw

CS3MESH4EOSC has received funding from the European Union's Horizon 2020 Research and Innovation programme under **Grant Agreement No. 863353**.