



WLCG – **Big** Data, **Open** Data

Visit from IHEP

13 June 2014

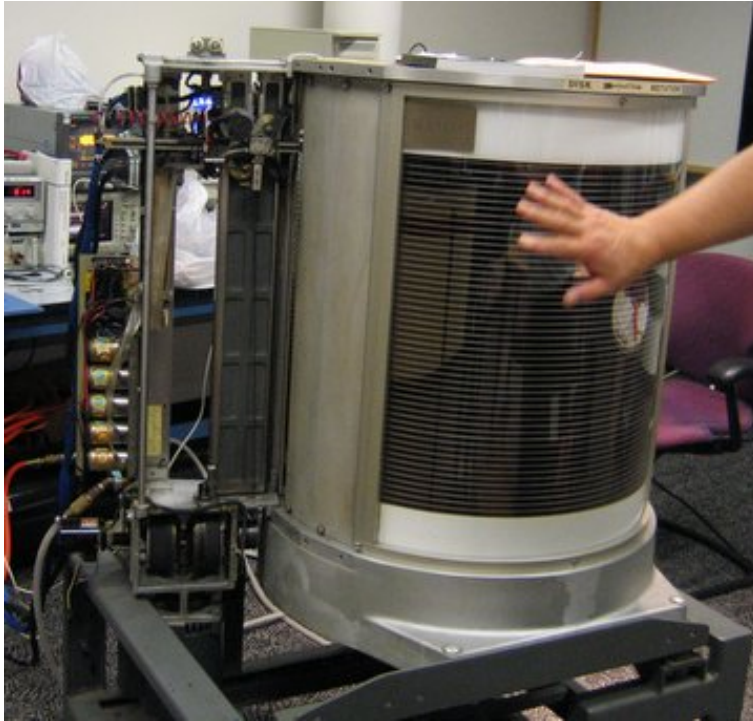
Jamie.Shiers@cern.ch

Project Manager, HEP Data Preservation (DPHEP)

Introduction

- CERN stores around **100,000 TB** (100PB) of physics data > 2/3 of which is from the LHC
- This will grow to around **5EB** (5,000,000 TB) and beyond in the next 1-2 decades
- Aside from its direct **scientific** potential, it also has significant value for **educational** purposes
- We need to **preserve** it – and the ability to (re-)use it, now and in the long-term
- This is now both **affordable** as well as **technically possible**

IBM 350 RAMAC



1956, 5 Mch, 8 Kch/s IO

PDP DECtape



1970, 144K 18_bit words

Options

- Ignore problem: we'd like to but....

~300K tapes were 'archived'... .. ~150K were manually mounted....



Before migration exercise of 1992/1993

..and then copied to Redwoods....



After migration exercise of 1992/1993

Why build an LHC?

THE STANDARD MODEL

	Fermions			Bosons	
Quarks	u up	c charm	t top	γ photon	Force carriers
	d down	s strange	b bottom	Z Z boson	
Leptons	ν_e electron neutrino	ν_μ muon neutrino	ν_τ tau neutrino	W W boson	
	e electron	μ muon	τ tau	g gluon	

Higgs^{*}
boson

***Yet to be confirmed**

Source: AAAS

BEFORE!





The diagram illustrates the data flow architecture at CERN. It is divided into two main horizontal sections. The top section, representing the surface, shows the 'CERN Computer Centre' with several server racks. A network of dotted lines connects these racks to various experimental sites located underground. The bottom section, representing the underground area, shows the paths of four major experiments: LHCb, ATLAS, CMS, and ALICE. Each experiment is connected to the surface network by a vertical line of dots, representing a data link. The data flow rates for each experiment are specified in pink boxes. The overall background features a stylized landscape with mountains and a lake.

Data flow to permanent storage: 4-6 GB/sec

CERN Computer Centre

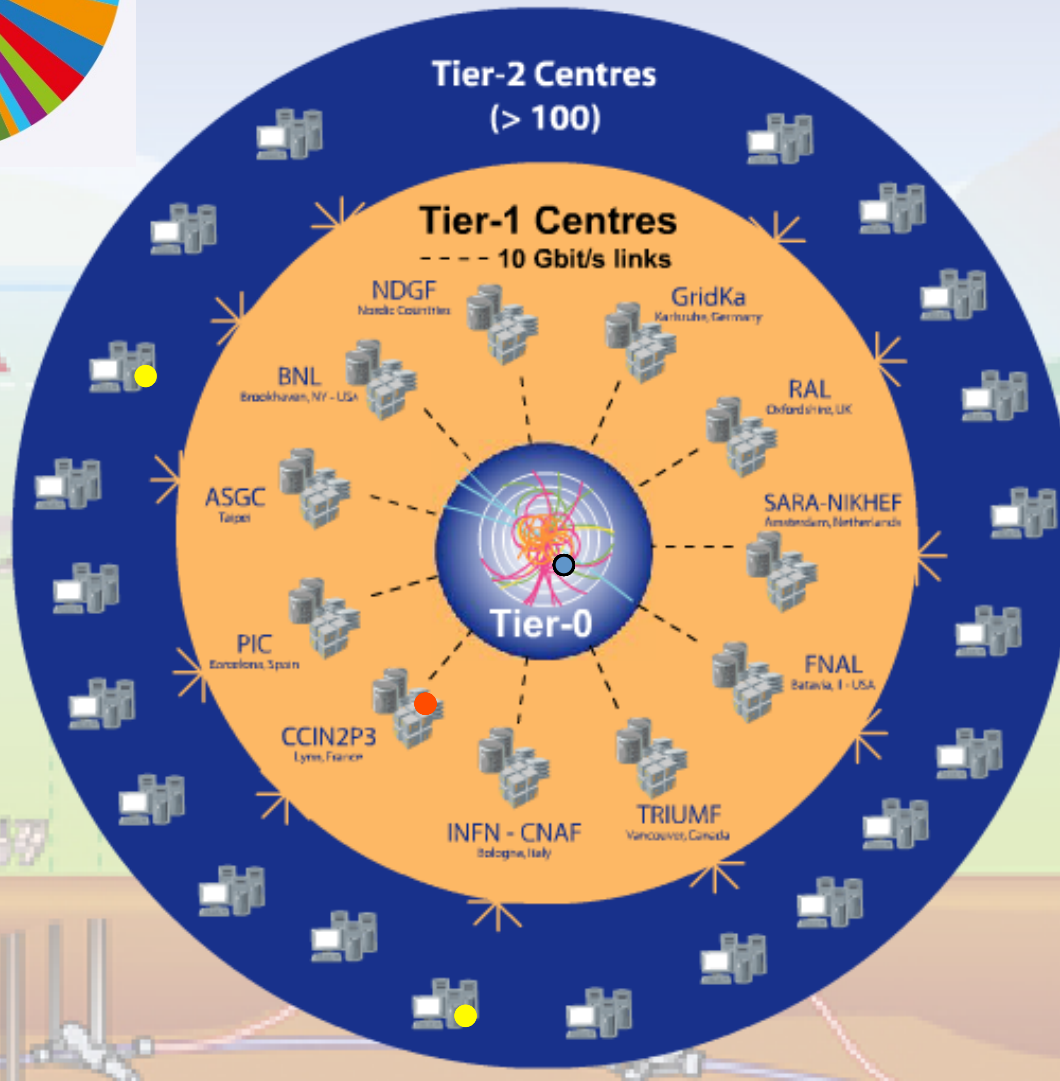
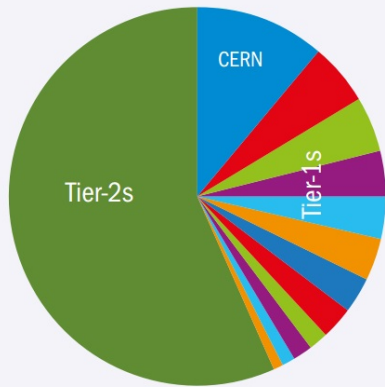
LHCb ~ 200-400 MB/sec

ATLAS ~ 1-2 GB/sec

ALICE ~ 1.25 GB/sec

CMS ~ 1-2 GB/sec

Tier 0 – Tier 1 – Tier 2



Tier-0 (CERN):

- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):

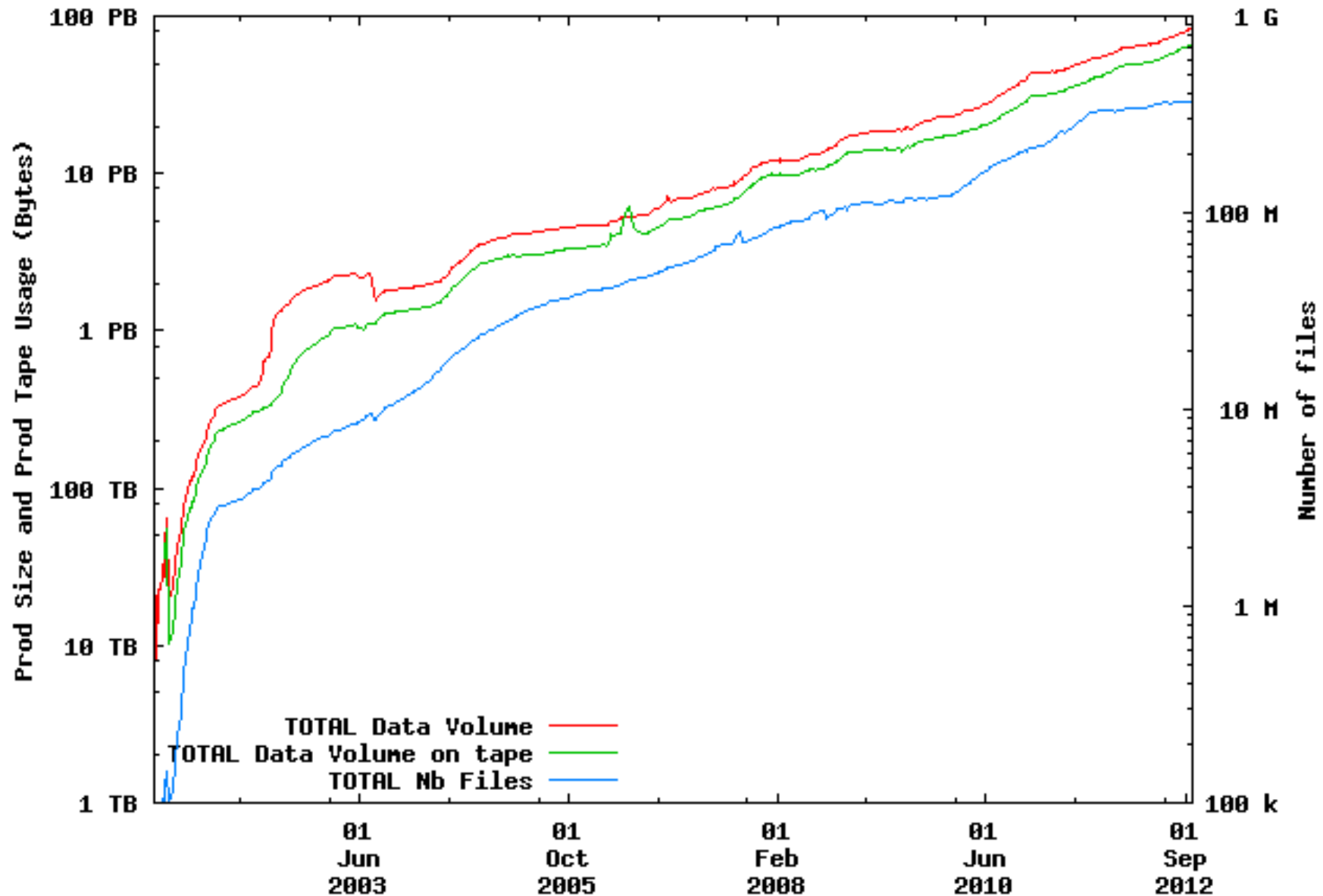
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):

- Simulation
- End-user analysis

CERN has ~100 PB archive

Experiments Production Data in CASTOR



Generated Sep 25, 2012 CASTOR (c) CERN/IT

LHC Data – Access Policies

Level (standard notation)	Access Policy
L0 (raw) (cf “Tier”)	Restricted even internally
L1 (1 st processing)	Large fraction available after “embargo” (validation) period
L2 (analysis level)	Specific (meaningful) samples for educational outreach: pilot project(s) on-going
L3 (publications)	Open Access (CERN policy)

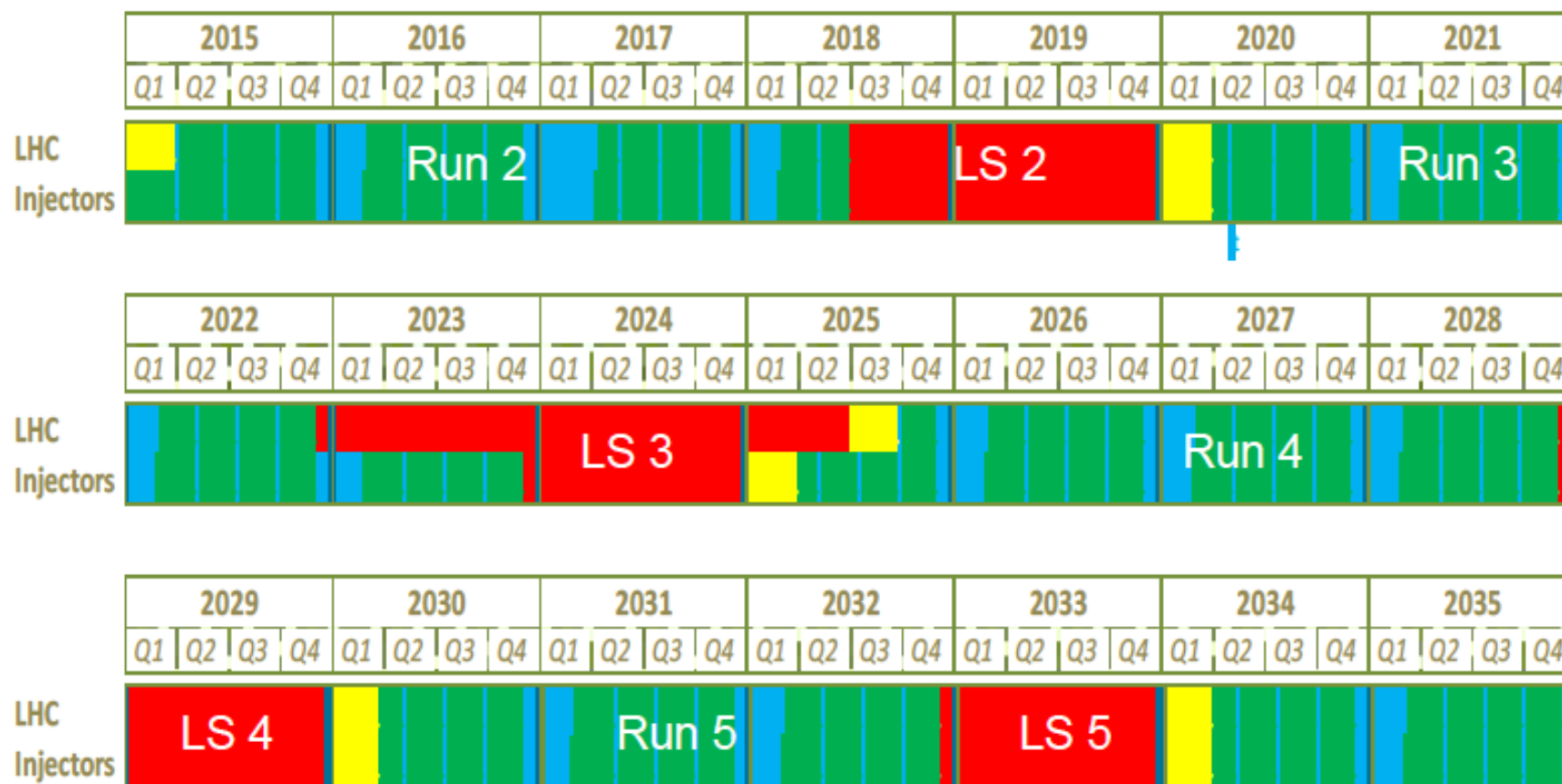
LHC schedule beyond LS1

Only EYETS (19 weeks) (no Linac4 connection during Run2)

LS2 starting in 2018 (July) 18 months + 3months BC (Beam Commissioning)

LS3 LHC: starting in 2023 => 30 months + 3 BC

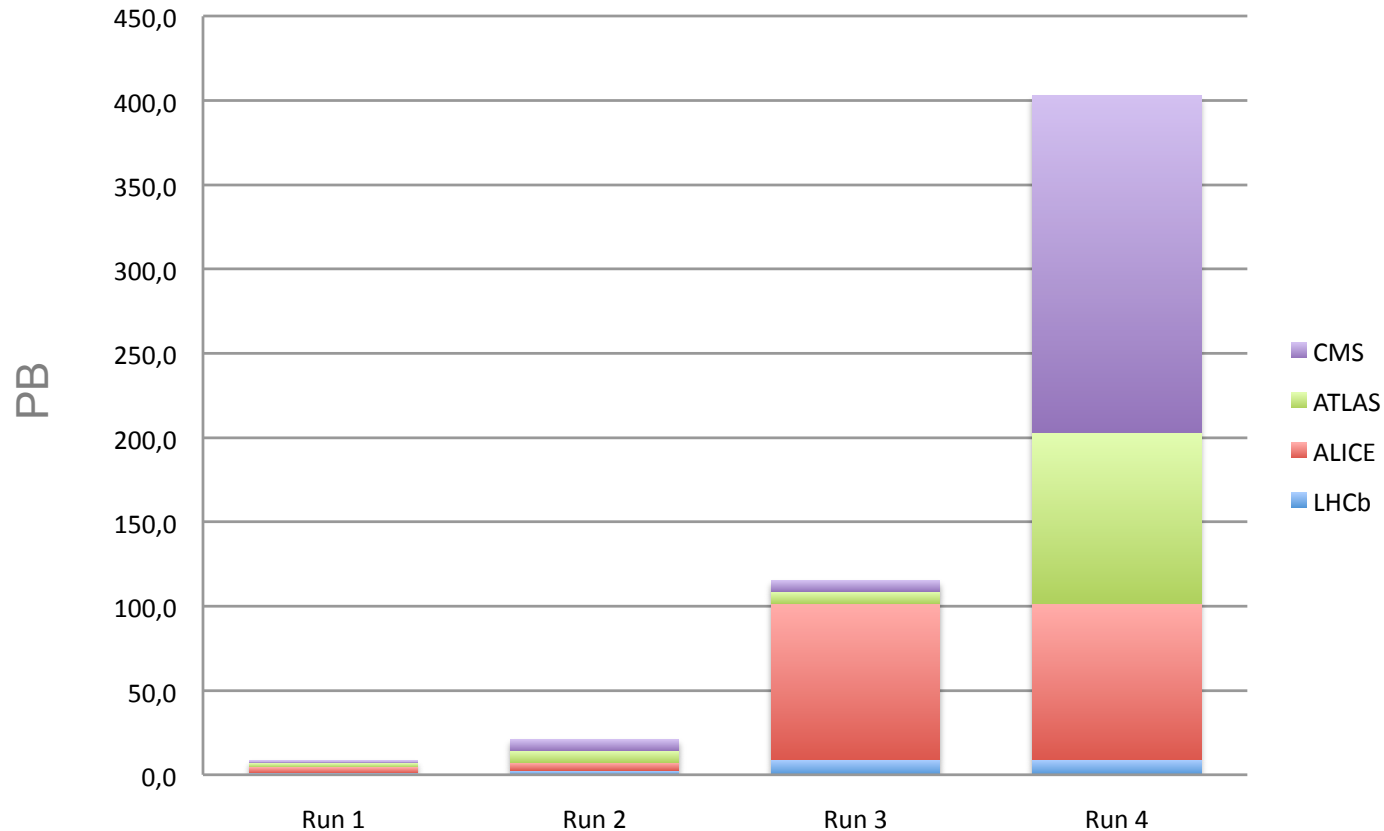
injectors: in 2024 => 13 months + 3 BC



LHC schedule approved by CERN management and LHC experiments spokespersons and technical coordinators
Monday 2nd December 2013



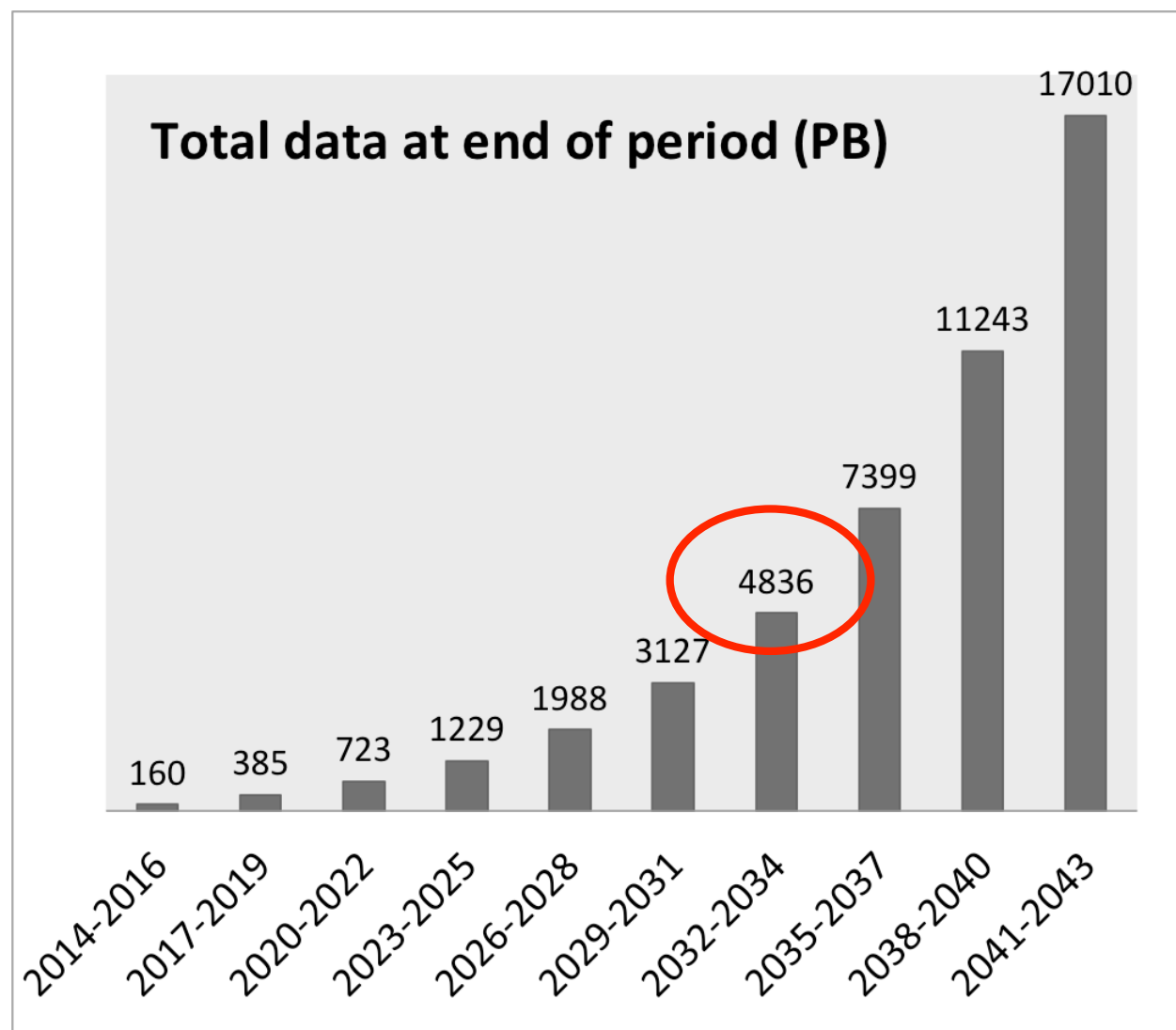
Data: Outlook for HL-LHC

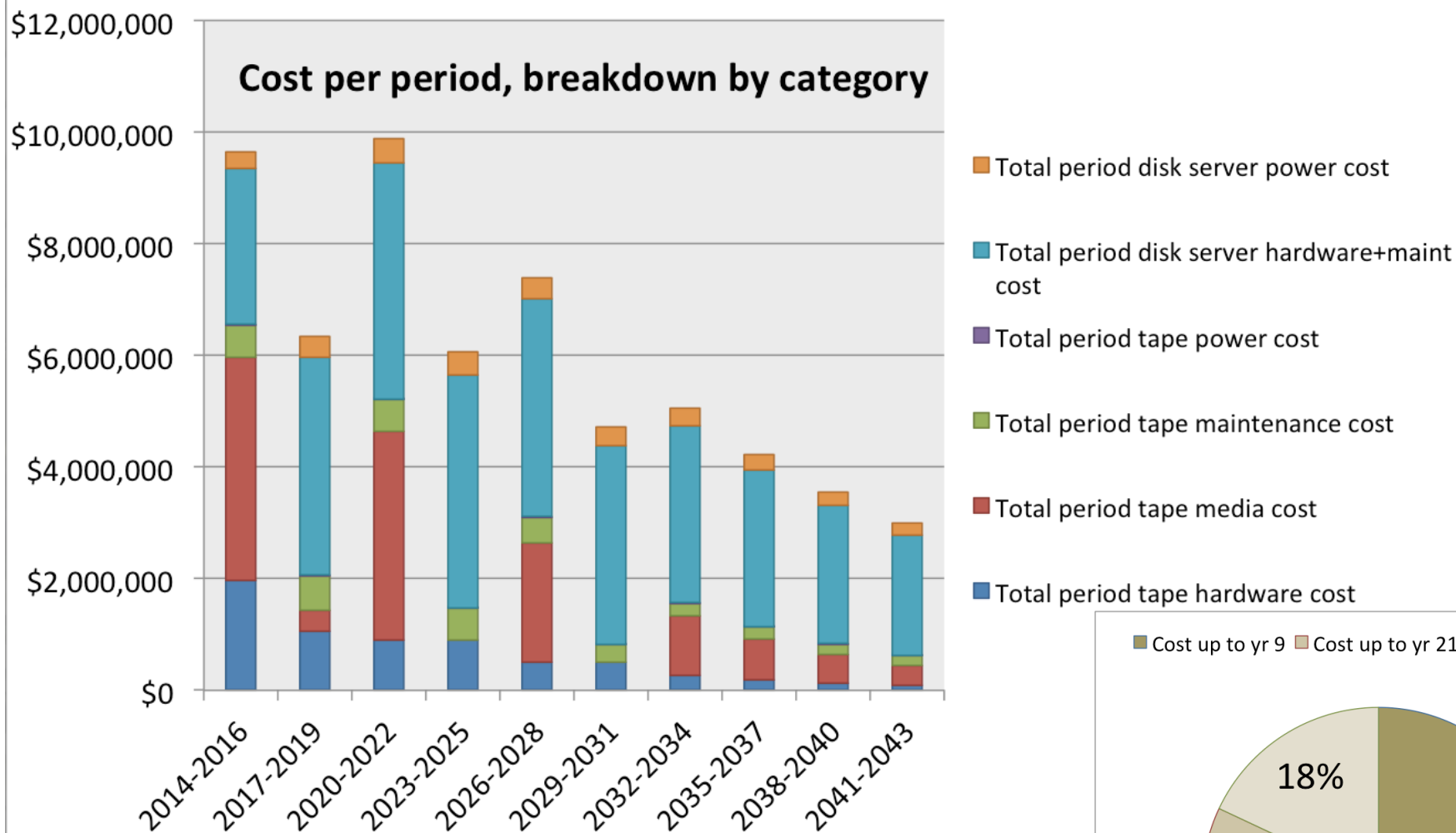


- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
 - To be added: derived data (ESD, AOD), simulation, user data...

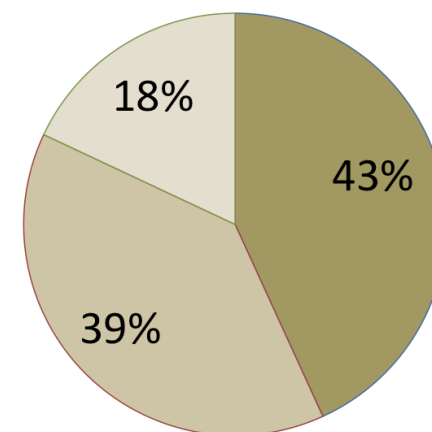
➤ **0.5 EB / year is probably an under estimate!**

Start with 10PB, then +50PB/year, then +50% every 3y (or +15% / year)

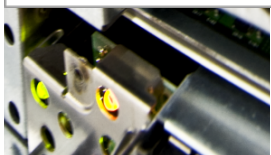




Cost up to yr 9 Cost up to yr 21 Cost up to yr 30



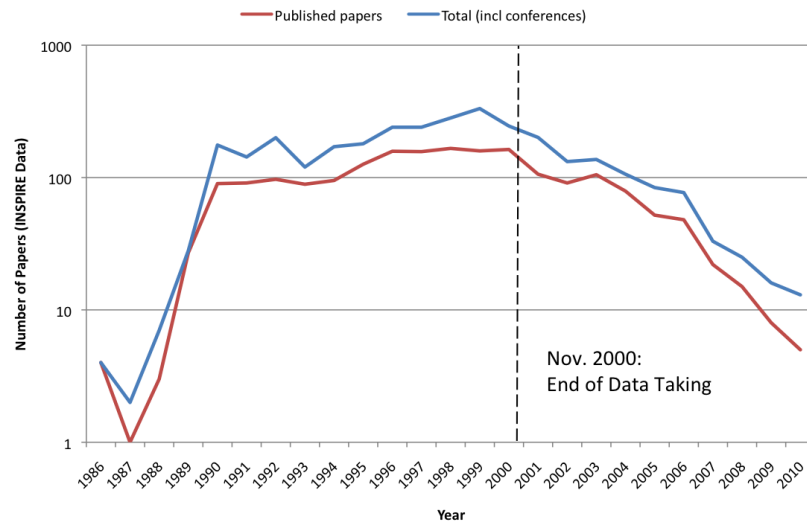
Total cost: ~59.9M\$
(~2M\$ / year)



2020 Vision for LT DP in HEP

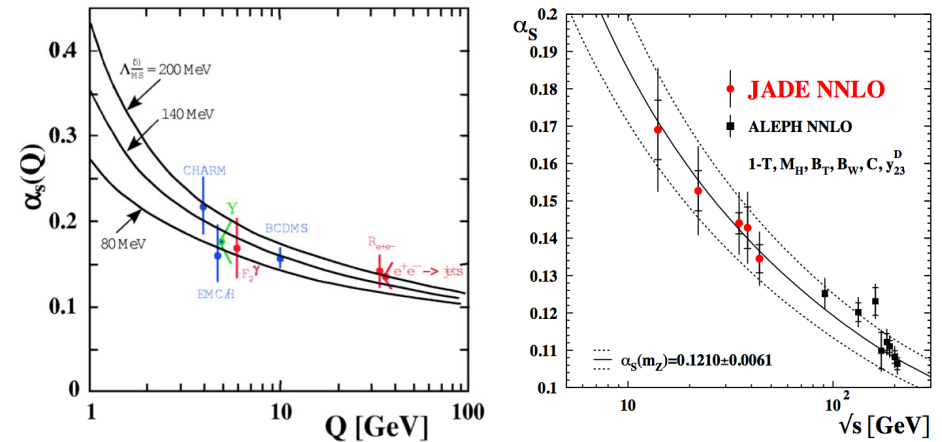
- Long-term – e.g. LC timescales: *disruptive change*
 - By 2020, all archived data – e.g. that described in DPHEP Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) access policies and possibilities to annotate further
 - Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards
 - **DPHEP portal**, through which data / tools accessed
- **Agree with Funding Agencies clear targets & metrics**

1 – Long Tail of Papers



3

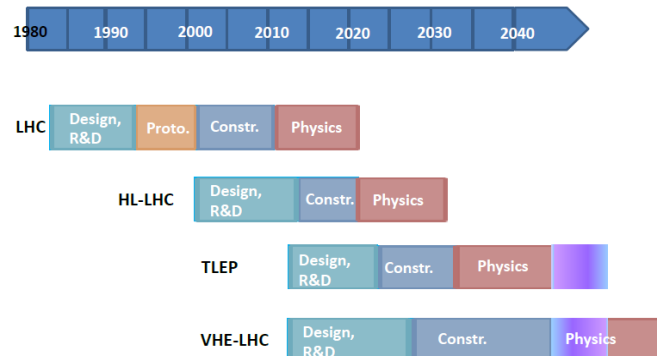
2 – New Theoretical Insights



4

3 – “Discovery” to “Precision”

possible long-term time line



Use Case Summary

1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

**Volume: 100PB + ~50PB/year
(+400PB/year from 2020)**

Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the Office of Science for research funding are required to include a Data Management Plan (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be shared and preserved or explains why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.
- Similar requirements from European FAs and EU (H2020)

1. DPHEP Portal

2. **Digital library** tools (Invenio) & services (CDS, INSPIRE, ZENODO) + related tools (HepData, RIVET, ...)
3. **Sustainable software**, coupled with advanced **virtualization** techniques, “snap-shotting” and **validation** frameworks
4. Proven bit preservation at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50
5. Open Data (“Open everything”)

DPHEP Portal – Zenodo like?

The screenshot displays the Zenodo.org website interface. At the top, there is a browser window with multiple tabs open, including 'D4.3_Quality_Tr...', 'EINFRA-1-2014', 'EINFRA-9-2015', 'INFRADEV-4-201...', 'PartBTemplate', 'Knowledge Exch...', 'DPHEP workshop', 'BNL www.bnl.gov/lhc', and 'ZENODO'. The address bar shows 'zenodo.org'. Below the browser window, the Zenodo homepage is visible. It features a search bar with a 'Search' button. A 'Filter by types' section is present. The 'Recent Uploads' section lists three items:

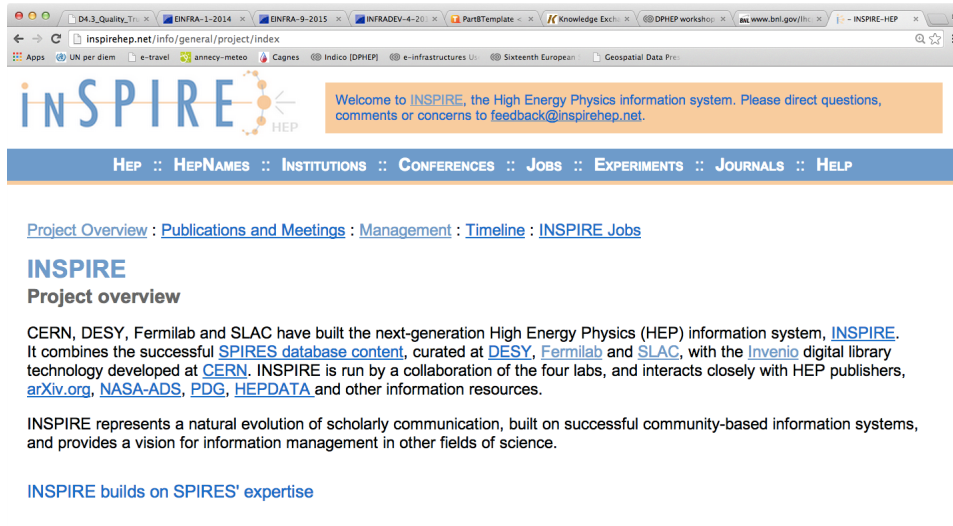
- 01 April 2014** | Journal article | Open access | [View](#)
The e-book phenomenon: a disruptive technology
Tom D. Wilson
The emergence of the e-book as a major phenomenon in the publishing industry is of interest, world-wide. The English language market, with Amazon.com as the major player in the market may have dominated attention, but the e-book has implications for many ...
- 10 November 2010** | Thesis | Open access | [View](#)
Χρόνιες Επιδράσεις του Καπνίσματος στη Λειτουργική Ικανότητα του Κυκλοφορικού Συστήματος Νεαρών Υγείων Ατόμων
Papathansiou, George ; Evangelou, Angelos
Εισαγωγή Το κάπνισμα αποτελεί τον σοβαρότερο ίσως παράγοντα κινδύνου μελλοντικής καρδιοαγγειακής νοσηρότητας και θνητότητας ενώ θεωρείται ως η κυριότερη αντιστρεπτή αιτία θανάτου. Το κάπνισμα συνδέεται με χρονότρωση καθυστέρηση λόγω δυσλειτουργίας του ...
Uploaded by [George](#) on 30 March 2014.
- 30 March 2014** | Report | Open access | [View](#)
Archaeobotanical remains from Mitchelstown and Ballynamona

On the right side, there is a 'GitHub integration' section with a GitHub logo and text: 'Want to preview the public beta of GitHub integration? Just [Sign In](#) with your GitHub account and [click here](#).' Below that is a 'New to ZENODO?' section with a list of features:

- Research. Shared. — all research outputs from across all fields of science are welcome!
- Citeable. Discoverable. — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- Community Collections — accept or reject uploads to your own community collections (e.g workshops, EU projects or your complete own digital repository).
- Funding — integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- Flexible licensing — because not everything is under Creative Commons.
- Safe — your research output is stored safely for the future in same cloud infrastructure as research data from CERN's Large Hadron Collider.
- DropBox integration — upload files straight from your DropBox.

Documentation projects with INSPIREHEP.net

- Internal notes from all HERA experiments now available on INSPIRE
 - A collaborative effort to provide “consistent” documentation across all HEP experiments – starting with those at CERN – as from 2015
 - (Often done in an inconsistent and/or ad-hoc way, particularly for older experiments)



The screenshot shows the INSPIRE website homepage. At the top, there's a navigation bar with links: HEP, HEPNames, INSTITUTIONS, CONFERENCES, JOBS, EXPERIMENTS, JOURNALS, and HELP. Below this is a section titled "Project Overview" with links to Publications and Meetings, Management, Timeline, and INSPIRE Jobs. The main content area is titled "INSPIRE Project overview" and contains text about the system's development by CERN, DESY, Fermilab, and SLAC, and its integration with the SPIRES database and Invenio digital library. It also mentions that INSPIRE represents a natural evolution of scholarly communication and provides a vision for information management in other fields of science. At the bottom, it states "INSPIRE builds on SPIRES' expertise".

ZEUS Internal Notes

10 records found

1. Inclusive-jet production in NC DIS with HERA II.

J. Tarron C. Glasman. ZEUS-IN-09-004.

[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)

[Detailed record](#) - [Similar records](#)

2. Three-subjet distributions in neutral current deep inelastic scattering.

E. Ron C. Glasman, J. Tarron. ZEUS-IN-09-003.

[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)

[Detailed record](#) - [Similar records](#)

3. 2009 Guide to Funnel: The ZEUS Monte Carlo Production Facility.

A. Parenti. ZEUS-IN-09-002.

[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)

[Detailed record](#) - [Similar records](#)

4. Automated calculation of radiative correction to electron-proton charged current DIS at HERA.

I. Marfin. ZEUS-IN-09-001.

[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)

[Detailed record](#) - [Similar records](#)





The Guidelines 2014-2015

Guidelines Relating to Data Producers:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.
2. The data producer provides the data in formats recommended by the data repository.
3. The data producer provides the data together with the metadata requested by the data repository.



Guidelines Related to Repositories (4-8):

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.
5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.
6. The data repository applies documented processes and procedures for managing data storage.
7. The data repository has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit work flows across the data life cycle.



Guidelines Related to Repositories (9-13):

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.
10. The data repository enables the users to discover and use the data and refer to them in a persistent way.
11. The data repository ensures the integrity of the digital objects and the metadata.
12. The data repository ensures the authenticity of the digital objects and the metadata.
13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.



Guidelines Related to Data Consumers (14-16):

- 14. The data consumer complies with access regulations set by the data repository.
- 15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.
- 16. The data consumer respects the applicable licences of the data repository regarding the use of the data.



DSA self-assessment & peer review

- Complete a self-assessment in the [DSA online tool](#). The online tool takes you through the 16 [guidelines](#) and provides you with support
- Submit self-assessment for peer review. The peer reviewers will go over your answers and documentation
- Your self-assessment and review will not become public until the DSA is awarded.
- After the DSA is awarded by the Board, the DSA logo may be displayed on the repository's Web site with a link to the organization's assessment.

Summary

1. DPHEP portal: build in collaboration with other disciplines, using existing, sustainable technologies
2. Digital libraries: continue existing collaborations
3. Sustainable “bit preservation” – certified repositories supplemented by HEP “best practices” – with resources reviewed along with other aspects of LHC programme
4. “Knowledge capture & preservation”: *still an area for (significant?) improvement*
5. Open “Big Data”: key to unlocking long-term re-use

Conclusions

- *We need to work together:*
 - Funding agencies, governments, policy makers, technology and service providers ...
 - As well as multiple disciplines (sciences, arts & humanities, e-government etc.)
- *to deliver sustainable services and solutions for long-term data and knowledge preservation*



Data Sharing in Time & Space

Challenges, Opportunities and Solutions(?)

Jamie.Shiers@cern.ch

Workshop on Best Practices for Data
Management & Sharing



International Collaboration for Data Preservation and
Long Term Analysis in High Energy Physics



Science & Technology
Facilities Council

Cultural, Economic and Societal Impacts of big science projects

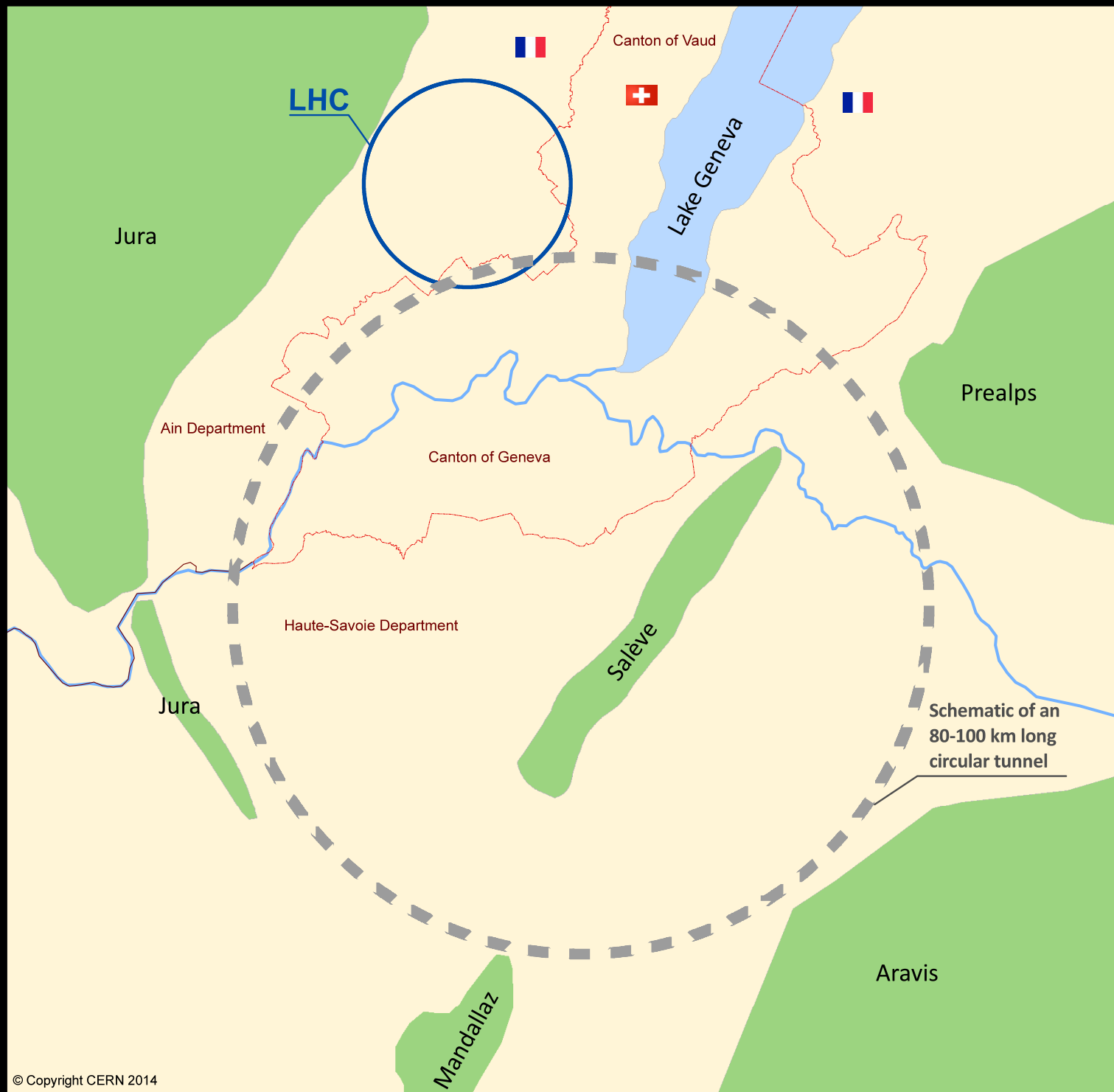
John Womersley

Chief Executive

Science and Technology Facilities Council

[Selected slides from Future Circular Colliders workshop]

10 February 2014



☒ Science case

Convince me that this project is scientifically excellent

☒ Project Plan

Convince me that you know what you are doing:
scope, costs and schedule are under control

☒ “Business case ”

Convince me that this is a good use of public money



Science & Technology
Facilities Council

TevatronImpact

A symposium celebrating extraordinary contributions to science, technology & society

June 11, 2012

Ramsey Auditorium

Fermilab

Batavia, Illinois, USA

1:00 p.m. Symposium

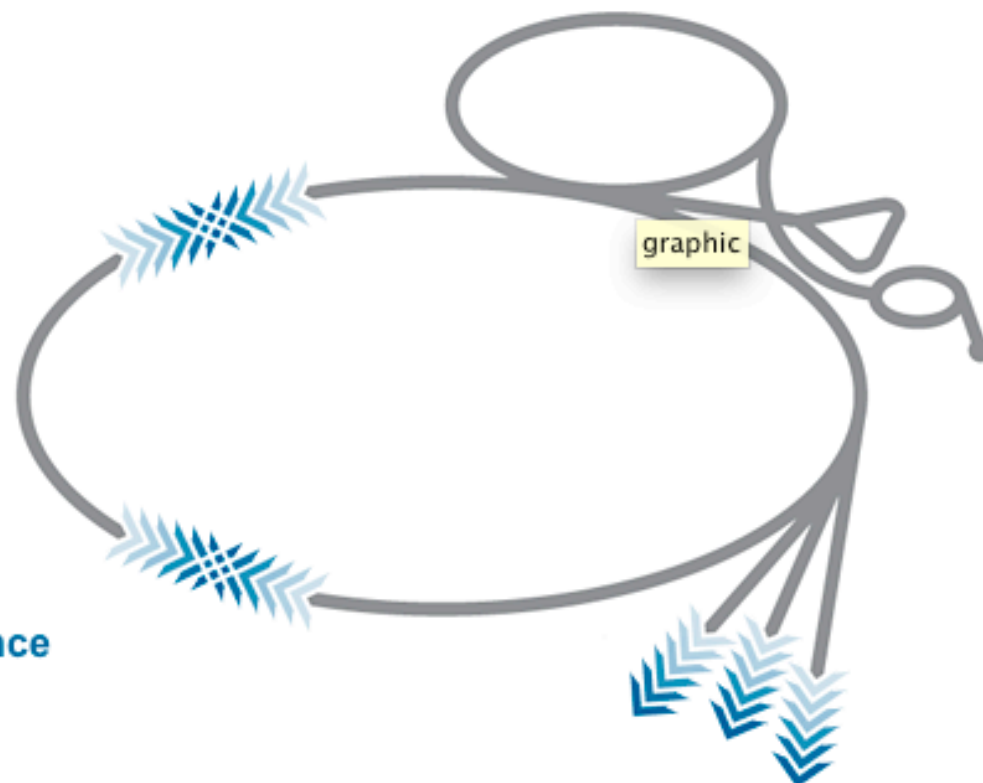
6:00 p.m. Reception

Featuring speakers honoring three decades of Tevatron history and a performance by Winifred Haun & Dancers

[Watch the symposium live](#)

Registration not required to attend

Please also join us for the [45th Fermilab Users' Meeting](#)
Showcasing recent results from Fermilab's experimental program
June 12–13, 2012



What did the Tevatron cost?

- Tevatron accelerator
 - \$120M (1983) = \$277M (2012 \$)
- Main Injector project
 - \$290M (1994) = \$450M (2012 \$)
- Detectors and upgrades
 - Guess: 2 x \$500M (collider detectors) + \$300M (FT)
- Operations
 - Say 20 years at \$100M/year = \$2 billion
- Total cost = **\$4 billion**



PhD Student Training

- Value of a PhD student
 - \$2.2M (US Census Bureau, 2002) = \$2.8M (2012 \$)
- Number of students trained at the Tevatron
 - 904 (CDF + DØ)
 - 492 (Fixed Target)
 - 18 (Smaller Collider experiments)
 - 1414 total
- Financial Impact = **\$3.96 billion**



Balance sheet

- 20 year investment in Tevatron ~ \$4B
- Students \$4B
- Magnets and MRI \$5-10B } ~ \$50B total
- Computing \$40B

Very rough calculation – but confirms our gut feeling that investment in fundamental science pays off

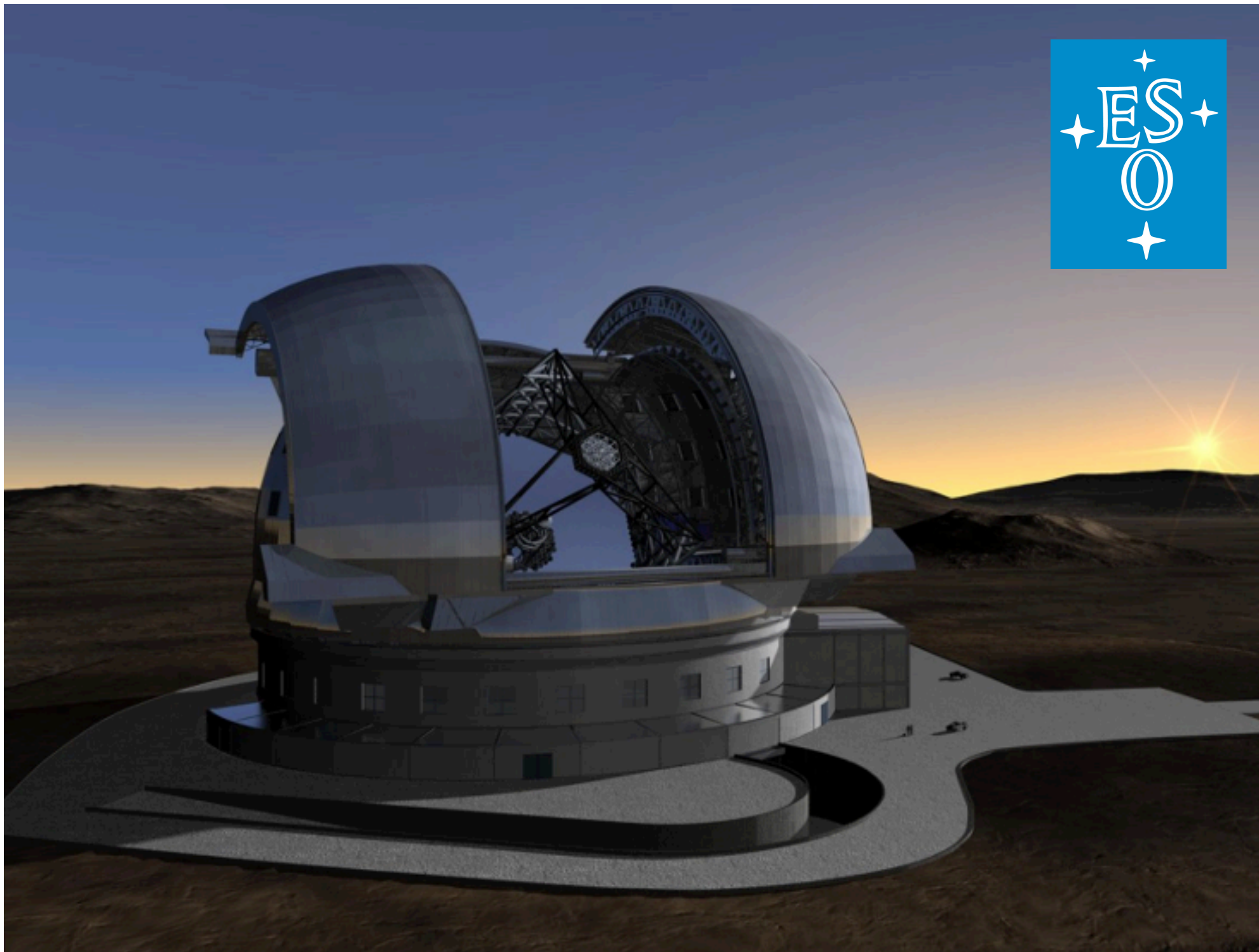
I think there is an opportunity for someone to repeat this exercise more rigorously

cf. STFC study of SRS Impact

<http://www.stfc.ac.uk/2428.aspx>



Science & Technology
Facilities Council











Big Data

Global Internet traffic in 2013
1.2 Zettabytes

SKA Phase I data per year (2023)
16 Zettabytes