

# Preservation, re-use and (open) access to High-Energy Physics data (or lack thereof)

HEP and its data  
What is the trouble?  
A possible way forward



Salvatore Mele  
CERN

Costs, Benefits and Motivations  
for Digital Preservation  
Nice - November 28<sup>th</sup> 2008

# High-Energy Physics (or Particle Physics)



HEP aims to understand how our Universe works:

- by discovering the most elementary constituents of matter and energy
- by probing their interactions
- by exploring the basic nature of space and time

In other words, try to answer two eternal questions:

- "What is the world made of?"
- "What holds it together?"

Build the largest scientific instruments ever to reach energy densities close to the Big Bang; write theories to predict and describe the observed phenomena

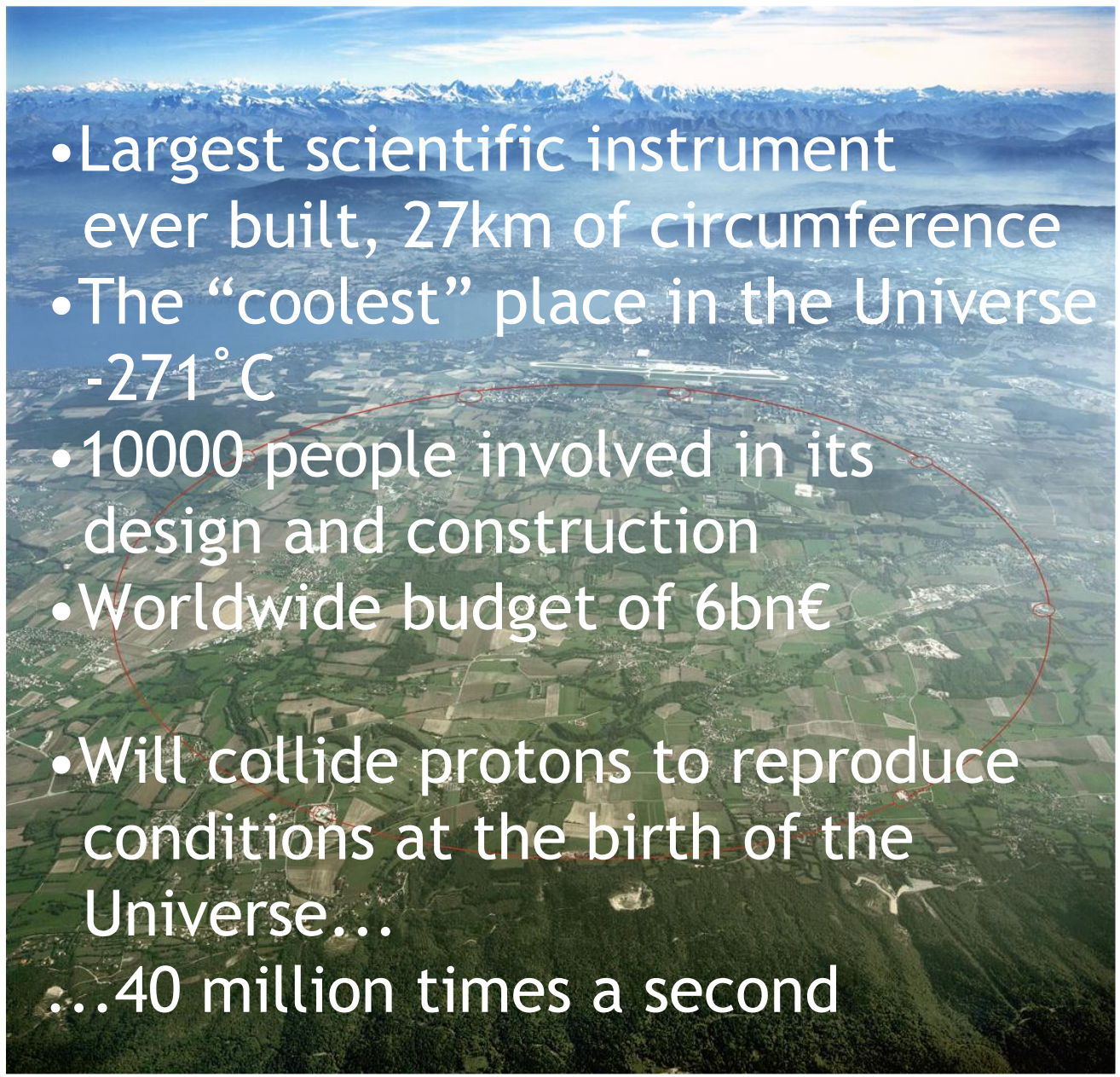
# CERN: European Organization for Nuclear Research (since 1954)

- The world leading HEP laboratory, Geneva (CH)
- 2500 staff (mostly engineers, from Europe)
- 9000 users (mostly physicists, from 580 institutes in 85 countries)
- 3 Nobel prizes (Accelerators, Detectors, Discoveries)
- Invented the web    
- Just switched on the 27-km (6bn€) LHC accelerator
- Runs a 1-million objects Digital Library

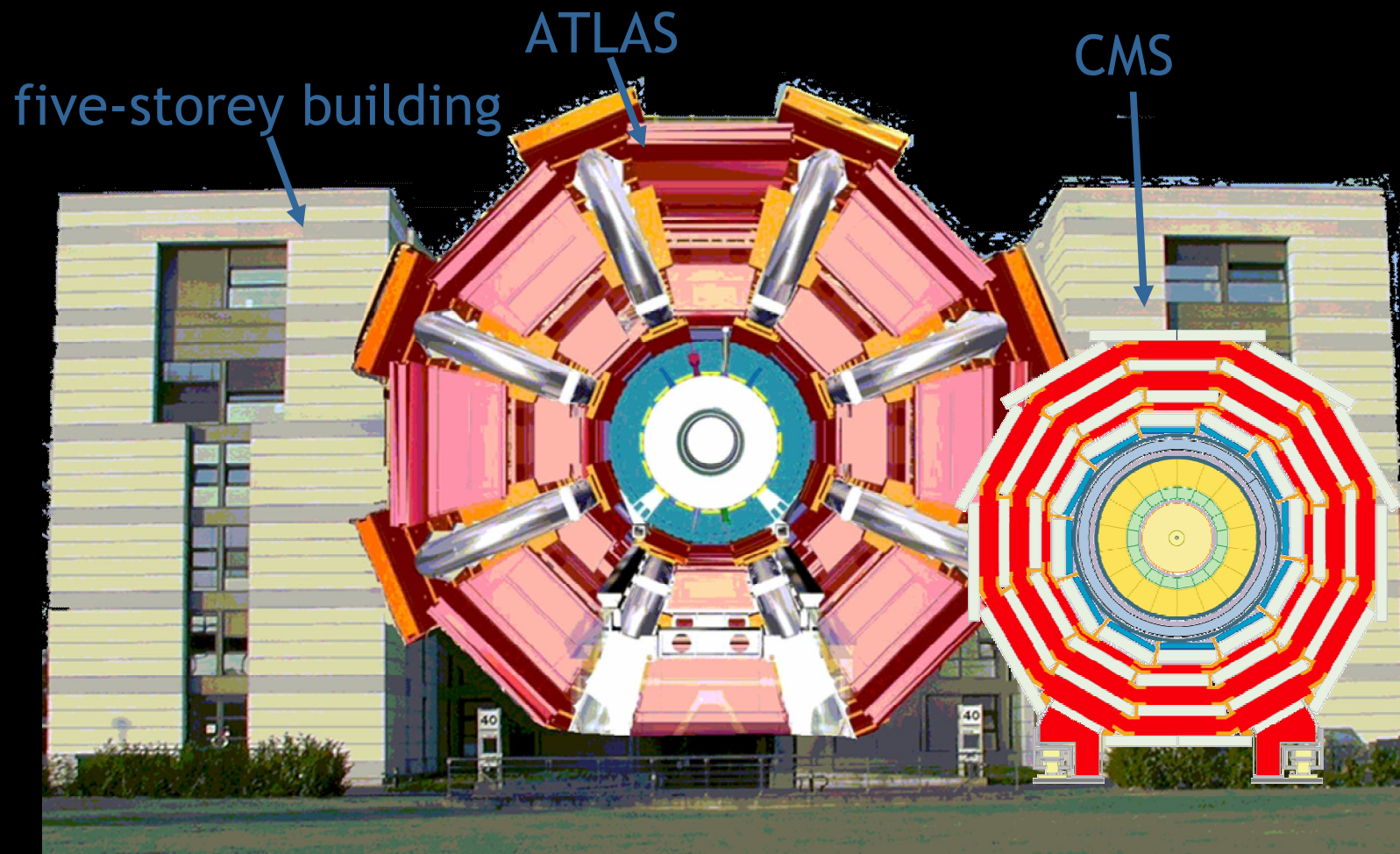
CERN Convention (1953): *ante-litteram* Open Access mandate

“... the results of its experimental and theoretical work shall be published or otherwise made generally available”

# The Large Hadron Collider

- 
- Largest scientific instrument ever built, 27km of circumference
  - The “coolest” place in the Universe  
-271 °C
  - 10000 people involved in its design and construction
  - Worldwide budget of 6bn€
  - Will collide protons to reproduce conditions at the birth of the Universe...  
...40 million times a second

The LHC experiments:  
about 100 million “sensors” each  
[think your 6MP digital camera...  
...taking 40 million pictures a second]

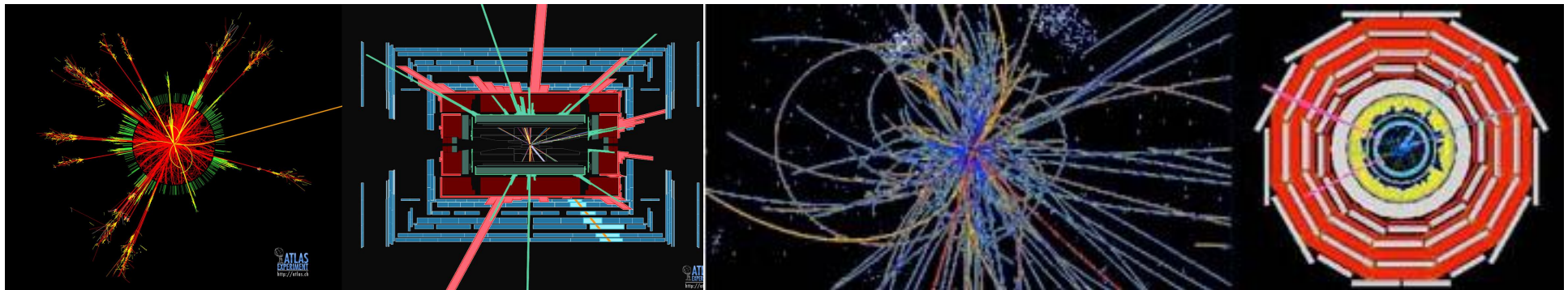


# The LHC data

- 40 million events (pictures) per second
- Select (on the fly) the ~200 interesting events per second to write on tape
- “Reconstruct” data and convert for analysis:  
“physics data” [inventing the grid...]

---

(x4 experiments x15 years)	Per event	Per year
Raw data	1.6 MB	3200 TB
Reconstructed data	1.0 MB	2000 TB
Physics data	0.1 MB	200 TB

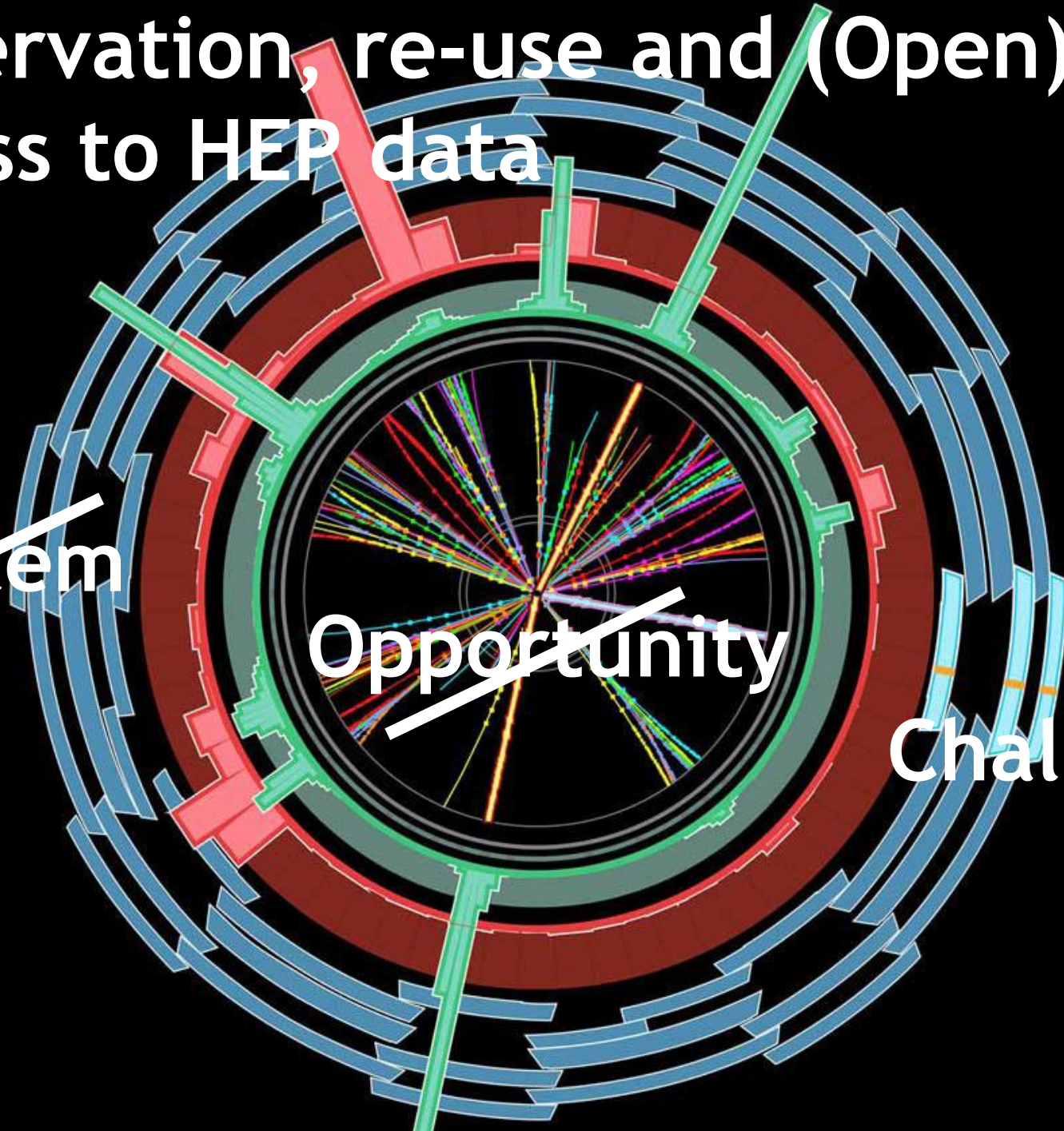


Preservation, re-use and (Open)  
Access to HEP data

~~Problem~~

~~Opportunity~~

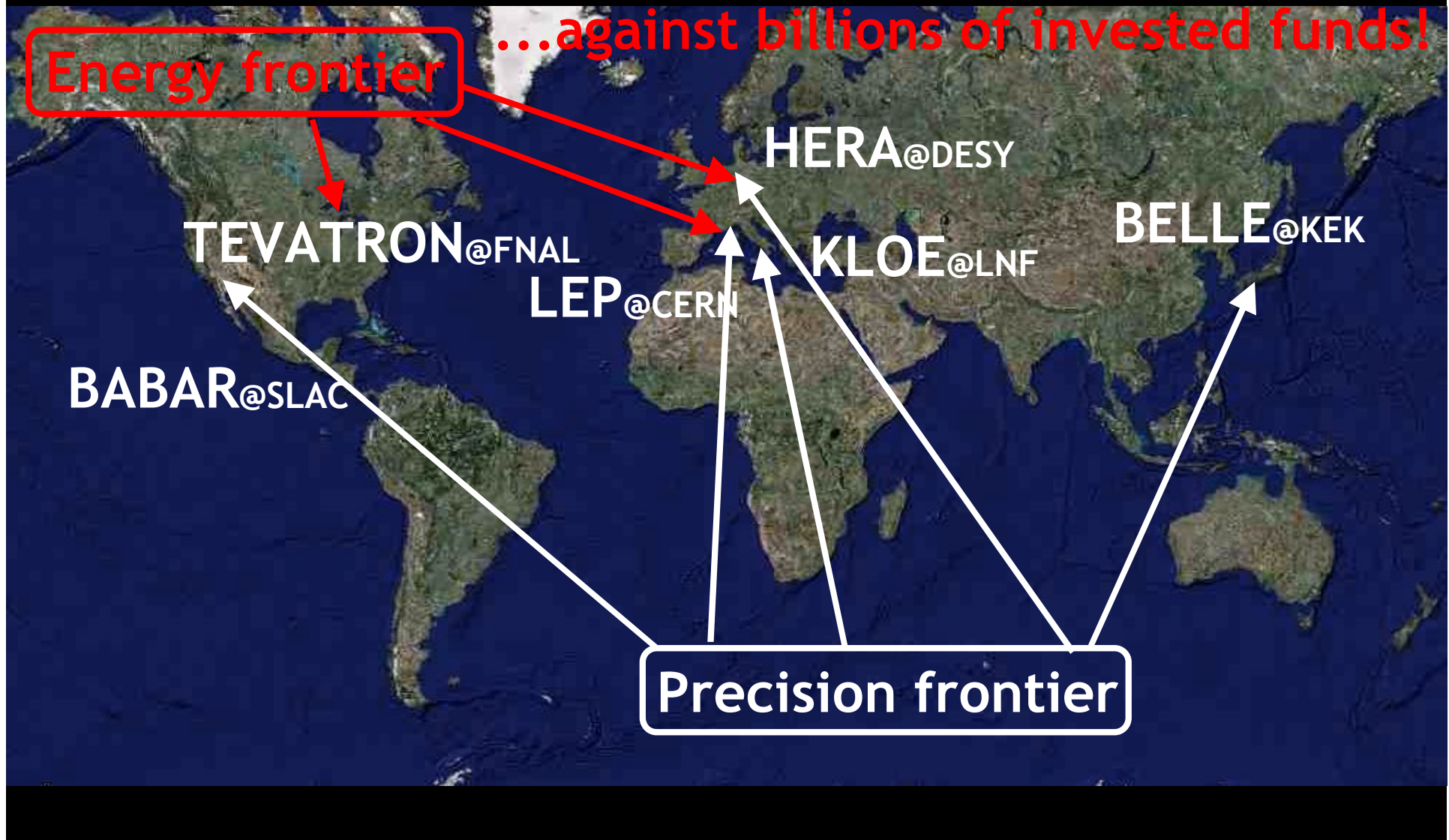
Challenge



# Some other HEP facilities (recently stopped or about to stop)

No real long-term archival strategy...

...against billions of invested funds!





# Why shall we care?

- We cared producing these data in first instance
- Unique, extremely costly, non-reproducible
- Might need to go back to the past (it happened)
- A peculiar community (the web, arXiv, the grid...)
- “Worst-case scenario” complex data, no preservation... “If it works here, will work in many other places”

# Preservation, re-use and (open) access continua (who and when)

- The same researchers who took the data, after the closure of the facility (~1 year, ~10 years)
- Researchers working at similar experiments at the same time (~1 day, week, month, year)
- Researchers of future experiments (~20 years)
- Theoretical physicists who may want to re-interpret the data (~1 month, ~1 year, ~10 years)
- Theoretical physicists who may want to test future ideas (~1 year, ~10 years, ~20 years)



# Data preservation, circa 2004

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-PH-EP/2004-024  
June 8, 2004

Studies of Hadronic Event Structure  
in  $e^+e^-$  Annihilation  
from 30 GeV to 209 GeV  
with the L3 Detector

The L3 Collaboration

Submitted to *Physics Reports*

arXiv:hep-ex/0406049v1 18 Jun 2004

Unacceptable in the 21st century!

$D$	at $\sqrt{s} = 194.4$ GeV	at $\sqrt{s} = 200.2$ GeV	at $\sqrt{s} = 209.2$ GeV
0.000-0.016	$34.894 \pm 1.033 \pm 0.772$	$33.486 \pm 1.015 \pm 0.586$	$33.374 \pm 1.015 \pm 0.599$
0.016-0.032	$9.185 \pm 0.517 \pm 0.296$	$9.290 \pm 0.510 \pm 0.537$	$8.491 \pm 0.510 \pm 0.511$
0.032-0.048	$4.744 \pm 0.377 \pm 0.218$	$4.744 \pm 0.411 \pm 0.218$	$4.744 \pm 0.411 \pm 0.218$
0.048-0.064	$3.138 \pm 0.341 \pm 0.159$	$3.175 \pm 0.341 \pm 0.159$	$2.949 \pm 0.341 \pm 0.159$
0.064-0.080	$1.75 \pm 0.284 \pm 0.108$	$1.75 \pm 0.284 \pm 0.108$	$1.75 \pm 0.284 \pm 0.108$
0.080-0.096	$1.13 \pm 0.274 \pm 0.077$	$1.13 \pm 0.274 \pm 0.077$	$1.13 \pm 0.274 \pm 0.077$
0.096-0.112	$0.73 \pm 0.254 \pm 0.058$	$0.73 \pm 0.254 \pm 0.058$	$0.73 \pm 0.254 \pm 0.058$
0.112-0.128	$0.48 \pm 0.234 \pm 0.043$	$0.48 \pm 0.234 \pm 0.043$	$0.48 \pm 0.234 \pm 0.043$
0.128-0.144	$0.31 \pm 0.214 \pm 0.032$	$0.31 \pm 0.214 \pm 0.032$	$0.31 \pm 0.214 \pm 0.032$
0.144-0.160	$0.19 \pm 0.194 \pm 0.023$	$0.19 \pm 0.194 \pm 0.023$	$0.19 \pm 0.194 \pm 0.023$
0.160-0.176	$0.12 \pm 0.174 \pm 0.017$	$0.12 \pm 0.174 \pm 0.017$	$0.12 \pm 0.174 \pm 0.017$
0.176-0.192	$0.08 \pm 0.154 \pm 0.013$	$0.08 \pm 0.154 \pm 0.013$	$0.08 \pm 0.154 \pm 0.013$
0.192-0.208	$0.05 \pm 0.134 \pm 0.009$	$0.05 \pm 0.134 \pm 0.009$	$0.05 \pm 0.134 \pm 0.009$
0.208-0.224	$0.03 \pm 0.114 \pm 0.007$	$0.03 \pm 0.114 \pm 0.007$	$0.03 \pm 0.114 \pm 0.007$
0.224-0.240	$0.02 \pm 0.094 \pm 0.005$	$0.02 \pm 0.094 \pm 0.005$	$0.02 \pm 0.094 \pm 0.005$
0.240-0.256	$0.01 \pm 0.074 \pm 0.004$	$0.01 \pm 0.074 \pm 0.004$	$0.01 \pm 0.074 \pm 0.004$
0.256-0.272	$0.00 \pm 0.054 \pm 0.003$	$0.00 \pm 0.054 \pm 0.003$	$0.00 \pm 0.054 \pm 0.003$
0.272-0.288	$0.00 \pm 0.034 \pm 0.002$	$0.00 \pm 0.034 \pm 0.002$	$0.00 \pm 0.034 \pm 0.002$
0.288-0.304	$0.00 \pm 0.014 \pm 0.001$	$0.00 \pm 0.014 \pm 0.001$	$0.00 \pm 0.014 \pm 0.001$
0.304-0.320	$0.00 \pm 0.004 \pm 0.000$	$0.00 \pm 0.004 \pm 0.000$	$0.00 \pm 0.004 \pm 0.000$
0.320-0.336	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$
0.336-0.352	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$
0.352-0.368	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$
0.368-0.384	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$
0.384-0.400	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$	$0.00 \pm 0.000 \pm 0.000$
First Moment	$0.0387 \pm 0.0023 \pm 0.0047$	$0.0435 \pm 0.0028 \pm 0.0037$	$0.0429 \pm 0.0029 \pm 0.0033$
Second Moment	$0.0056 \pm 0.0010 \pm 0.0016$	$0.0064 \pm 0.0010 \pm 0.0021$	$0.0064 \pm 0.0012 \pm 0.0020$

Table 52: Differential distribution for  $D$ -parameter at  $\sqrt{s} = 194.4, 200.2$  and  $209.2$  GeV. The first uncertainty is statistical, the second systematic.

140 pages of tables

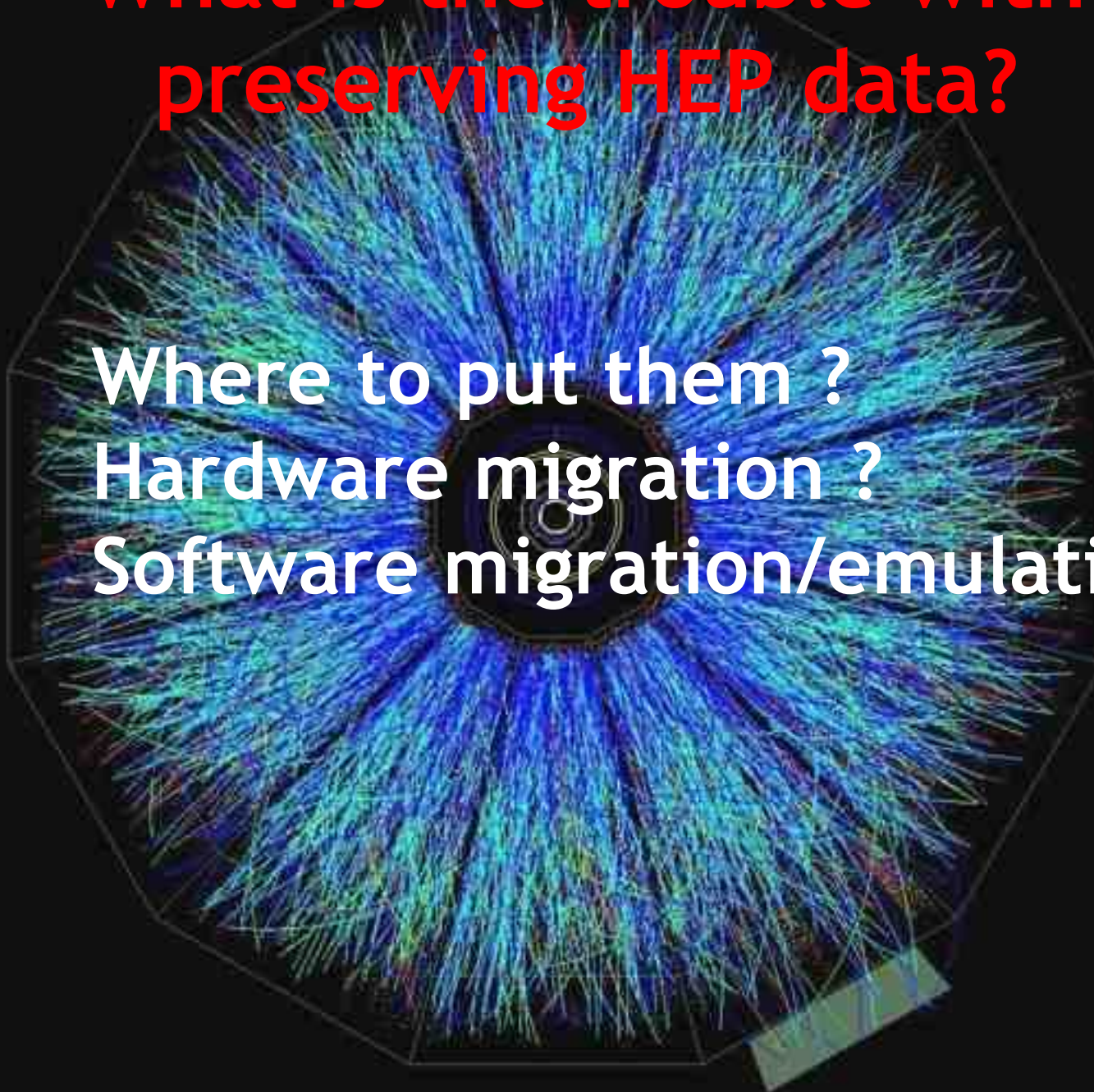
81

# What is the trouble with preserving HEP data?

Where to put them ?

Hardware migration ?

Software migration/emulation?





**What is the trouble with  
preserving HEP data?**

**Where to put them ?**

**Hardware migration ?**

**Software migration/emulation?**

# HEP, Open Access & Repositories

- HEP is decades ahead in thinking Open Access:
  - Mountains of paper preprints shipped around the world for 40 years (at author/institute expenses!)
  - Launched arXiv (1991), archetypal Open Archive
  - >90% HEP production self-archived in repositories
  - 100% HEP production indexed in SPIRES (community run database, first WWW server on US soil)
- OA is second nature: posting on arXiv before submitting to a journal is common practice
  - No mandate, no debate. Author-driven.
- HEP scholars have the tradition of arXiving their output (helas, articles) somewhere

**Data repositories would be a natural concept**

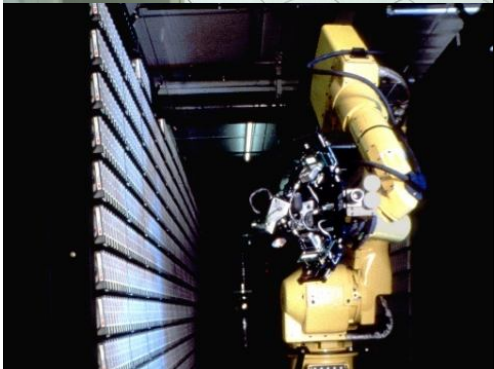


**What is the trouble with  
preserving HEP data?**

**Where to put them ?**

**Hardware migration ?**

**Software migration/emulation?**



# Life-cycle of previous-generation CERN experiment L3 at LEP

- 1984 Begin of construction
- 1989 Start of data taking
- 2000 End of data taking
- 2002 End of *in-silico* experiments
- 2005 End of (most) data analysis

## Storage and migration of data at the CERN computing centre

- |      |          |         |           |       |
|------|----------|---------|-----------|-------|
| 1993 | ~150'000 | 9track  | → 3480    | 0.2GB |
| 1997 | ~250'000 | 3480    | → Redwood | 20GB  |
| 2001 | ~25'000  | Redwood | → 9940    | 60GB  |
| 2004 | ~5'000   | 9940A   | → 9940B   | 200GB |
| 2007 | ~22'000  | 9940B   | → T1000A  | 500GB |



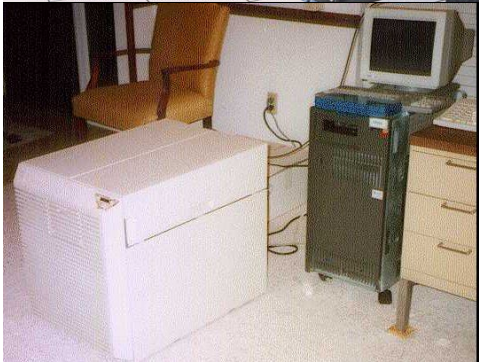
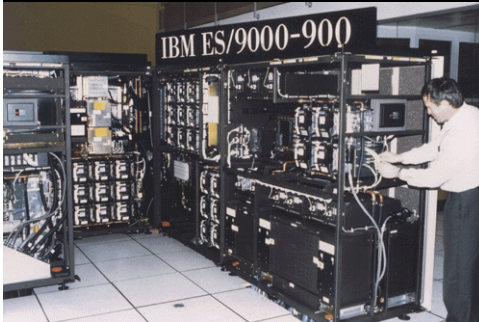


**What is the trouble with  
preserving HEP data?**

**Where to put them ?**

**Hardware migration ?**

**Software migration/emulation?**



# Life-cycle of previous-generation CERN experiment L3 at LEP

1984 Begin of construction

1989 Start of data taking

2000 End of data taking

2002 End of *in-silico* experiments

2005 End of (most) data analysis

## Computing environment of the L3 experiment at LEP

1989-2001 VAX for data taking

1986-1994 IBM for data analysis

1992-1998 Apollo (HP) workstations

1996-2001 SGI mainframe

1997-2007 Linux boxes

# Life-cycle of previous-generation

## CERN experiment L3 at LEP

1984 Begin of construction

1989 Start of data taking

2000 End of data taking

2001 End of *in-situ* experiments

2005 End of (most) data analysis

## Computing environment of the L3 experiment at LEP

1989-2001 VAX for data taking

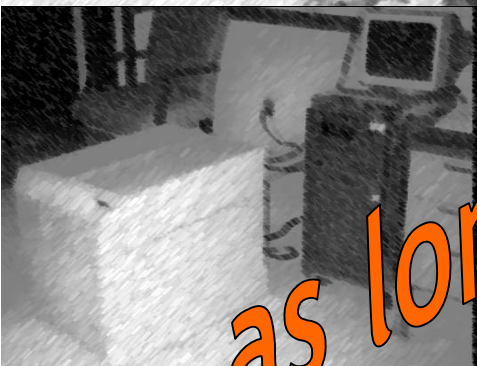
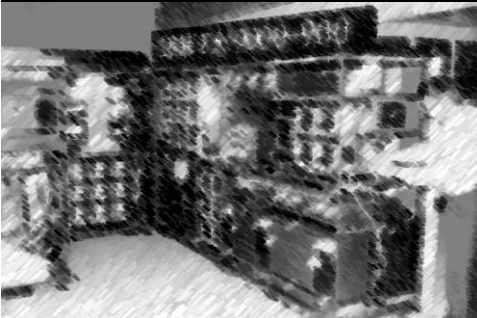
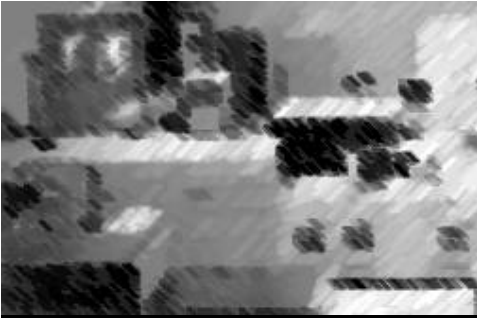
1989-1994 IBM for data analysis

1992-1998 Apollo (HP) workstations

1996-2001 SGI mainframe

1997-2007 Linux boxes

All is good... as long as the experiment runs!  
(Après moi le déluge)





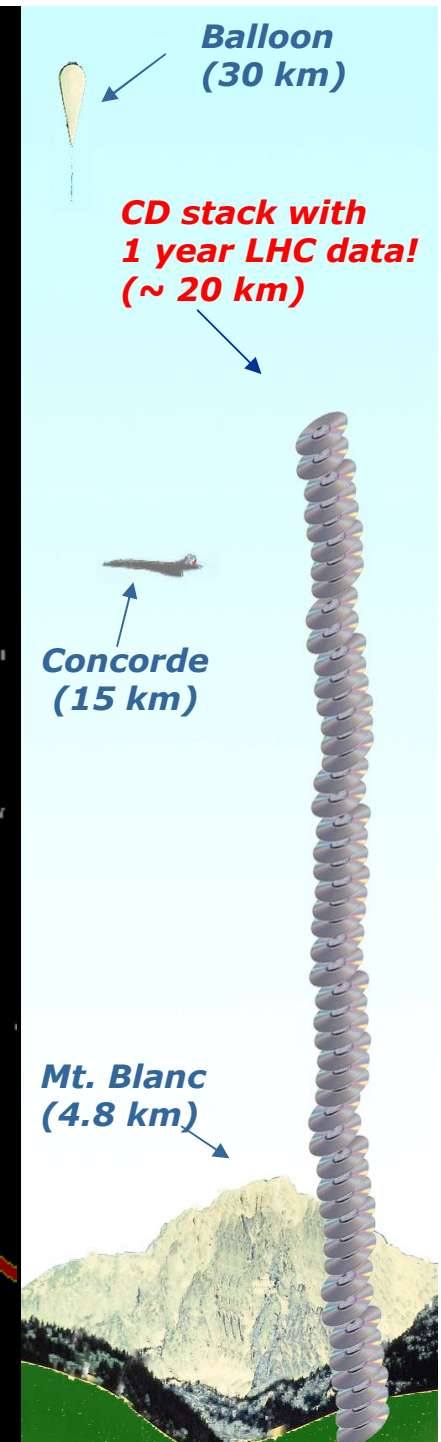
**What is the trouble with  
preserving HEP data?**

~~Where to put them ?  
Hardware migration ?  
Software migration/emulation?~~

**The HEP data !**

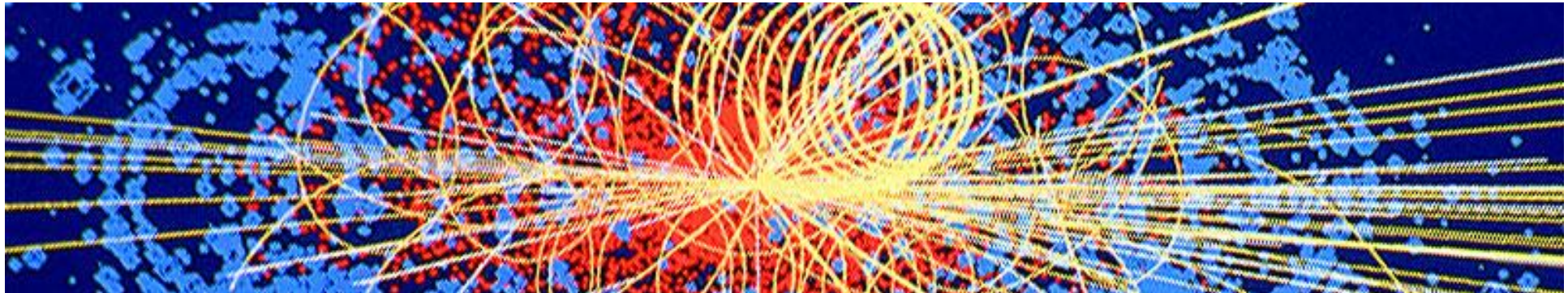
# Preserving HEP data?

- The HEP data model is highly complex. Data are traditionally not re-used as in Astronomy or Climate science.
- Raw data → calibrated data → skimmed data → high-level objects → physics analyses → results.
- All of the above duplicated for *in-silico* experiments, necessary to interpret the highly-complex data.
- Final results depend on the grey literature on calibration constants, human knowledge and algorithms needed for each pass...oral tradition!
- Years of training for a successful analysis



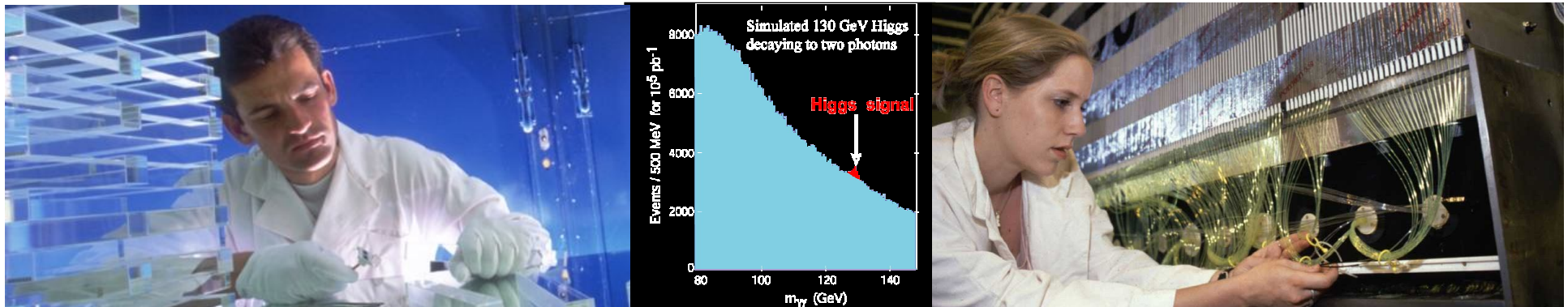
# Need a “parallel way” to publish/preserve/re-use/OpenAccess

- In addition to experiment data models, elaborate a parallel format for (re-)usable high-level objects
  - In times of need (to combine data of “competing” experiments) this approach has worked
  - Embed the “oral” and “additional” knowledge
- A format understandable and thus re-usable by practitioners in other experiments and theorists
- Start from tables and work back towards primary data
- How much additional work? 1%, 5%, 10%?



# Issues with the “parallel” way

- A small fraction of a big number gives a large number
- Activity in competition with research time
- 1000s person-years for parallel data models need enormous (impossible?) academic incentives for realization ...or additional (external) funds
- Need insider knowledge to produce parallel data
- Address issues of (open) access, credit, accountability, “careless measurements”, “careless discoveries”, reproducibility of results, depth of peer-reviewing
- A monolithic way of doing business needs rethinking



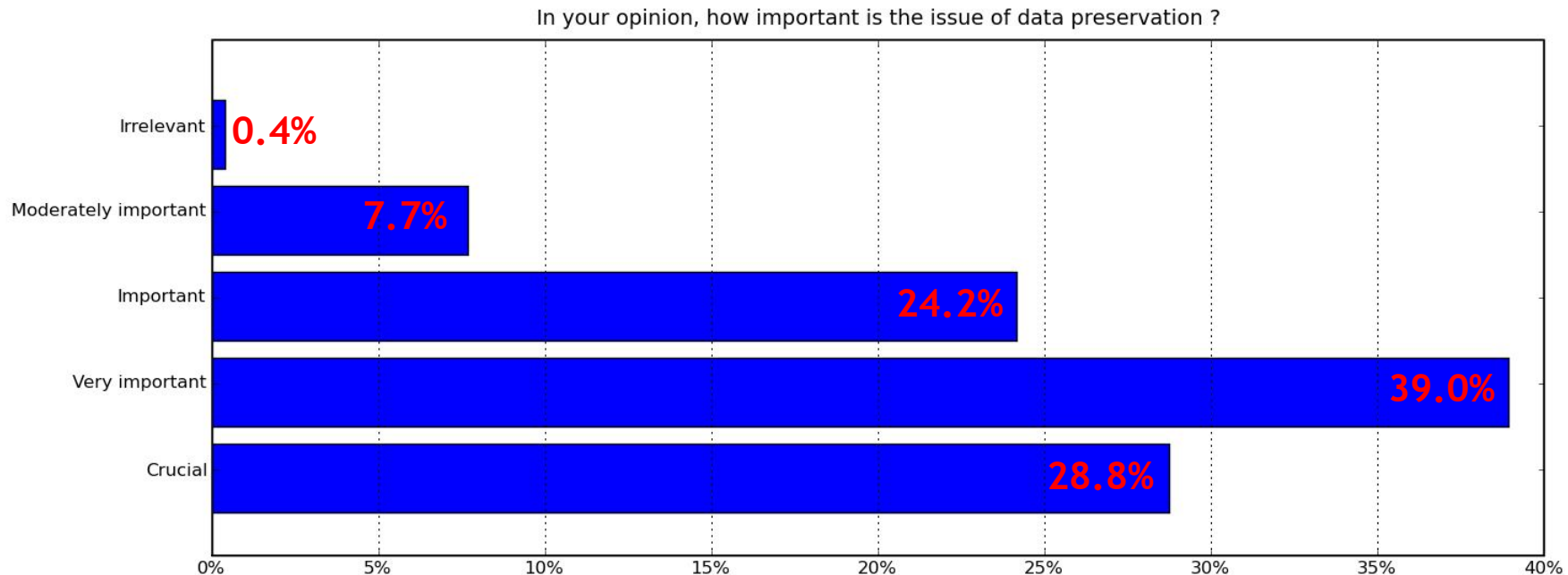
# Preservation, re-use and (Open) Access to HEP data... first steps!

- Outgrowing an institutionalized state of denial
- A difficult and costly way ahead
- An issue which starts surfacing on the agenda
  - Part of an integrated digital-library vision
  - CERN joined the Alliance for Permanent Access
  - CERN part of the FP7 PARSE.Insight project
- PARSE.Insight case study on HEP preservation:
  - HEP tradition: community-inspired and community-built (e-)infrastructures for research
  - Understand the desires and concerns of the community
  - Chart a way out for data preservation in the “worst-case scenario” of HEP, benefiting other communities
  - Debates and workshops planned in next months



# HEP Community Survey - First findings

500+ scientists answering a survey in its 1st week. 50% from Europe, 25% from the US. 34% left their e-mail address to hear more. 17% wants to be interviewed to say more!



- **92%**: data preservation is important, very important or crucial
- **40%**: thinks important HEP data have been lost
- **45%**: access to old data would have improved my scientific results
- **74%**: a trans-national infrastructure for long-term preservation of all kinds of scientific data should be built

**Two dozen more questions about do's and dont's in HEP preservation**

# Conclusions

- HEP spearheaded (Open) Access to Scientific Information: 50 years of preprints, 16 of repositories  
... but data preservation is not yet on the radar
- Heterogeneous continua to preserve data for
- No insurmountable technical problems
- The issue is the data model itself
  - (Primary) data intelligible only to the producers
  - A monolithic culture need a paradigm shift
  - Preservation implies massive person-power costs
- **Need cultural consensus and financial support**

**Preservation is appearing on the agenda...**

**Powerful synergy between CERN and  
the Alliance for Permanent Access & PARSE. Insight  
Exciting times are ahead!**

