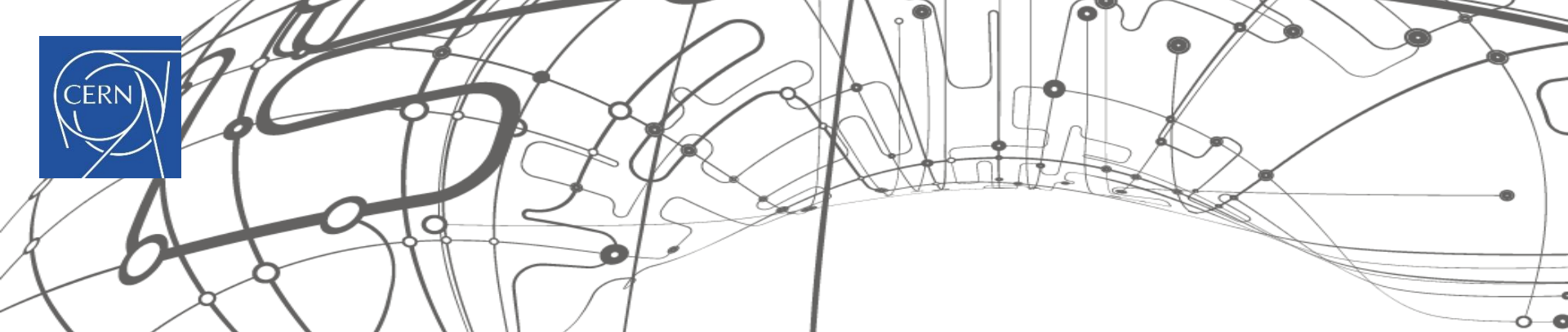


Big Data for Big Discoveries

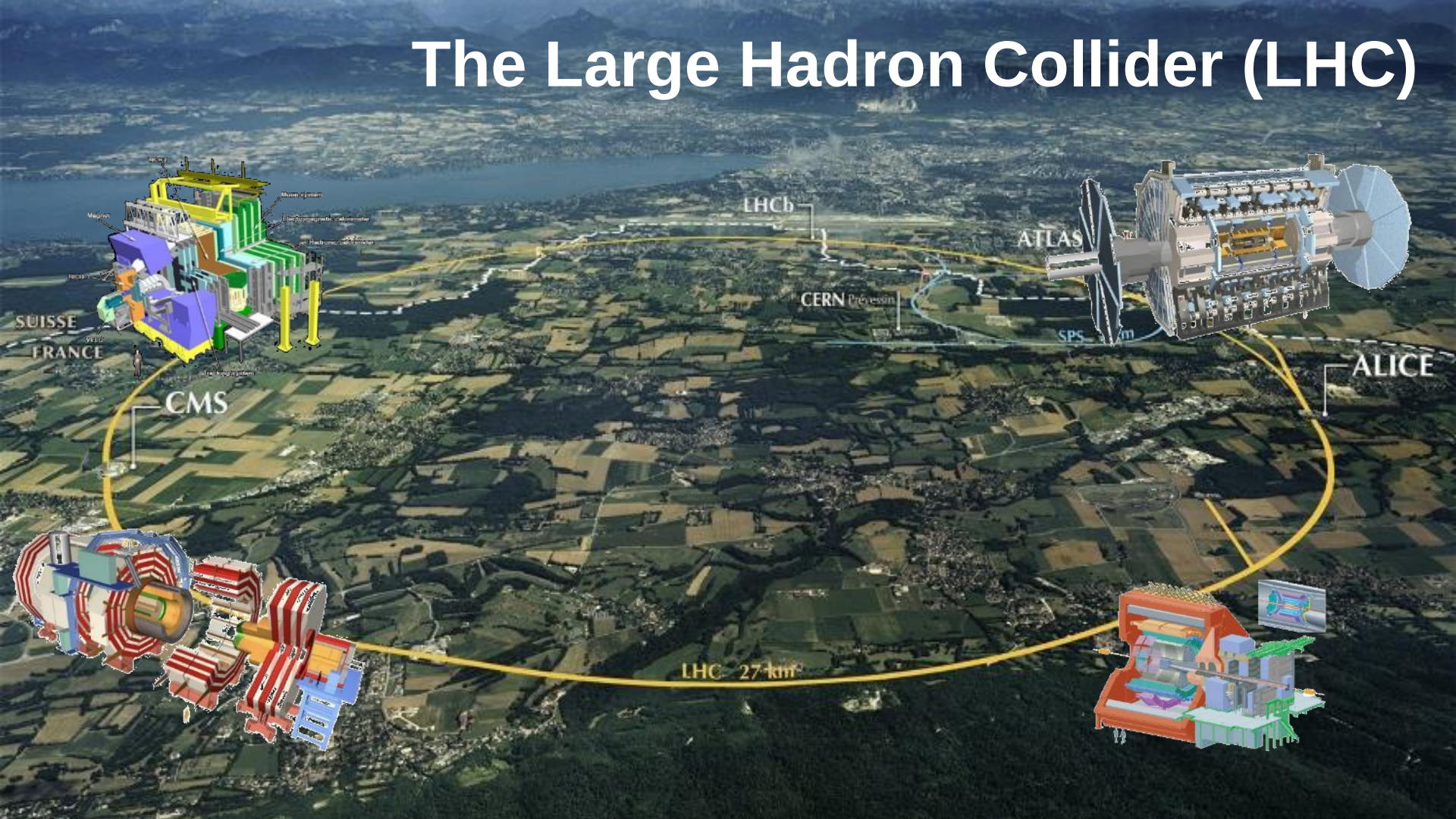
How the LHC looks for Needles by Burning Haystacks

Alberto Di Meglio

CERN openlab Head

DOI: 10.5281/zenodo.45449,

The Large Hadron Collider (LHC)



Burning Haystacks

The Detectors: 7000 tons “microscopes”
150 million sensors
Data generated 40 million times per second

➔ **Peta Bytes / sec !**

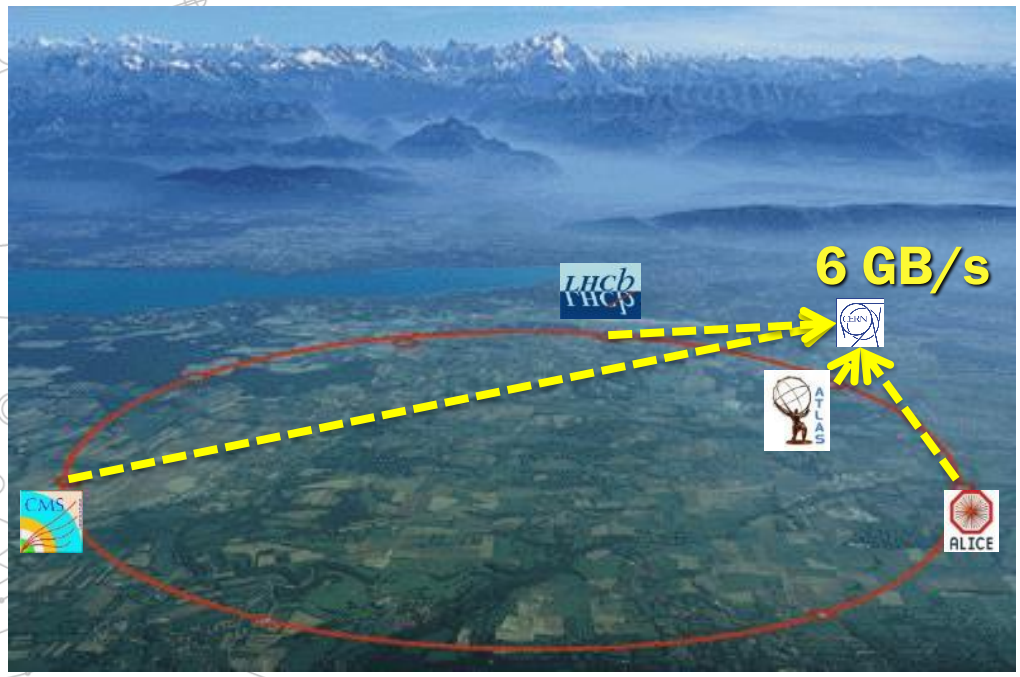
Low-level Filters and Triggers
100,000 selections per second

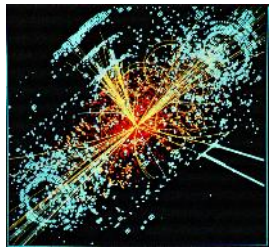
➔ **Tera Bytes / sec !**

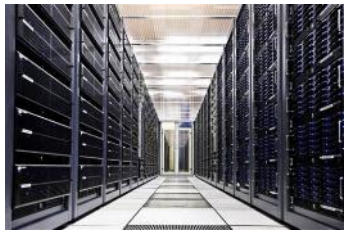
High-level filters (HLT)
100 selections per second

➔ **Giga Bytes / sec !**

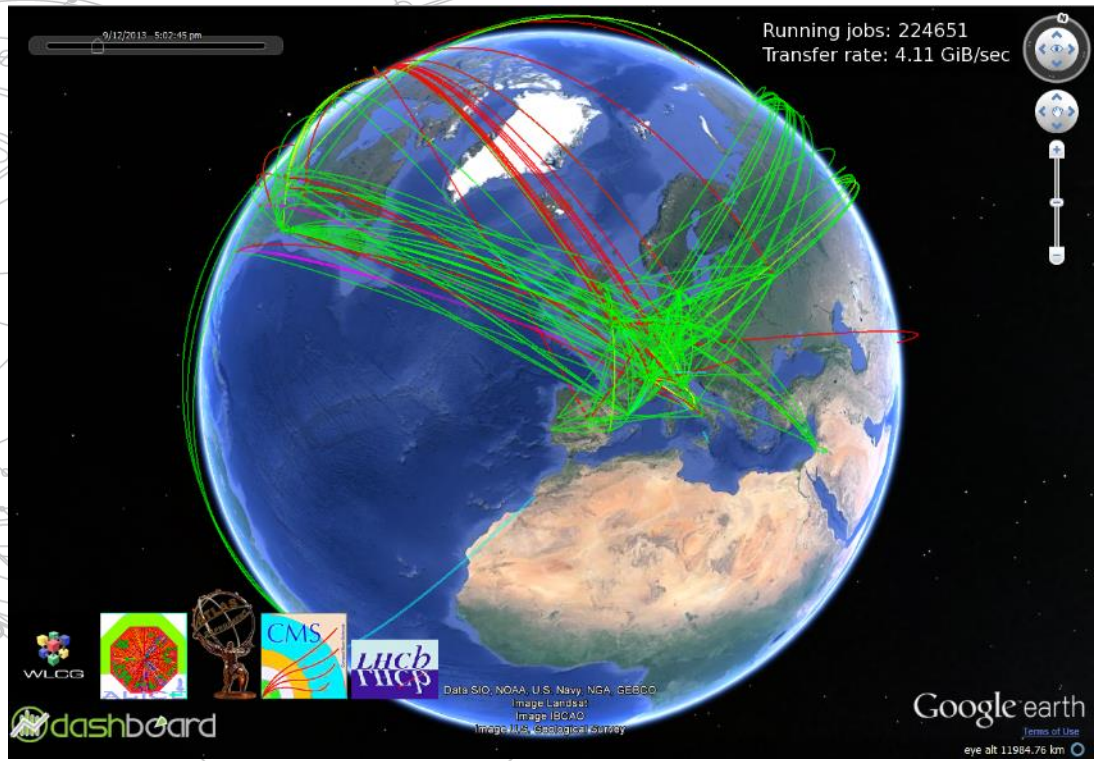
Storage, Reconstruction, Simulation, Distribution

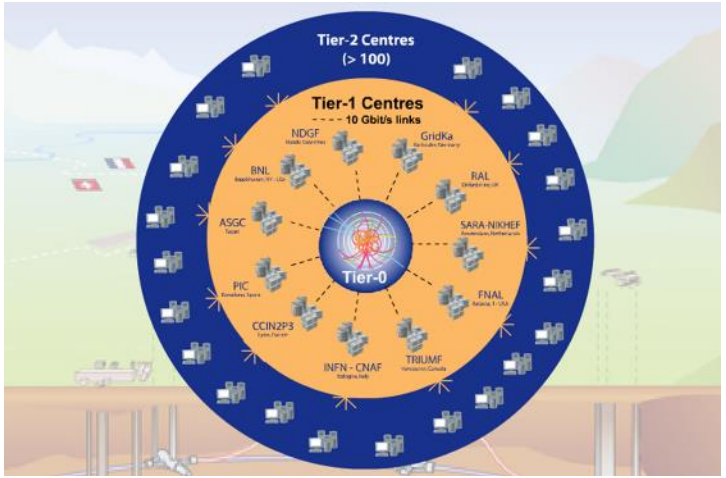






Worldwide LHC Computing Grid





Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (12 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (68 Federations, ~140 centres):
- Simulation
- End-user analysis

- 525,000 cores
- 450 PB

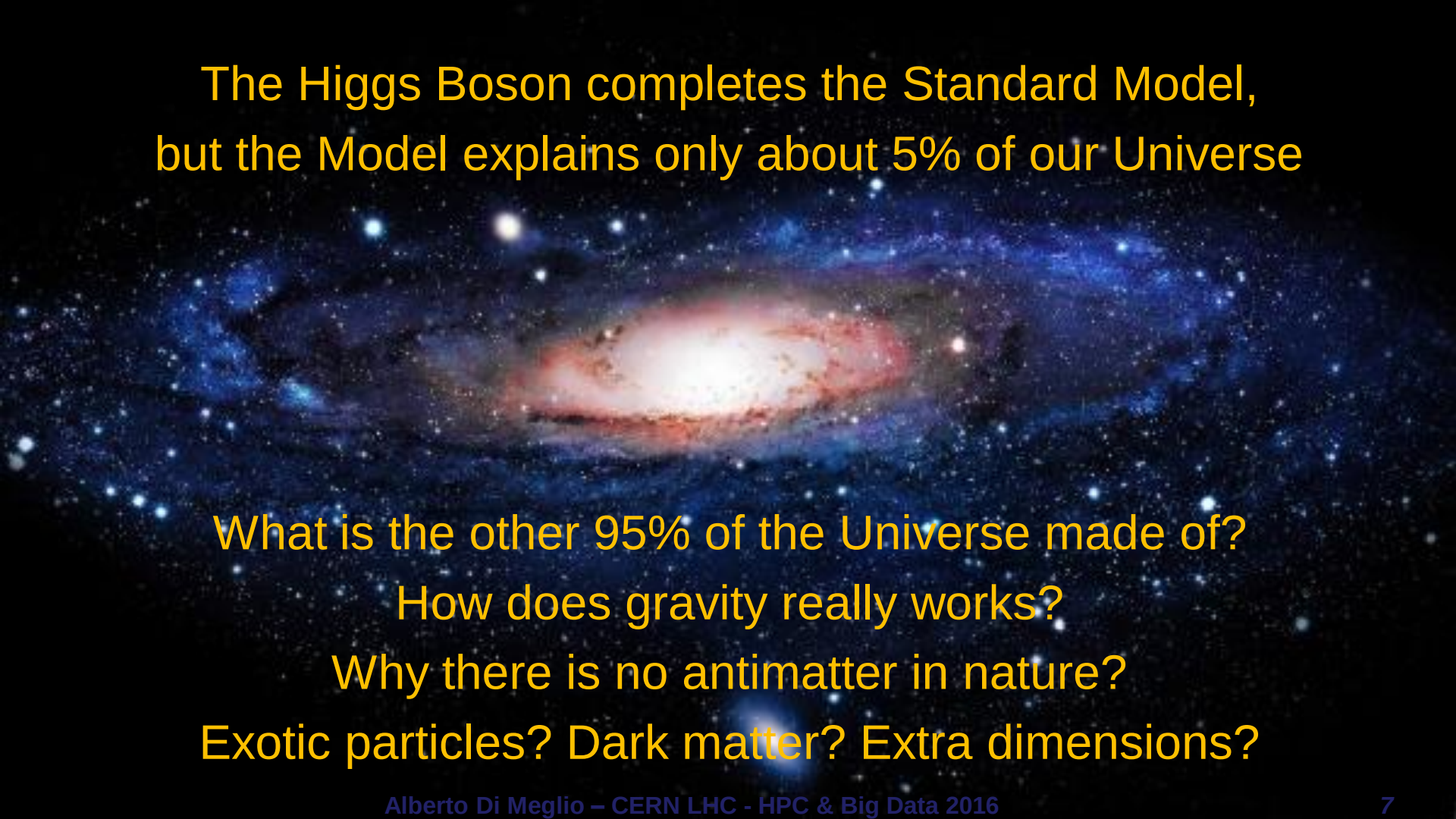
1 PB/s of data generated by the detectors

Up to 30 PB/year of stored data

A distributed computing infrastructure
of half a million cores working 24/7

An average of 40M jobs/month

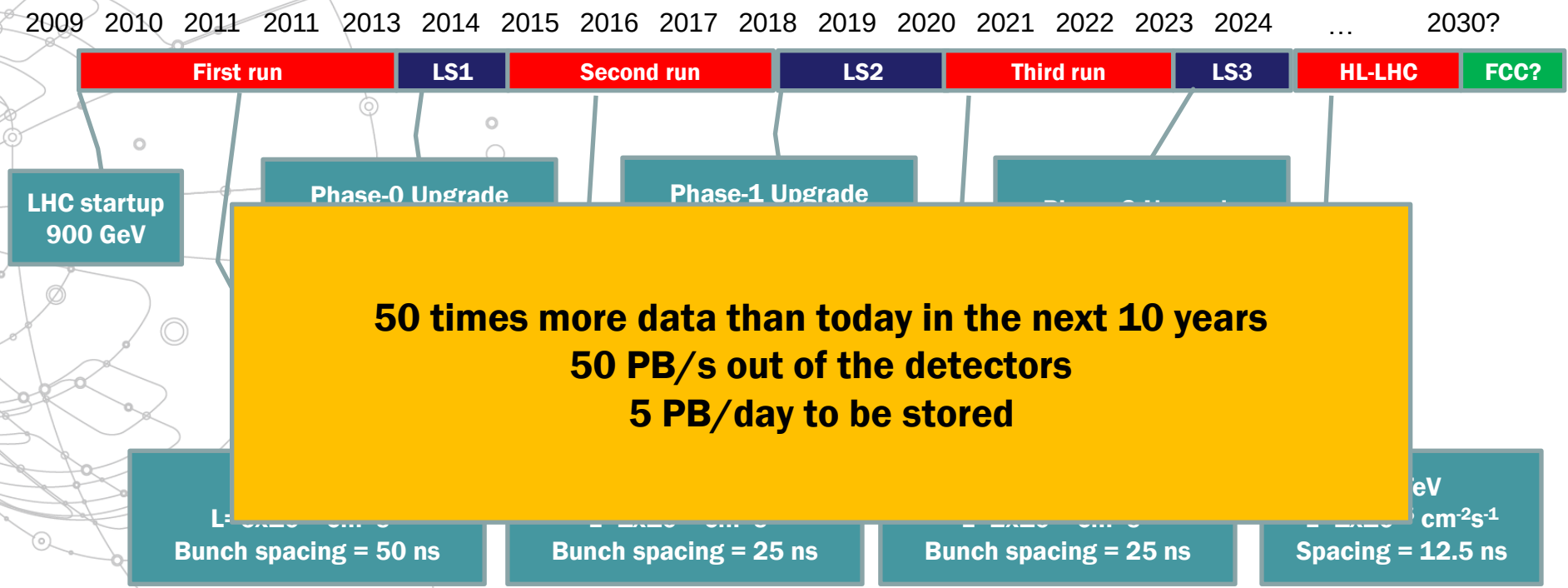
An continuous data transfer rate of 4-6 GB/s
across the Worldwide LHC Grid (WLCG)



The Higgs Boson completes the Standard Model,
but the Model explains only about 5% of our Universe

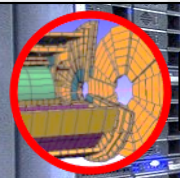
What is the other 95% of the Universe made of?
How does gravity really works?
Why there is no antimatter in nature?
Exotic particles? Dark matter? Extra dimensions?

LHC Schedule

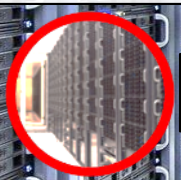


Information Technology Research Areas

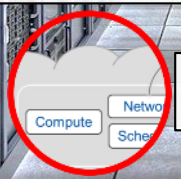
Data acquisition and filtering



Computing platforms, data analysis, simulation

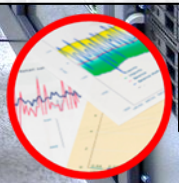


Data storage and long-term data preservation



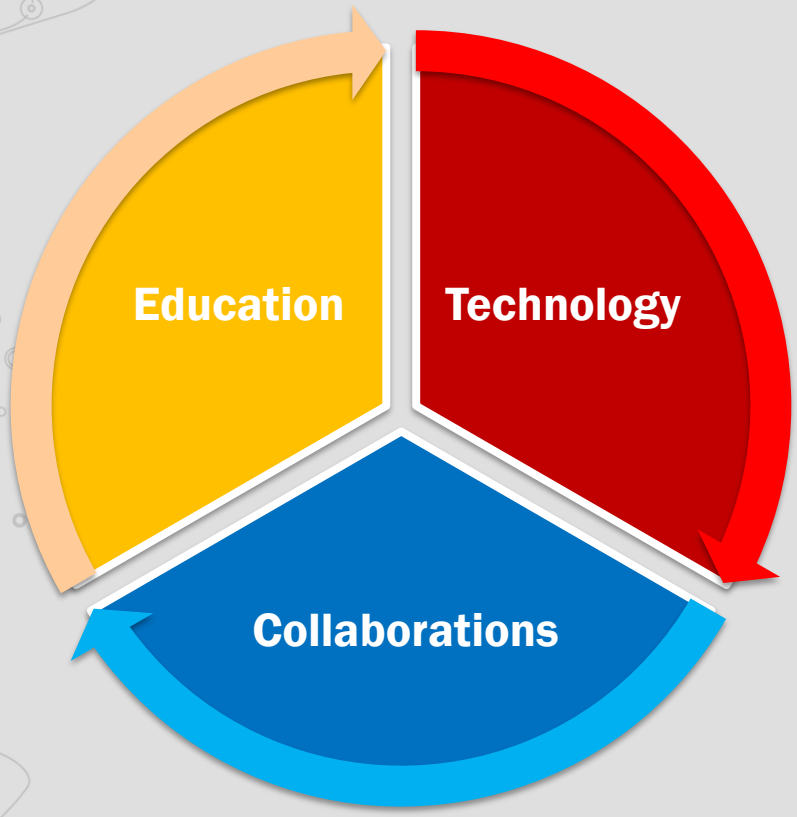
Compute provisioning (cloud)

Networks



Data analytics

How can we Address the Challenges?



Flat Budget

Technology Evolution

Computing

Storage

Infrastructure

Data Analytics

- More performance, less cost
- Redesign of DAQ
  - Less custom hardware, more off-the-shelf CPUs, fast interconnects, optimized SW
- Redesign of SW to exploit modern CPU/GPU platforms (vectorization, parallelization)
- New architectures (e.g. automata)

Technology Evolution

Computing

Storage

Infrastructure

Data Analytics

- Reduce data to be stored
  - Move computation from offline to online
- More tapes, less disks
  - Dynamic data placements
- New media
  - Object-disks, shingled
- Persistent meta-data, in-memory ops
  - Flash/SSD, NVRAM, NVDIMM

Technology Evolution

Computing

Storage

Infrastructure

Data Analytics

- Economies of scale
- Large-scale, joint procurement of services and equipment
- Hybrid infrastructure models
  - Allow commercial resource procurement
  - Large-scale cloud procurement models are not yet clear

Technology Evolution

Computing

Storage

Infrastructure

Data Analytics

- Reduce analysis cost, decrease time
- More efficient operations of the LHC systems
  - Log analysis, proactive maintenance
- Infrastructure monitoring, optimization
- Data reduction, event filtering and identification on large data sets (~10PB)

Collaborations

HEP Community

CERN

WLCG Worldwide LHC Computing Grid

CMS

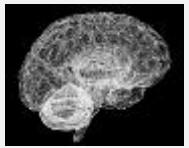
ALICE

LHCb

ATLAS

Laboratories, academia, research

EIROforum



Industrial Collaboration Frameworks

CERN openlab

Idea^{s}

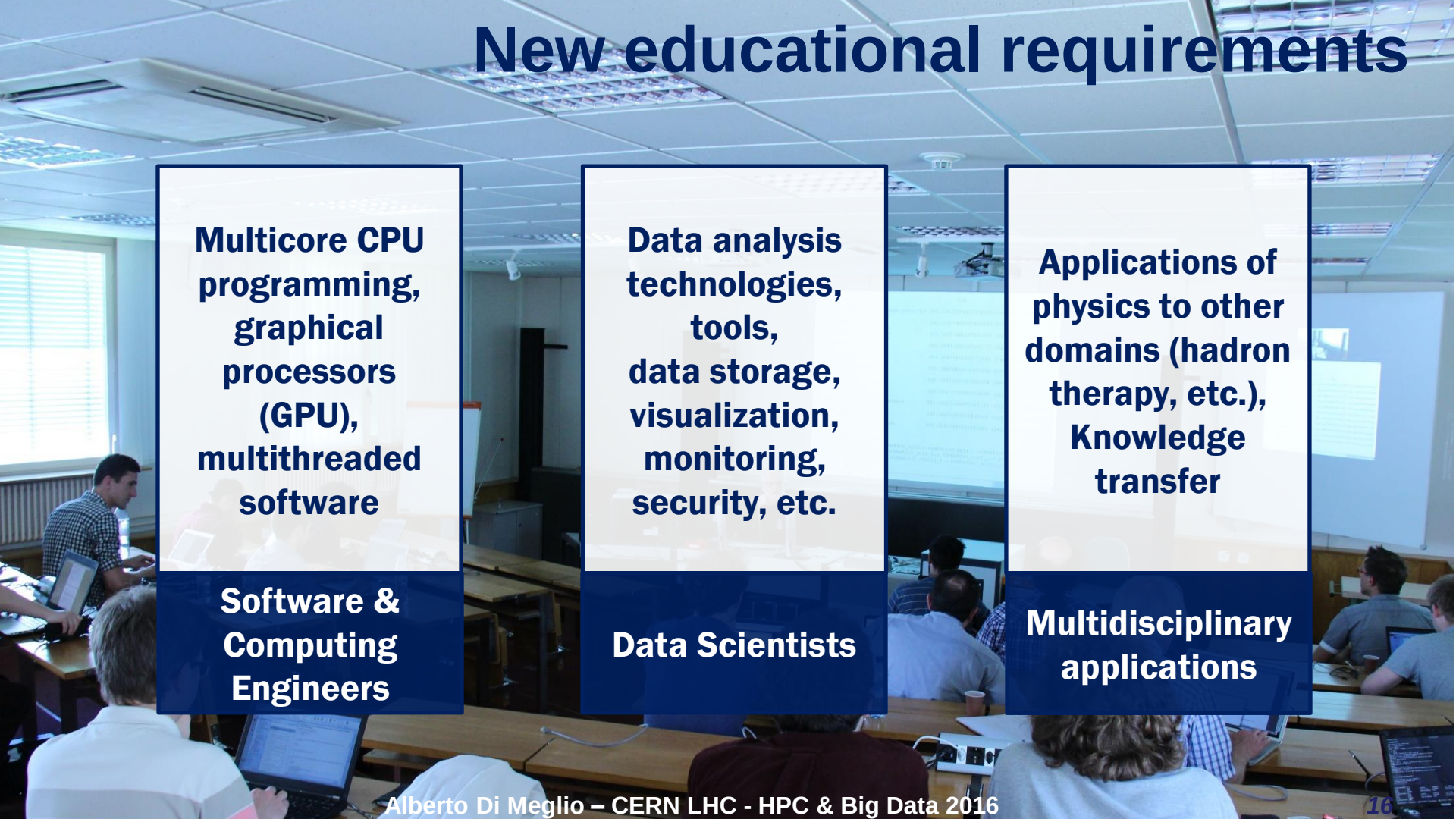
Joint Infrastructure Initiatives

HELIX NEBULA THE SCIENCE CLOUD

PICSE

INDIGO - DataCloud

New educational requirements



Multicore CPU programming, graphical processors (GPU), multithreaded software	Data analysis technologies, tools, data storage, visualization, monitoring, security, etc.	Applications of physics to other domains (hadron therapy, etc.), Knowledge transfer
Software & Computing Engineers	**Data Scientists**	**Multidisciplinary applications**

Take-Away Messages

- LHC is a long-term project
  - Need to adapt and evolve in time
- Look for cost-effective solutions
  - Continuous technology tracking
  - Aim for "*just as good as necessary*" based on concrete scientific objectives
- Collaboration and education
  - People is the most important asset to make this endeavour sustainable over time

EXECUTIVE CONTACT

Alberto Di Meglio, CERN openlab Head

alberto.di.meglio@cern.ch

TECHNICAL CONTACT

Maria Girone, CERN openlab CTO

Maria.girone@cern.ch

Fons Rademakers, CERN openlab CRO

fons.rademakers@cern.ch

COMMUNICATION CONTACT

Andrew Purcell, CERN openlab Communication Officer

Andrew.purcell@cern.ch

Mélissa Gaillard, CERN IT Communication Officer

melissa.gaillard@cern.ch

ADMIN CONTACT

Kristina Gunne, CERN openlab Administration Officer

kristina.gunne@cern.ch