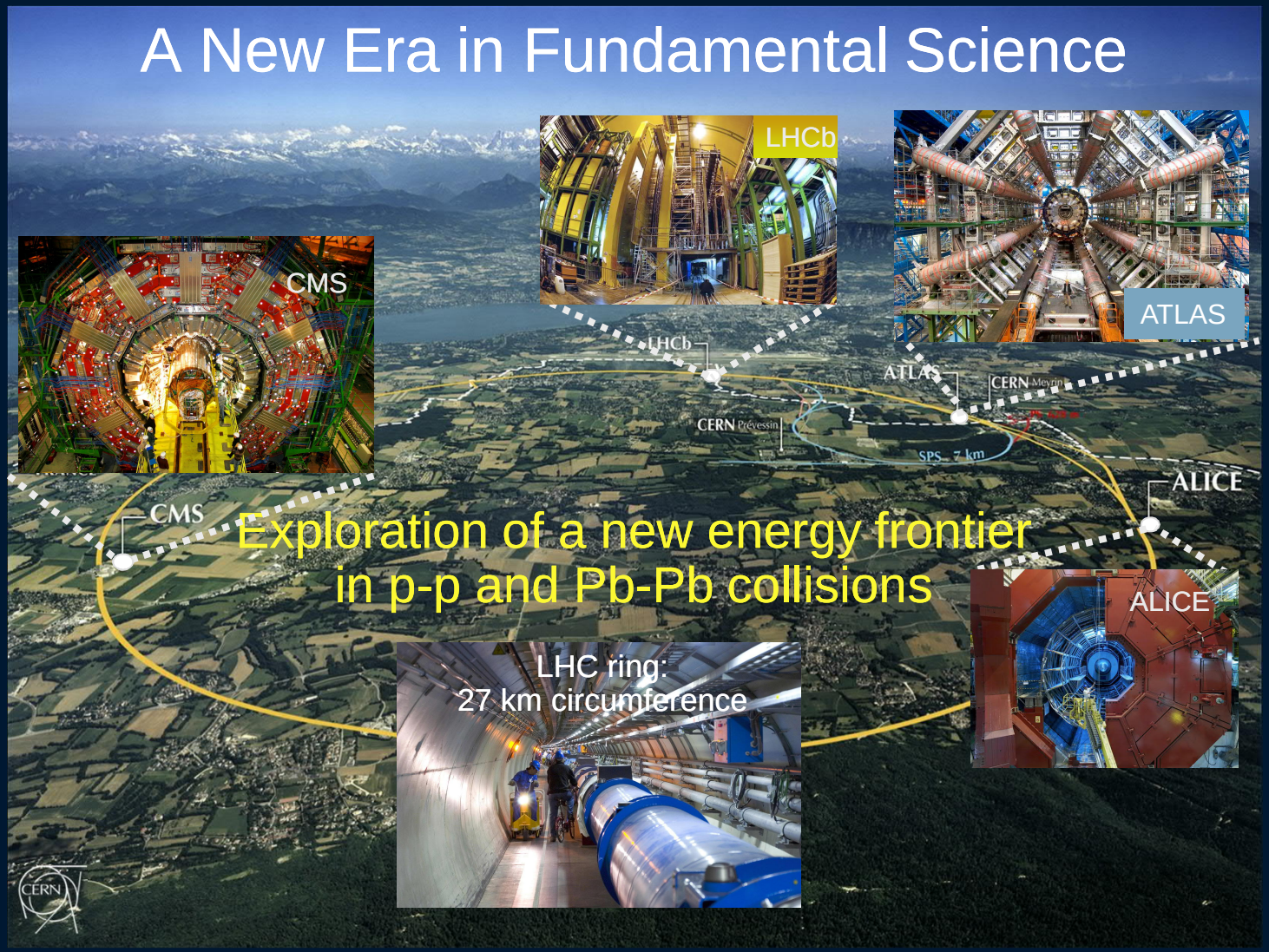

CERN

Open data management at CERN

Bob Jones
CERN
Bob.Jones <at> cern.ch

A New Era in Fundamental Science



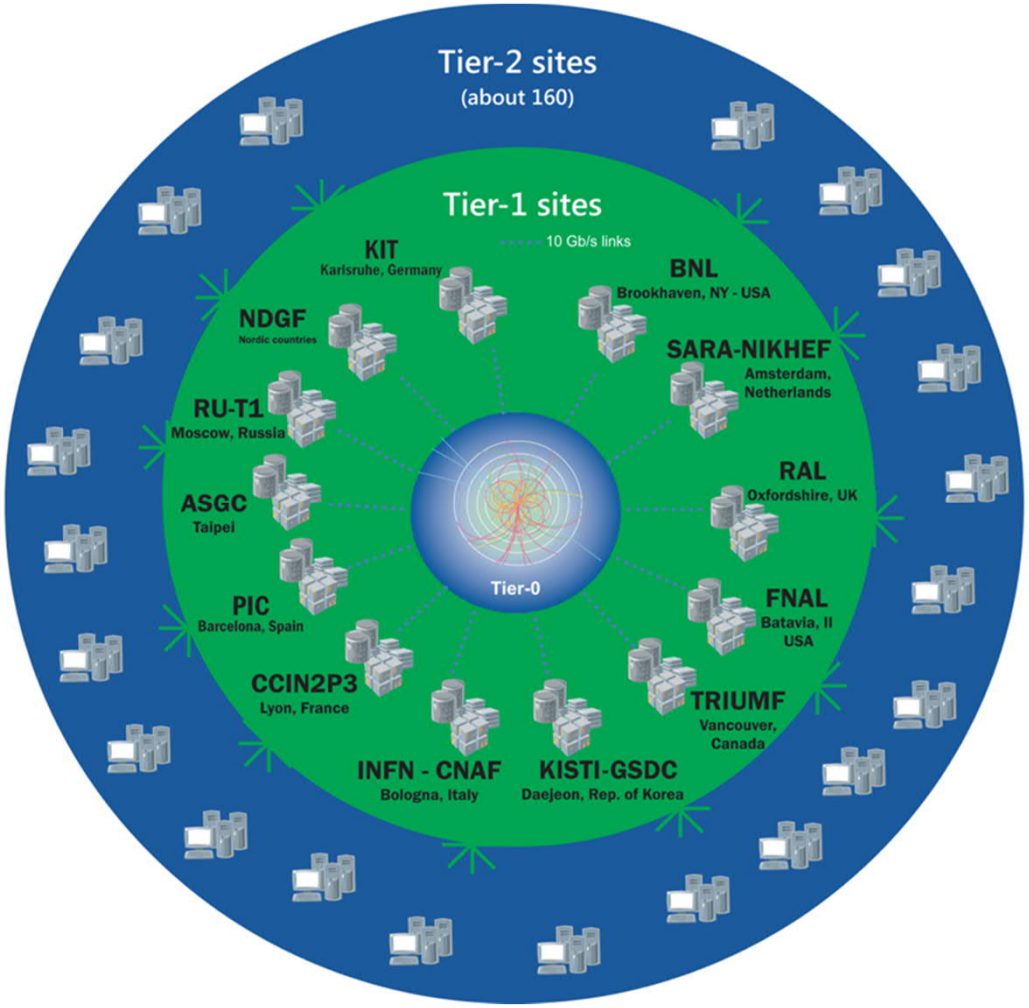
CERN

The Worldwide LHC Computing Grid

Tier-0 (CERN): dat
recording,
reconstruction and
distribution

Tier-1:
permanent storage,
re-processing,
analysis

Tier-2:
Simulation,
end-user analysis



**nearly 170 sites,
40+ countries**

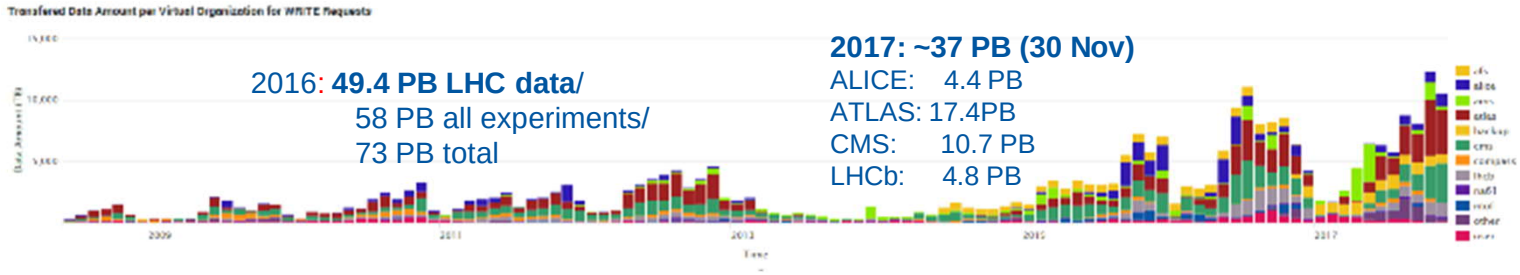
700 PB of storage

2 million jobs/day

WLCG:
An International collaboration to distribute and analyse LHC data

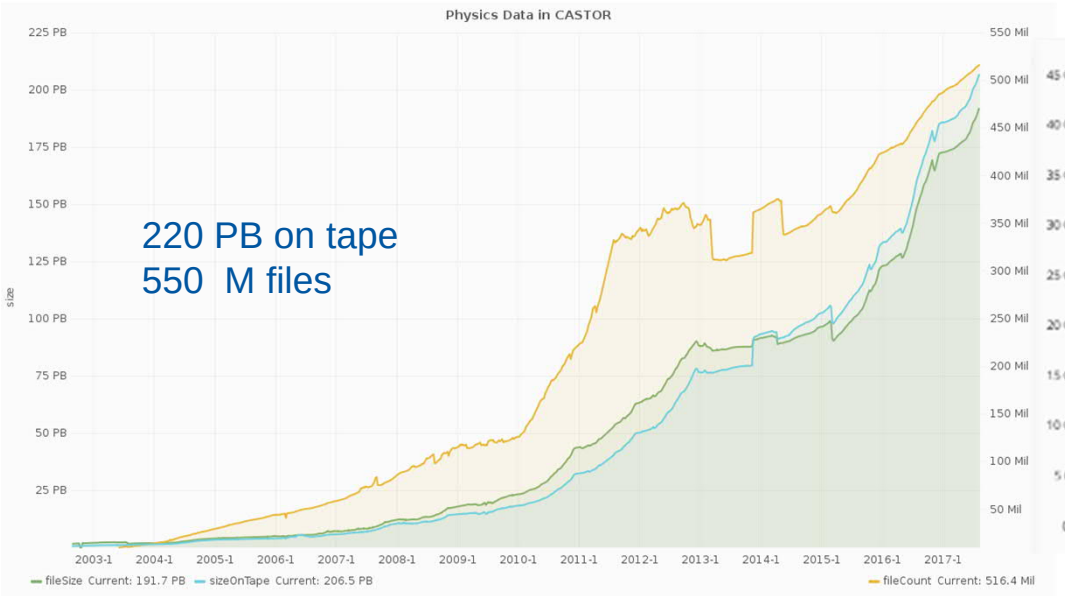
Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists

WLCG Data 2016-17

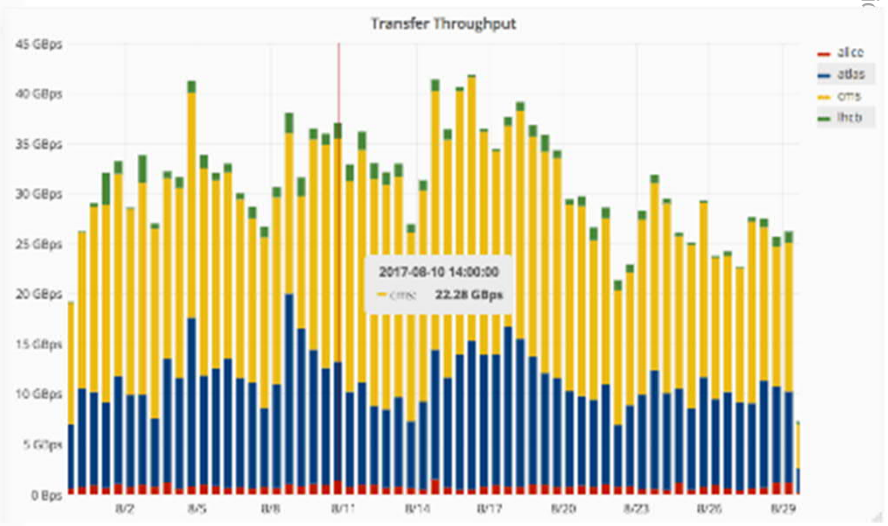


2016: **49.4 PB LHC data/**
58 PB all experiments/
73 PB total

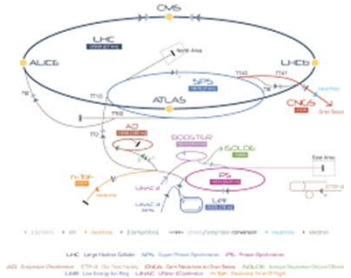
2017: ~37 PB (30 Nov)
ALICE: 4.4 PB
ATLAS: 17.4PB
CMS: 10.7 PB
LHCb: 4.8 PB



220 PB on tape
550 M files



Open Data at CERN



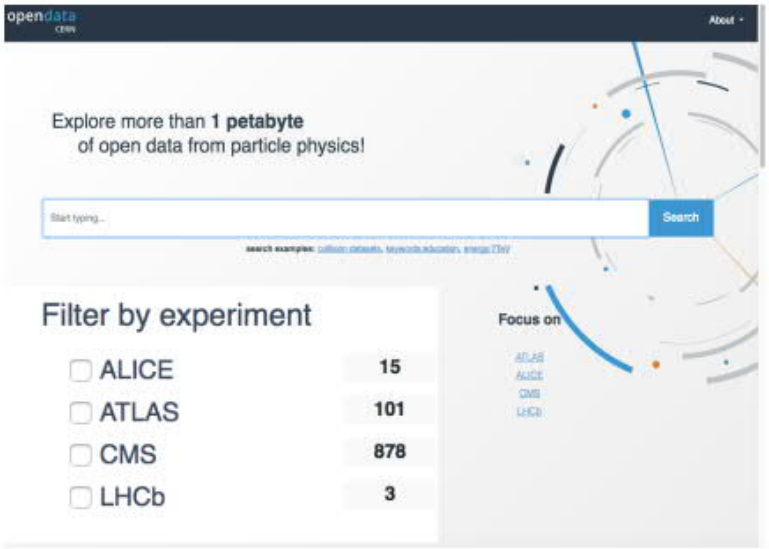
- The 4 main LHC experiments have approved **Open access** policies whereby (increasing) fractions of their data are made available after suitable “embargo periods”
  - These refer to “*derived data*” + documentation + s/w and environment

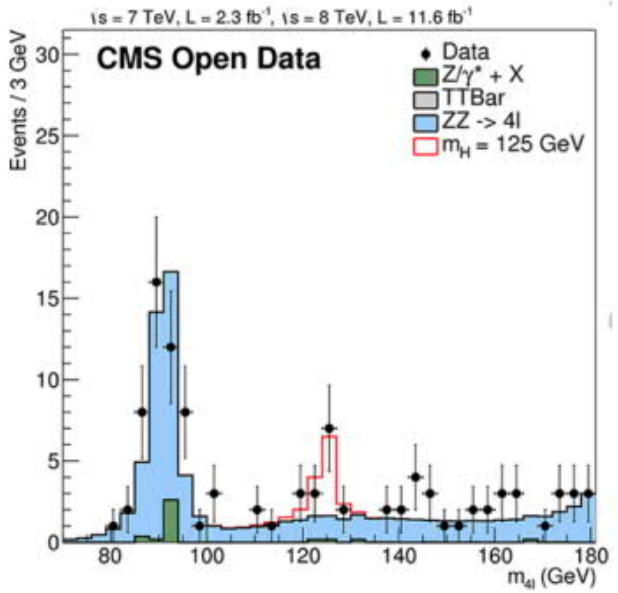
- But LHC data volume is already >200PB
  - Expected to reach ~10(-100)EB during HL-LHC
  - We need to **preserve** all of this (but not all is **Open**)

LHC: Open Data

http://opendata.cern.ch/

- Service was launched in November 2014
  - CMS 2012 open data release
    - 1PB of collision and MC data, example analyses, VM
- The service aims at publishing complex data in the open, enabling the community to **conduct preservation** in the open.
- **Standardizing the information** so it can be understood (by humans and machines) in the future.
- High interest for research and **education**





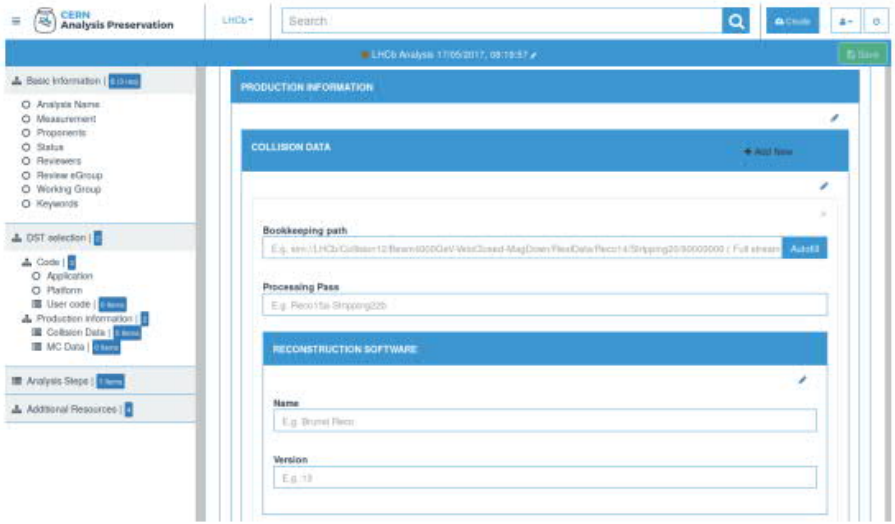
Jet Substructure Studies with CMS Open Data

Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, Jesse Thaler

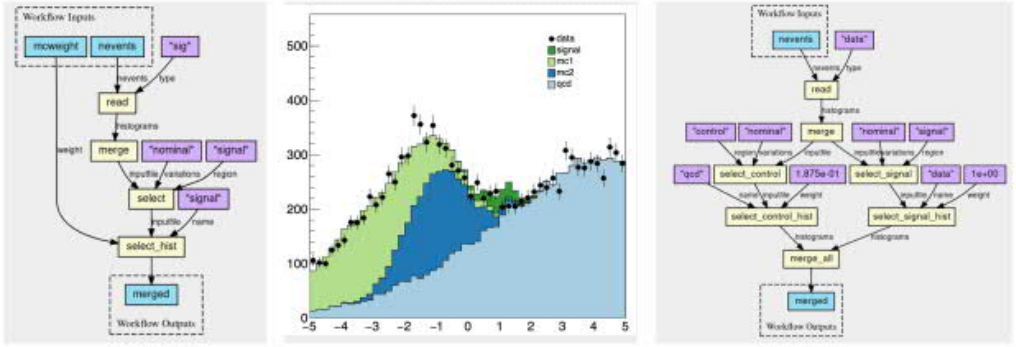
(Submitted on 19 Apr 2017 (v1), last revised 28 Sep 2017 (this version, v3))

CERN Analysis Preservation and Reusable analyses

- CAP : preserve analysis
  - Command-line client to ease submission through REST API
  - Import software from GitLab
  - Connections to collaboration databases to profit from existing information
  - http://analysispreservation.cern.ch/



- REANA : improve workflow
  - Run research data analyses on containerised compute clouds
  - REANA v0.1.0 developer preview released
  - Support for CWL workflows widely used in life sciences
  - ROOT use case examples
  - http://reana.io/



sig mc data

CERN as a Trusted Digital Repository

- We believe **ISO 16363 certification** will allow us to implement best practices and ensured for the long-term.

- **Scope**: Scientific Data and CERN's Digital Memory

- **Timescale**: complete prior to 2020

Artefactual Systems

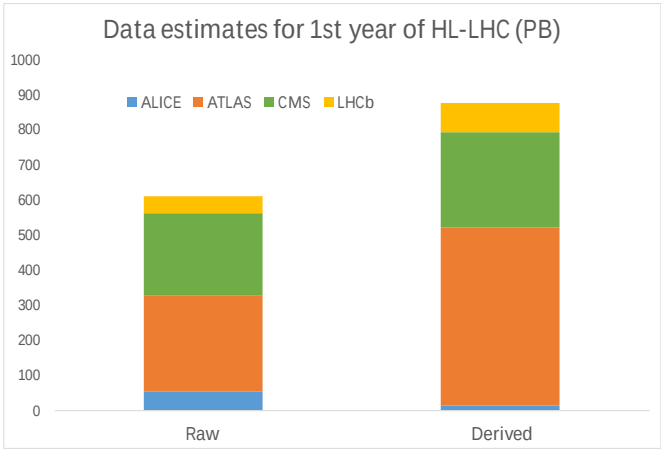
Challenges: LHC Run3 and Run4 Scale

Raw data volume for LHC increases exponentially and with it processing and analysis load

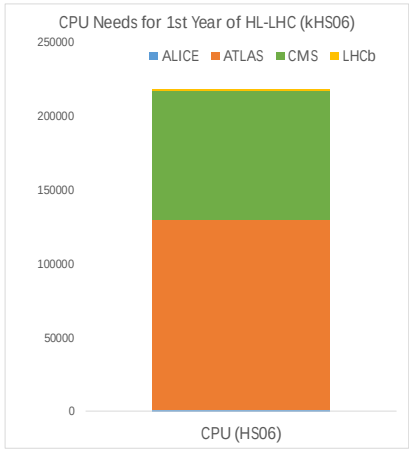
Technology at ~20%/year will bring x6-10 in 10-11 years

Estimates of resource needs at HL-LHC x10 above what is realistic to expect from technology with reasonably constant cost

Technology revolutions are needed



Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB



CPU:
- x60 from 2016

Evolution of Computing – Community White Paper*

A powerful backbone for data transfer and data storage in a few data lakes.

In line with EIROForum paper on Federated Scientific Cloud.

Use of heterogenous computing resources including HPC and dedicated processors.

Ease transition to heterogeneous structure by exploiting commonalities.

Evolution of Computing discussed with Users and Funding Agencies including joint usage of infrastructure.

Agreement with SKA on collaborating in computing efforts.

* http://hepsoftwarefoundation.org/activities/cwp.html

Prototype to be developed in context of ESCAPE H2020 project (2019-2021)

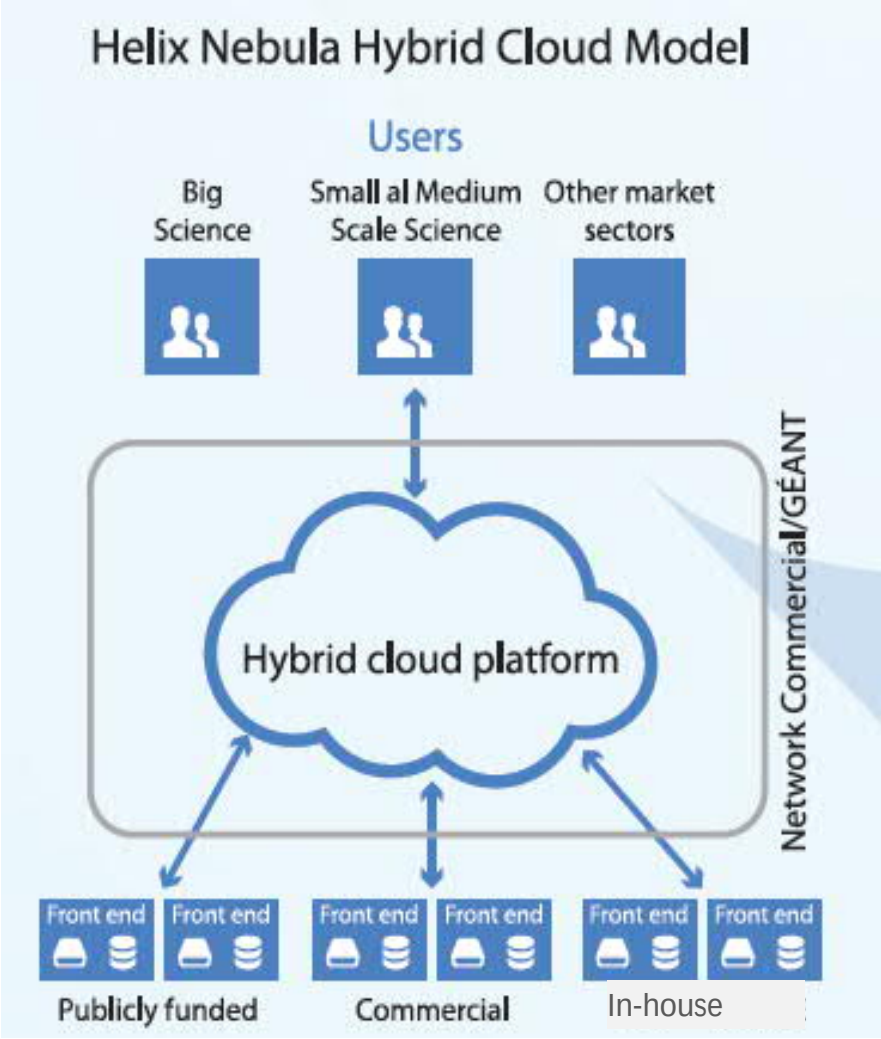
- address the stewardship of data handled by research infrastructures according to the FAIR principles and in line with the objectives of Open Science
- ensure the connection of the research infrastructures to the EOSC

The Hybrid Cloud Model

Brings together

- research organisations,
- data providers,
- publicly funded e-infrastructures,
- commercial cloud service providers

In a hybrid cloud with procurement and governance approaches suitable for the dynamic cloud market



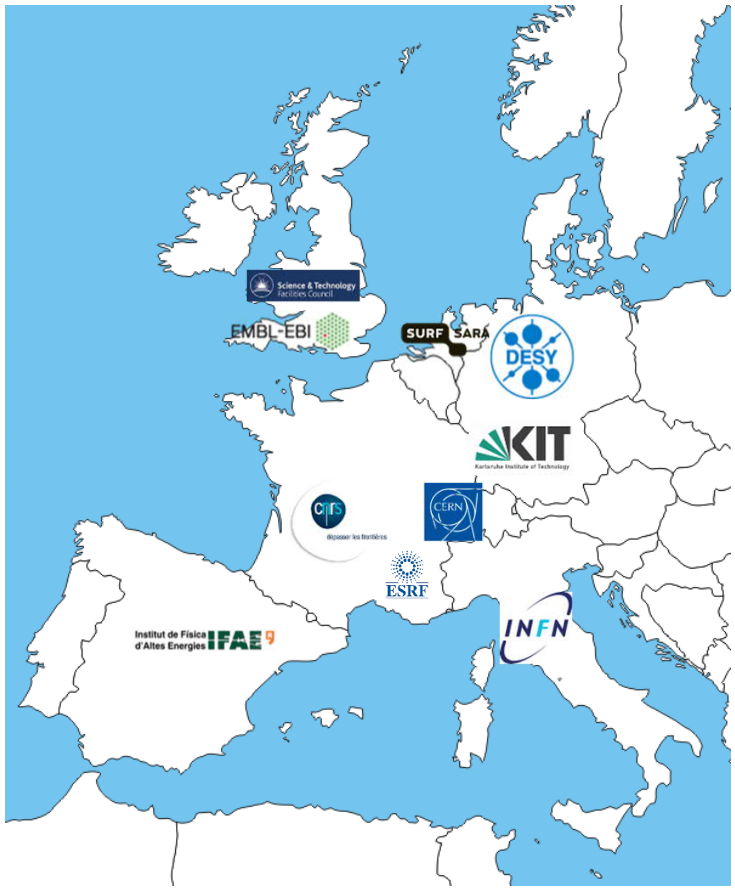
Helix Nebula Science Cloud
Joint Pre-Commercial Procurement

Procurers: CERN, CNRS, DESY, EMBL-EBI, ESRF, IFAE, INFN, KIT, STFC, SURFSara
Experts: Trust-IT & EGI.eu

Resulting IaaS level services support use-cases from many research communities

High Energy Physics | Astronomy | Life Sciences | Photon/Neutron Sciences | Long Tail of Science

Deployed in a hybrid cloud combining procurers data centres, commercial cloud service providers, GEANT network and eduGAIN fed. identity mgmt.

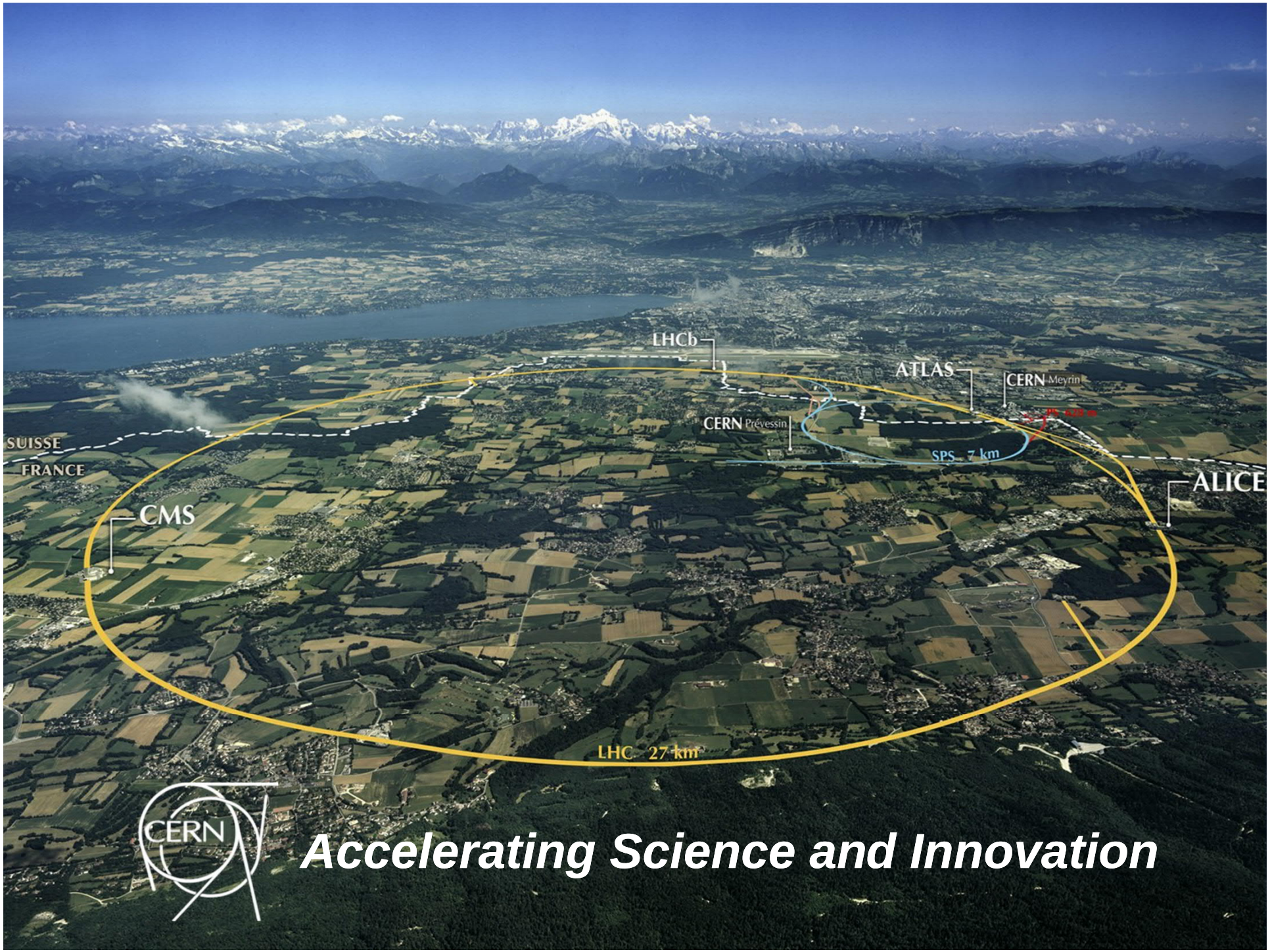


Co-funded via H2020 Grant Agreement 687614

Project results at the end of the year.
Public session will be held at CERN on 29 Nov'18

Webcast and more info:
www.hnscicloud.eu


e 2 0
u 1 8
· a t
The European Open Science Cloud
Launch Event
EOSC
23 November 2018,
10:00 – 13:30 hrs
University of Vienna
Library, main reading room


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC 27 km
CERN
Accelerating Science and Innovation