

Data Preservation in High Energy Physics

Cristinel DIACONU
CPP Marseille & DESY



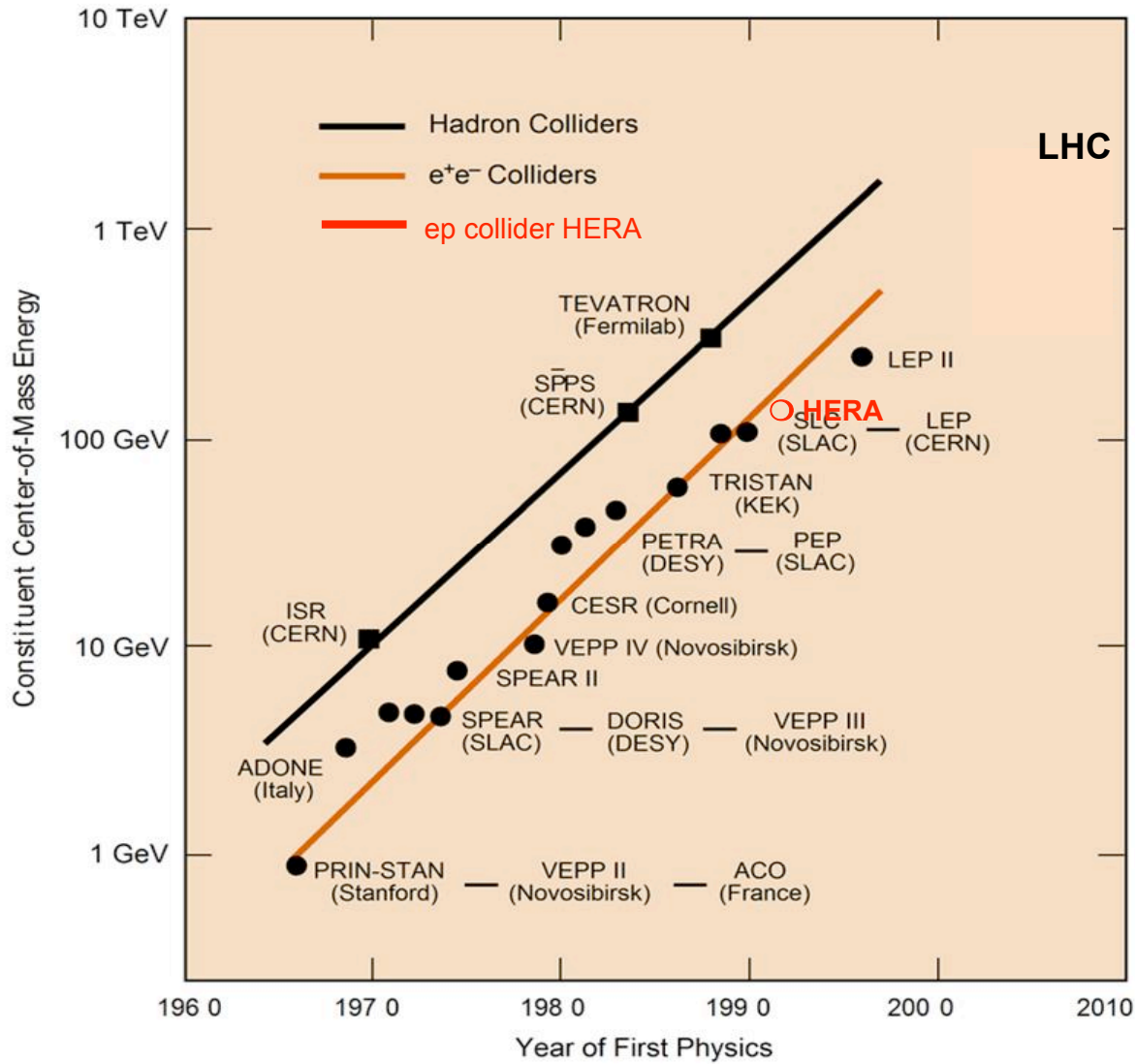
Introduction

- Digital Data production explodes, storage capacities cannot follow
- Digital Data is fragile, the preservation must be addressed :
 - Technology, access, economics, workflow ...
- Task forces already in place to address this issue in a generic way
 - e.g. Blue Ribbon, APA, DPC, eSciDir...
- Scientific Data is a major component of the ongoing efforts (complexity)
- **Data Preservation in HEP?**

<http://www.alliancepermanentaccess.eu>

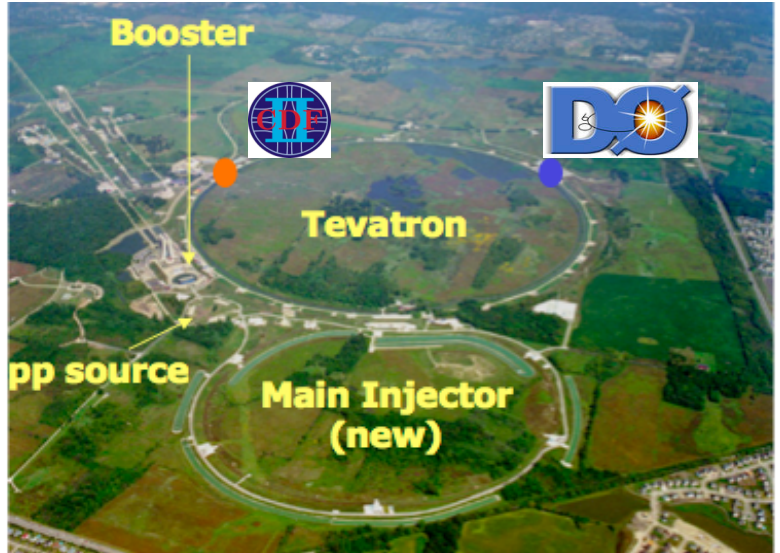
<http://brtf.sdsc.edu>
(intermediate report and references)

High Energy Physics

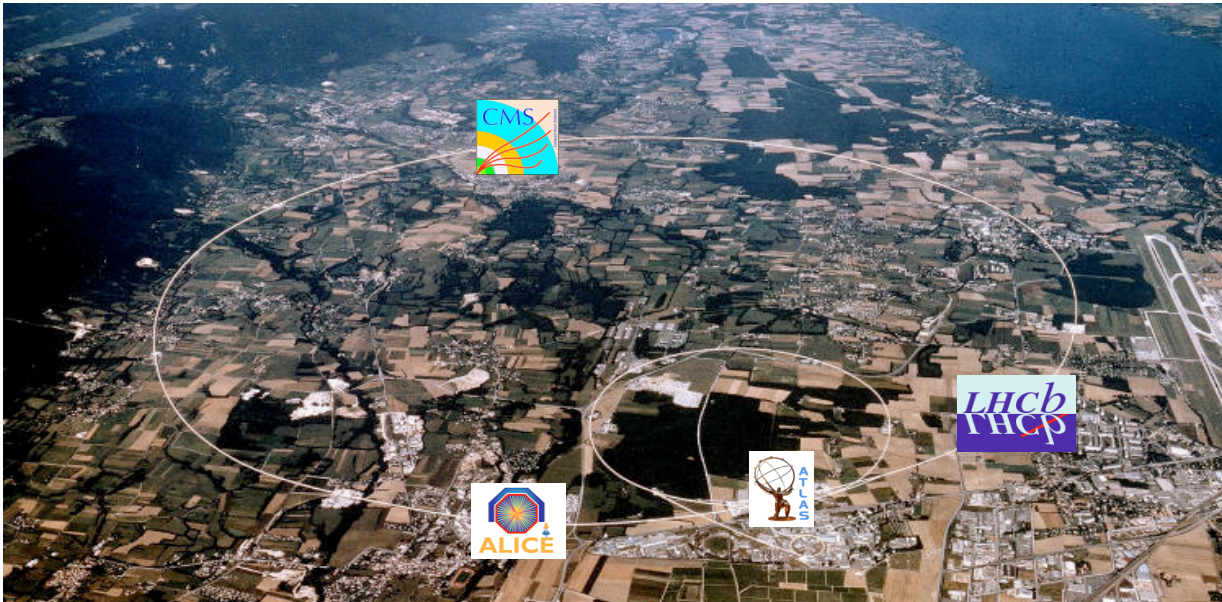


Energy frontier probed with complex experimental installation

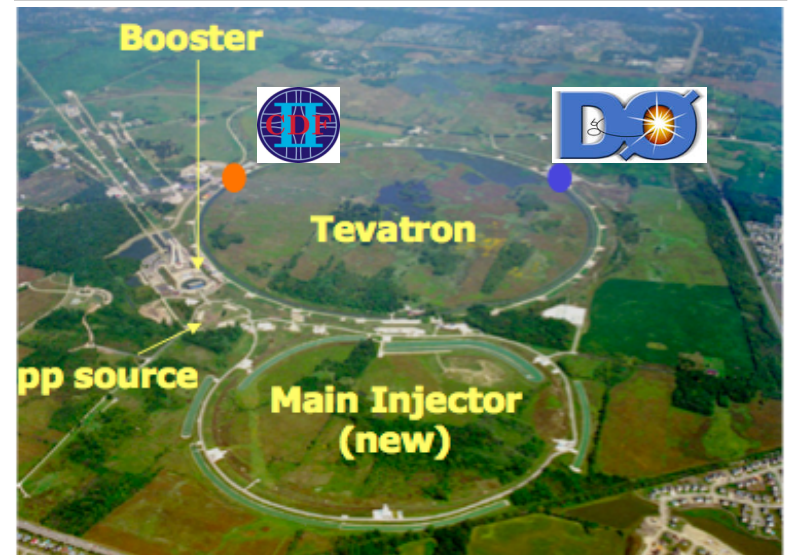
High Energy Physics Data are Unique



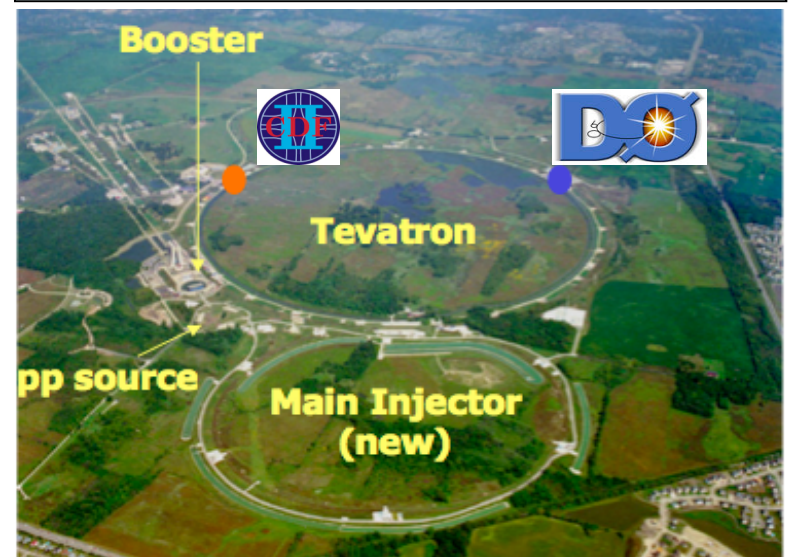
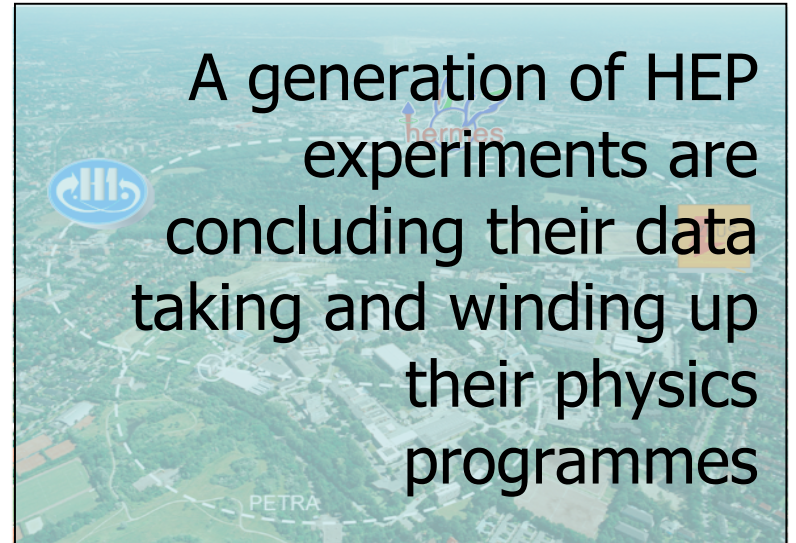
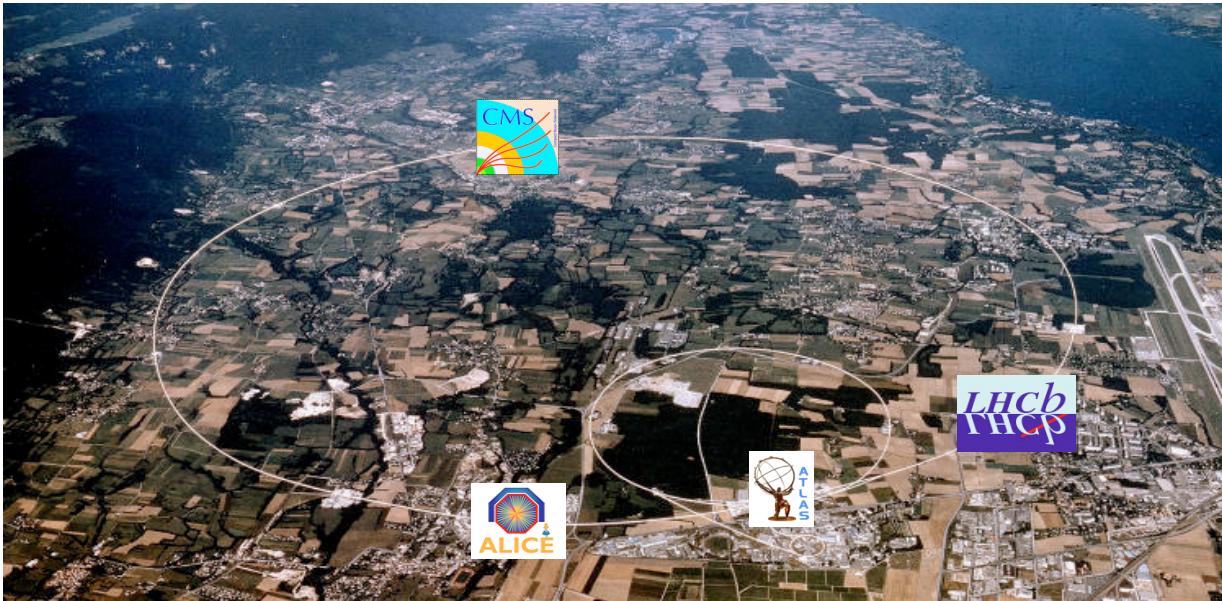
High Energy Physics Data are Unique



A generation of HEP experiments are concluding their data taking and winding up their physics programmes

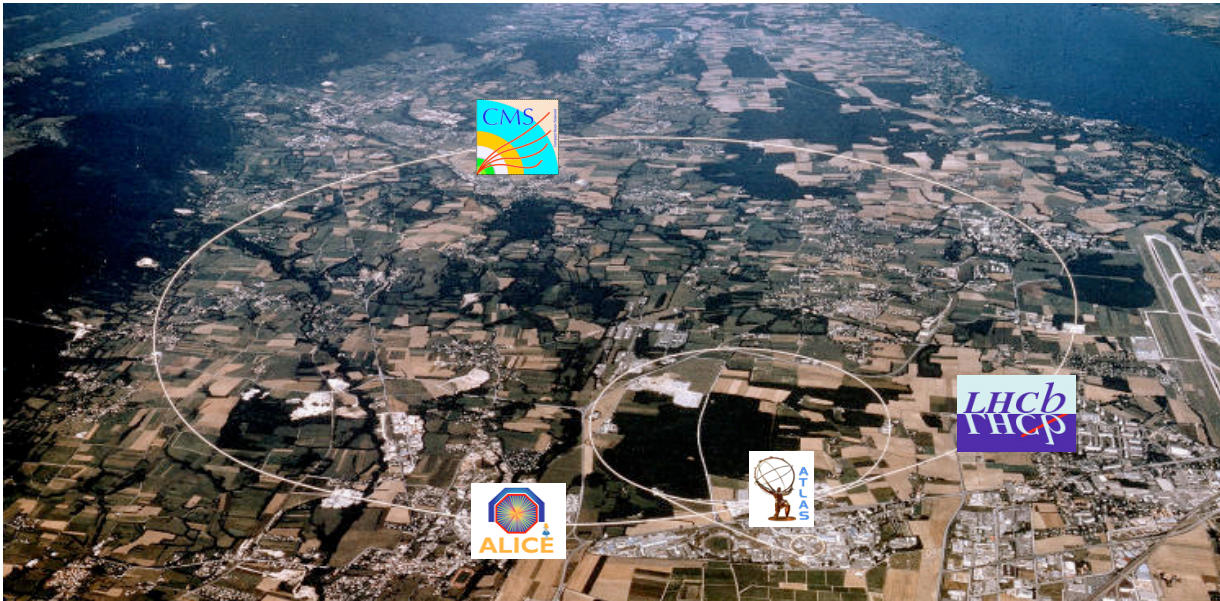


High Energy Physics Data are Unique



The experimental data from these experiments still has much to tell us, from the existing analyses still to be completed..

High Energy Physics Data are Unique

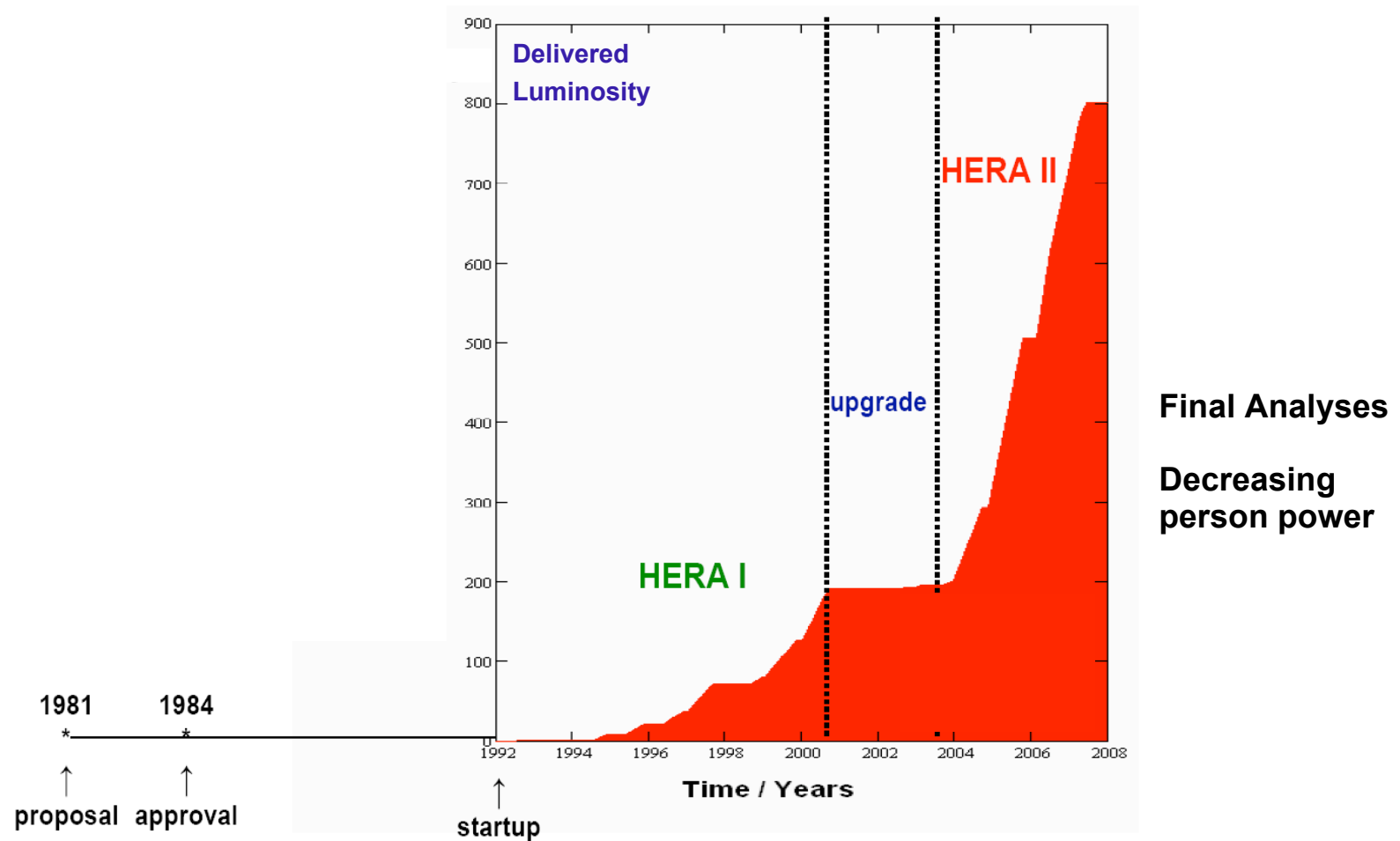


The experimental data from these experiments still has much to tell us, from the existing analyses still to be completed..

A generation of HEP experiments are concluding their data taking and winding up their physics programmes

..but they may also contain things we do not yet know, which may only come to light at a later date via LHC data or a new theory

Time profile of data collection in HEP



Multi-decade endeavours

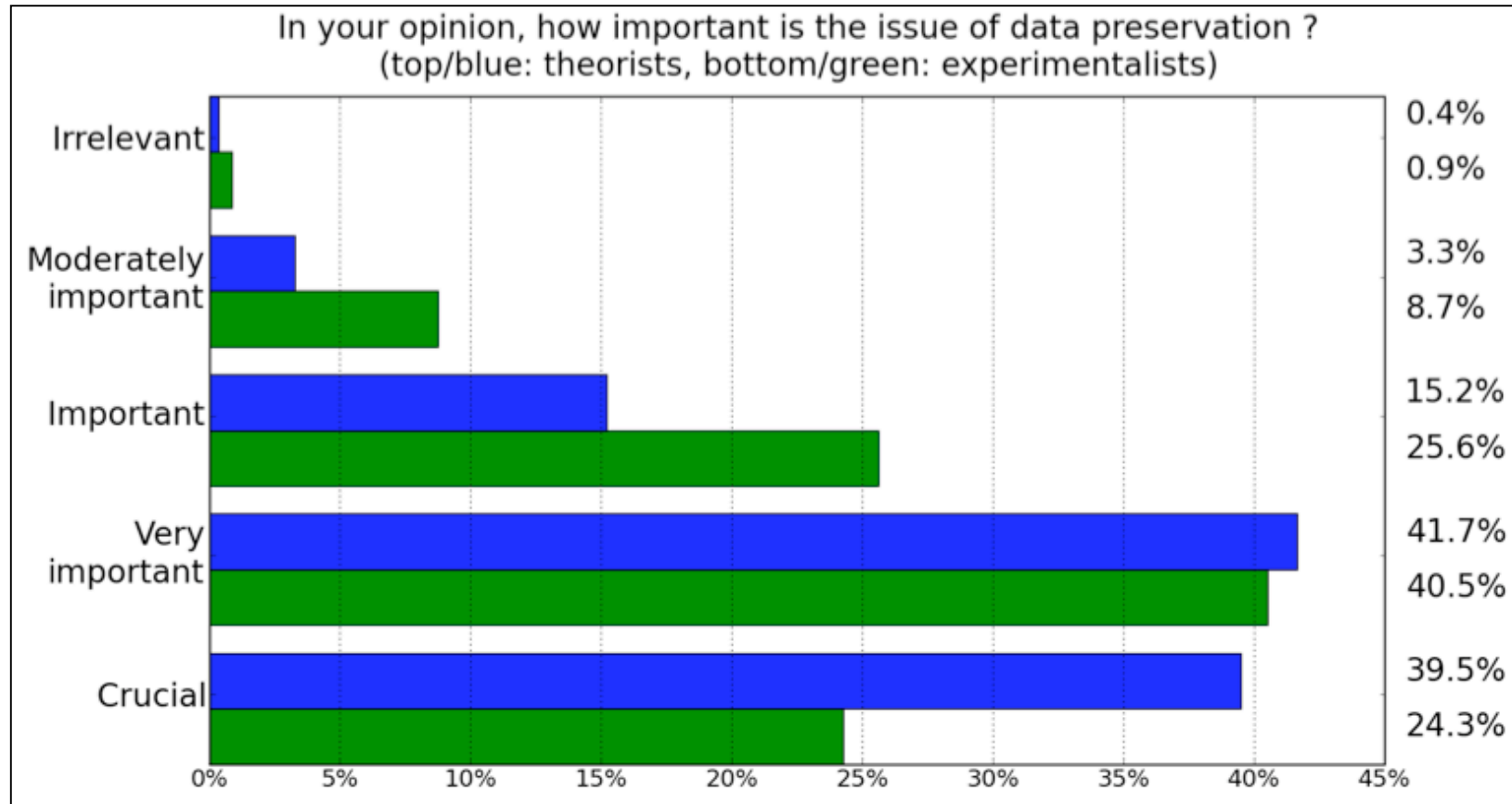
Data analysis is may take several years (reprocessing)

Why Should We Preserve HEP Data?

- It is unique and is the product large investments:
 - it has a potential for other/new physics studies
- We may want to re-do previous measurements
 - Increased precision, reduced systematics
 - New and improved theoretical calculations / MC models
 - Newly developed analysis techniques
- We may want to perform new measurements
 - At energies and processes where no other data are available (or will become available in the future)
- Investigate if new phenomena found today
 - Go back and check in the old data

PARSE.Insight Survey

S.Mele



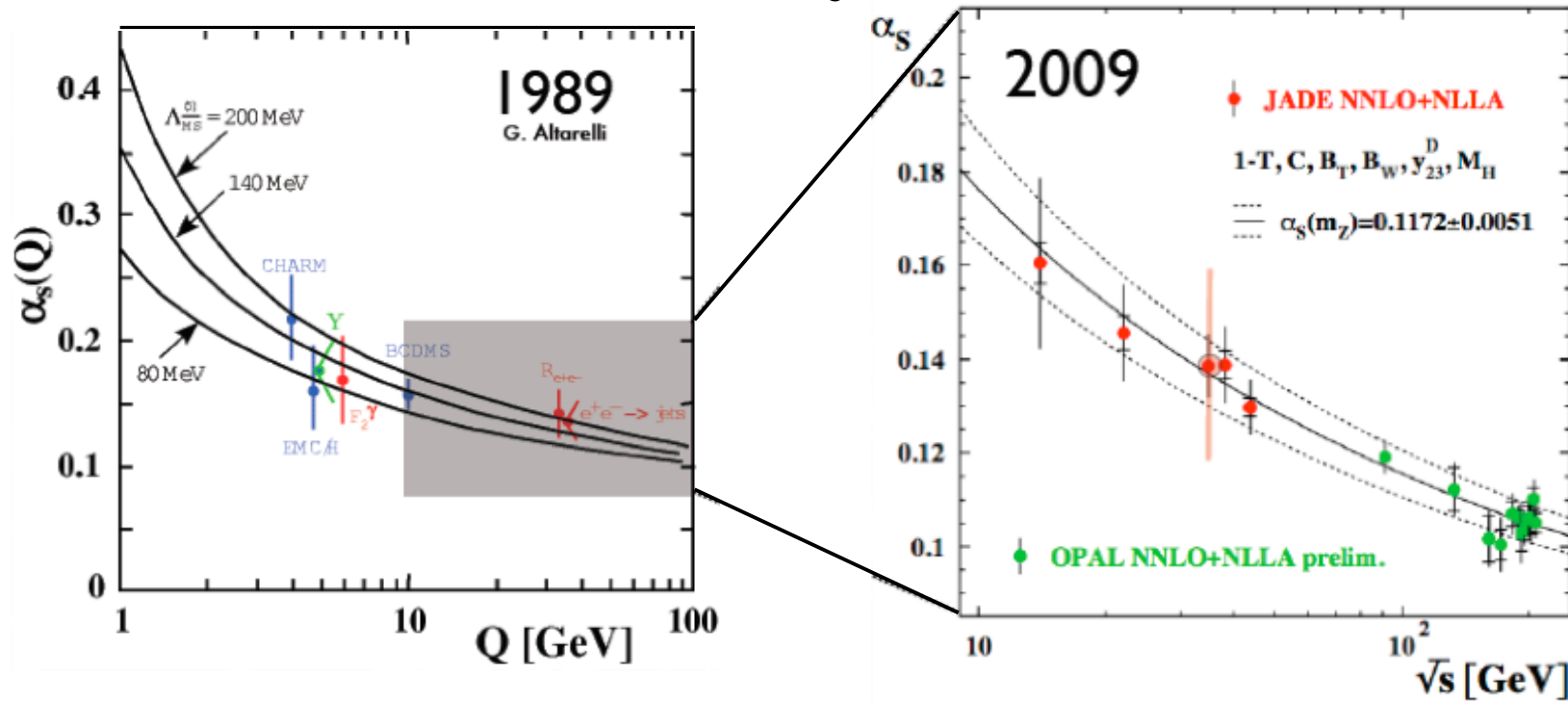
A largely supported idea in the community (90%)

The Challenge of Preserving HEP Data

- HEP has little or no tradition or clear current model of long term conservation of data in a meaningful and useful way
- The preservation of and supported long term access to the data is generally not part of the planning, software design or budget of a HEP experiment
 - The main assumption has probably been that the data will always be superseded by the next experiment: but this is not always the case!
 - Another (sometimes wrong) assumption is that the physics potential is exhausted at the end of the program
- There is also little tradition of useful open access of HEP data beyond the walls of the original collaboration
 - This is clearly a difficult prospect, with many issues like control, correctness and reputation of the experiment, not to mention a lack of portability and the state of the documentation

Re-analysis of JADE Data from PETRA

Precision measurements of α_s and tests of asymptotic freedom

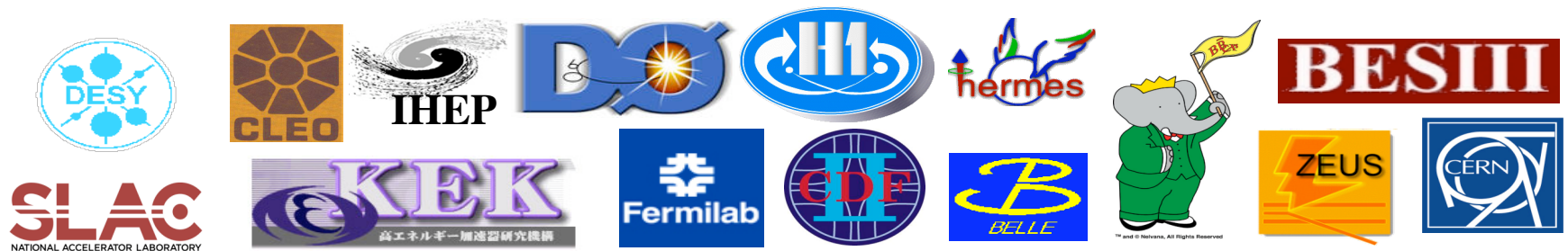


In NLO QCD: $\alpha_s(35 \text{ GeV}) = 0.14 \pm 0.02$
No indication of a running α_s signature

In re-summed NNLO QCD: $\alpha_s(M_Z) = 0.1172 \pm 0.0051$
Significant evidence of running α_s and asym. freedom

An Inter-Experimental Study Group

- Start a common enterprise between experiments and associated computing centres: a “sampling” of collider experiments ee,pp,ep
 - First contacts made and meetings held autumn 2008
- *International Steering Committee*
 - Made up of the Spokespersons of the HEP experiments and the Directors of the associated computing centres
- *International Advisory Committee*
 - Chaired by J. Dorfan and S. Bethke
- First Workshop at DESY in January 2009
 - Initial discussions, to share ideas and to see the current picture
- Follow up workshop at SLAC in May 2009
 - To achieve some concrete goals and converge on a set of recommendations to ICFA in form of a blueprint for data preservation



DESY Workshop, January 2009



First Workshop on Data Preservation and Long Term Analysis in HEP

DESY, Hamburg, Germany
Mon 26th - Wed 28th January 2009

Objectives of the Workshop
Review the physics objectives of data persistency in HEP
Exchange information on the analysis model used by HEP experiments
Address the hardware and software persistency issue
Review the funding programs and other existing international initiatives
Converge to a common set of recommendations for future experiments

<http://indico.cern.ch/conferenceDisplay.py?confId=42722>

Local Organizing Committee:
Christine Dicuon (CPH/DESY)
Tobias Hees (DESY)
Volker Gehrmann (DESY-IT)
David South (TU Dortmund)
Krzysztof Wozniak (DESY-IT)

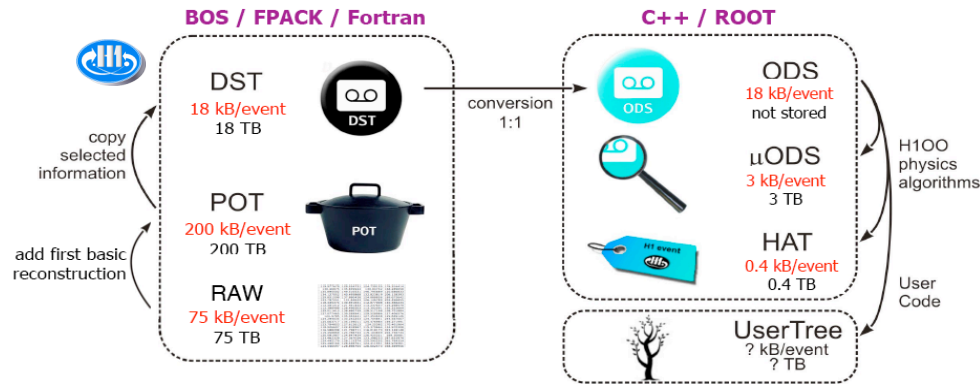
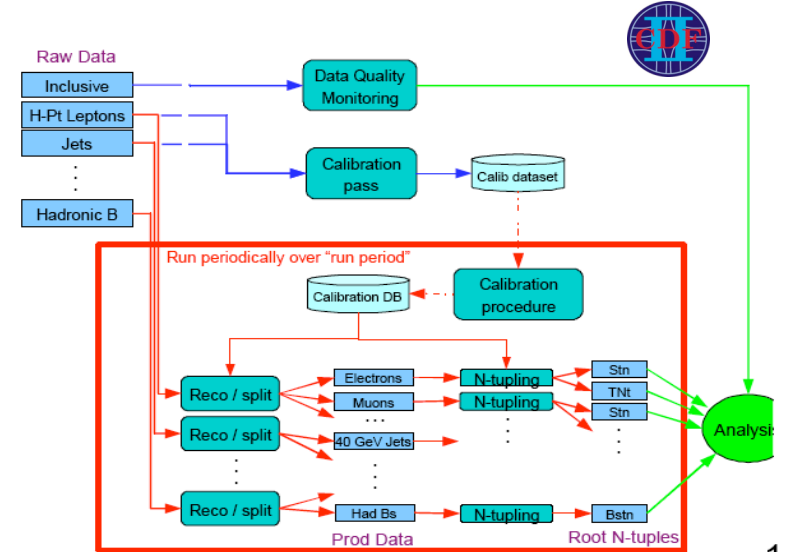
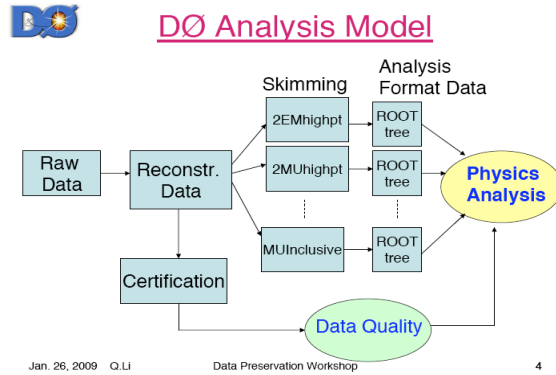
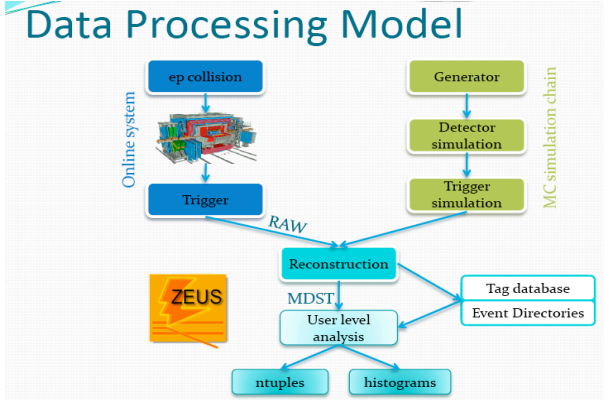
DESY-IT: Volker Gehrmann (DESY)
RI: Christine Dicuon (CPH/DESY)
DESY: Tobias Hees (DESY)
FRAL/DOE: Andrew Robinson (CERN)
FRAL/IT: Volker Gehrmann (DESY)
DO: Daniel Casper (FNAL), Charles Wood (FNAL)
CERN: Benjamin Kropf (DESY), Robert Soper (FNAL)
SHARIT: Geng Chen (DESY)
SEI/IT: Wang Wang (HUST)
KEK/IT: Takashi Sekino (KEK)
SLAC: Nina Yamashita (SLAC), Tom Erbacher (Hawaii)
SLAC/IT: Richard Nayak (SLAC)
ALICE: Francois Le Flander (ALICE)
ALICE/IT: David Soper (CERN)
CERN/IT: Juraj Konecny (CERN)
CERN/INFSE: Salvatore Hahn (CERN)



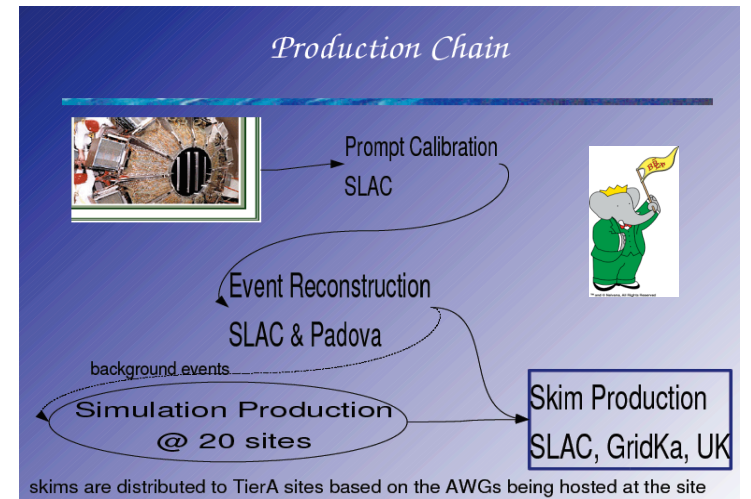
- Workshop held at DESY with ~50 participants and lots of useful discussion
- Two days of pre-prepared talks, third morning of dedicated discussions
- Conference webpage:

<http://indico.cern.ch/conferenceDisplay.py?confId=42722>

HEP Data Analysis Models



- Familiar descriptions of data analysis chain, from reconstruction to analysis level
 - RAW → POT → DST → *ntuple*



Present HEP Experiments

- **Data Format and Volume**

- Differences appear early on: US experiments use skims at early stage: different ntuples for different physics working groups
- Event sizes vary from a few to 100 kB/event; total size of expected data and MC to conserve **0.5 to 10 PB**

- **Software**

- Reconstruction C++ / C / Fortran; Simulation GEANT 3 (Fortran) or 4 (C++)
- ROOT and C++ almost universal as analysis level software (Belle: Fortran)

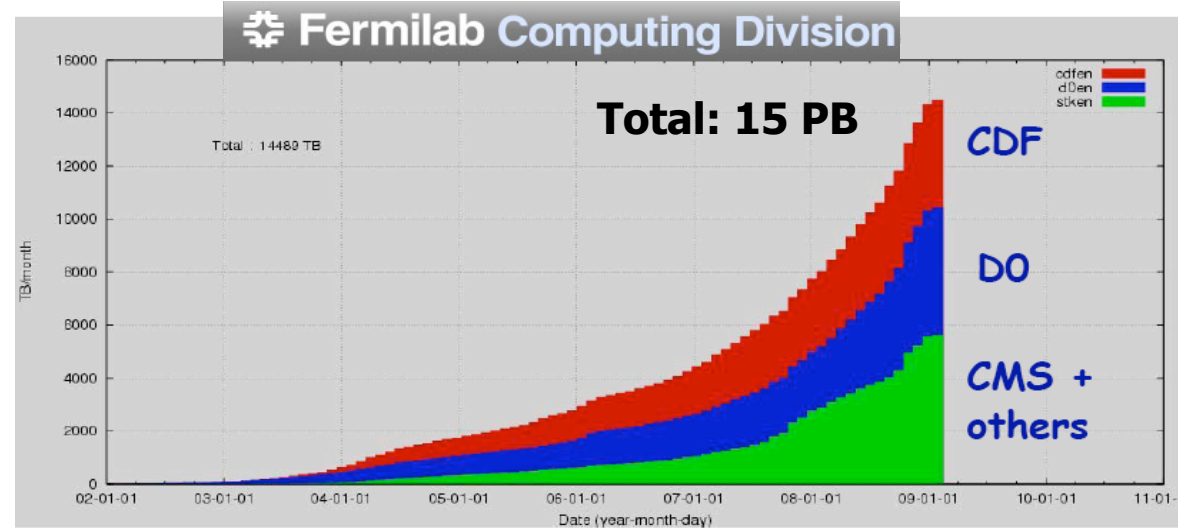
- **Reprocessing**

- Done by many experiments but (for example) CDF plans not to

- **Simulation**

- MC production generally done on GRID, analysis on local farm ($\sim 10^3$ CPUs)

The Challenge of Handling HEP Data

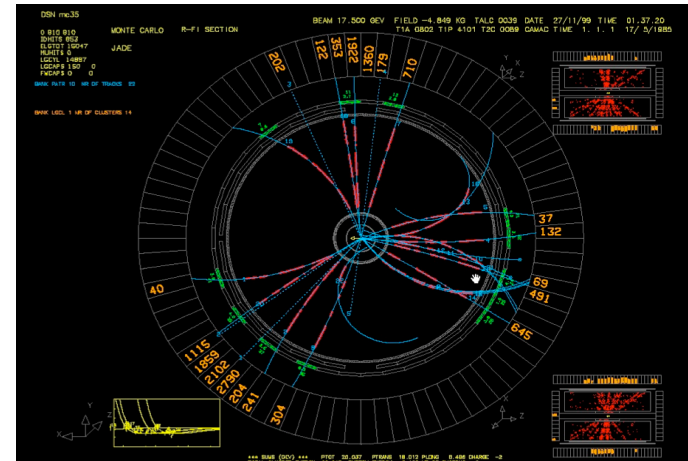


The Challenge of Handling HEP Data

- Massive data traffic, storage and migration is within the scope of all HEP computing centres
- The conservation of tapes is not equivalent to data preservation
 - Older tapes are often not accessible after 2-3 years
- The distribution of data complicates the task
- There is a grey area between the experiments and the computing centres concerning the long term preserved data
 - Missing Hardware: unreadable tapes
 - Software for accessing the data is usually under the control of the experiments

Past Experiences of Data Preservation: PETRA

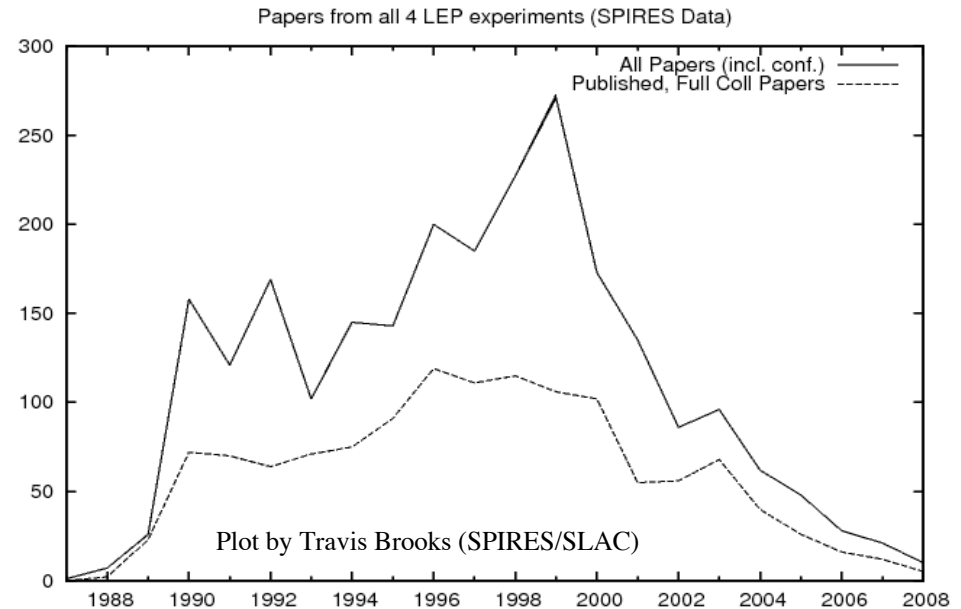
- Recent re-analysis of JADE data is such a case (S. Bethke, J. Olsson et al.)
 - Conversion of old data format done in 2005 and 2008
 - Successful revitalisation and validation of
 - complete JADE reconstruction and simulation software and event display
 - Involved conversion, translation and some rewriting of original code



- There were several interesting JADE anecdotes along the way, including:
 - A hand-typed recovery of a luminosity / calibration file from a (green) paper copy found in Jan's DESY office
 - An old version of BOSlib79 found at the Tokyo computing centre
 - Original JADE MC 9-track tapes found in a Heidelberg University cupboard
- Only through careful documentation will we avoid such things

Past Experiences of Data Preservation: LEP

- There is still a significant number of new LEP publications, approaching 10 years after the end of taking and even after the official end of collaborations



- However, no coherent approach was attempted, no project to preserve the full data analysis capabilities
 - Several analyses still alive (ALEPH laptop model, Higgs group high level data)
 - But in general the preserved data are lacking in standardisation and have limited, model dependent usage
- There is an imminent danger that the LEP data will be absorbed in the "digital black hole" in a few years, if nothing is done

Past Experiences of Data Preservation

- It is likely that most older HEP experiments have in fact simply lost the data
 - ...in some cupboard, in the basement, or trashed...
- For the few known preserved data examples, in general the exercise has not been a planned initiative by the collaboration, but a push by knowledgeable people
- The task in hand is to provide a coherent set of guidelines for future experiments to ensure the longevity of our data

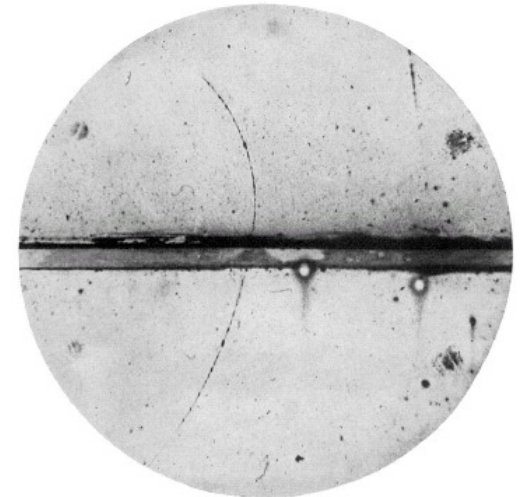


FIG. 1. A 63 million volt positron ($H_p = 2.1 \times 10^6$ gauss-cm) passing through a 6 mm lead plate and emerging as a 23 million volt positron ($H_p = 7.5 \times 10^6$ gauss-cm). The length of this latter path is at least ten times greater than the possible length of a proton path of this curvature.



Goal of the data preservation: reuse at a later stage

... correctly



Needs some planning

"Don't be ridiculous Caruthers, you must have mistranslated it. How can it possibly say, King Ramases@www.ram2.com?"

Plans for a Systematic Approach to Data Preservation in HEP

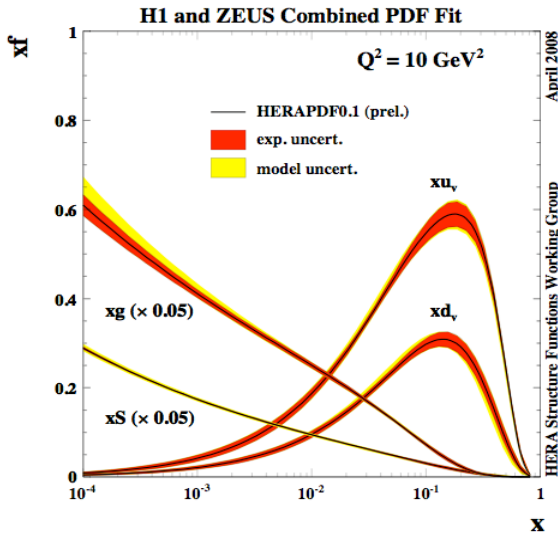
- Physics Cases for Data Preservation in HEP
 - Survey of possible benefits from data preservation
 - Including business models
 - Including links with other research fields
- Preservation Models
 - Input from ee, ep, pp experiments
 - Priorities, costs and benefits, links to technology
- Collaborations, Governance, and Data Access Policies
 - Including contacts with general initiatives
- Technologies and Facilities
 - Survey and assessment of existing infrastructures in HEP and their adaptability to data preservation demand
 - Reflection on the impact of new technologies on the data preservation methods

Possible Future Use Cases

- Reanalysis of the old data for better precision
 - Better systematics, new simulations
 - Investigate observations in more recent experiments
- Combination of similar (“sister”) experiments
 - Gain in precision
- New experiments compare/combined with data from the previous generation
 - Energy dependence, cross-checks
- Data used in global analyses, based on new theories

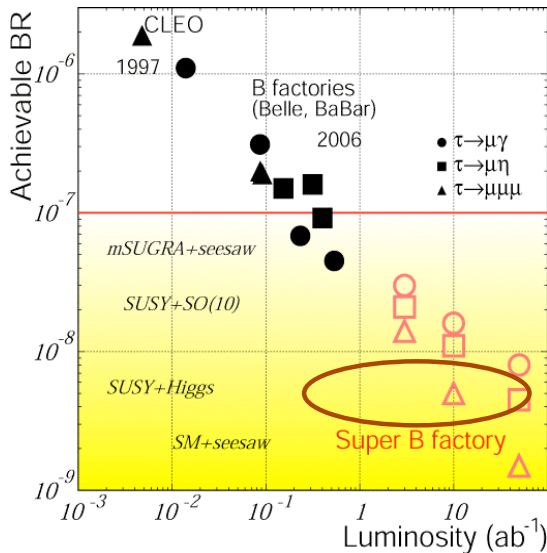
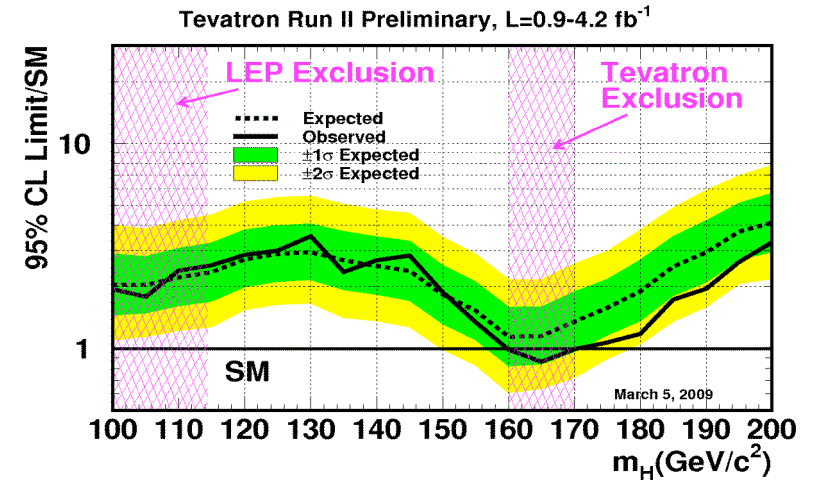
- The costs and benefits of such cases need careful evaluation

Physics Case for Data Preservation



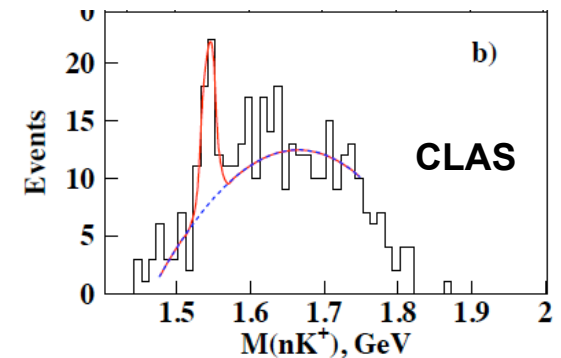
- The ep collisions recorded at HERA are a unique data set, unlikely to be superseded in the near future (LHeC?)

- The pp collision data from Tevatron will also provide a contingency for LHC data, as well as a lower energy point



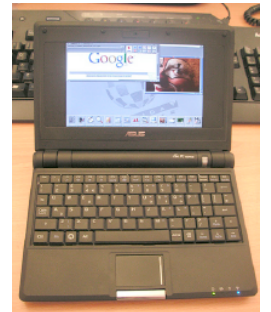
- As already shown, ee data may still provide future interest, e.g. for comparison to Super-B, in particular if data sets can be combined (BaBar + Belle in trials now)

...surprises can occur at lower energies too



Models for Data Preservation

- The HEP models could follow one of the three directions already discussed elsewhere (DPC handbook)
 - **Technology preservation**
 - Freeze the hardware : limited capability, one day it will fall apart however
 - **Technology emulation**
 - Prepare it once (?), migrate the “middleware”
 - **Continuous migration**
 - Follow technology changes (adjust, redesign, recompile etc....)



Models for Preservation

D.South

Minimum Level of Preservation	
0	RAW data
1	Reconstruction Simulation Database considerations? Commercial software?
2	DST
3	Ntuple / analysis level data (and MC?) <i>production</i>
4	Existing ntuple / analysis level
5	Combined analysis with a (for example) H1+ZEUS "ep ntuple"
6	Outreach : very simple format

The basic level to conserve

Essentially frozen, but reconstruction software still compiles, so changes are possible...

A new simulation: can it use old reconstruction (issue of F vs C++)?

DST level expects no further development, this is the final version

Rolling model, fluid preservation from here up: gives regular verification of full chain

Fixed ntuple, "all" analysis level info

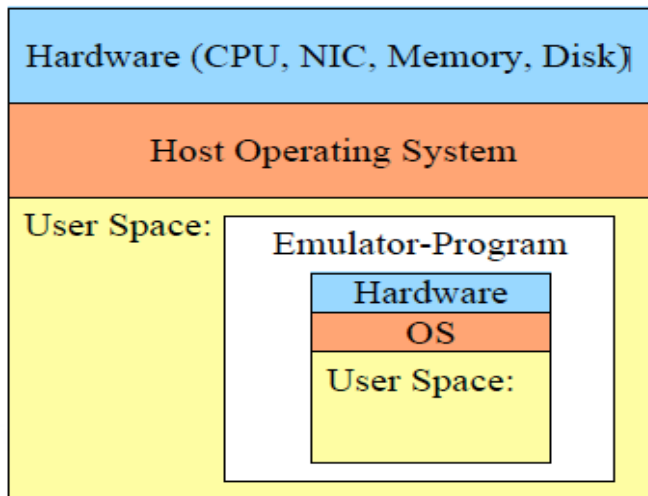
Common format ntuple (repository?)

Not enough for full analysis(?), but rather for open access / outreach

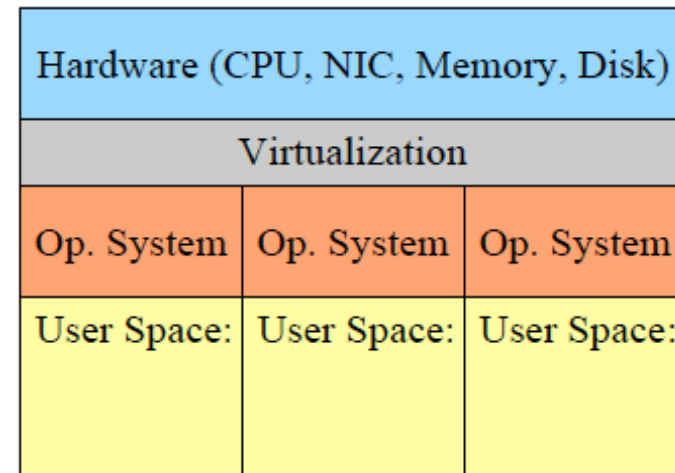
Use of Virtualisation / Emulation techniques ?

Emulation and virtualisation

Emulation



Virtualisation



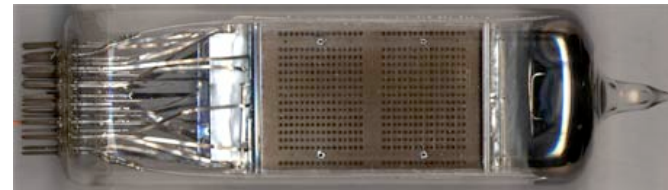
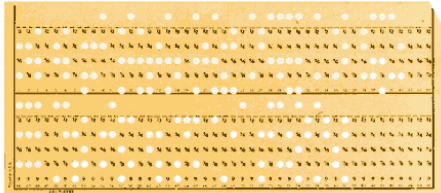
Y.Kemp

An different operating system can be “preserved”
Can a HEP computing environement also be preserved this way?

Virtualisation

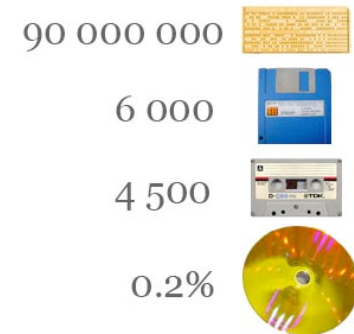
- The computing problem for a hibernated experiment is (should be) better defined than for a running experiment, but the computing performance increases
 - Change of computing paradigm?
 - The computing model, nowadays distributed most of the time, tend to become “local”
- Virtual computing seem a promising solution to avoid technological hick-ups
 - There is no miracle: work is needed to prepare the HEP computing models for virtual computing
 - Babar pioneers this technique (H. Neal)
 - SL4 software compiled on SL5 VM, work in progress for real scale demonstration
- Full scale virtual HEP models are a challenge and cannot overcome the necessary post-mortem “clean-up” work
- Most likely the technology and the virtual models for DP are complementary

Data migration



Data Migration and new technologies

- Continuous migration of archived data works if:
 - The next generation media costs half as much
 - All media is robot managed (no shelved tapes!)
 - Migration is possible on a short timescale
- Future mass storage media: SSD (Flash..)
 - More reliable than disk or tape
 - Requires less power than disk (x10)
 - Currently more expensive (x20), but coming down
 - Density increasing with respect to disks (smaller footprint)
- Consistent data migration policy still to be defined
 - Data loss and risk analysis
 - Geographical placement of the data sets
 - Uniform framework for scientific data management

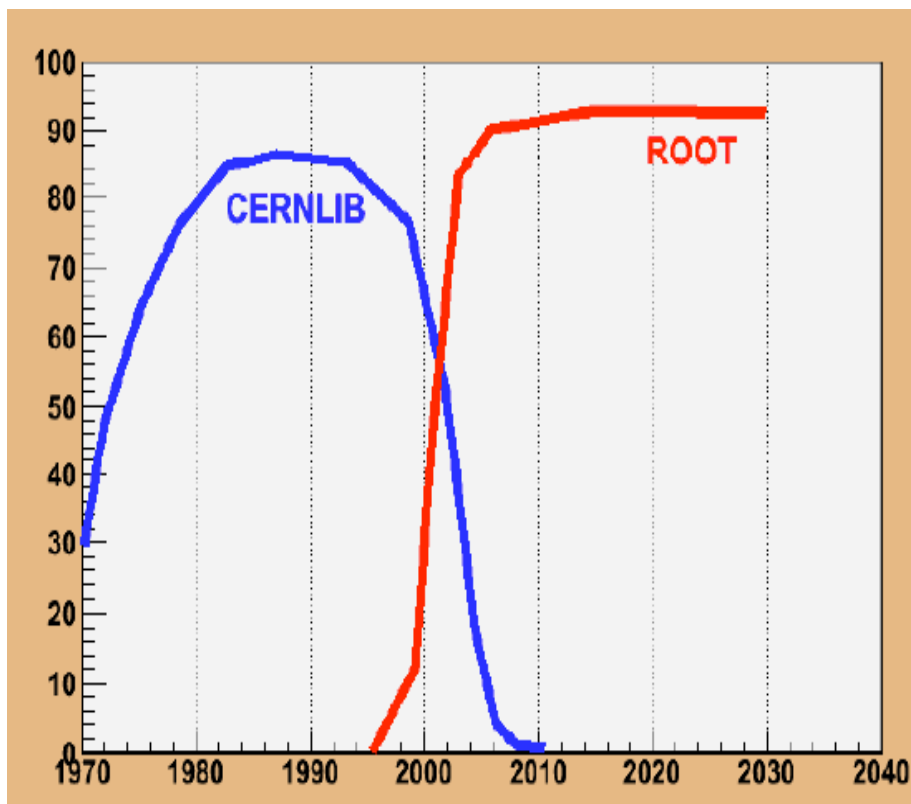


**Total cost of data migration =
double current costs:
 $1 + 1/2 + 1/4 + 1/8 .. = 2$**

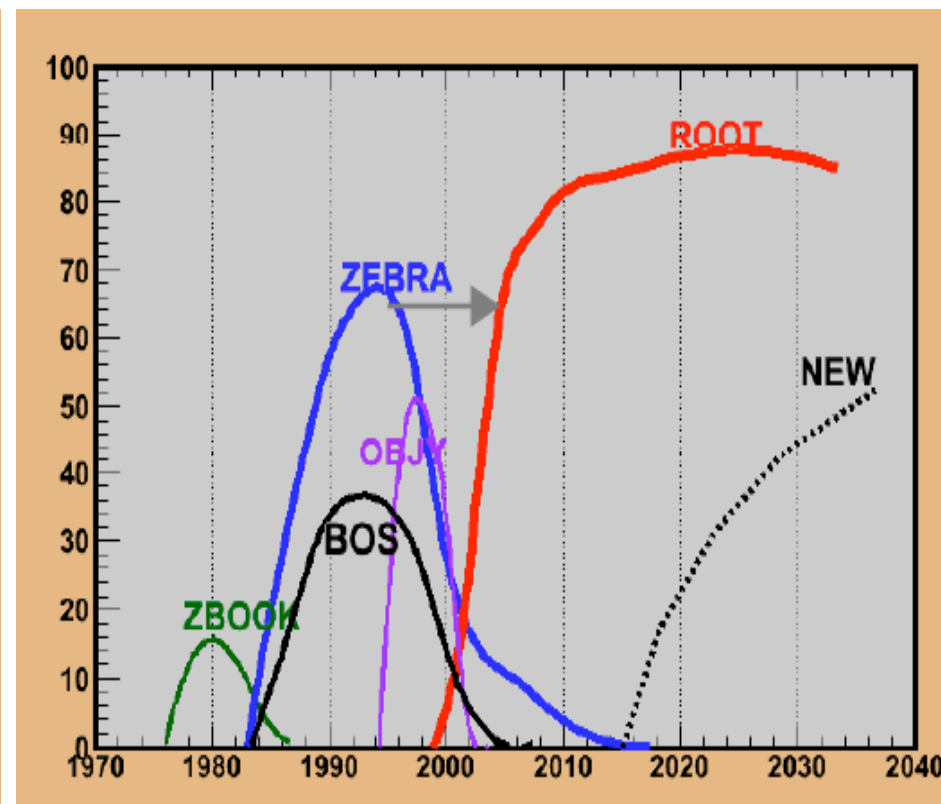


Analysis software

Libraries



I/O



R. Brun

Root offer the needed coherence in the next few decades
But many other dangers: comercial, “ghosts” etc.

Data Access and Supervision in Long Term

- Data preservation costs money and it needs organisation
 - HEP Organisations need to take into account appropriate budget planning
- HEP data preservation makes sense if the appropriate supervision structures are defined
 - Long term evolution of the HEP collaborations need to be initiated during the active lifetime
 - Data access and publication procedures for a hibernated collaboration should be defined
 - Data stewardship in computing centres must be defined taking into account the preservation models and included in the economical considerations of the DPHEP projects

Data preservation: a simple model

Define

- Supervision structure
 - Long term collaborative and governance aspects
- Data access and publication protocols
 - Workflow: use generic models?
- R&D project
 - Objectives, resources, time profile

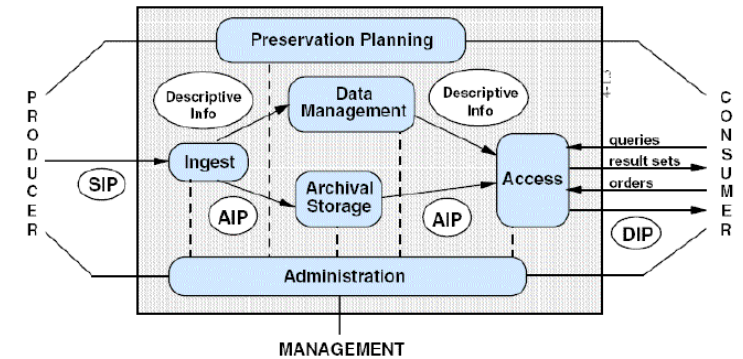
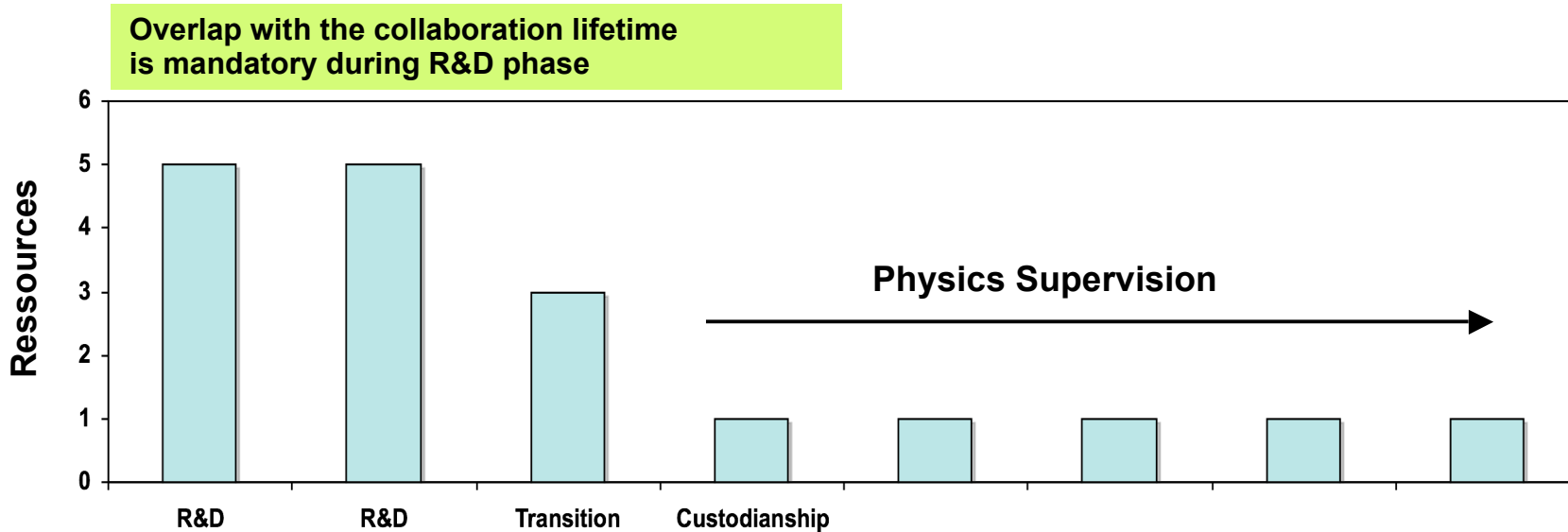



FIGURE 2.1: **The OAIS Reference Model**
<http://public.ccsds.org/publications/archive/650x0b1.pdf>, Page 4-1.
 Source: Consultative Committee for Space Data Systems January 2002.



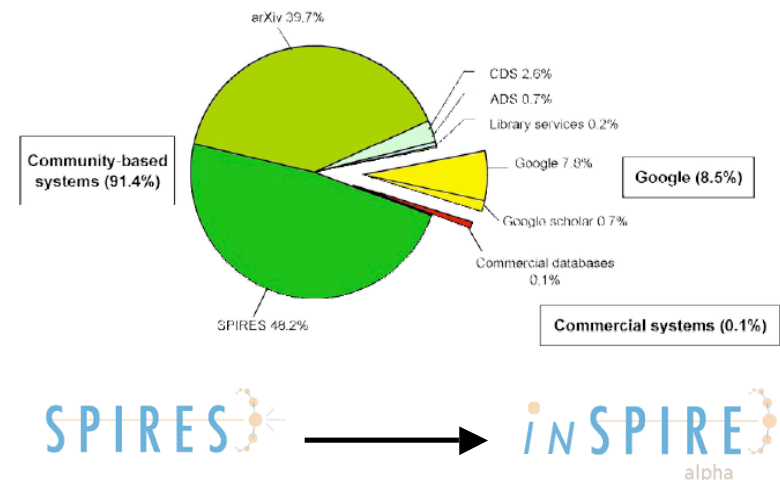
Global Solutions in HEP

- Common repositories for data exists for other domains (Protein Data Bank, Astronomical Virtual Observatory, etc.)
 - Common data format(s)
 - Agreement on data access policies
 - Supervision process
- Published HEP data is well retained
 - Journals, HEP Data base, PDG
 - SPIRES soon to transform into INSPIRE, whereby they will offer custodianship of more than scientific articles
- Is a common HEP repository a feasible idea?
 - Abstraction level should be uniform: a big challenge



HEPDATA: REACTION DATA Database
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the Durham Database Group (UK) with help from the COMPAS group (Russia.) and is updated at regular intervals.

SURVEY OF OVER 2000 PHYSICISTS
Which HEP information system do you use the most?



T.Brooks

Summary

- HEP data is a long term investment and contains a true potential for physics results beyond the collaborations lifetime
- A study group has been formed to reflect on data preservation and long term analysis in HEP: <http://dphep.org>
- The aim of this initiative is to provide a written document to ICFA containing guidelines on this subject

Next workshop is at SLAC, May 26-28 2009.

Thanks: David South, Homer Neal, Gregory Dubois-Felsman, François Le Diberder, Richard Mount, Volker Guelzow, Martin Gasthuber, Tobias Haas, Krzysztof Wrona, Volker Guelzow, Amber Boehnlein, Takashi Sasaki, Qizhong Li, Eric Varnes, Bogdan Lobodzinski, Benno List, Jan Olsson, Frederic Hemmer, Dmitri Ozerov, Rene Brun, Marcello Maggi, Salvatore Mele, Andre Holzner, Gang Chen, Stephen Wolbers, Jonathan M. Dorfan, Fabio Hernandez, Martin Gasthuber, Daniel Riley, David Corney, Peter Igo-Kemenes, Matthias Schroder, Janusz Szuba, Yves Kemp, Travis Brooks, Serguei Levonian, Gunar Schnell, Fabio Pasian, Siggie Bethke

Backup

DPLTA Workshop Agenda (January 26-28, 2009)

Reports from Experiments: Data Analysis and Computing Models

Conveners: Homer Neal; Robert Roser (Fermilab)
(Seminar 4 (EVO): 09:00 - 12:20)

Computing Centres and Technologies

Conveners: Volker Guelzow (Unknown); Frederic Hemmer (CERN)
(Seminar 4 (EVO): 14:00 - 17:00)

Workshop Discussions: ee Experiments (Parallel)

(Seminar 4 (EVO): 17:00 - 18:00)

Workshop Discussions: ep Experiments (Parallel)

(Seminar 3a: 17:00 - 18:00)

Workshop Discussions: pp Experiments (Parallel)

(Seminar 5: 17:00 - 18:00)

Reports from Past Experiences of Data Preservation

Conveners: Tobias Haas; Takashi Sasaki
(Seminar 4 (EVO): 09:30 - 12:45)

Open Access and Long Term Collaborative Governance

Conveners: Richard Mount; Siegfried Bethke (Max-Planck-Institut fur Physik)
(Seminar 4 (EVO): 14:00 - 17:05)

Workshop Discussions: Options for Long Term Data Analysis (Summary Discussions)

Conveners: Cristinel Diaconu (Faculte des Sciences de Luminy)
(Seminar 4 (EVO): 09:30 - 12:40)

- 2.5 days, 24 talks, 3 parallel sessions, 8 organised discussions