Challenges and opportunities for the realization of a federated, international life science data lake

Davide Salomoni (davide@infn.it)
INFN
LifeWatch Scientific Community Meeting
Rome, 27-29/5/2019

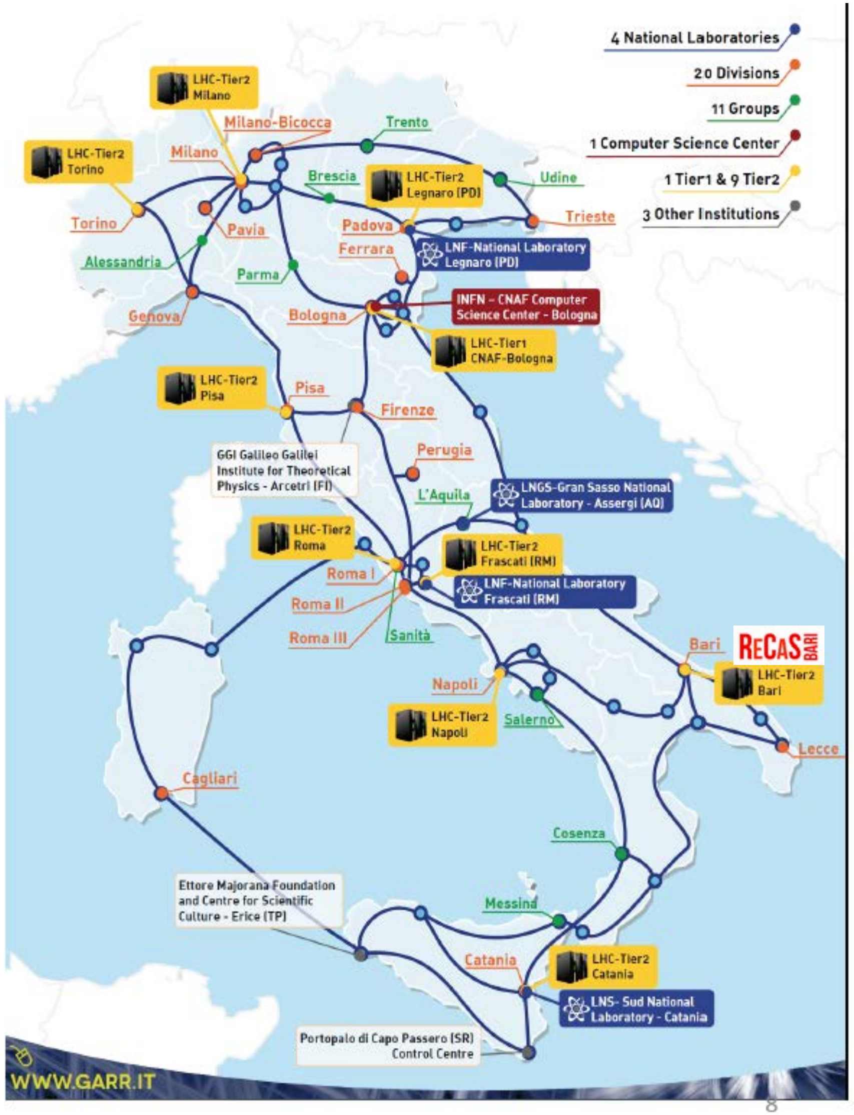
About me

- I am a physicist by education, but since 1991 playing with computers, networks, data and IT in general.
- I worked for several years abroad (USA, the Netherlands) in public and private companies. Since 2005 back in Italy at INFN.
- I am the coordinator or contributors to several Cloud or Big Data-related projects involving multi-disciplinary communities.
- I manage the Software Development & Distributed Systems group at the INFN National Computing Center, located in Bologna.

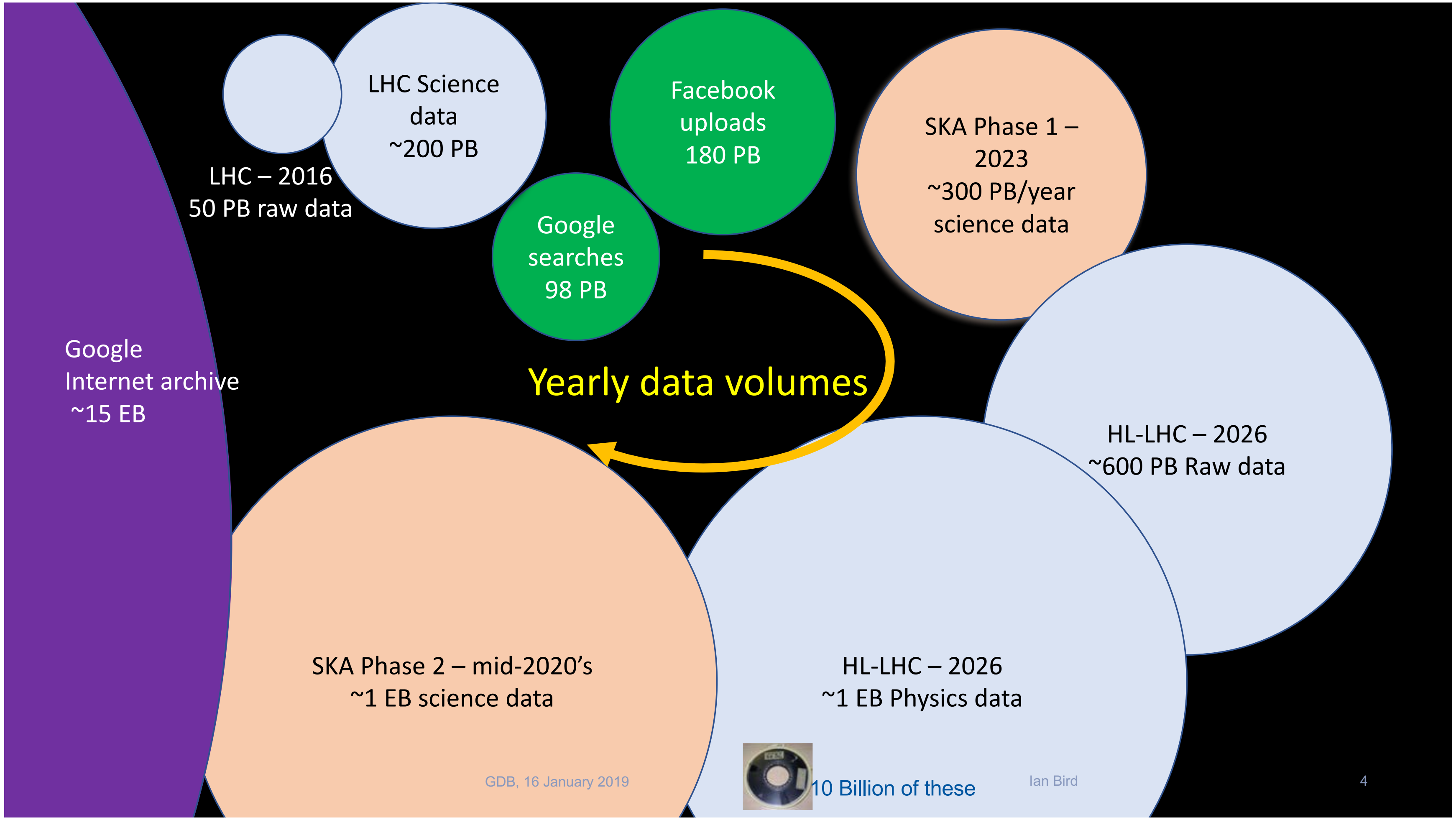
This presentation is my own view about what useful developments for our scientific stakeholders might be, for what regards distributed infrastructures applied to (for example) life science.

INFN (National Institute for Nuclear Physics) – www.infn.it

- A long tradition in **state-of-the-art distributed IT technologies**, from the first small clusters to Grid and Cloud-based computing.
- INFN is not interested in computing per-se, but as an essential way to **support its research and mission**.
- For the past 10 years, this mainly meant supporting the experiments @ CERN (LHC), although the scope is now widening very quickly to other communities.
- Currently, INFN operates:
  - 9 medium size centers (Tier-2s in the LHC Computing Grid lingo)
  - 1 large Tier-1 center, at CNAF (Bologna) – certified ISO-27001
- All the INFN centers are connected with 10-100 Gbit/s dedicated connections through the GARR network.
- Collectively, our main centers have about 65,000 CPU cores, 50PB of enterprise-level disk space, 60PB of tape storage.



Typical data volumes



LHC Science data ~200 PB

LHC – 2016 50 PB raw data

Facebook uploads 180 PB

Google searches 98 PB

SKA Phase 1 – 2023 ~300 PB/year science data

Google Internet archive ~15 EB

Yearly data volumes

HL-LHC – 2026 ~600 PB Raw data

SKA Phase 2 – mid-2020's ~1 EB science data

HL-LHC – 2026 ~1 EB Physics data

10 Billion of these

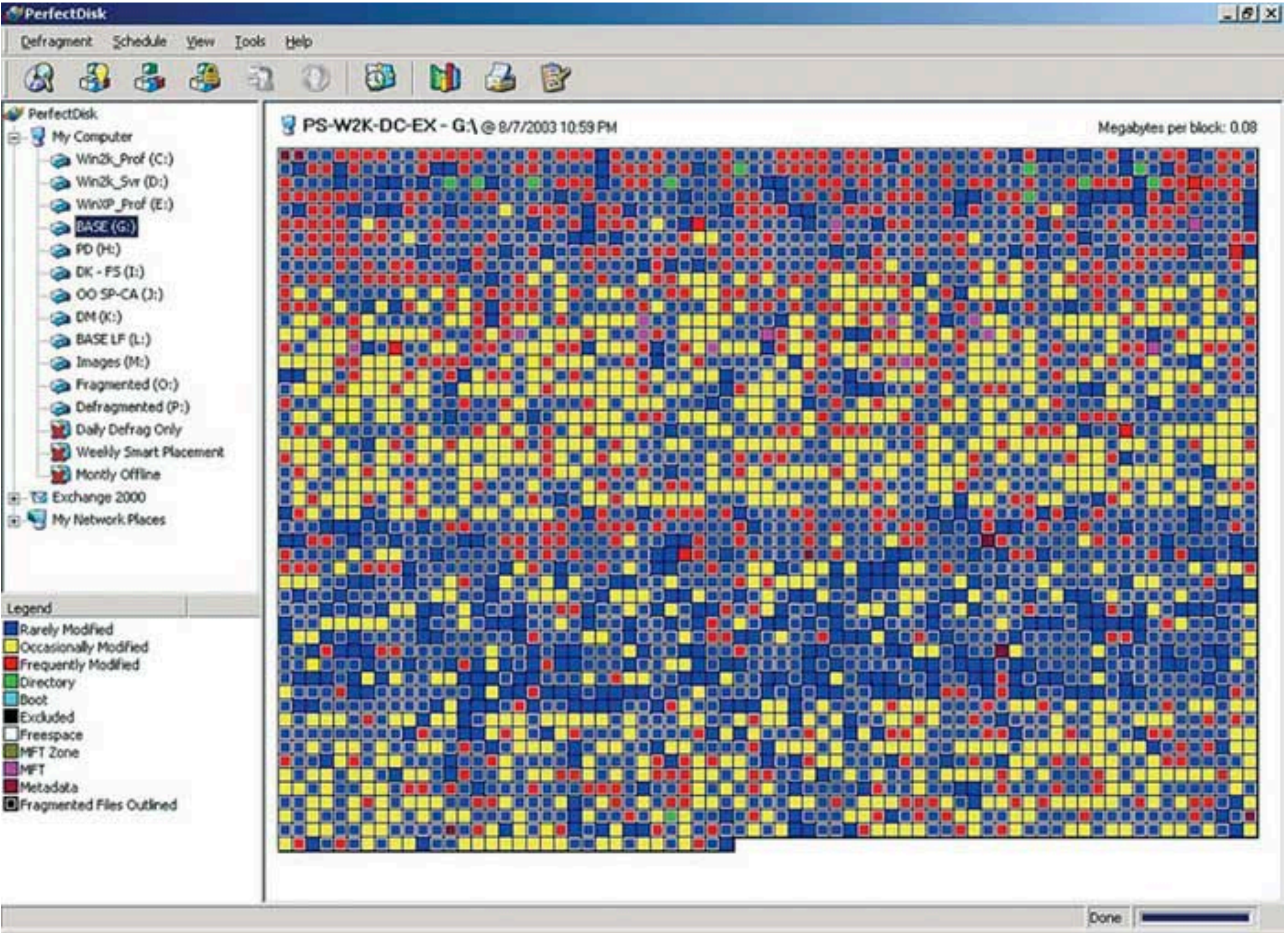
GDB, 16 January 2019

Ian Bird

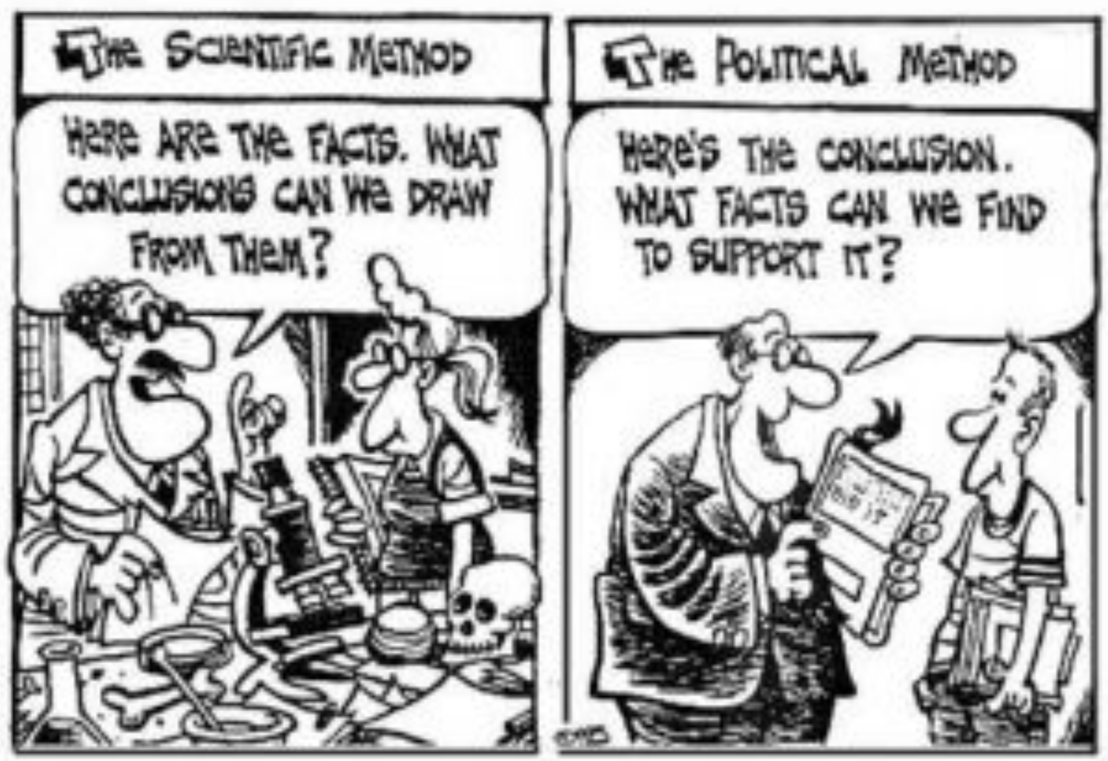
4

“A federated *data lake*”?

Challenge #1: fragmentation



Challenge #2: lose of focus


THE SCIENTIFIC METHOD
HERE ARE THE FACTS. WHAT CONCLUSIONS CAN WE DRAW FROM THEM?
THE POLITICAL METHOD
HERE'S THE CONCLUSION. WHAT FACTS CAN WE FIND TO SUPPORT IT?

Opportunities!

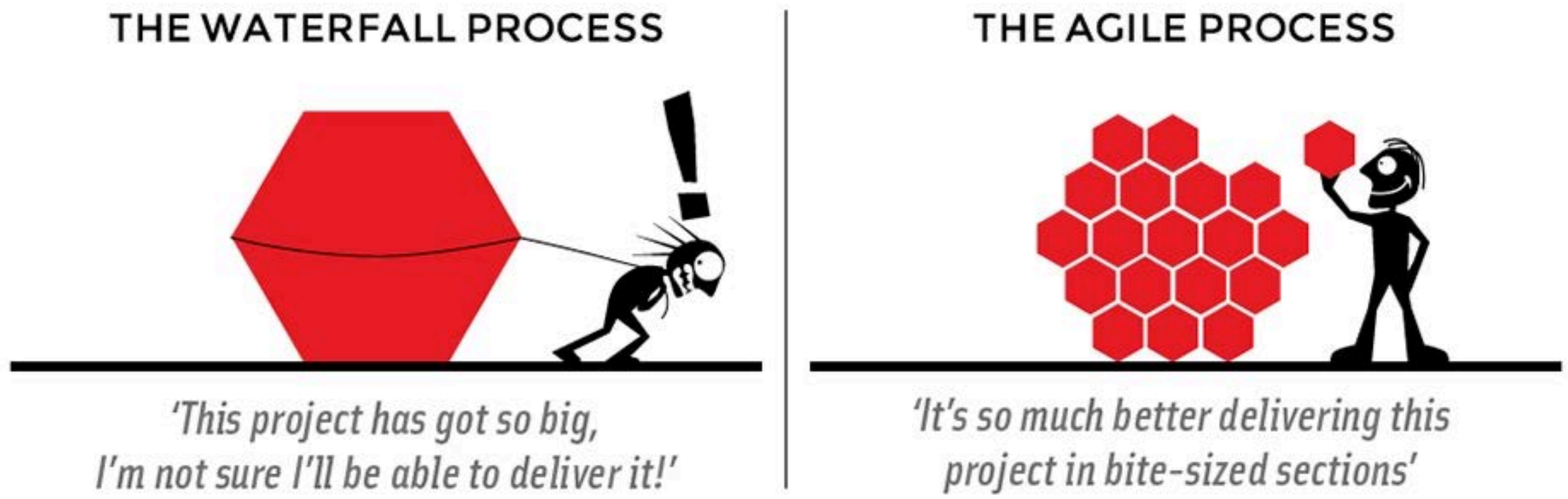
1. The impressive use cases that demand solutions that are not readily available in locked-in, smallish, general purpose infrastructures.
2. The multi-year experience that we as community have in delivering open, distributed computing solutions for science.
3. The technologies that make solutions concretely possible and their constant development.
4. Passion!

Old questions?

- Is this the usual debate about “**Research Infrastructures**” (community specific) vs. “**e-Infrastructures**” (general purpose, or at least theoretically so)?
- Or is it about a top-down vs. a bottom-up approach?

NO. Let’s focus on what our stakeholders are asking us instead:
Integrated Solutions

A general method


THE WATERFALL PROCESS
‘This project has got so big, I’m not sure I’ll be able to deliver it!’
THE AGILE PROCESS
‘It’s so much better delivering this project in bite-sized sections’

https://blog.ganttpro.com/en/waterfall-vs-agile-with-advantages-and-disadvantages/

How do we build a “scientific data lake”?

- … which must be unfragmented, agile, extensible, inclusive of existing solutions and know-how?

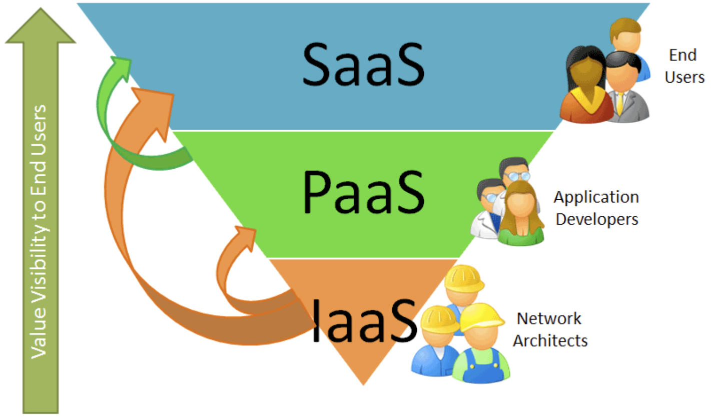
But, first of all, what is a ”data lake”?

Let’s tackle if from two angles:

1. **Service-oriented**: it is a **cloud of data services**, where open access and open science are key words.
2. **Support-oriented**: It is realized out of a **backbone** composed of a limited set of data centers offering resources and know-how.
   - In particular, I believe that know-how (not only money or political pressure) is a critical component to decree the success of a solution vs. its irrelevance.

The added value

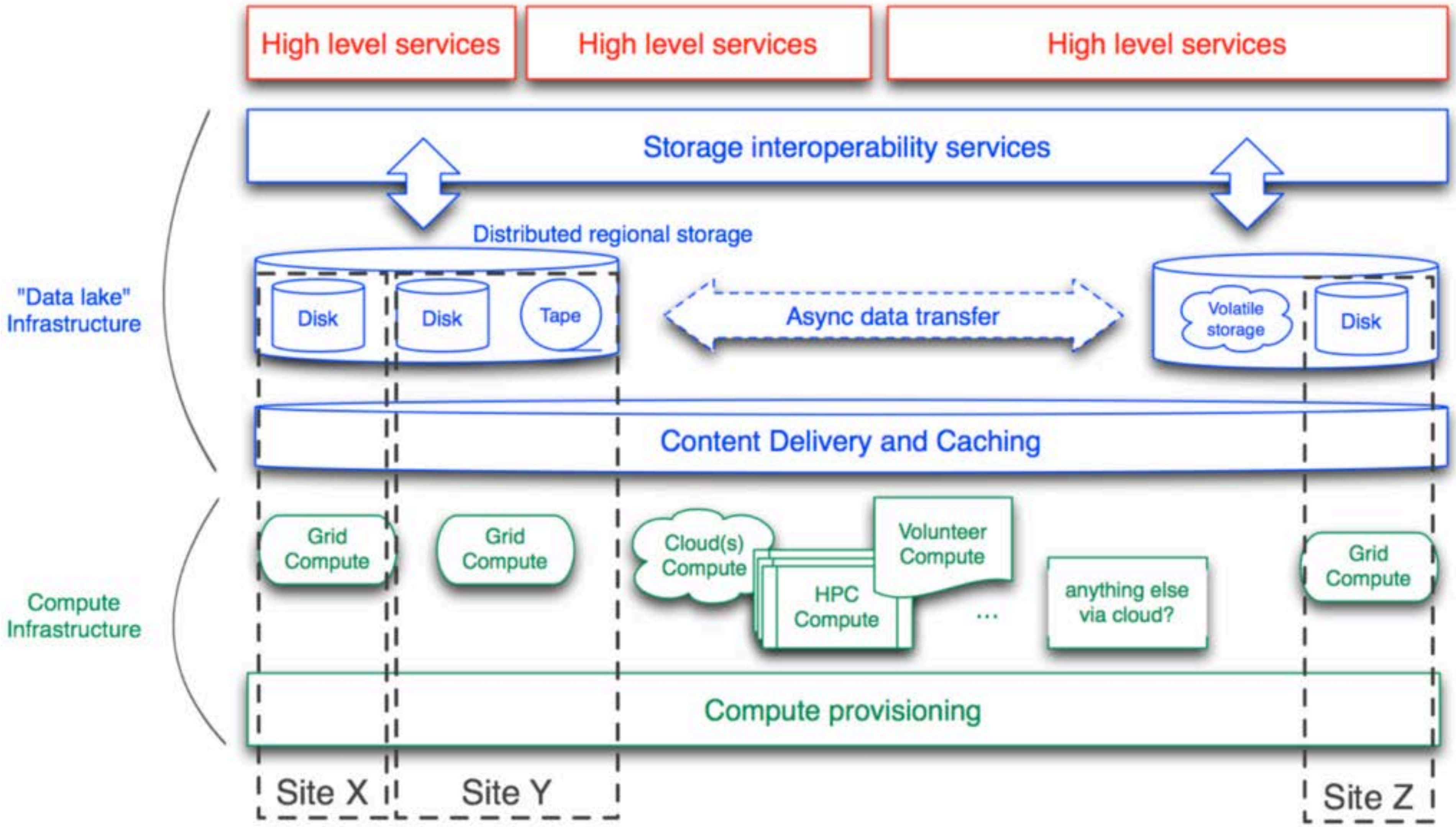
- Focus on **real added value solutions**, not on infrastructure or hardware – which nevertheless must be obviously available in some form.
- Focus on **bridging gaps across scientific domains and technology through open standards**, not on silos: research, education & agility.
- Focus on **progressive peer-to-peer agreements with larger entities** beyond the “lake”.



The water ecosystem as a planning model

- **The Pond**: a “single” center.
- **The Lake**: a backbone-centered federation centered on specific needs *but* using general solutions as much as possible.
- **The River**: the conduits to more general upper infrastructures.
- **The Sea**: a large, multi-purpose, many-stakeholders resource / solution set (e.g., the EOSC?)
- **The Ocean**: a worldwide collection of solutions.

Architecturally


High level services

High level services

High level services

"Data lake" Infrastructure

Storage interoperability services

Distributed regional storage

Disk

Disk

Tape

Async data transfer

Volatile storage

Disk

Content Delivery and Caching

Compute Infrastructure

Grid Compute

Grid Compute

Cloud(s) Compute

HPC Compute

Volunteer Compute

...

anything else via cloud?

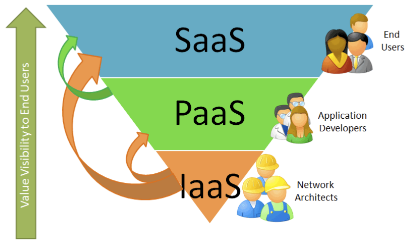
Grid Compute

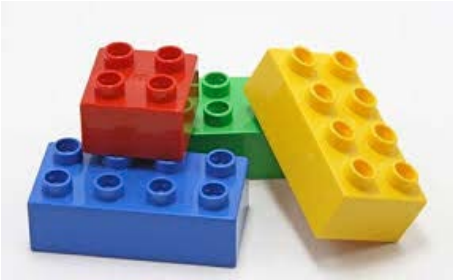
Compute provisioning

Site X

Site Y

Site Z





Some key technical points for the data lake (1)

- **Infrastructure as Code**: focus on problem-oriented, code-based dynamic solutions that program the infrastructure, not the other way around. We could also call this solution co-design.
- **Event-driven processing**, i.e. reaction to changes in data sets or in general to resource availability.
- **Intelligence as a Service,** e.g. Machine/Deep Learning as a Service, but also Competence Centers to analyze and build bespoke solutions.
- **Caching and linked data.** Compute & data locality is not guaranteed in data lake, therefore an effective content-delivery service is needed.
- **Service and data replication and data reproducibility** across the multiple backbone data centers.

Some key technical points for the data lake (2)

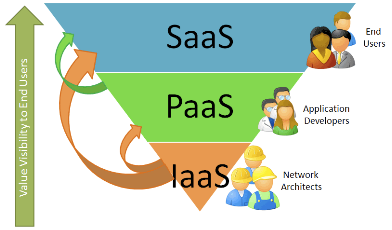
- **Data life cycle management**, including data QoS & effective data management plans.
- **User-level (or user-friendly) workflows** to overcome technology barriers for non IT-expert scientists.
- **Connection to multiple data sources** such e-infrastructures, HPC centers, opportunistic resources, devices, storage systems, data sets, sync & share services. This should happen through an open and technology neutral view of infrastructures.
- **Integration with smaller compute centers** or to **commercial providers** if advantageous.
- **Integration with national infrastructures**, and/or possibly also with super-national ones.

Some starting points for datalake-related solutions

- INDIGO-DataCloud, https://www.indigo-datacloud.eu
- DEEP-Hybrid DataCloud, https://deep-hybrid-datacloud.eu
- eXtreme-DataCloud, http://www.extreme-datacloud.eu
- ESCAPE, https://www.escape2020.eu

In summary

- It is naïve to think that, on the one hand, silos-based, proprietary, solutions and, on the other hand, general purpose monoliths with too many stakeholders will address the required solutions to handle the explosion of data production and related analysis.
- Transparency, support of *de jure* and *de facto* standards, provider-agnostic modular solutions are the way to go.
- We have the know-how and the technology building blocks to define and build a federated data lake model using open solutions.



- This data lake should be focused on creating a backbone of data services which will be used to satisfy concrete use cases and connecting via peer agreements with other resources, if useful.
- I know, I did not talk here about sustainability, optimization of resource usage, economies of scale, joint procurements, etc.! ☺

Thanks for listening – time for questions!

(now or later; my email: davide@infn.it)