CERN openlab

Big Data Platforms for Multi-Disciplinary Research

UNIGE-CERN Workshop on Life Sciences

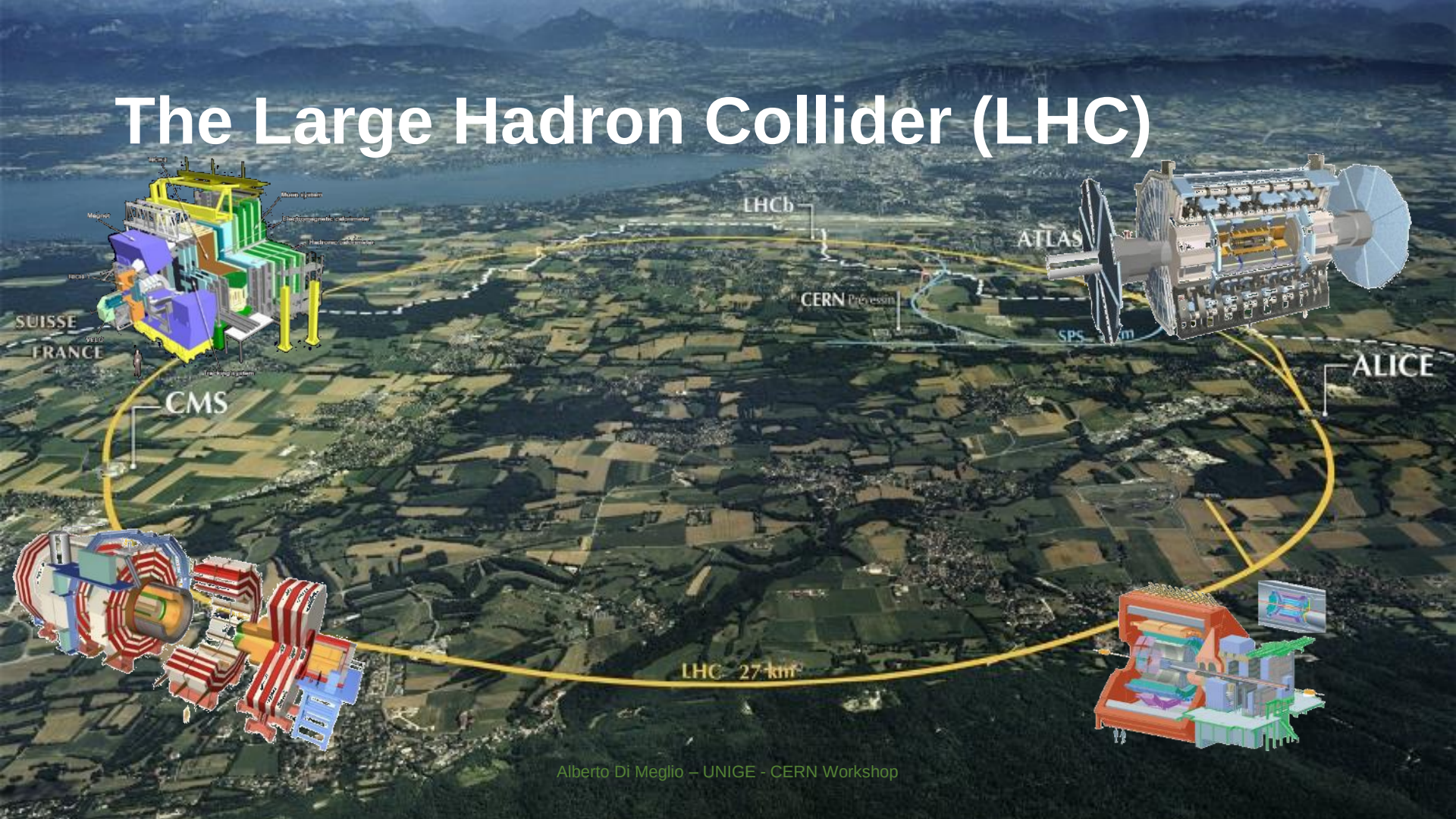
Alberto Di Meglio

16/06/2017

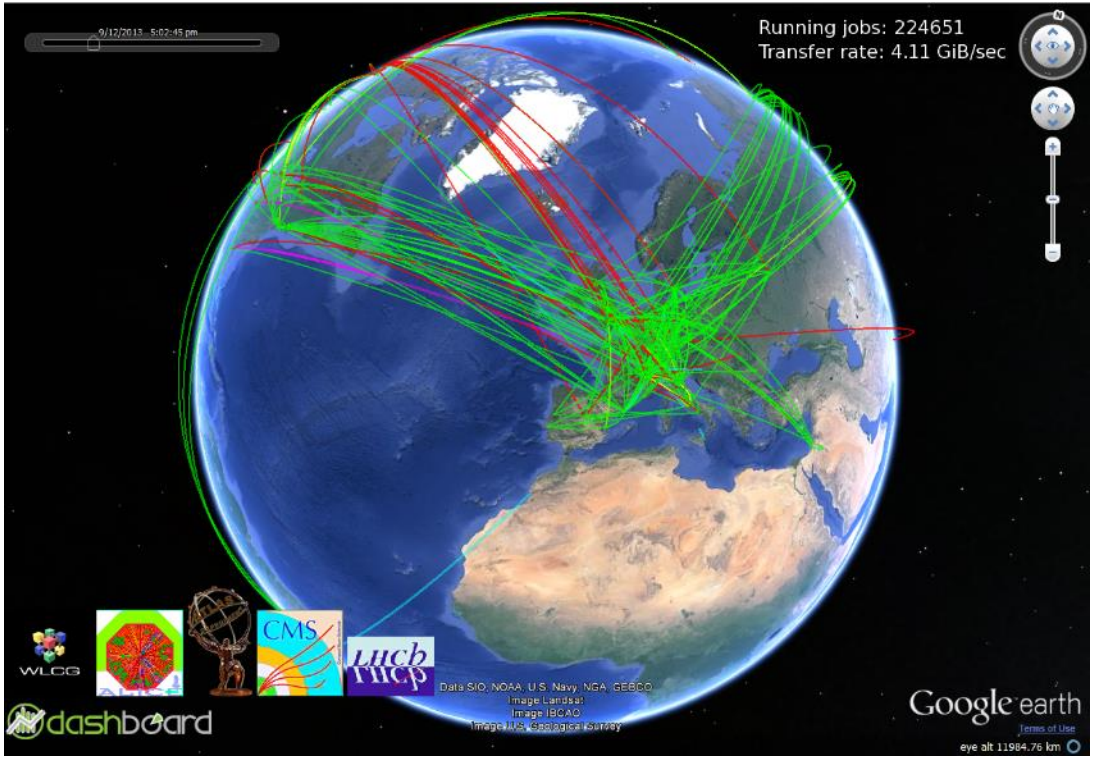
HEP and Life Science Computing

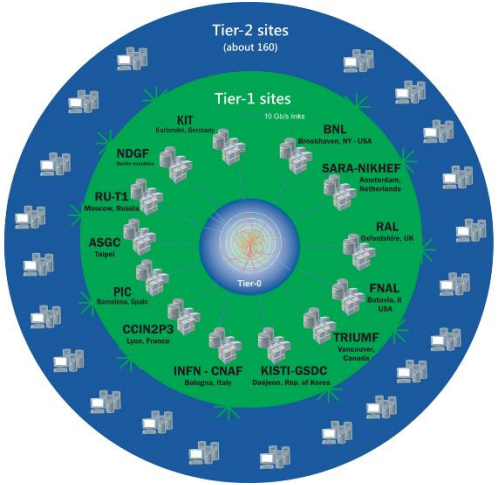
Attacking a Castle - Utagawa Yoshitora (active 1840-1880) c.1864. Oban Triptych.

The Large Hadron Collider (LHC)



Worldwide LHC Computing Grid





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

Tier-1 (14 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (72 Federations, ~149 centres):
- Simulation
- End-user analysis

- 660,000 cores
- 690 PB

How is CERN involved in Computing for Life Science

A Short history of Grid Computing

Already in the past 15 years, the efforts to build the current HEP infrastructure through various generations of EC-funded projects have produced interactions and given access to many cross-disciplinary projects

CERN and the HEP community have built a reputation about their expertise in structuring and moving forward data-driven research thanks to technical skills and the experience of the community on collaboration, governance models, standardization

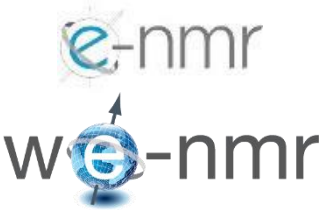
A number of research projects have been deployed on the grid showing potential, although the “business model” was never really sustainable

Examples of Infrastructure projects

Longstanding tradition of participation and support for IT infrastructures for health research since 2002 in the EGEE, EGI, EMI projects

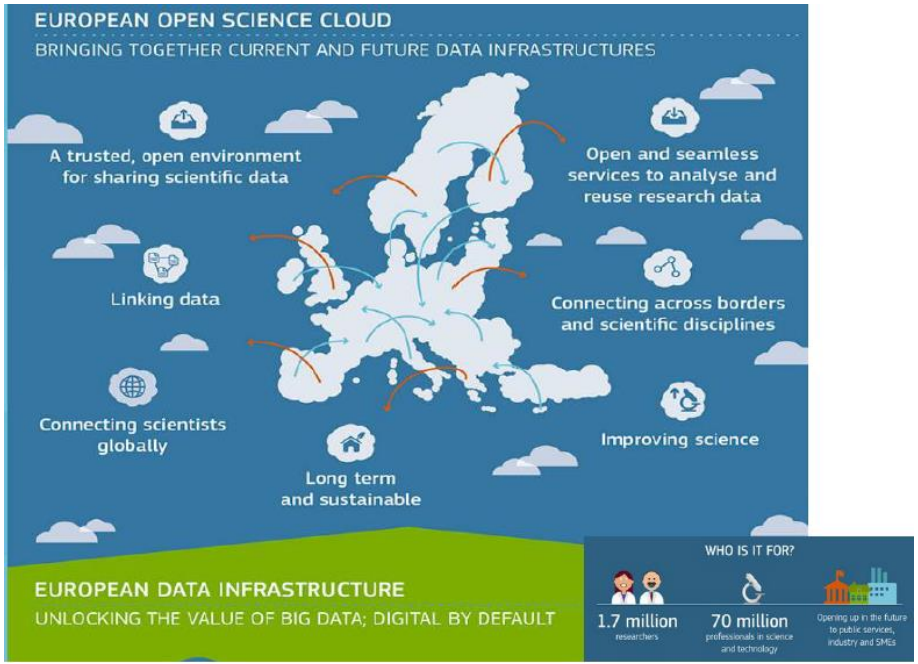
Health-e-Child

Mammogrid



Wisdom I and II (In-Silico Docking on Malaria), GPS@, Xmipp_MlrefIne, GATE, CDSS, gPTM3D, SiMRI 3D, and many other projects and applications

Open Science Clouds



The objective of CERN's participation in the work programme is to develop policies, technologies and services that can support the Organization's scientific programme, promote open science and expand the impact of fundamental research on society and the economy.

European Open Science Cloud can provide the context for future projects

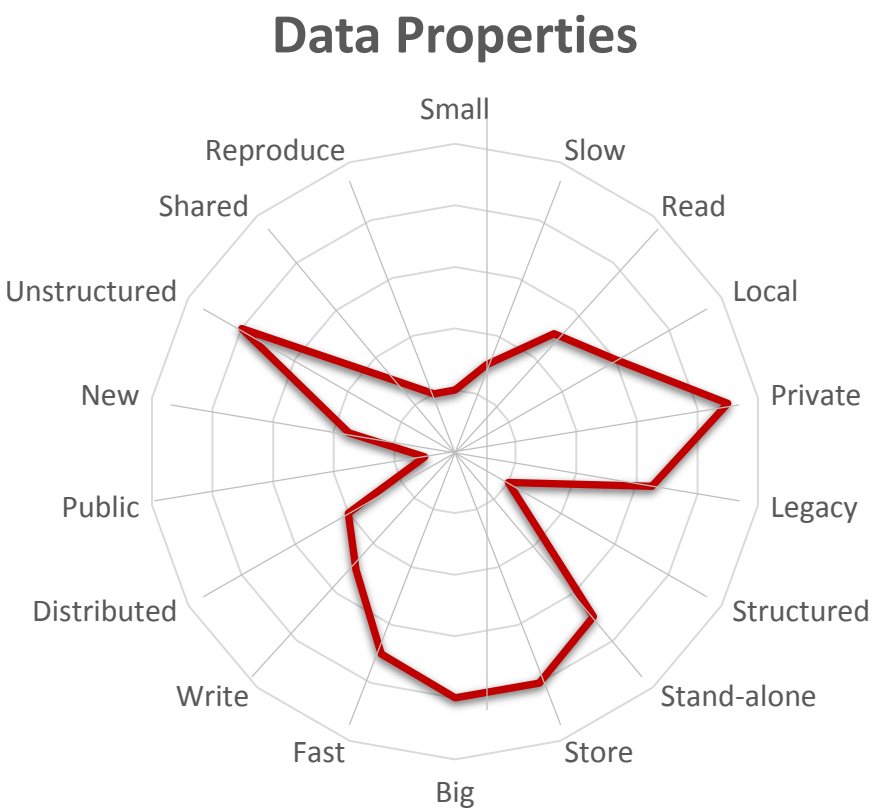
However, a major question remains:

how to really make it accessible and usable by scientific communities at large

“Big Data” and Multi-Disciplinary Distributed Research Platforms

Methods for preventing measles - Utagawa Fusatane, 1858 -- UCSF Japanese Woodblock Print Collection

It’s all about Data



Big > ~10s of TB over a (your-definition-of) short period of time

Slow/Fast is not an absolute value

Local/Distributed: cost and logistics of (secure) data storage and transfers

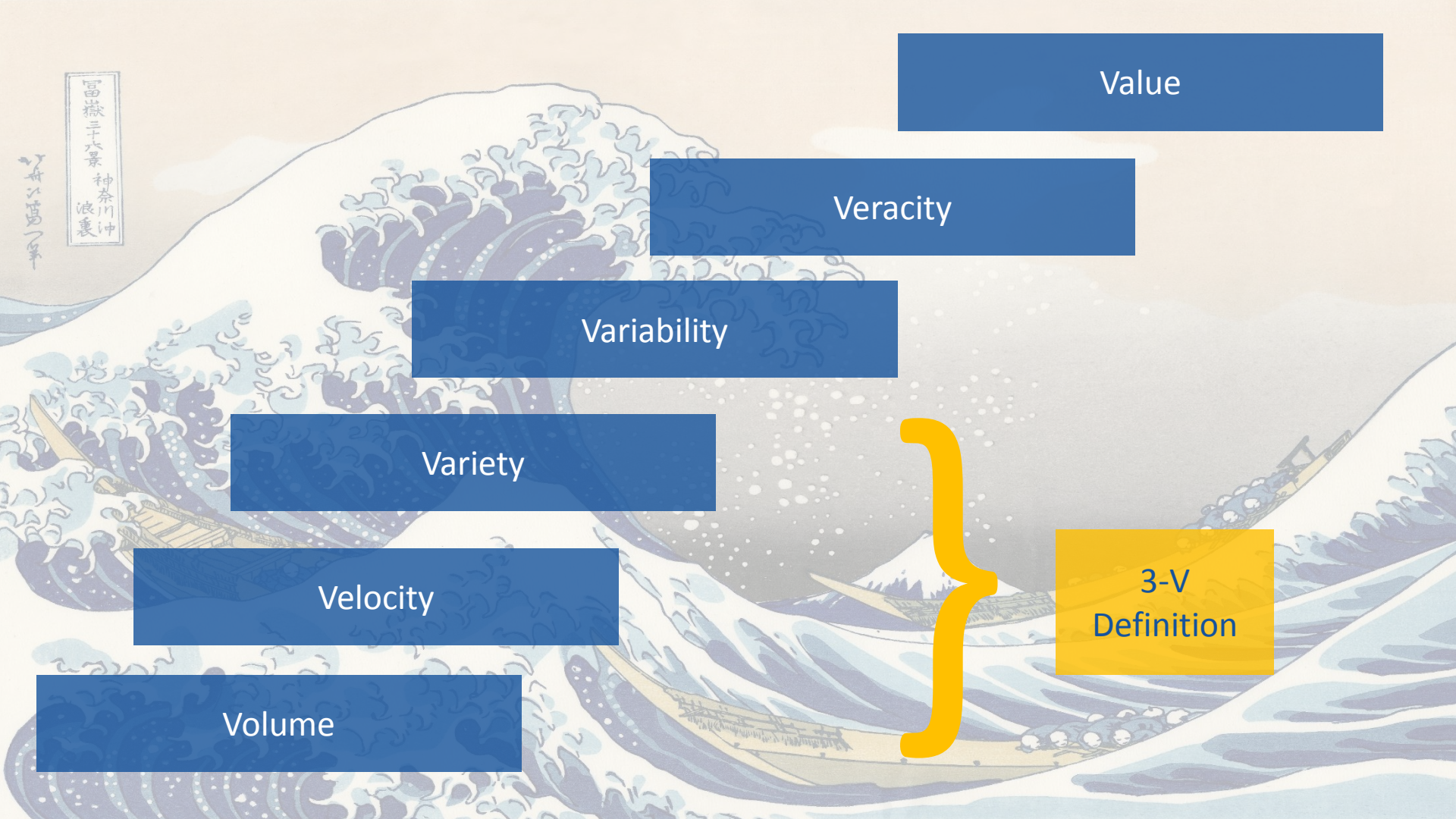
Public/Private: most characterizing property of healthcare applications

Legacy: not only the data itself (format), but how it is stored and accessed (Excel, SQL, etc.)

Unstructured ~ 80% (estimated)

Stand-alone/Shared is about interoperability and standards compliance

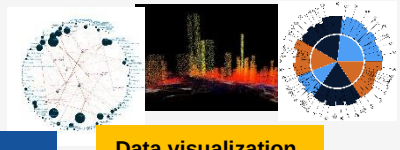
Store/Reproduce: how much must be stored, for how long, how much can be reproduced? Not the same as Raw or Processed data


Value
Veracity
Variability
Variety
Velocity
Volume
3-V
Definition

Information

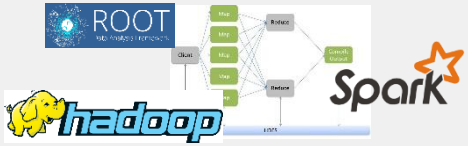
Applications designed for “big data”

Software

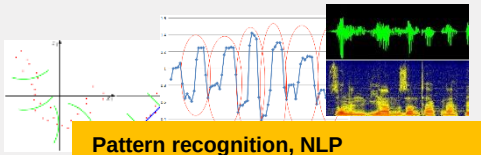


Data visualization,

Value
Veracity
Variability
Variety
Volume
Velocity

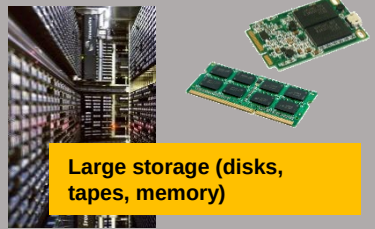


Data analysis
and analytics platforms



Pattern recognition, NLP
machine learning, predictive analytics

Platform



Large storage (disks, tapes, memory)

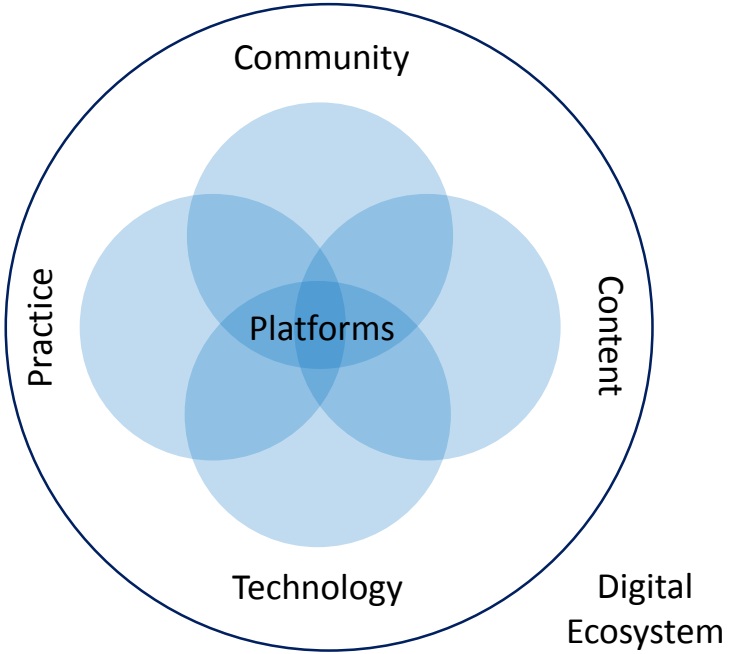
Fast data acquisition systems

Flexible networks

Distributed Computing and Data Grids, Clouds, HPC, Crowd Computing

Infrastructure

Platforms as Enablers of Ecosystem Effects



As scale grows, a sound digital ecosystem is generated by four main elements and a way for those elements to interact on common terms

Platforms are the unifying services at the intersections that enable

commoditization

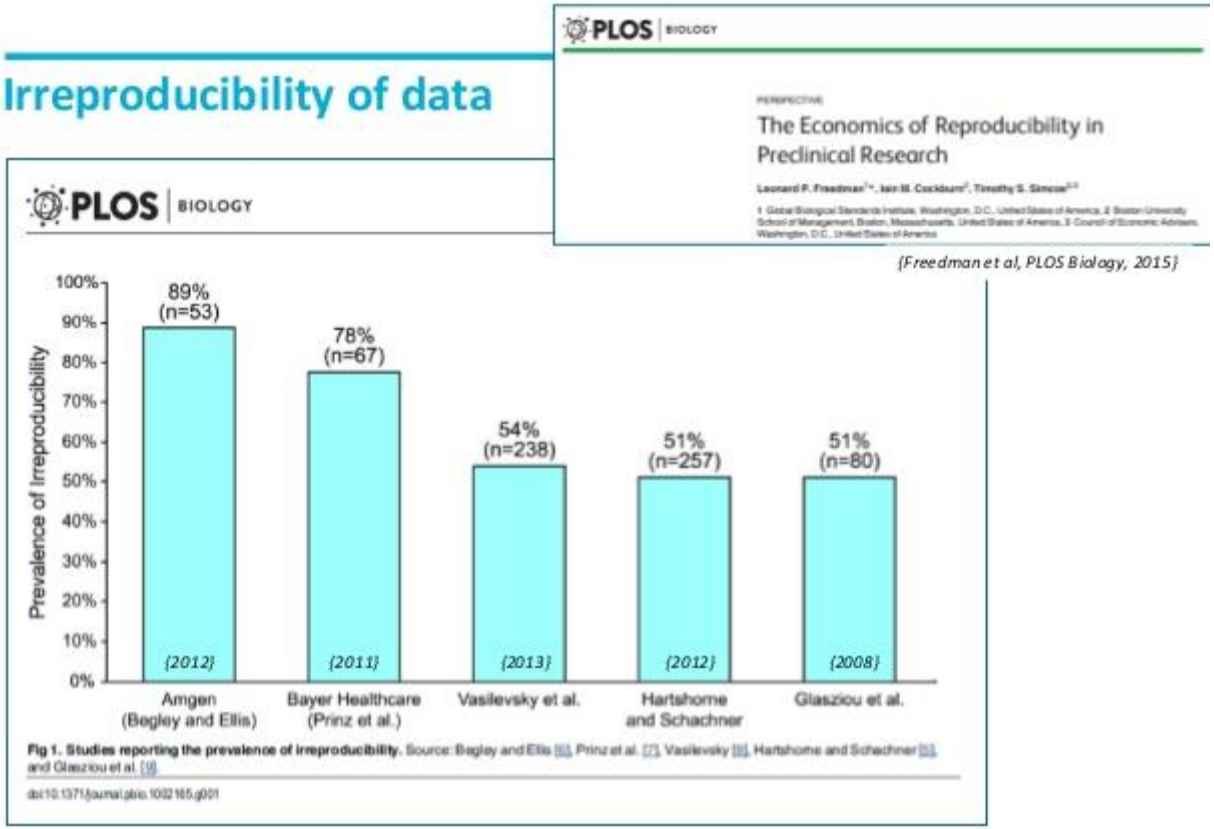
best practices

aggregation and integration

share and reuse of data

collaborations, etc.

Irreproducibility of data



{Freedman et al, PLOS Biology, 2015}

Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [9], and Glasziou et al. [10].

doi:10.1371/journal.pbio.1002165.g001

Ecosystem Effect

The Garden as a Metaphor for Platforms

Concerts, games, suggestion box (Events, feedback)

People walking, playing, fishing (Large communities)

Kiosks (Third-party services and extensions, Marketplace)

Opening and closing times, don't walk on the grass, no fishing on Tuesdays (Policies)

How many people came today? What did they do? (Statistics, trends, predictions)

The Mayor (Funding Agencies)

Park manager, guards (Governance)

Different plants and flowers (Commodity services)

Insects and animals (Small communities, micro services)

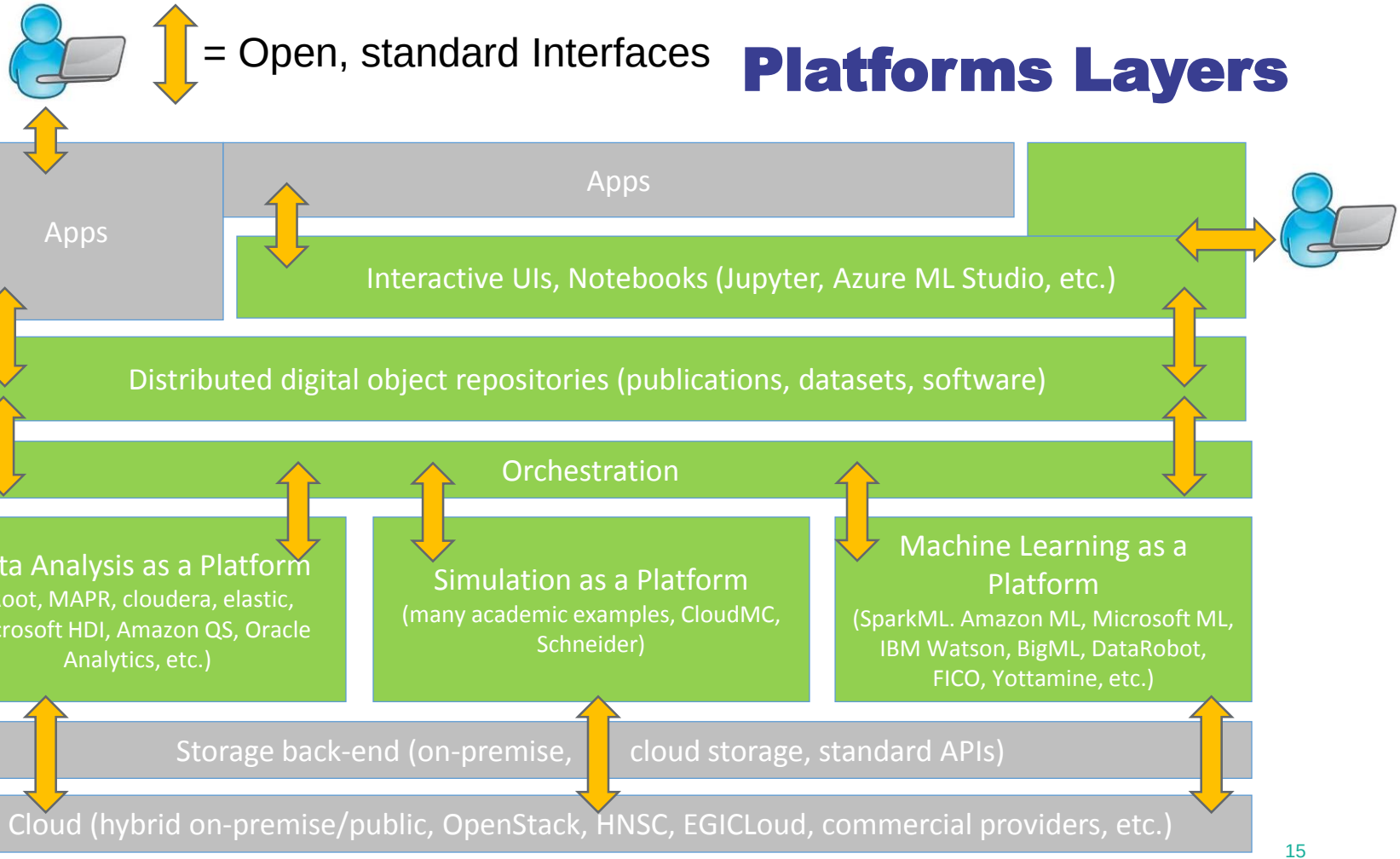
Seeds (Pilot projects)

Weeds (Failed projects)

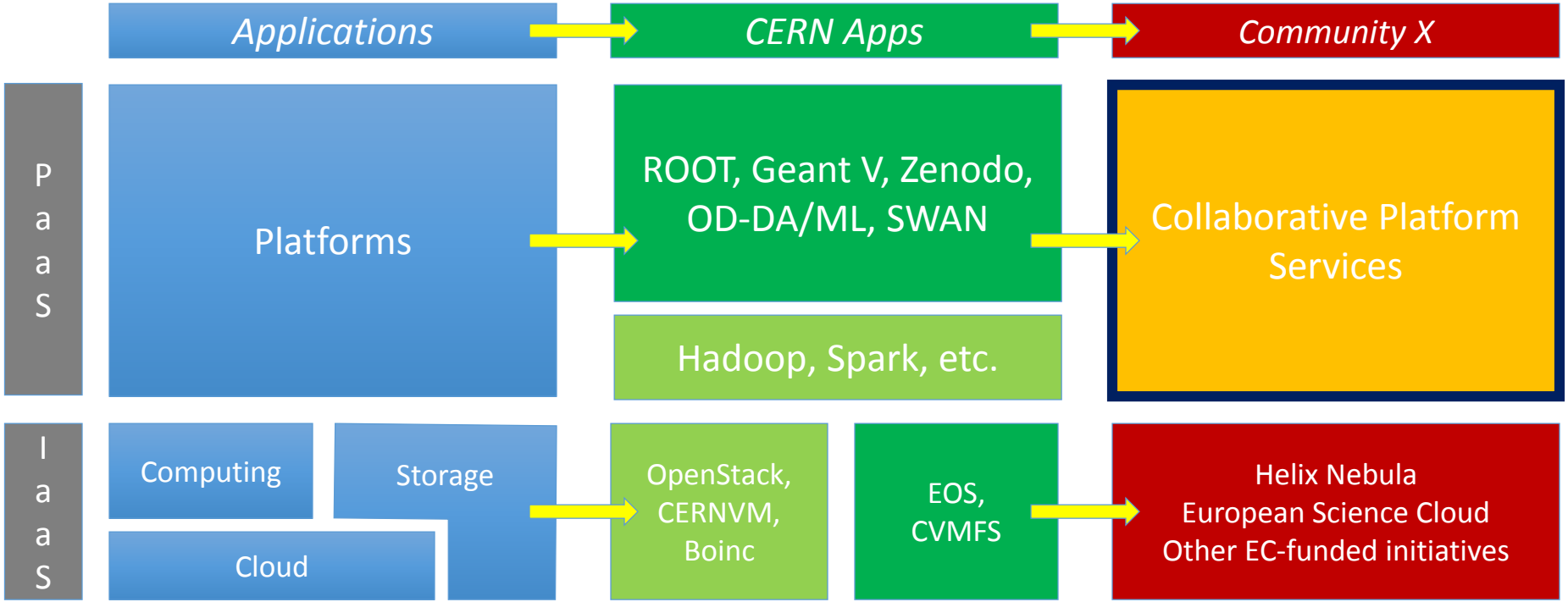
Sellers of fertilizers, equipment, furniture, electricity, water (Infrastructure providers)

Gardeners, botanists (Technical experts)

Platforms Layers



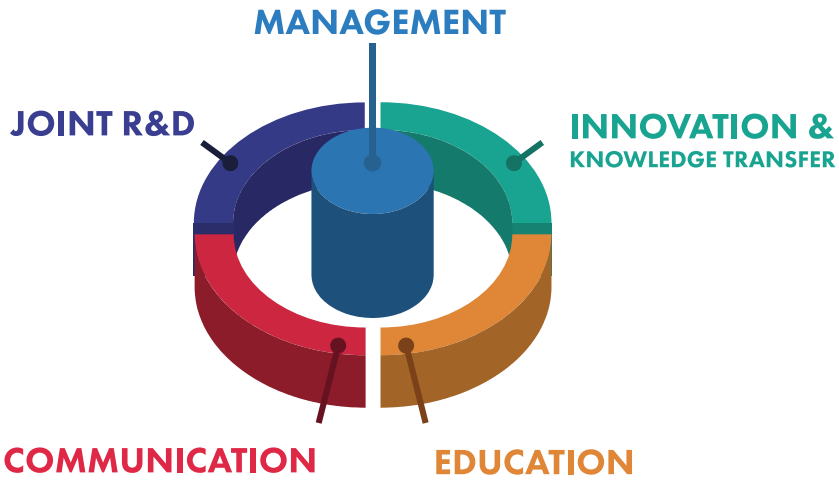
End-to-End Architecture



CERN OPENLAB'S MISSION

Our recipe for success

- **Evaluate** state-of-the-art technologies in a challenging environment and improve them.
- **Test** in a research environment today technologies that will be used in many business sectors tomorrow.
- **Train** the next generation of engineers/researchers.
- **Promote** education and cultural exchanges.
- **Communicate** results and reach new audiences.
- **Collaborate** and exchange ideas to create knowledge and innovation.



Founding Members:

CERN
King's College London (Dep. Twin R&GE)
SIDRA Research Centre (Qatar)
Intel

Gene@Scale

Design and implementation of an architecture-neutral, large-scale genomic research collaborative platform

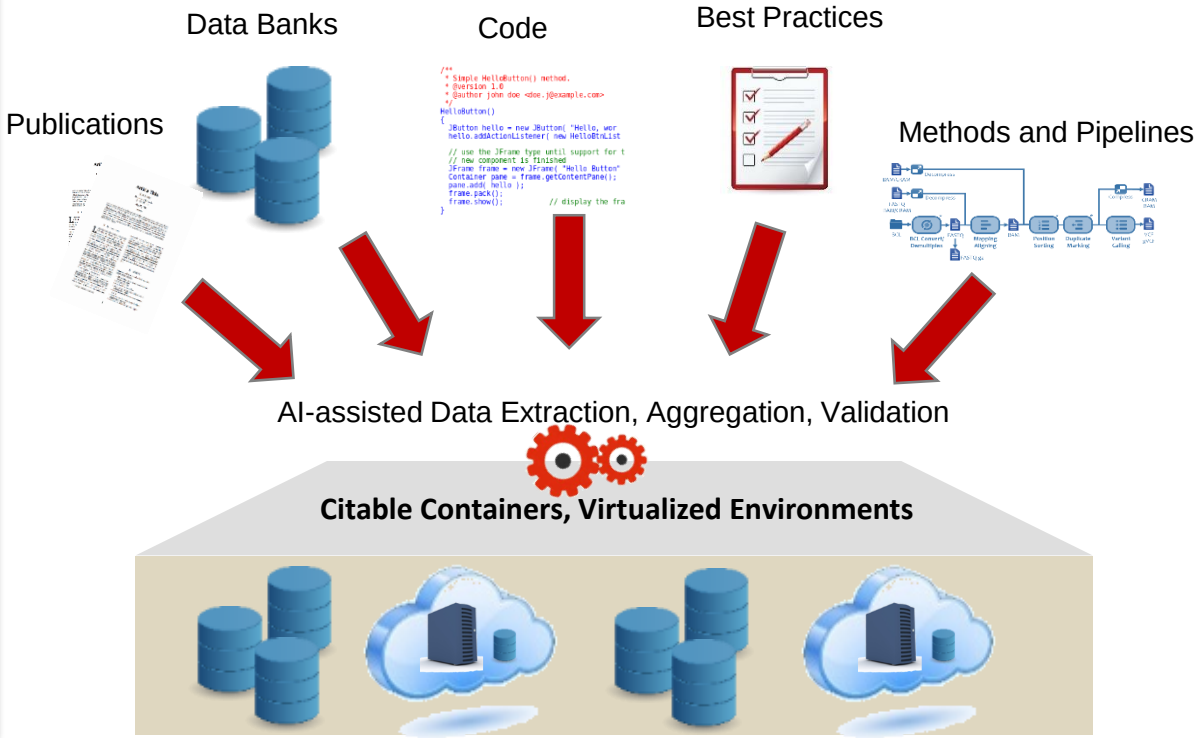
Cloud backend able to ingest and aggregate different types of data and formats, support for public and private data

Emphasis on ease of installation, reproducibility, sharing through citable containers and VEs

Best-of-Breed UX with support for AI-based data search and retrieval, text-mining

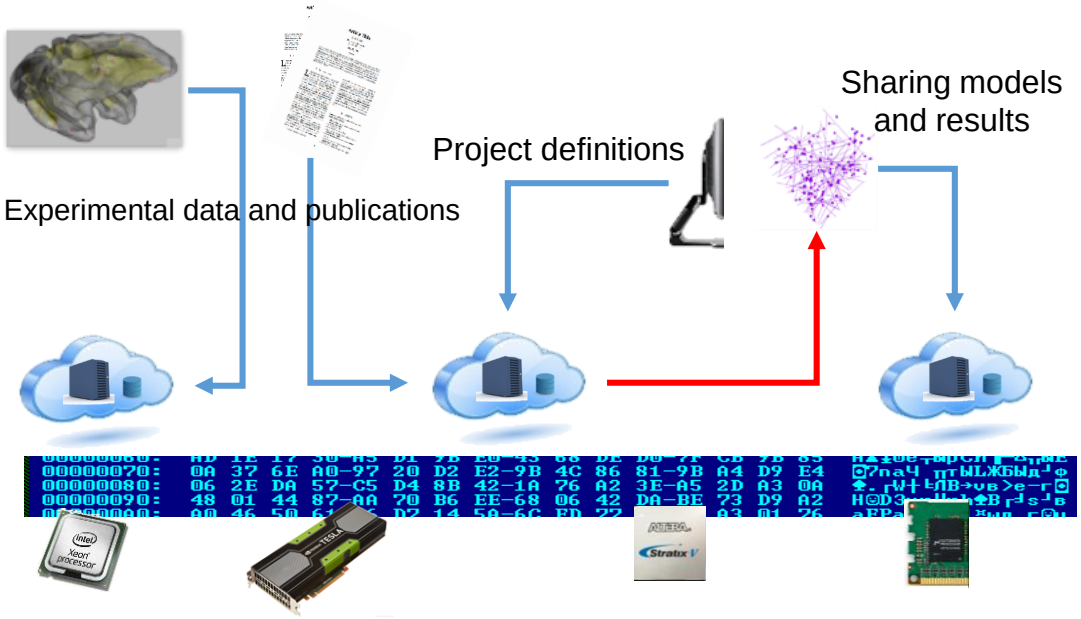
Current status: Prototype, focus on benchmarking of different tools for CNV discovery

Seeking collaborations: developers, domain experts, testers, use cases



Founding Members:

CERN
Newcastle University (Neuroinformatics Institute)
Innopolis University
Kazan University
Intel



BioDynaMo

Design and implementation of an architecture-neutral, large-scale biological development simulation platform

Cloud backend able to dynamically scale or shrink with the simulated model

Integration and sharing of experimental data (e.g. functional MRI images), textual publications, models, results

Cheaper to operate compared to more sophisticated HPC solutions and more expandable

Current status: Prototype, initial development of a software stack optimized for multi-core architectures

Seeking collaborations: developers, domain experts, testers, use cases

Founding Members:

CERN
Scimpulse Foundation
Intel

BIGHealth

Design and implementation of an architecture-neutral, large-scale systems biology platform

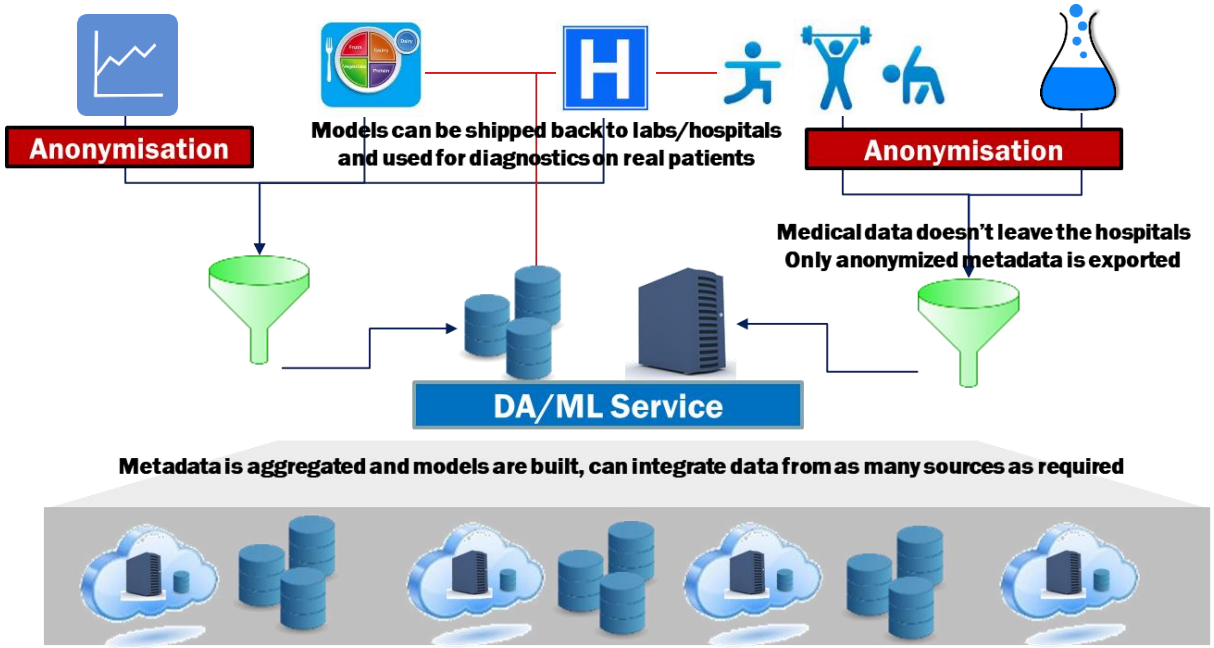
Cloud backend able to ingest different types of data and formats

Law-compliant, self-checking anonymization system

Cheaper to operate compared to more sophisticated HPC solutions and more expandable

Current status: Design

Seeking collaborations: developers, domain experts, testers, use cases



Conclusions

8 Views of Omi - #5. Returning Boats, Yabase – Utagawa Hiroshige, ca. 1834

Take-Away Messages

CERN and HEP have been pioneers of “big data”-based research

- Still many challenges ahead

Many other research communities facing similar scalability challenges

- Despite the differences, many valuable opportunities for collaboration and joint R&D exist

Industrial solutions can be used

- But usually scientific research needs are a few years ahead of what is possible today

Big Data has potential, but even more important is to make research more collaborative, reproducible and effective

QUESTIONS?

CONTACTS

ALBERTO DI MEGLIO
CERN openlab Head
alberto.di.meglio@cern.ch

MARIA GIRONE
CERN openlab CTO
maria.girone@cern.ch

FONS RADEMAKERS
CERN openlab CRO
fons.rademakers@cern.ch

ANDREW PURCELL
CERN openlab Communications Officer
andrew.purcell@cern.ch

KRISTINA GUNNE
CERN openlab Administration/Finance Officer
kristina.gunne@cern.ch

www.cern.ch/openlab