Data Preservation and Long Term Analysis in HEP

Study Group for Data Preservation and
Long Term Analysis in High Energy Physics

http://www.dphep.org

Digital data is fragile

"Digital information lasts forever -- or five years, whichever comes first."

Jeff Rothenberg, RAND Corp.

- Task forces already in place to address this issue in a generic way (standards)
  - e.g. Blue Ribbon, APA, DPC, eSciDir, …

http://www.alliancepermanentaccess.eu
http://brtf.sdsc.edu
(intermediate and final report and references)

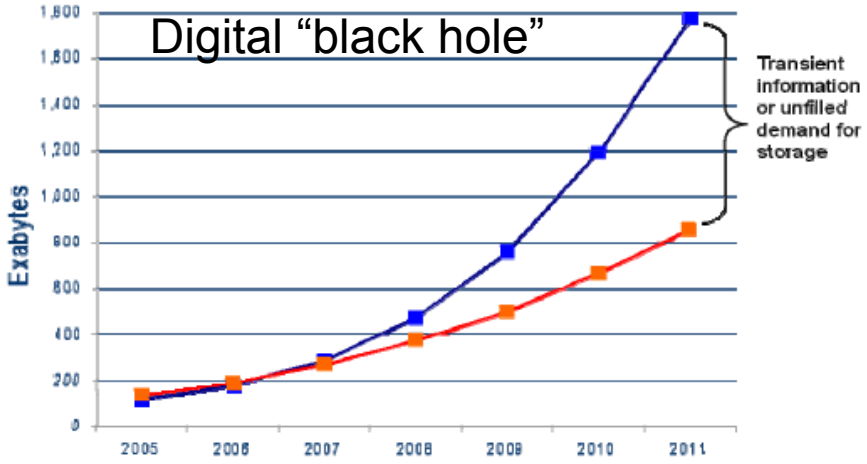


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

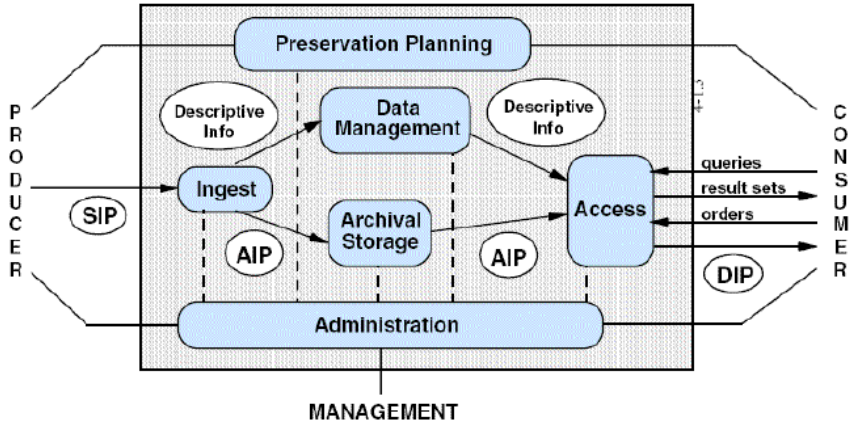


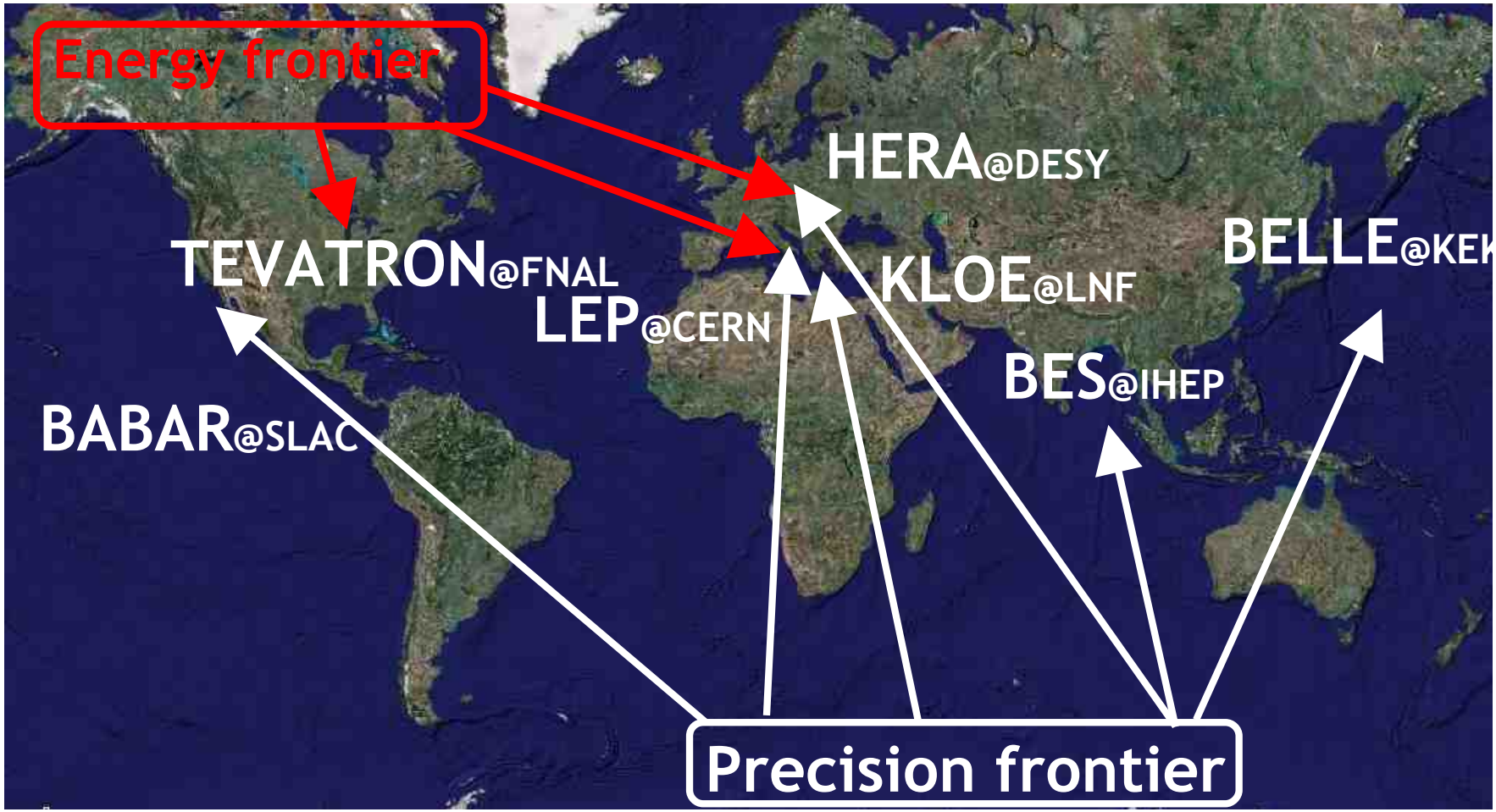
FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

- Scientific Data is a major component of the ongoing efforts (complexity)
- Some scientific fields are well advanced : astrophysics

Data in High Energy Physics

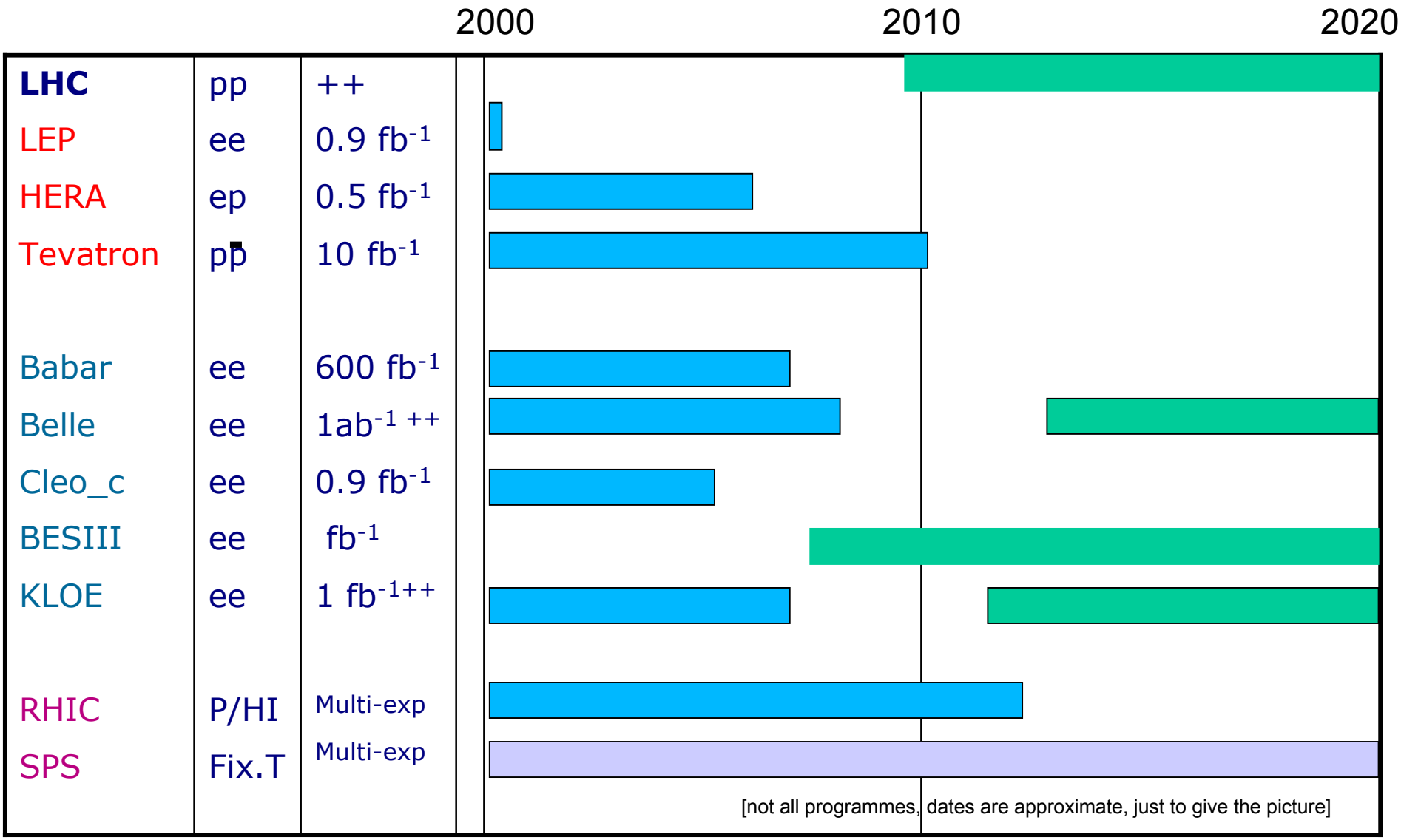
The pre-LHC landscape

colliders



+many other experiments fixed target, neutrino etc.
(CERN gray book: ~900 experiments)

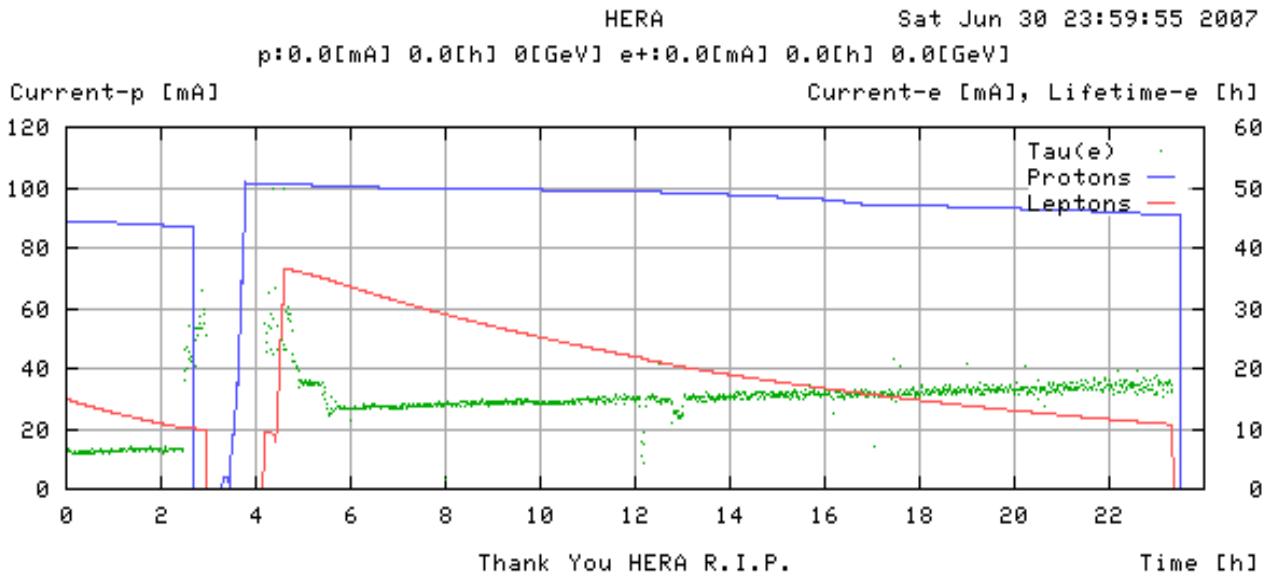
HEP experimental programs in +-10 years

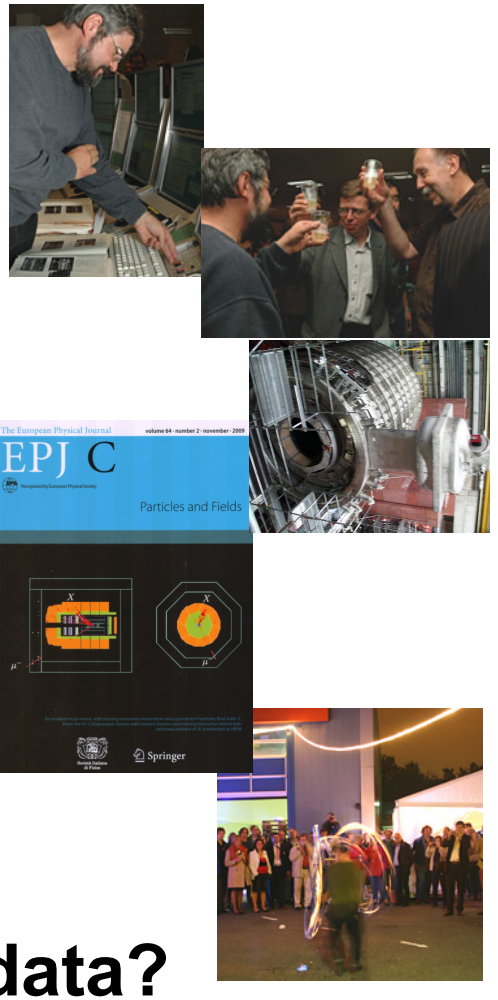


Unique data sets available: what is their fate?

After the end of the data taking:

do a party, dismantle detector, finalise the analyses, re-do a party : ~5 years



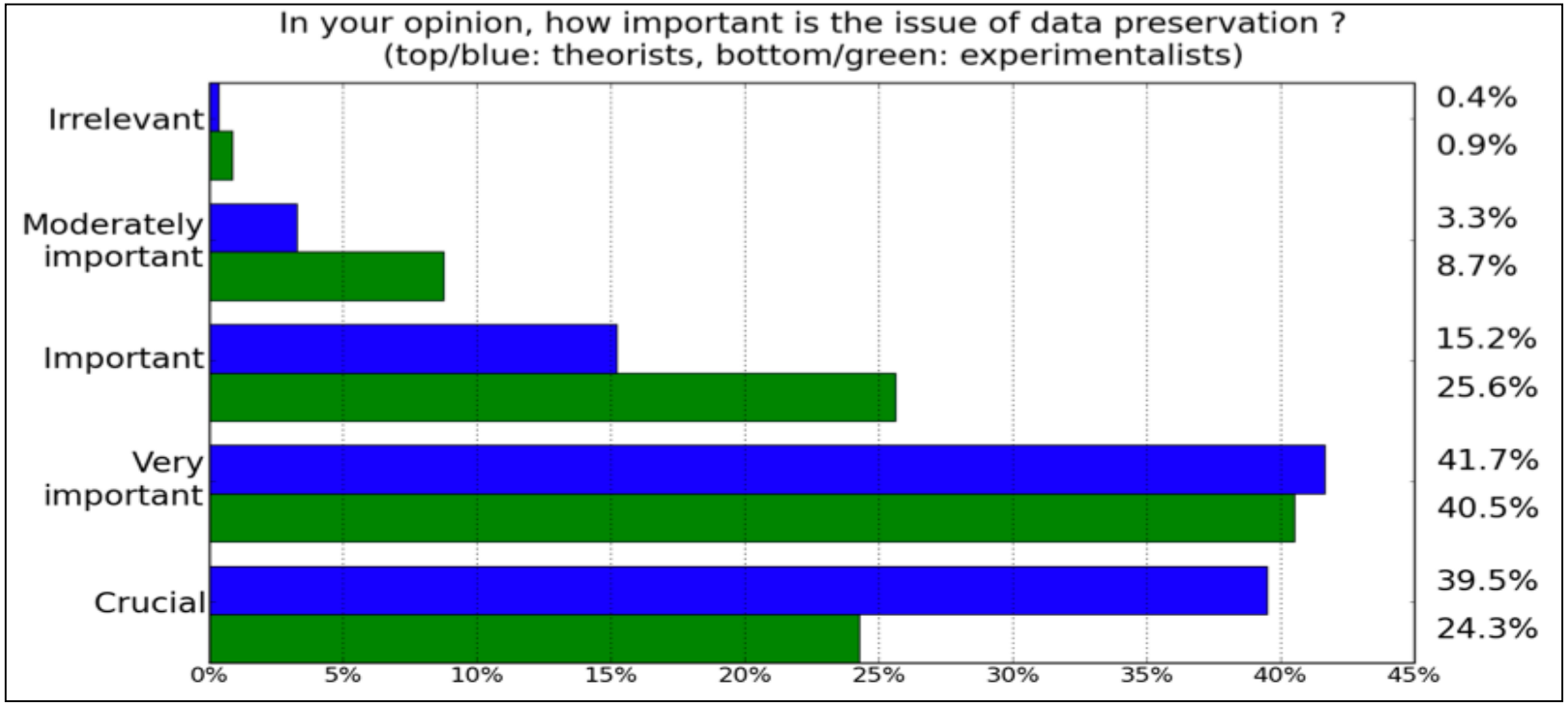


…what do you do with the data?

One idea…

Data Preservation: support in the HEP community

http://arxiv.org/abs/0906.0485
PARSE.Insight is financed by the European Commission and run at CERN

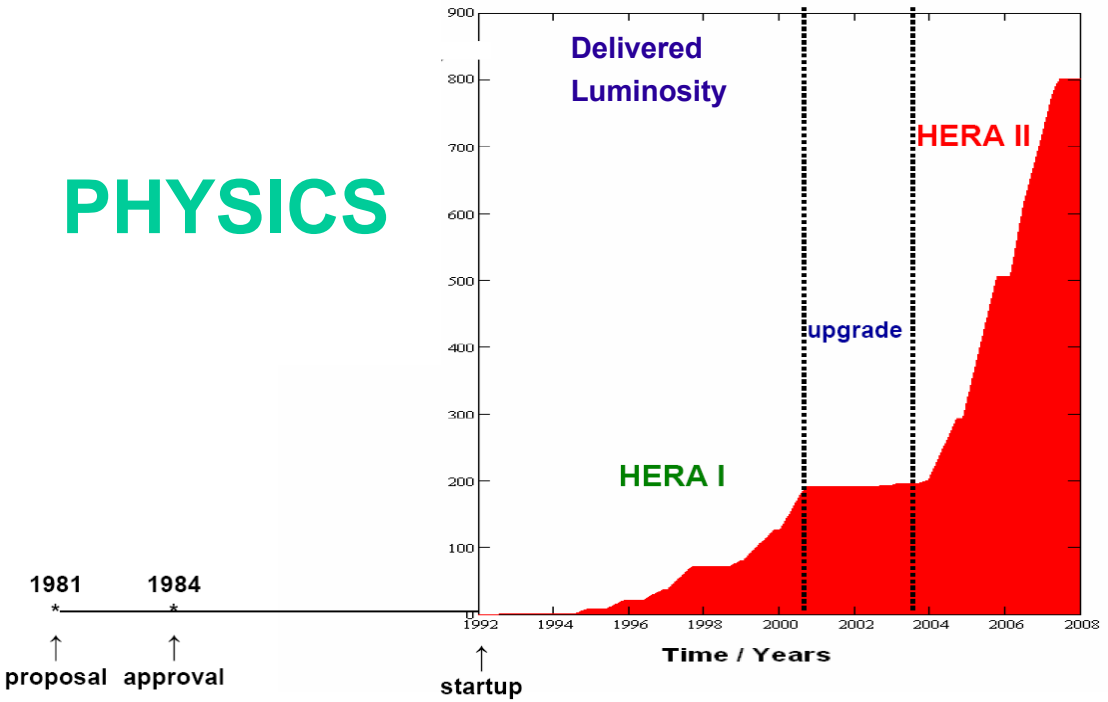


70%: very important or crucial
However, no coherent strategy exists: in general, HEP data is lost

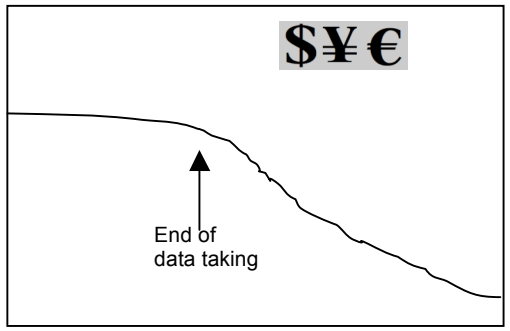
Why is difficult to preserve HEP data?

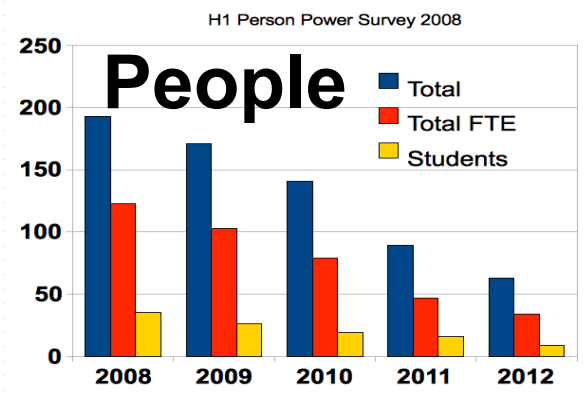
- Good physics is collected at the end, but:
- The resources decrease after the end of data taking
  - Dedicated resources need to be planned

PHYSICS



Funding





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…
  - (see later an exception)

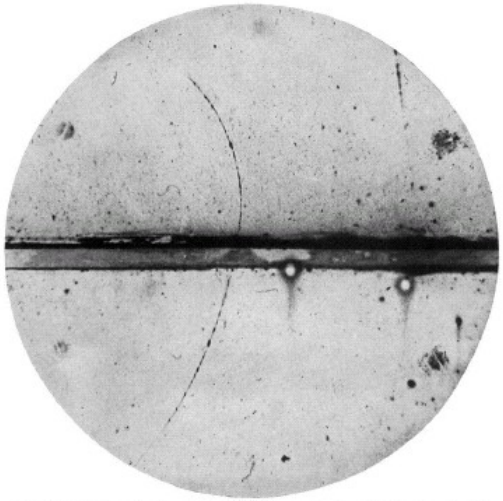


FIG. 1. A 63 million volt positron ($H\rho = 2.1 \times 10^5$ gauss-cm) passing through a 6 mm lead plate and emerging as a 23 million volt positron ($H\rho = 7.5 \times 10^4$ gauss-cm). The length of this latter path is at least ten times greater than the possible length of a proton path of this curvature.

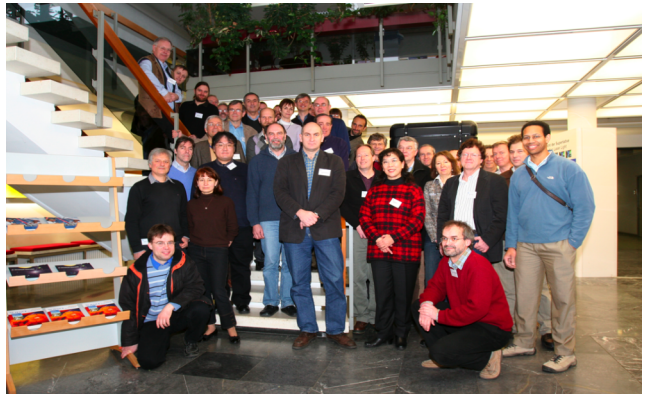
- For the few known preserved data examples, in general the exercise has **not** been a **planned** initiative by the collaboration, but a push by knowledgeable people

- The task in hand is to provide a **coherent set of guidelines for future experiments** to ensure the longevity of our data

International Study Group on HEP Data Preservation

- Collider Experiments
  - e^+e^-, ep, $p\bar{p}$
- Computing Centers
- Contacts with funding agencies

- About 50 contact persons



Coordination

Chair: Cristinel Diaconu (DESY/CPPM)

Working Groups Convenors:

Physics Case	François Le Diberder (SLAC/LAL)
Preservation Models	David South (DESY) , Homer Neal (SLAC)
Technologies	Stephen Wolbers (FNAL), Yves Kemp (DESY)
Governance	Salvatore Mele (CERN)

International Steering Committee

DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Stefan Soldner-Rembold (Manchester)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masanori Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (SLAC/LAL)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)
CLEO: David Asner (Carleton)
STFC: John Gordon (RAL)
JLAB: Graham Heyes (JLAB)

International Advisory Committee

Chairs: Jonathan Dorfan (SLAC) and Siegfried Bethke (MPI Munich)
Advisers: Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo)

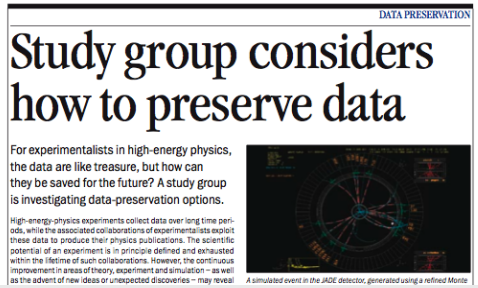
Activity

- Study Group Initiated in September 2008
- Workshops in 2009: DESY, SLAC, CERN (Dec. 7-9)
  - 30-40 participants, experiments represented
  - Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields (astrophysics)
- Report for ICFA released in november 2009 arXiv:0912.0255

DPHEP DESY

DPHEP @ SLAC NATIONAL ACCELERATOR LABORATORY

DPHEP CERN

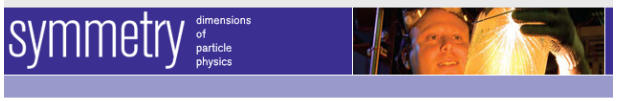

DATA PRESERVATION

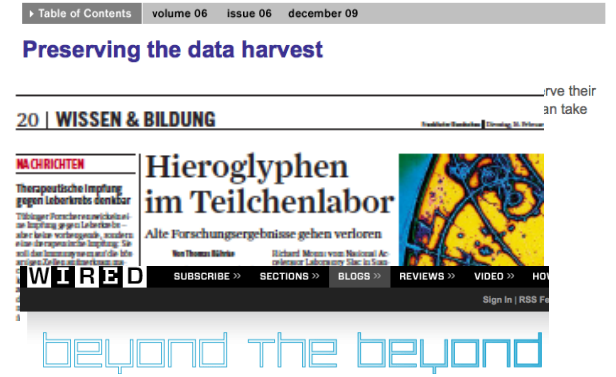
Study group considers how to preserve data

For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

High-energy-physics experiments collect data over long time periods, while the associated collaborations of experimentalists exploit these data to produce their physics publications. The scientific potential of an experiment is in principle defined and exhausted within the lifetime of such collaborations. However, the continuous improvement in areas of theory, experiment and simulation – as well as the advent of new ideas or unexpected discoveries – may reveal

A simulated event in the JADE detector, generated using a refined Monte





DPHEP-2009-001
July 30, 2009

Data Preservation in High-Energy Physics

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

http://dphep.org

Abstract

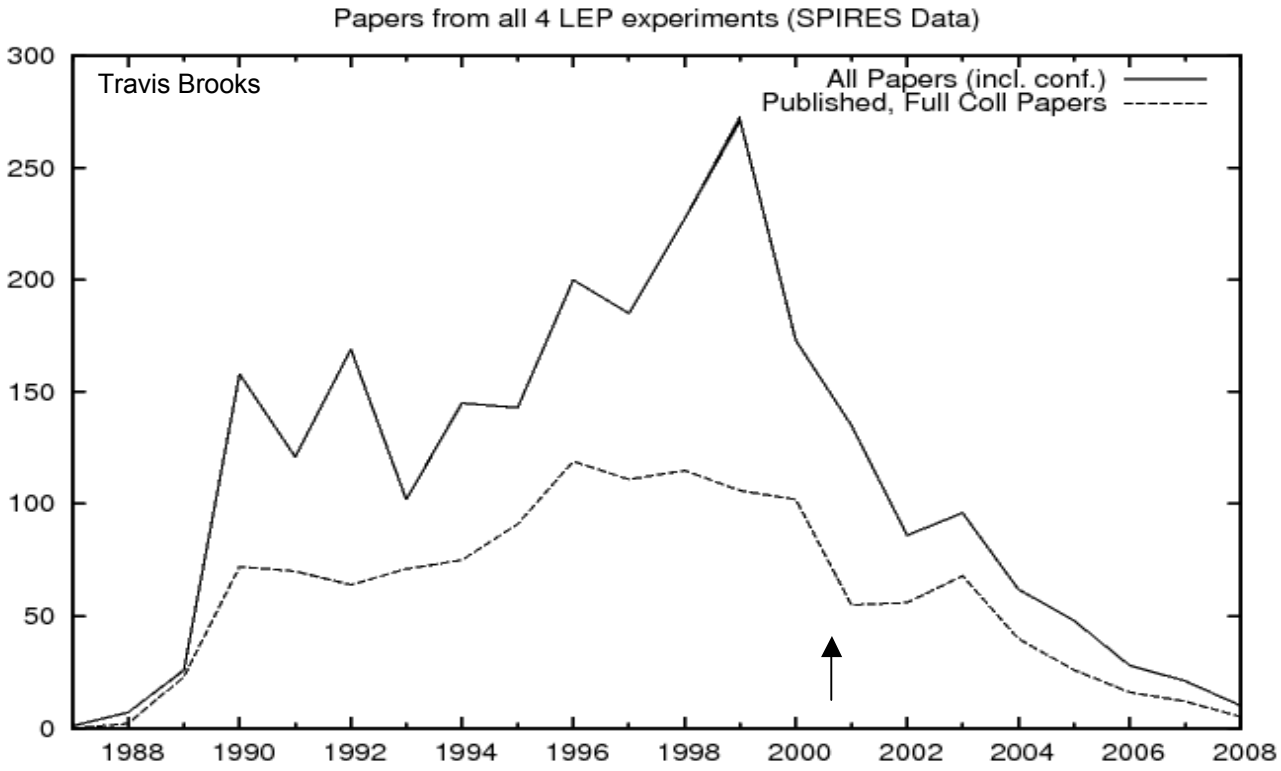
Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened at the end of 2008 and held two workshops, at DESY (January 2009) and SLAC (May 2009). This document is an intermediate report to the International Committee for Future Accelerators (ICFA) of the reflections of this Study Group.

Physics Case

- Collected data sets are mostly unique and have a true scientific potential
  - Long term completion and extension of the physics program
  - Cross collaborations
  - Data re-use
  - Scientific training, education, outreach

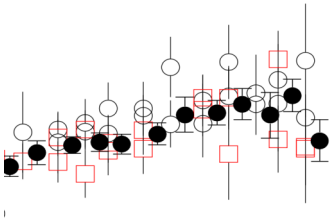
Physics Case I

- Long term completion and extension of the physics program



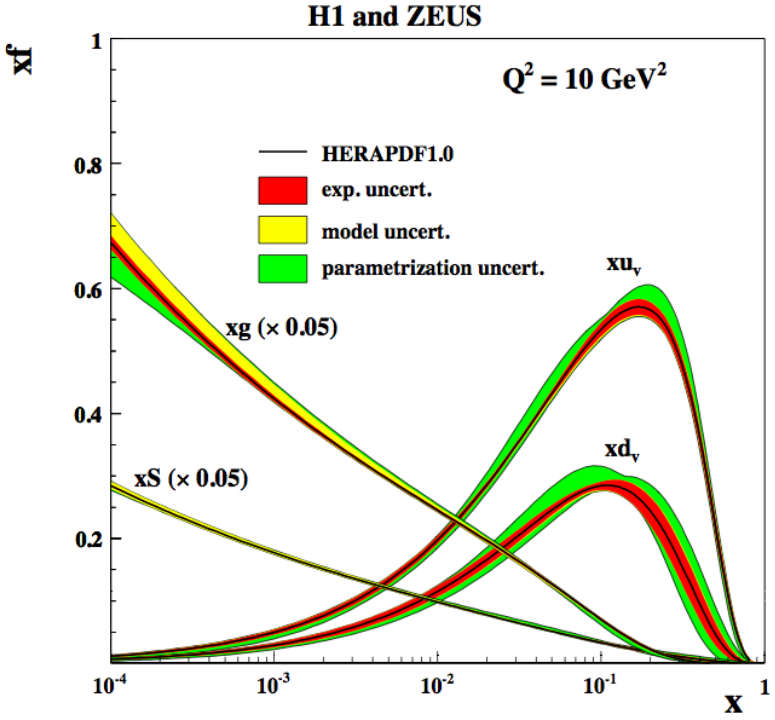
Physics subjects are published after the end of collisions/collaborations
5-10% of the papers are finalized in the “archival mode”

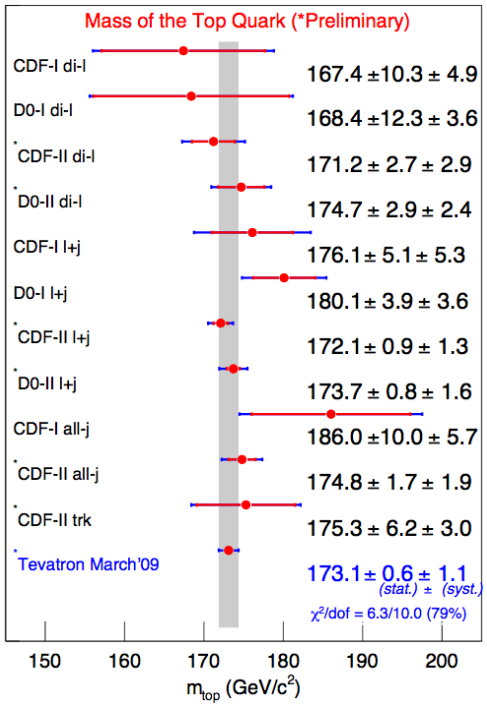
Physics Case II



- Cross collaborations

Already exist at LEP, Tevatron, HERA, Babar+Belle (in progress)

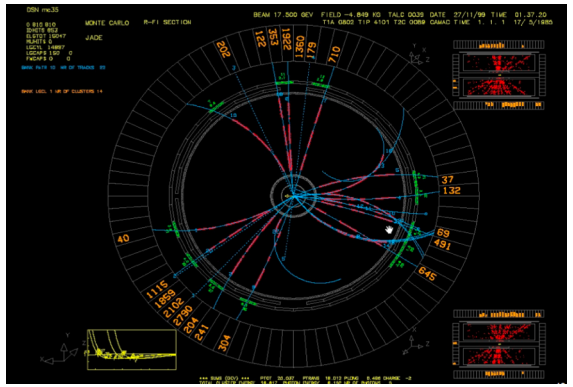




Preserved data would make possible more combined analyses across experiments

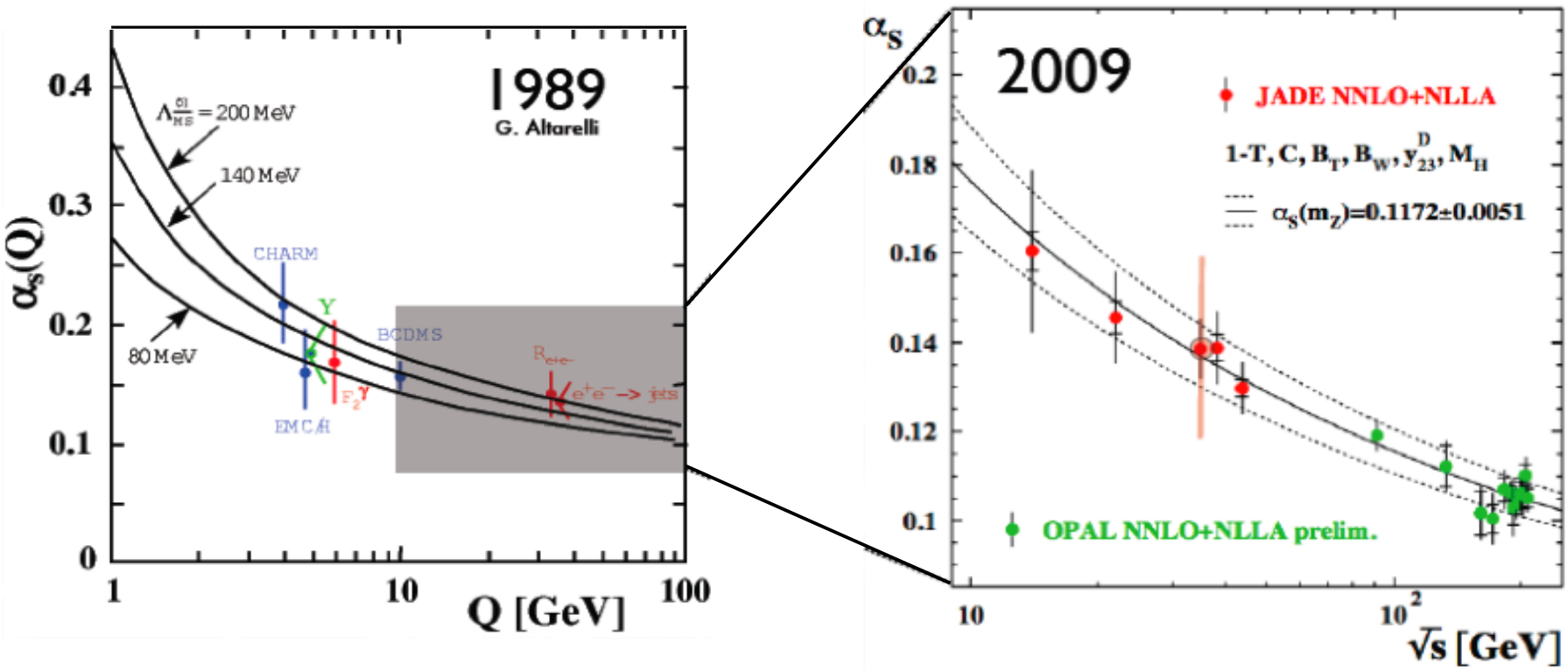
Physics Case III

- Data re-use
  - Improve precision on former measurements
  - apply new and improved theoretical predictions
  - check new physics in the old data samples
  - investigate discrepancies



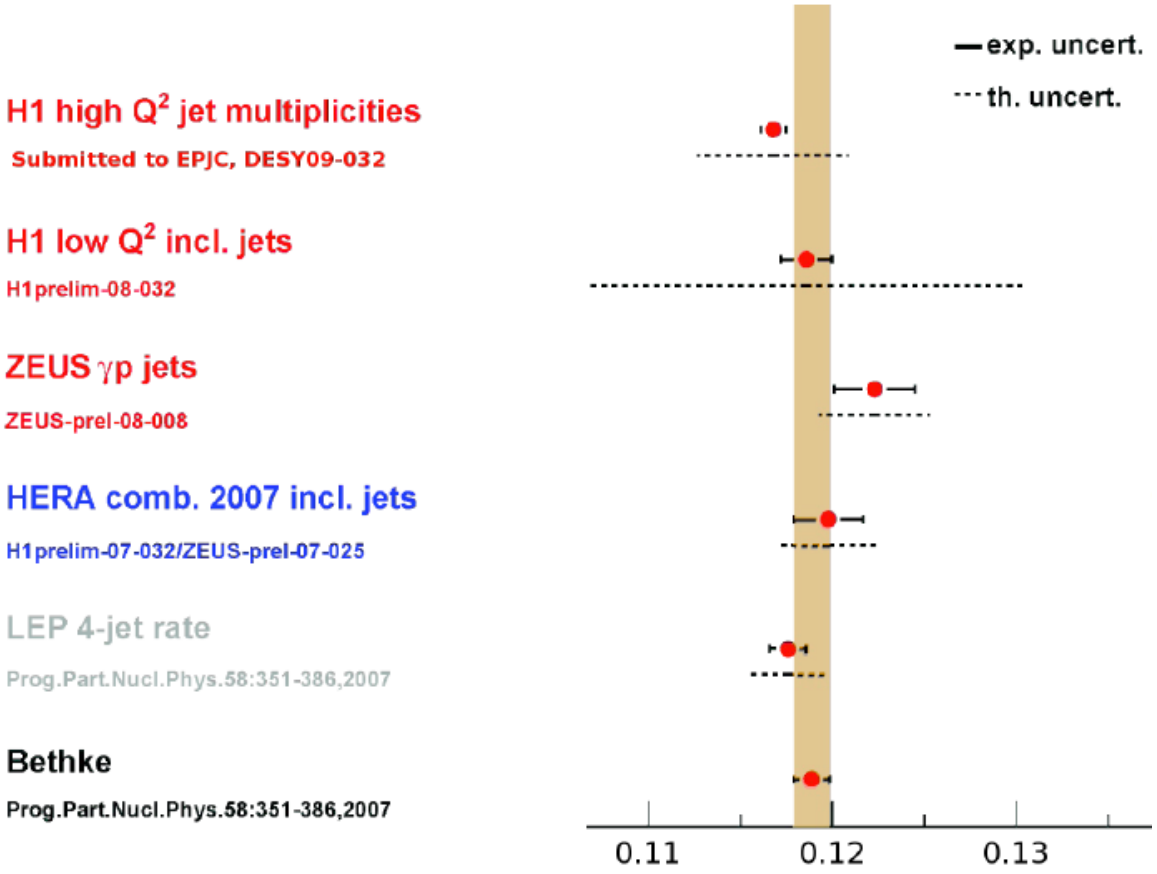
JADE: raw data preservation, software revitalisation individual initiative

10 publications



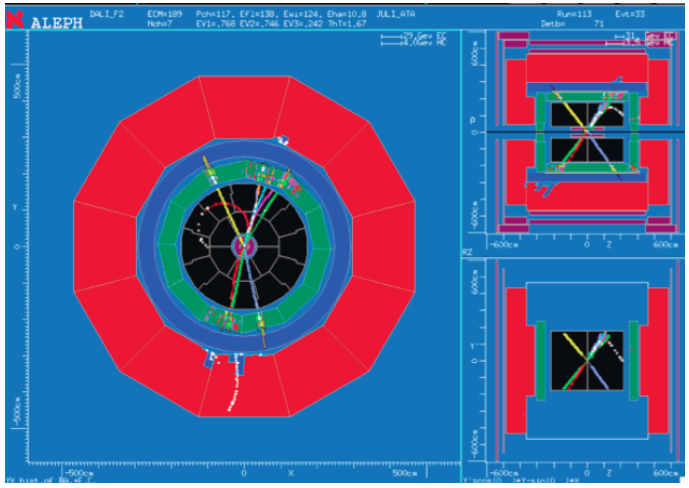
The history may well repeat itself….

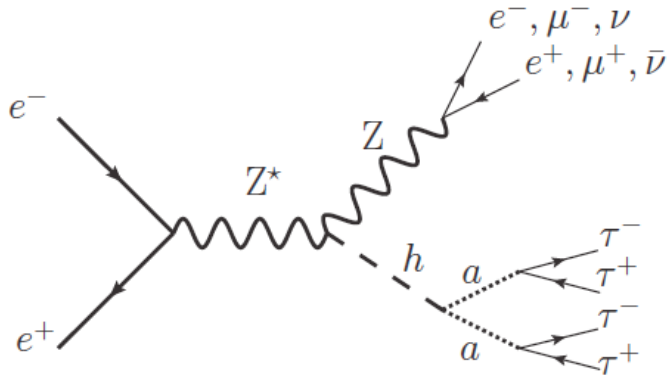
- ~10% of the measurements are dominated by non-experimental errors: theory, simulation



And actually just did….

- ALEPH analysis 2010: data/simulation rescued by a PhD student
- Unique physics case analysed 10 years after the end of collisions (5 years after the official end of the collaboration)

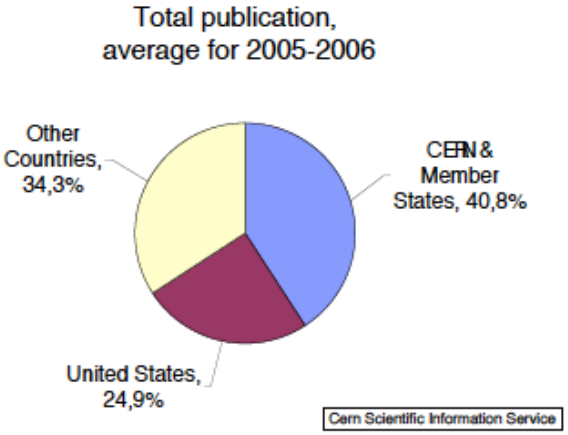


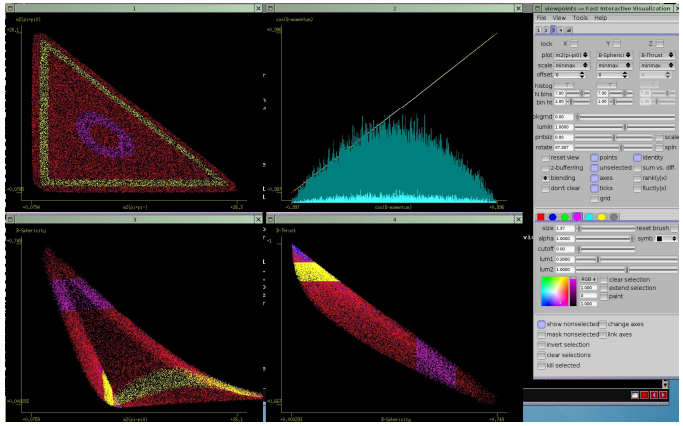


Physics Case IV

- Scientific training, education

Outreach



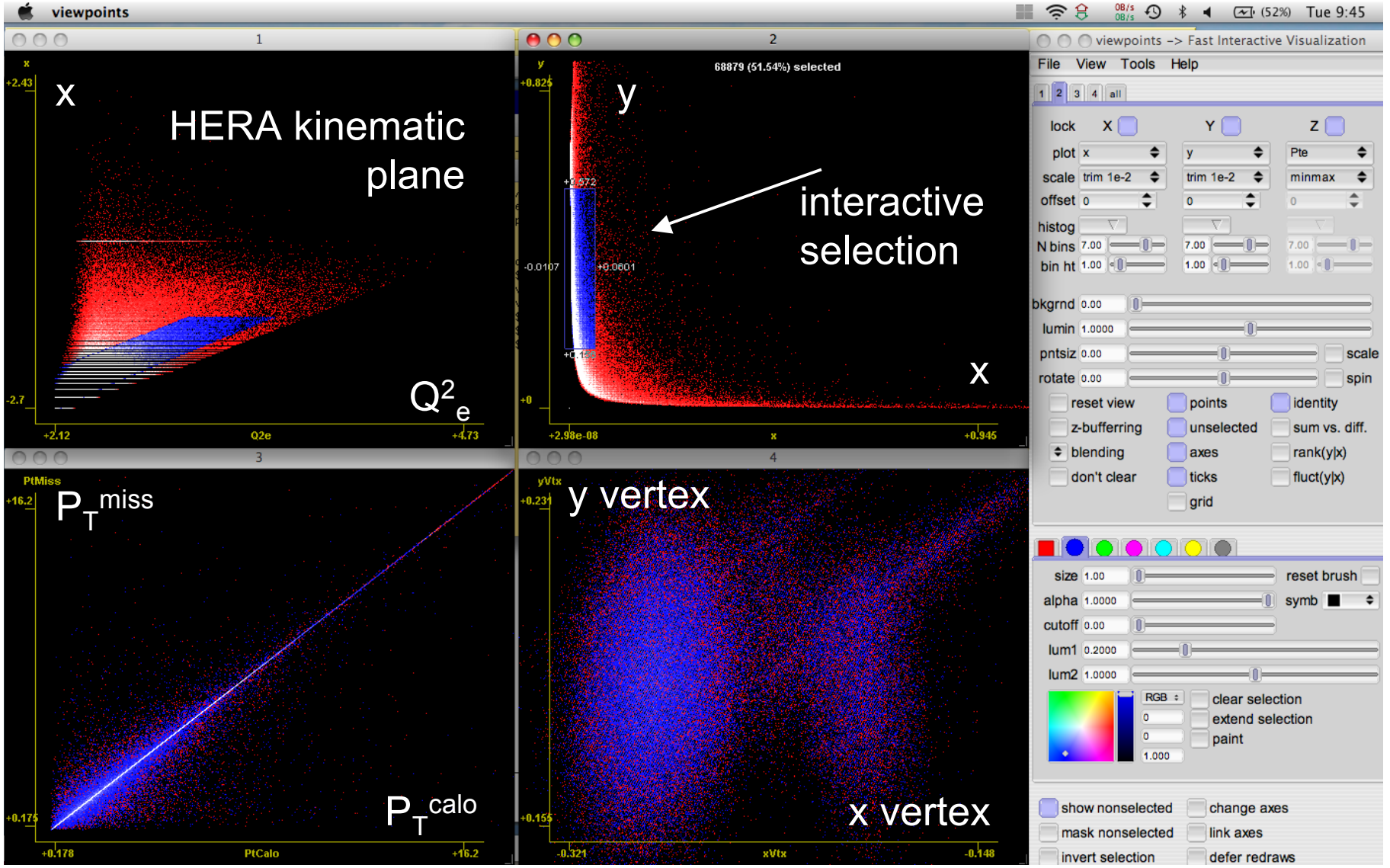


M. Bellis

Improve the overall high level education in HEP

Improve the connection of HEP-emerging countries to HEP data sets

Many attractive outreach tools available, like *Viewpoints* from NASA



Shown here is some H1 NC data, showing some basic HERA kinematics

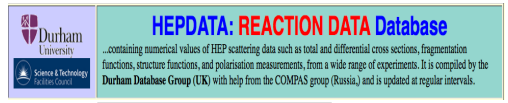
What is “HEP data”?

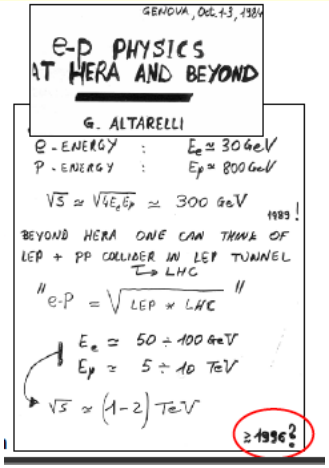
entropy

- Publications (journals, arxiv, spires, hepdata….)
- Digital information: event files, database
- Software: simulation, reconstruction, analysis, user
- Documentation: publications, notes, manuals, slides
- “Meta” information: news, messages
- Expertise (people)

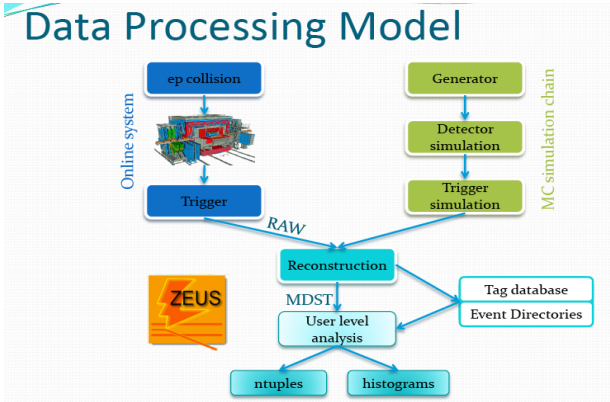
Journal of High Energy Physics
A refereed journal, written, run and distributed by electronic means

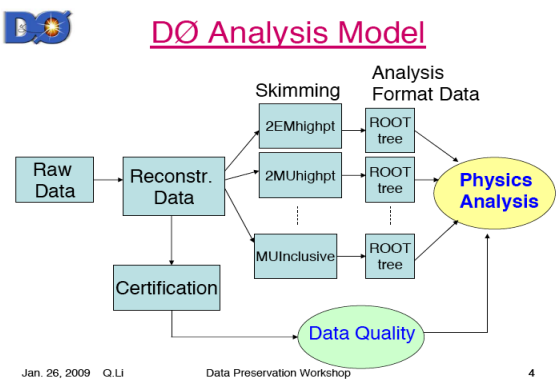
SPIRES → INSPIRE alpha

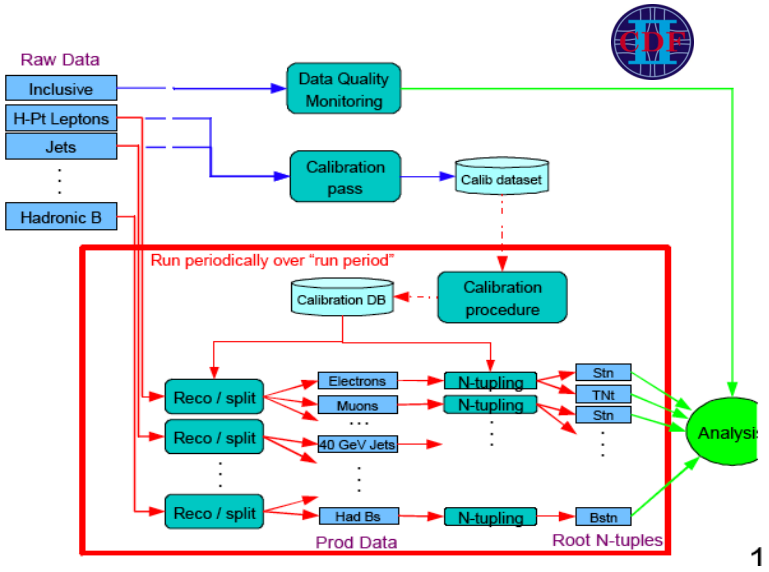


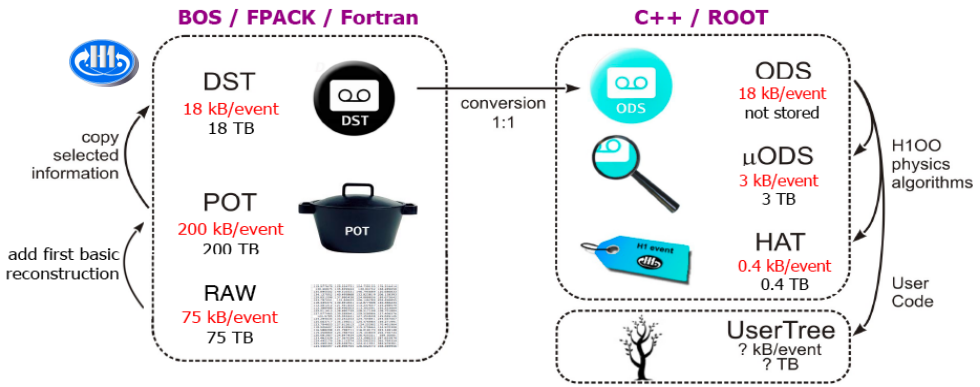


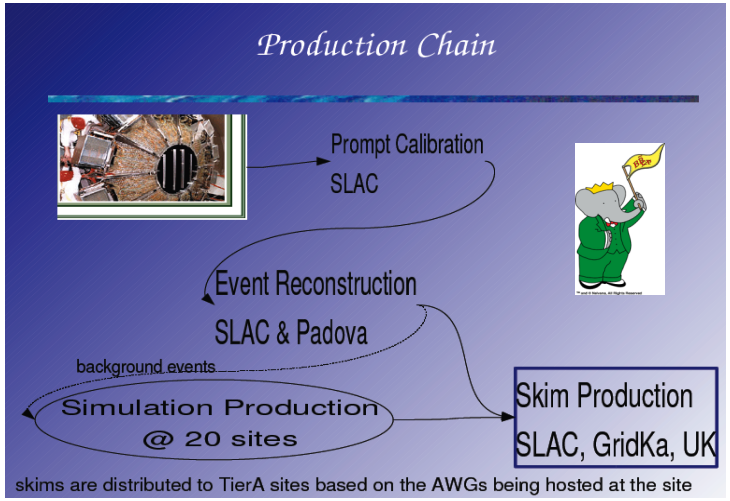
HEP Data Analysis Models











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

Models of Data Preservation

symmetry
dimensions of particle physics
A joint Fermilab/SLAC publication
VOLUME 06 ISSUE 06 DECEMBER 09

Models of Data Preservation

Preservation Model	**Use case**
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Cost, complexity, benefits

JADE
Babar
H1
ZEUS (3/4)

Each level implies an R&D project at experiment level

A word from archivists

Libraries do have an ongoing reflection on documents preservation

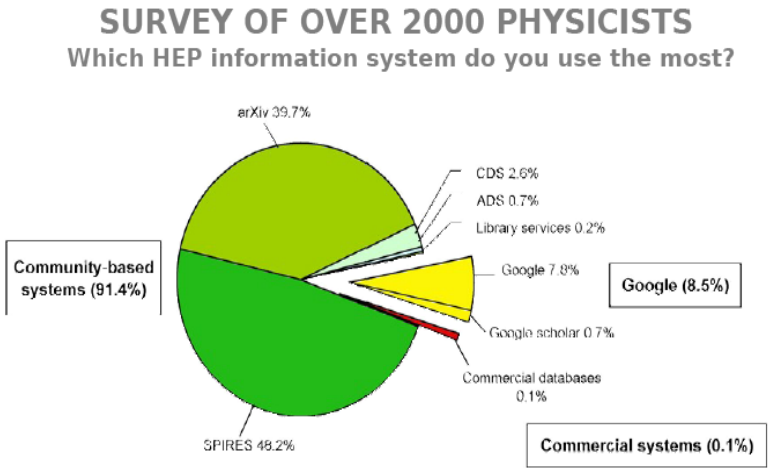
DOE Research and Development Records Schedule

Scientific Data:

- Raw data (all levels)
  - 10 year retention (N1-434-07-01, item 4c(12)
- Evaluated or Summarized data
  - Level 1: permanent retention (N1-434-96-9, item1B13a)
  - Level 2: 25-year retention (N1-434-96-9, item1B13b
  - Level 3: 10-year retention (N1-434-96-9, item1B13c)

Deken -- 2nd Workshop on Data Preservatio 42

Jean Marie Deken,
M.A., M.L.I.S., C.A.
Archivist and Head
SLAC Archives and History Office



SPIRES → INSPIRE alpha

Published papers
+meta-data
+experimental data?

Technological issues

- Computing centers are (in principle) able to store the data
  - 0.5 to 10 Pb /exp.
  - Total cost of data storage double current costs: 1 + 1/2 + 1/4 + 1/8 .. = 2
- Technological evolution and data migration
  - Software maintenance is the real issue
  - Preservation, emulation, migration
  - New possibilities: virtualization and cloud computing
- Interface with experiments needs to be defined
  - Procedures, agreements, resources
  - Supervision and custodianship of data sets, archival expertise

…from an email

Dear Dr. Diaconu,

In the tape storage area we still have 4132 tapes of type 3840 containing some HERA data (see attached information).

We do not have a functioning reading device anymore and the storage area was polluted recently, so it is likely that the tapes are damaged.

Would you like us to send you these tapes or should we destroy them directly?

Sincerely Yours,

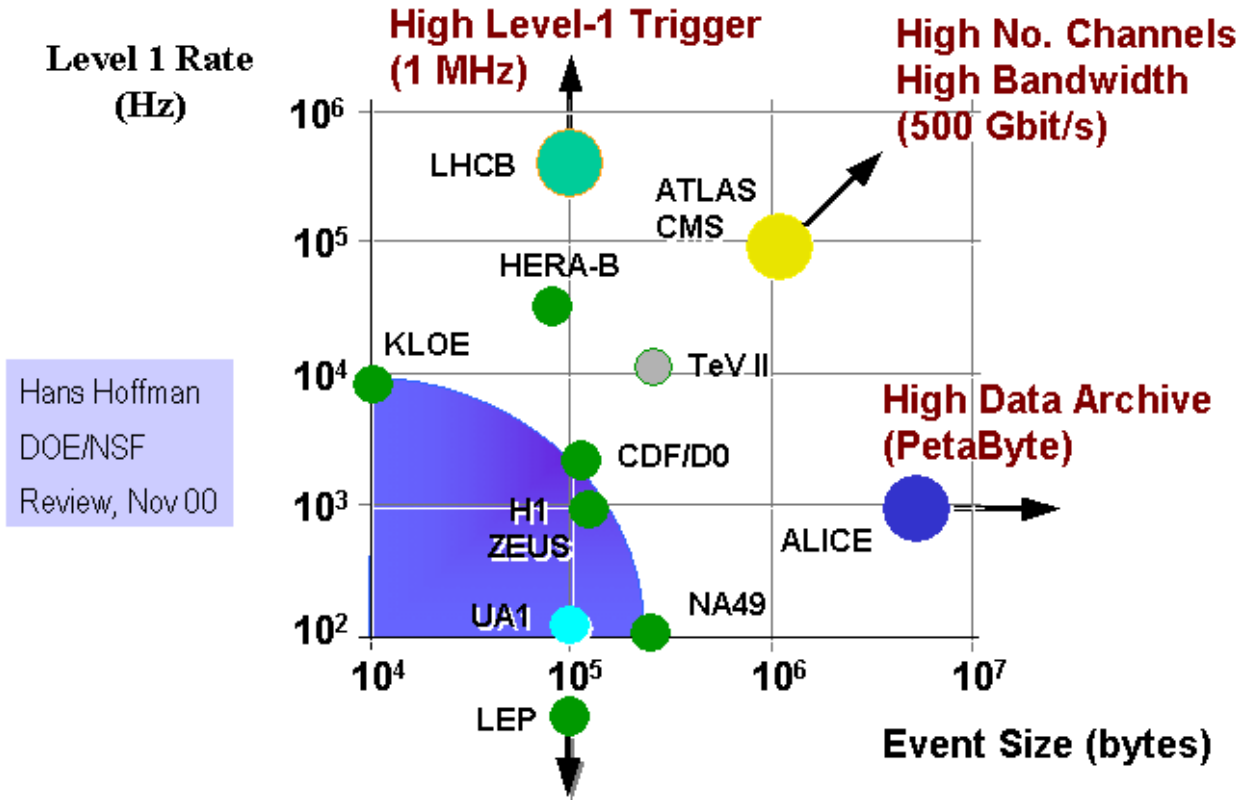
Tape administration service

[A large computing center]

BP3001 GAILLAC HERA1 3480
BP3002 GAILLAC HERA1 3480
BP3003 GAILLAC HERA1 3480
BP3004 GAILLAC HERA1 3480
BP3005 GAILLAC HERA1 3480
BP3006 GAILLAC HERA1 3480
BP3007 GAILLAC HERA1 3480
BP3008 GAILLAC HERA1 3480
BP3009 GAILLAC HERA1 3480
.....

A new job in large HEP computing centers: data archivist
+ adequate supervision (see later)

Data Samples: not a problem for h/w capacity



Storage technology should be comfortable by the end of the experiment
Migration should be carefully planned

Generic 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

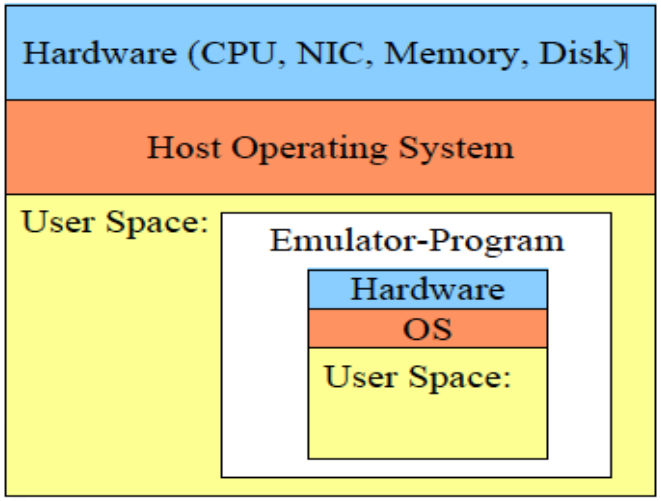
    Prepare it once (?), migrate the "middleware"
  - Continuous migration

    Follow technology changes (adjust, redesign, recompile etc….)

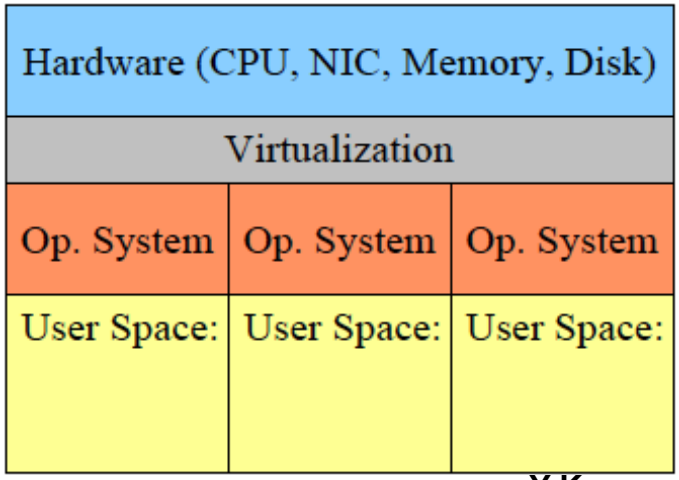
Emulation and virtualisation

One of the “killers”: the technology step (OS)

Emulation



Virtualisation



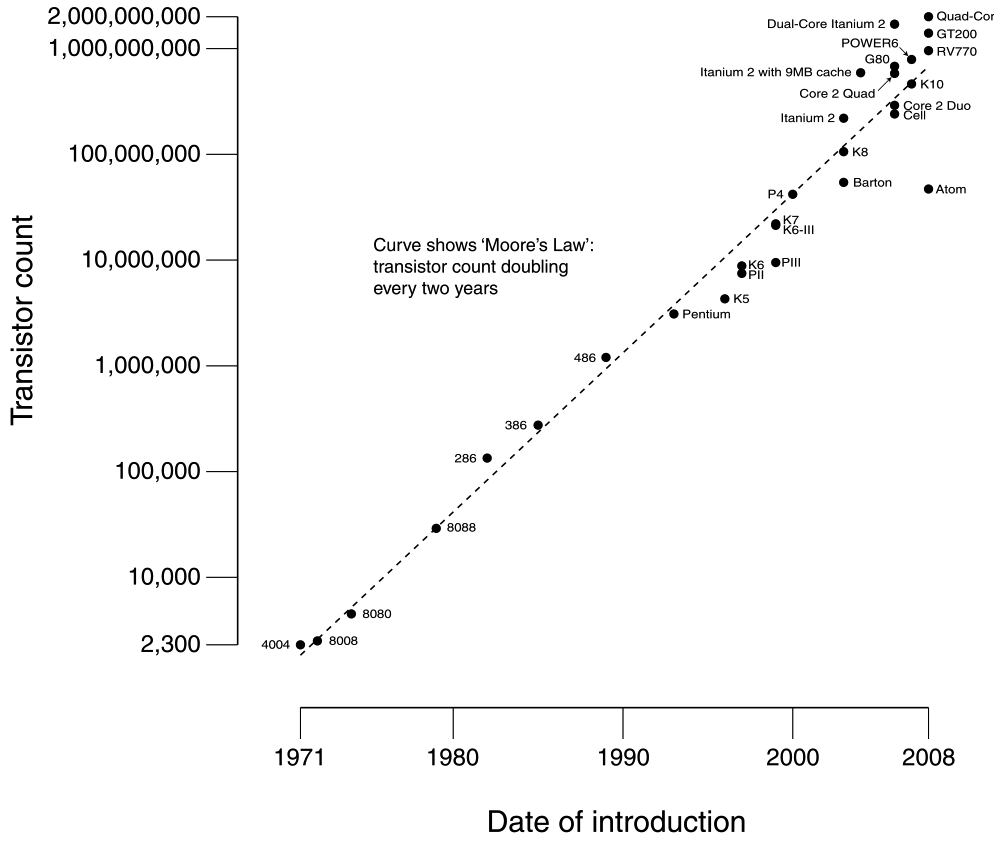
Y.Kemp

An different operating system can be “preserved”?
Can a HEP computing environment also be preserved this way?
Relation with “cloud computing”?
What is the lifetime of the industrial virtualisation techniques?

Computing power

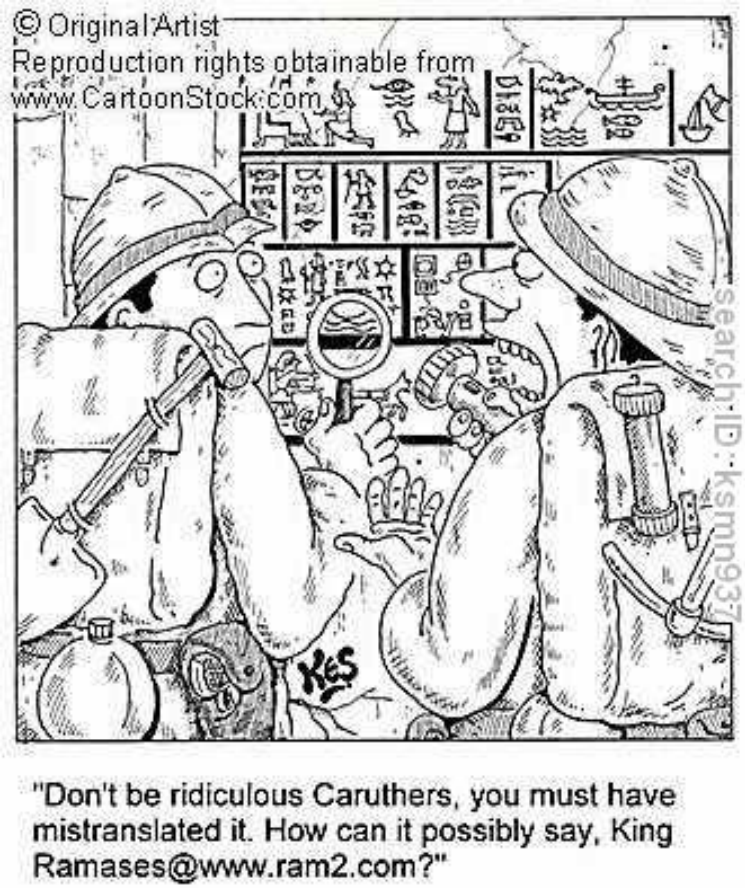
wikipedia

CPU Transistor Counts 1971-2008 & Moore's Law

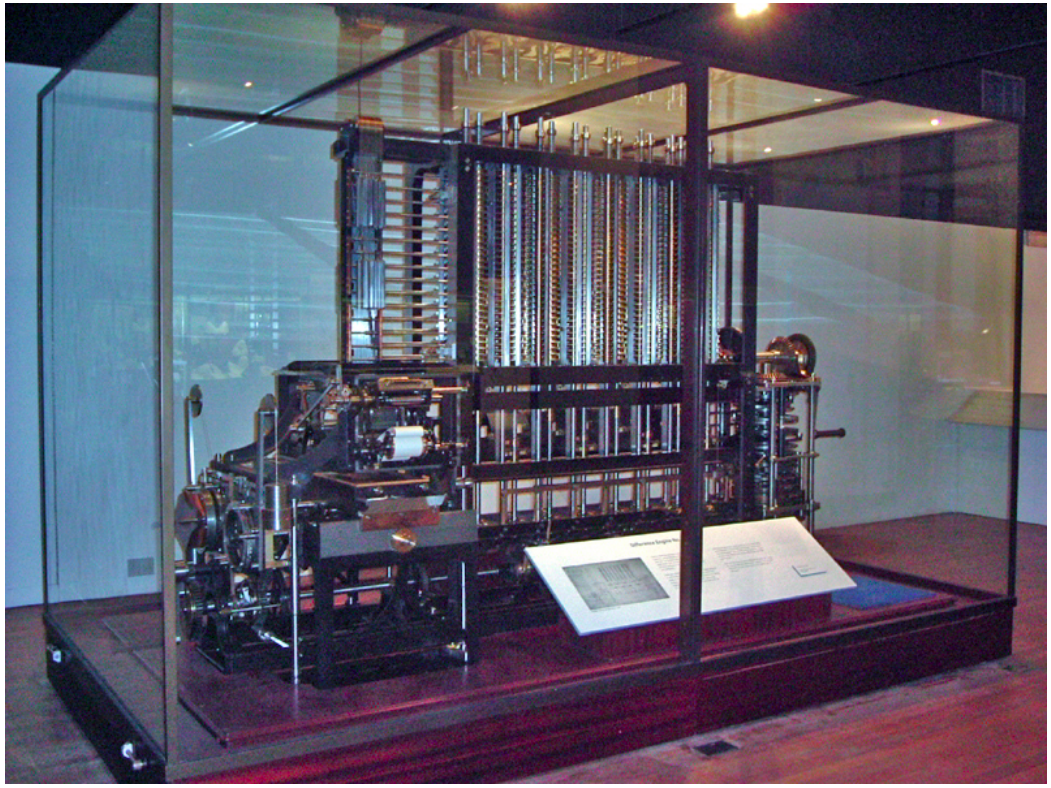


The archival system should be prepared to absorb the technological evolutions

Software/knowledge persistency



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

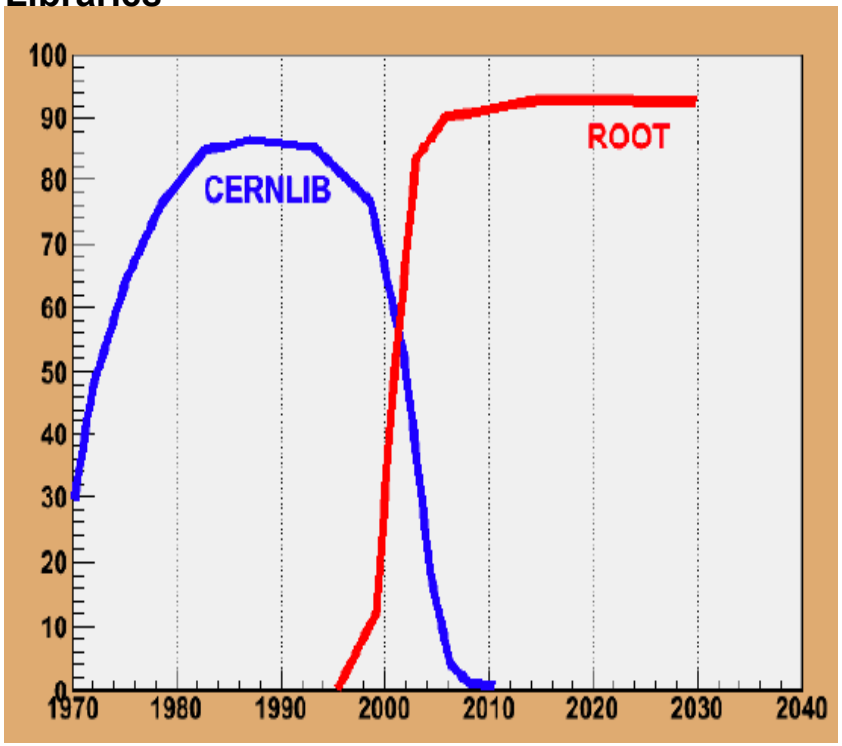


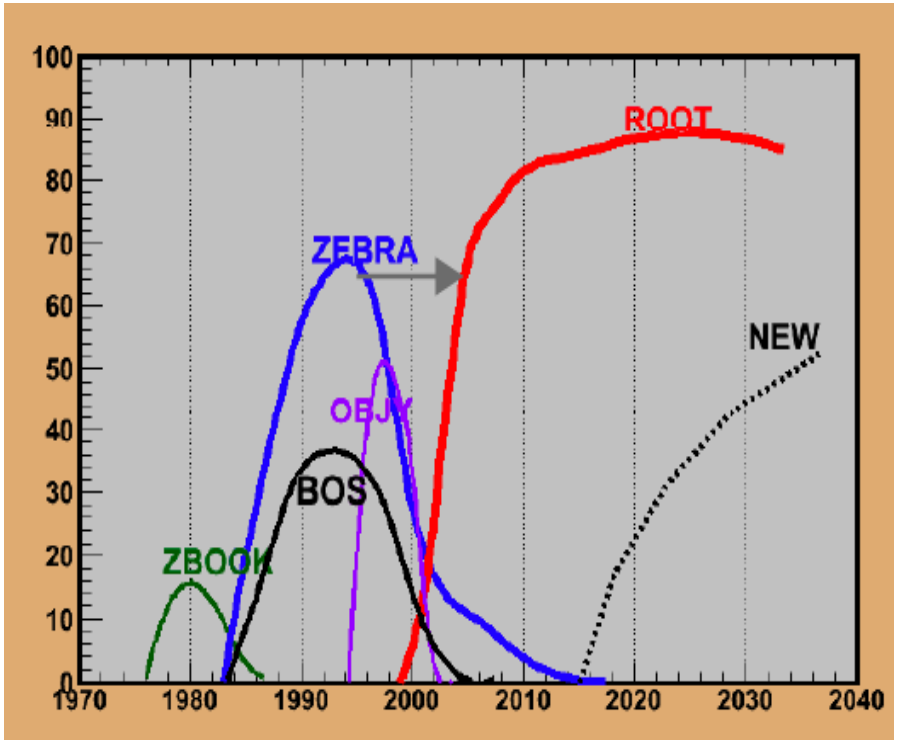
"Errors using inadequate data are much less than those using no data at all."
Charles Babbage

Analysis software

Software is a source of concern: maintenance, migration, validation

Libraries **I/O** **R. Brun**



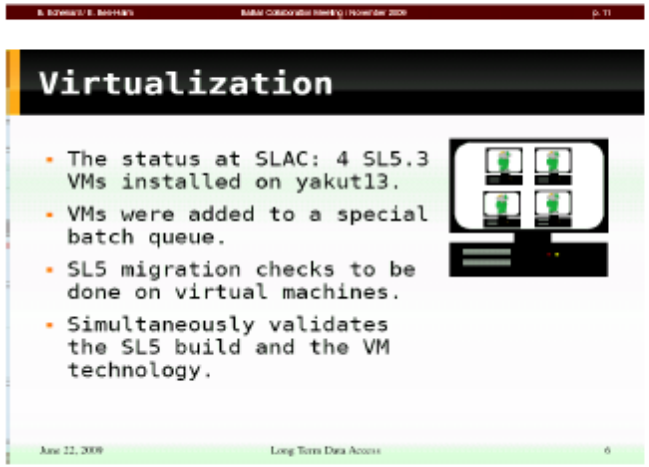


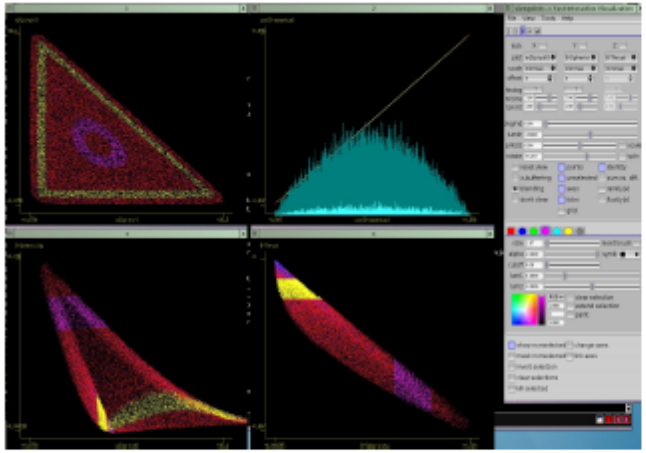
Root offer the needed coherence in the next few decades
Many other dangers: commercial, "ghosts" etc.

An example: Babar Archival project

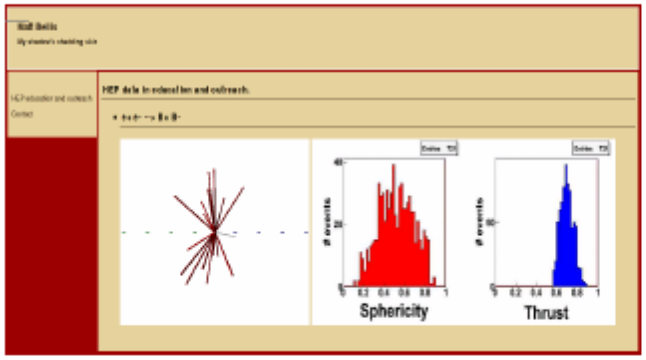
Important remark:
resources taken into account in the funding model of the analysis phase

System being commissioned, It worked very well
In production for 2012





Outreach tools/data already being used in classrooms



Also major advancements in the use of cloud computing

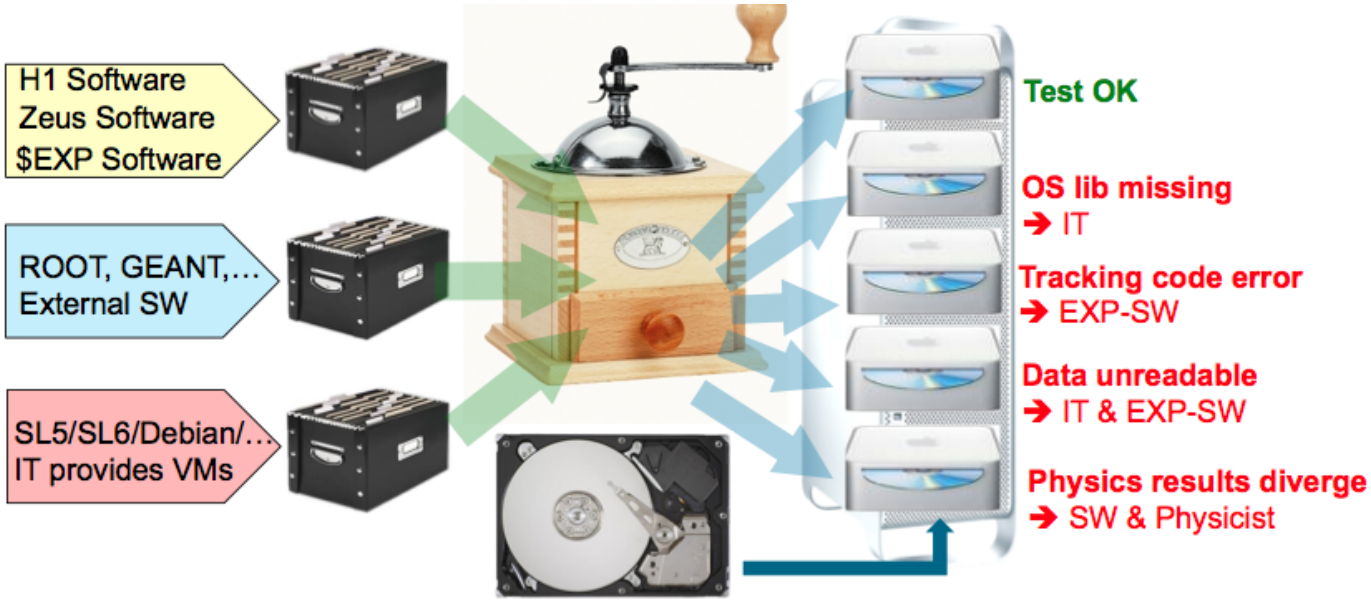
Similar activities at DESY/HERA

Towards generic solutions

A crucial aspect: ability to detect incoherence in absence of an intensive human survey

Work on standards and validation

DESY-IT project



Clear separation between providers of input.

Automated VM image generator provided centrally.

Tests defined by $EXP.
Test data store provided by IT.

Different VMs run SW and tests. Depending on results, different action needed.

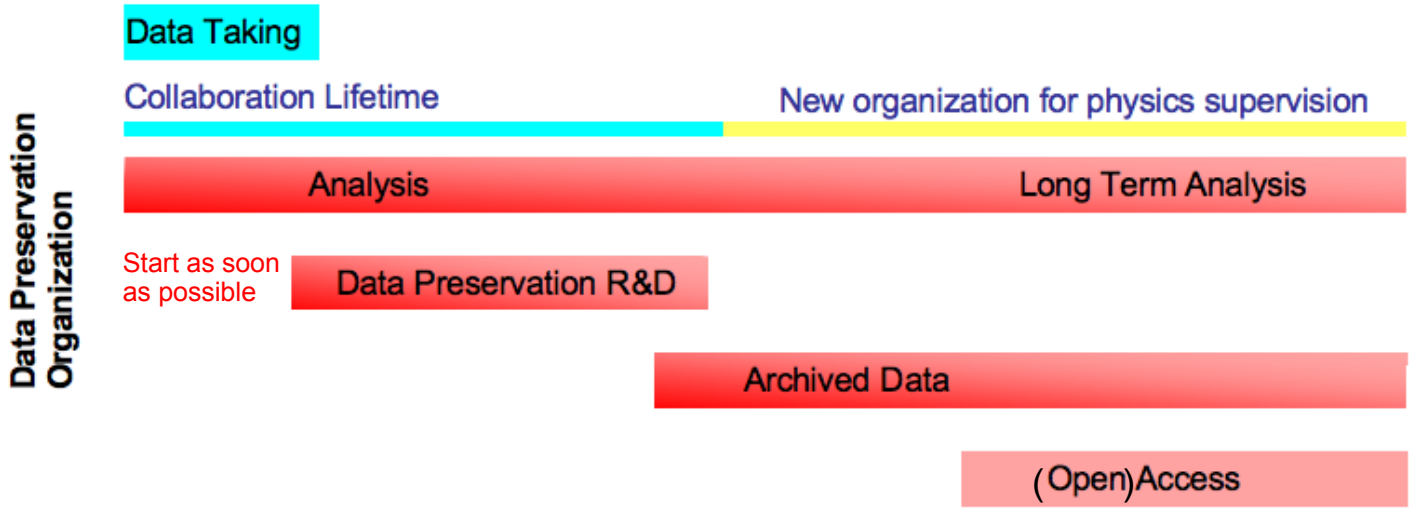
Name | PRC closed session | 29.4.2010 | Page 1

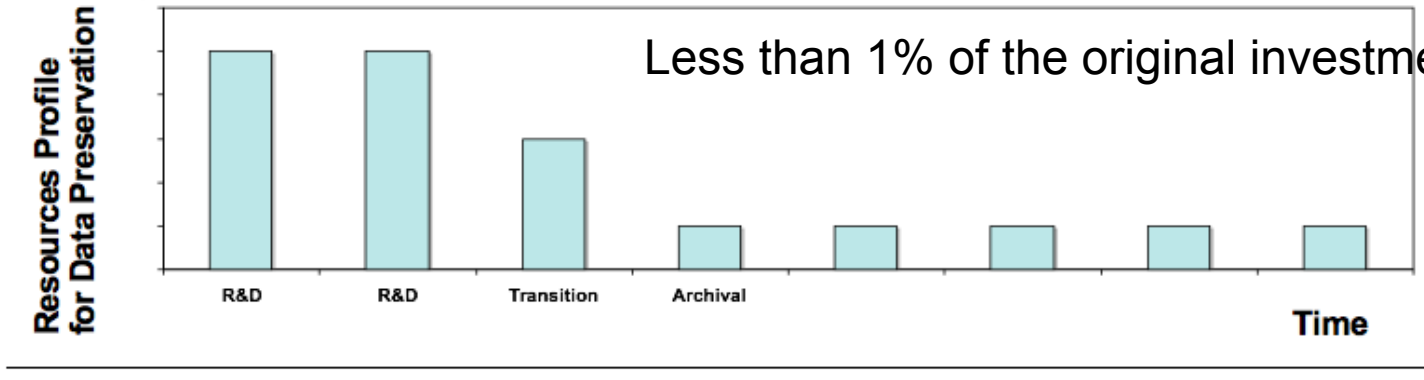
Governance

- Preserved data sets management
  - Scientific supervision of the preserved data sets
  - Authorship and Access to data
  - Channels to outreach and education
  - Endorsement: experiment, laboratory and funding agencies
  - HEP global solutions: common policy and standards

Transition scenario and resources

(experiment level)





Less than 1% of the original investment

Define a timescale?

Estimative costs/experiment: 3FTEs for the “surge” 0.5 FTE archival mode

Towards an International Organization

Estimative costs: 3 FTEs /3-5 years to make the structure sustainable

DPHEP 2009: Intermediate Recommendations

- ICFA document: A broad reflection on benefits and strategies, a few recommendations
  - Prioritization against other general issues in HEP (new experiments, funding, resources) is **not** addressed at this stage

1. An **urgent and vigorous action** is needed to ensure data preservation in HEP.
2. The preservation of the **full analysis capability** of experiments is recommended, including the preservation of reconstruction and simulation software.
3. an interface to the experiment know-how should be introduced: **data archivist position** in the computing centers.
4. The preservation of HEP data requires a **synergic action** of all stakeholders: experimental collaborations, laboratories and funding agencies.
5. An **International** Data Preservation **Forum** is proposed as a reference organisation.The Forum should represent experimental collaborations, laboratories and computing centres.

Feedback from HEP community

- Support from major labs expressed:
  - DESY, CERN, Fermilab, SLAC, IHEP,JLAB
- ICFA August 2009
  - Support data preservation in high energy physics
  - Endorse the International Study Group as an **ICFA subgroup**
  - Nominate a Chair of the subgroup (C.Diaconu 2009/2010)
- HEPAP (DOE and NSF) October 2009
  - "Data preservation would allow for reanalysis using new theory or experimental techniques and detailed combined analyses with new data. It could also be very useful for education and outreach activities.
  - An international organization could provide the necessary guidance and governance. "
- FALC january 2010
  - Positively received, in particular the educational aspects

The DPHEP perspective

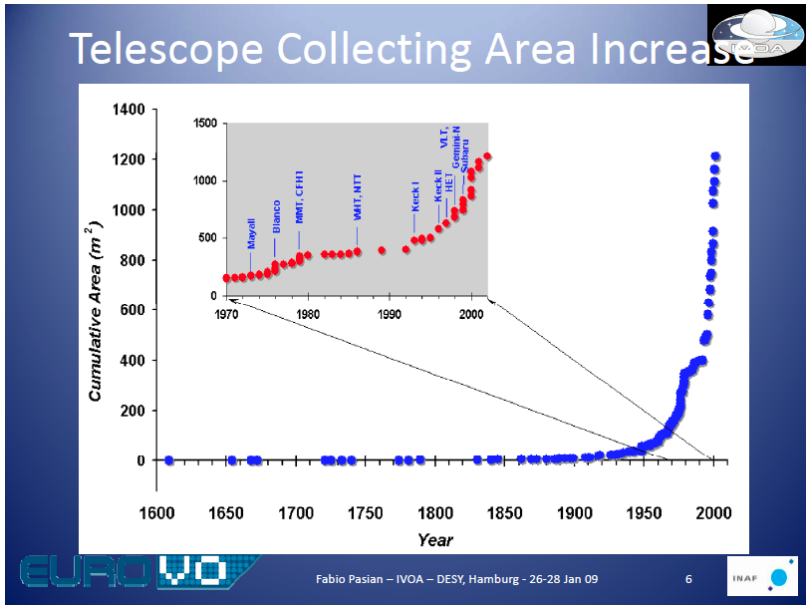
- DPHEP started in January 2009
- Intermediate report released in November 2009
  - Principles and problem setting
  - General recommendations (models, R&D, data archivist, International Organization, etc.)
- 2010: Produce a Blue Print
  - Workshop @ KEK, July 8-10, 2010
  - Documented research case, detailed experiments projects
  - Transverse activities: outreach and education, technology R&D
  - International Organization
  - Resources and funding schemes:

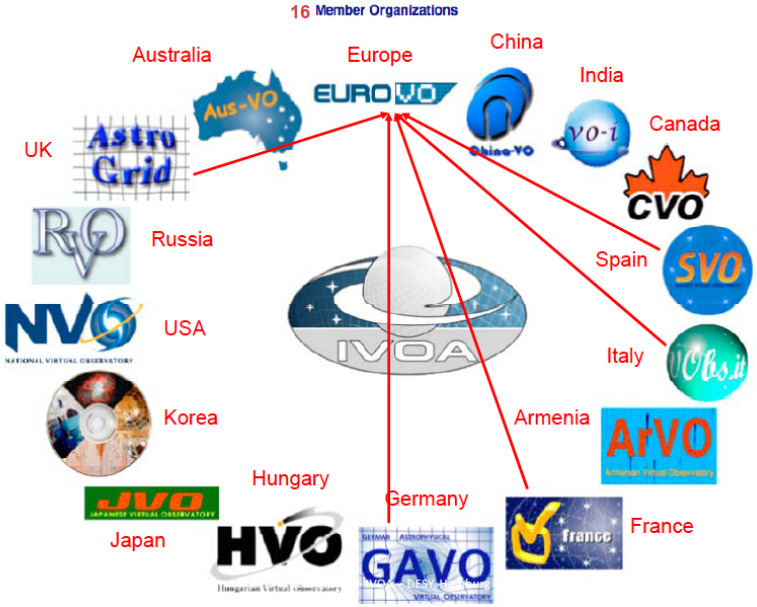
    Funding Agencies, Laboratories, International Programs

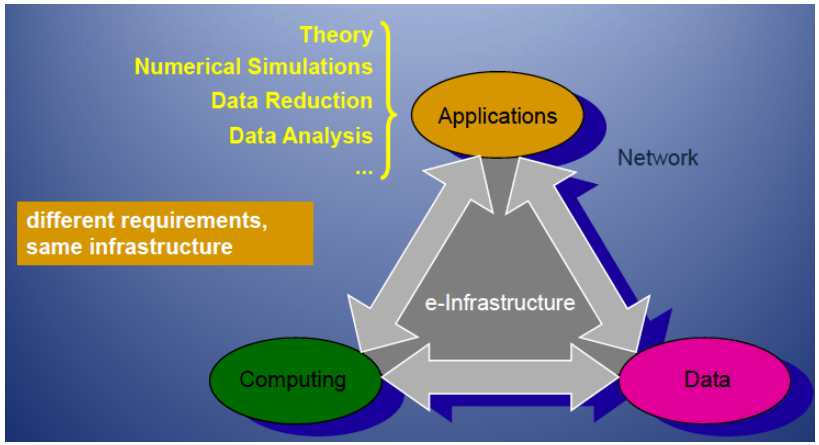
Conclusion and outlook

- Data preservation in HEP is important because:
  - It is based on a relevant physics case
  - It is timely, given the experimental situation and plans
  - Enhance the return on investment in the experimental facilities
  - It is most likely cost-effective, provides research at low cost
- Requires a strategy and well-identified resources
- International cooperation is the best way to proceed
  - **Unique** opportunity to build a coherent structure for the **future**

Virtual Observatories in Astrophysics







F.Pasian

- Data Archives Inter-operable
- Work on standards and access to
  - Data, simulation, mining techniques
- International, multi-experiment

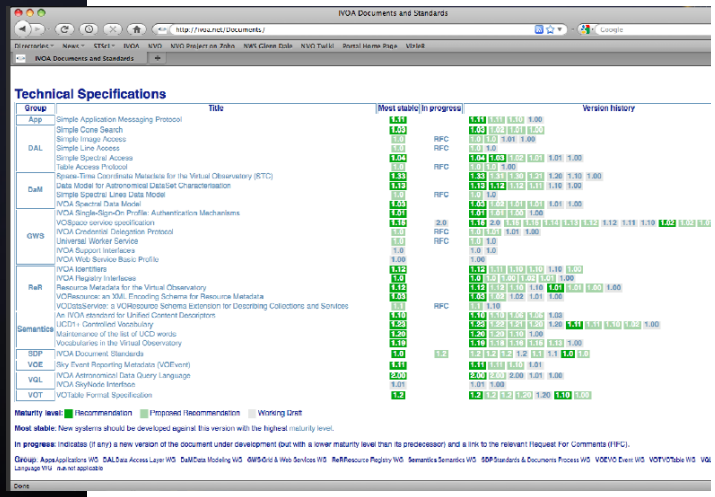
International Virtual Observatory for Astrophysics

Dr. Robert J. Hanisch
Director, US Virtual Astronomical Observatory
Space Telescope Science Institute
Baltimore, MD

- ~50 major data centers and observatories with substantial on-line data holdings
- ~10,000 data "resources" (catalogs, surveys, archives)
- data centers host from a few to ~100 TB each, currently ~1 PB total
- current growth rate ~0.5 PB/yr, expected to increase soon
- current request rate ~1 PB/yr
- for Hubble Space Telescope, data retrievals are 3X data ingest; papers based on archival data constitute 2/3 of refereed publications

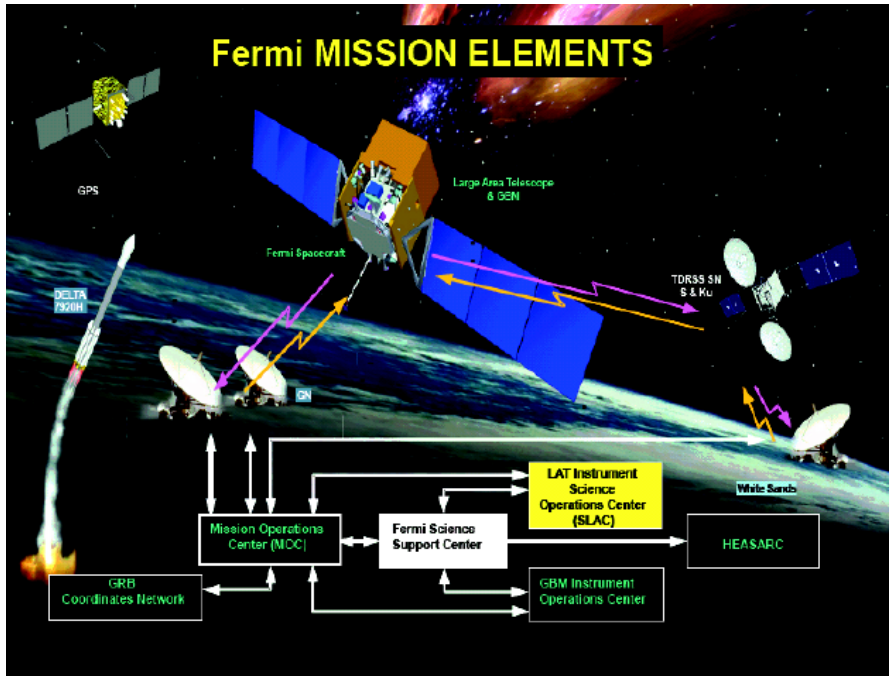
VO's and Standards

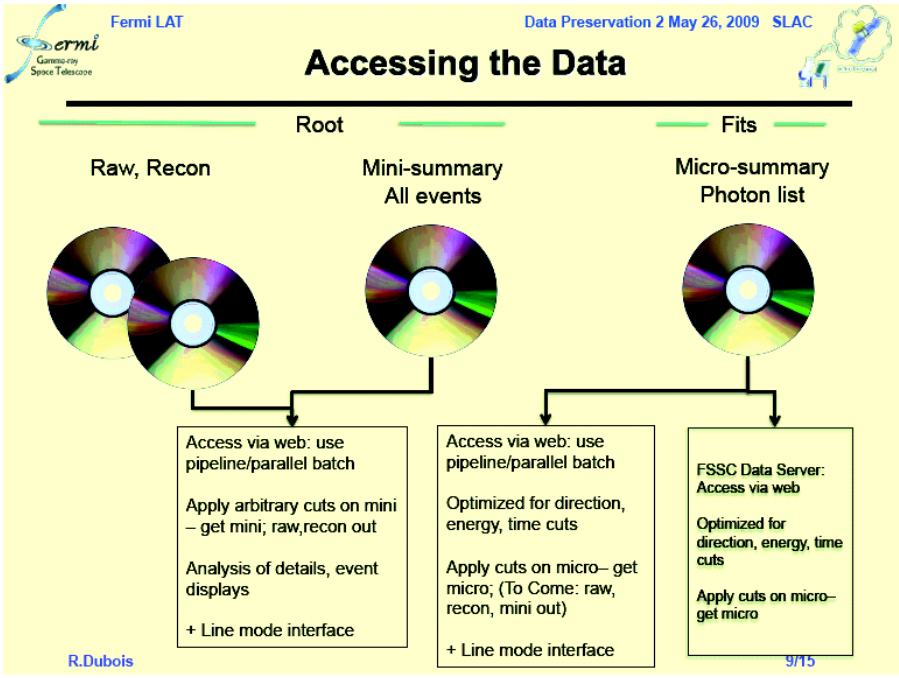
- IVOA began in June 2002 — http://ivoa.net
  - Self-organizing
  - No funds of its own, no dues; relies 100% on project participation
  - Rotating chair (18-month term)
- IVOA now has 17 member projects
  - Aggregate funding ~$50M (since inception)
  - Projects range from 2-3 people to ~20 FTE
- Forum for discussion and sharing of experience
- Twice per year "Interoperability" workshops bring together ~100 participants
- Adopted a standards process based on W3C
  - Note
  - Working Draft → Proposed Recommendation → Recommendation
  - IAU endorsement
  - See http://ivoa.net/Documents/



- Data management activities at major astronomy facilities are typically 3-5% of annual operating budget, including h/w, s/w, and staff. Staff accounts for ~85% of total.
- VO development and operations are ~20% additional to baseline data management costs (international aggregate)

Open access in astrophysics





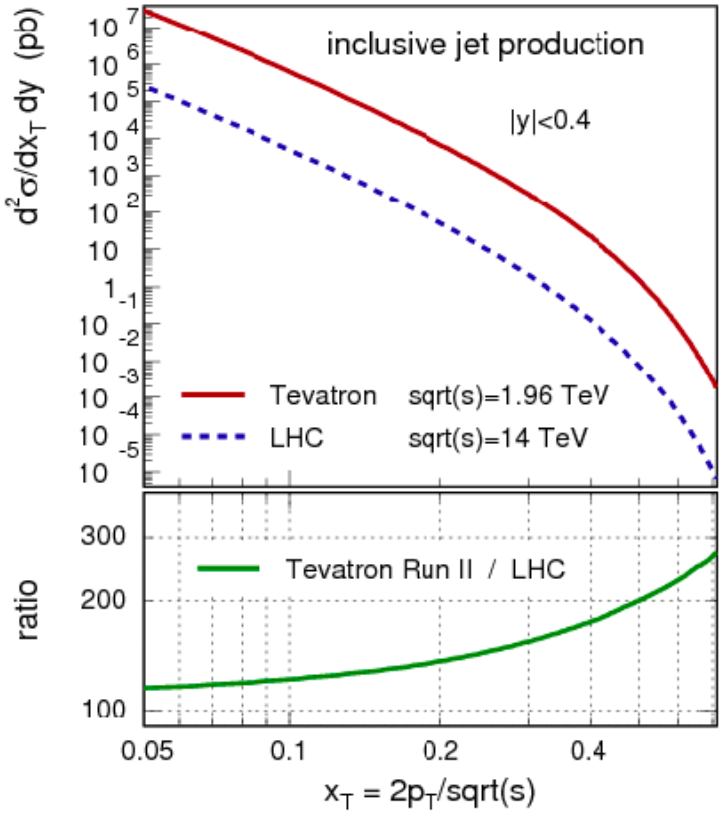
LAT Principal Investigator Peter Michelson added: **"The LAT team has made significant discoveries and significant progress in many areas. I expect that the collaboration will continue to come out with the most results, but I also expect others to make discoveries. Releasing this data is good for the project, good for the collaboration, and good for science."**

—Kelen Tuttle

SLAC Today, August 25, 2009

Another example: high x constraints from Tevatron

Inclusive Jets: Tevatron vs. LHC



PDF sensitivity:

→ Compare Jet Cross Section at fixed xT = 2pT / sqrt(s)

Tevatron (ppbar)

>100x higher cross section @ all xT

>200x higher cross section @ xT>0.5

LHC (pp)

- need more than 1600fb-1 luminosity to compete with Tevatron@8fb-1
- more high-x gluon contributions
- but more steeply falling cross sect. at highest pT (=larger uncertainties)

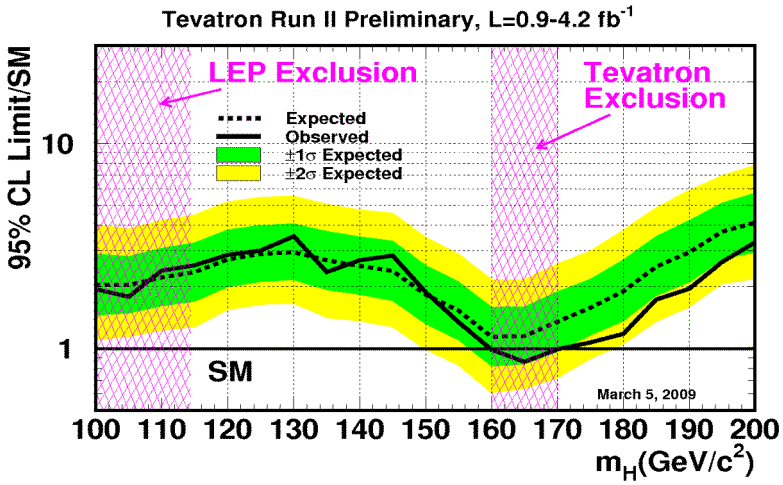
→ Tevatron results will dominate high-x gluon for some time ...

21

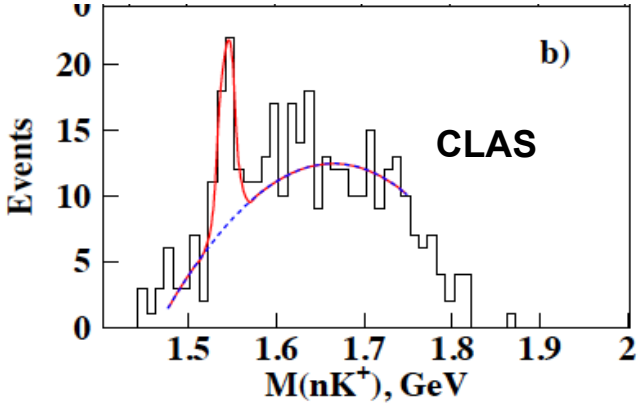
M. Wobisch

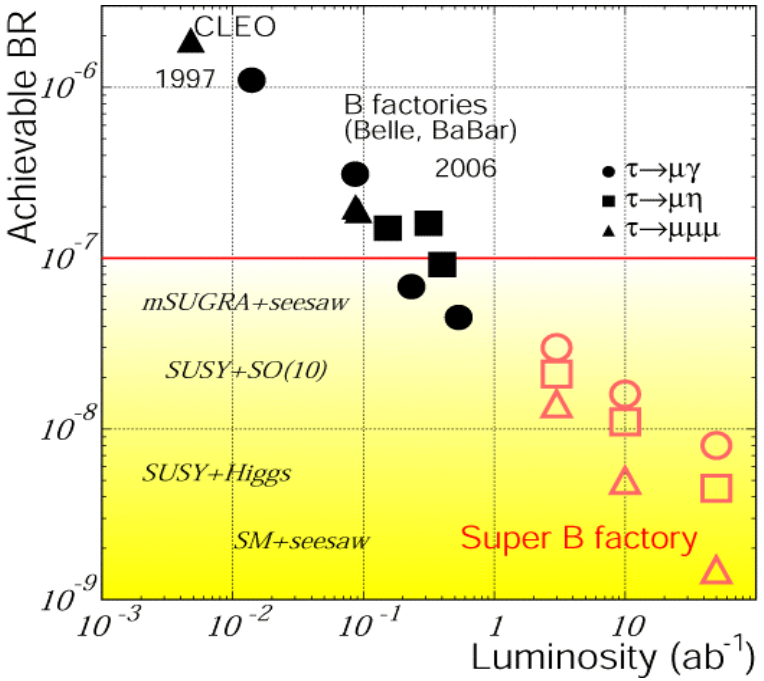
More examples: contingency with future programs

- Tevatron/LHC
- B- and SuperB-factories
- Low energy



...surprises can occur
at lower energies too

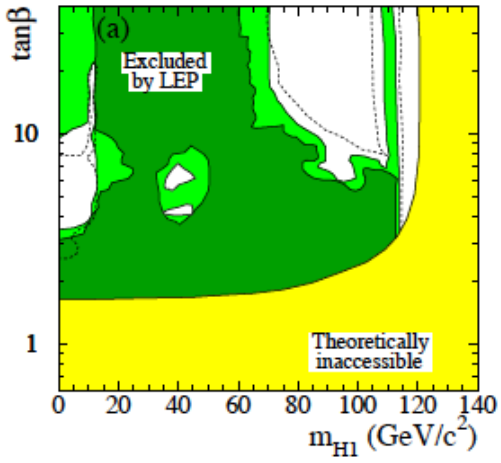




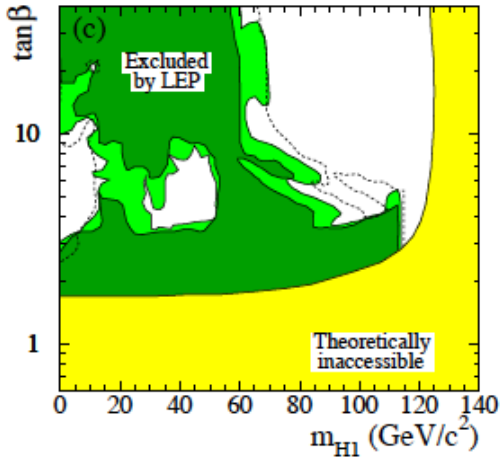
Excluded?

Analysis optimization?

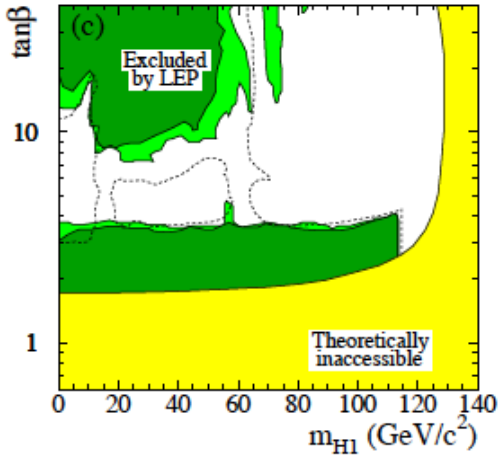
$m_t = 169.3\ \mathrm{GeV}/c^2$



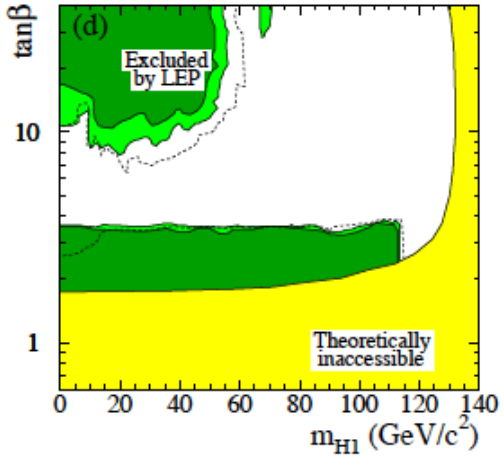
$m_t = 174.3\ \mathrm{GeV}/c^2$



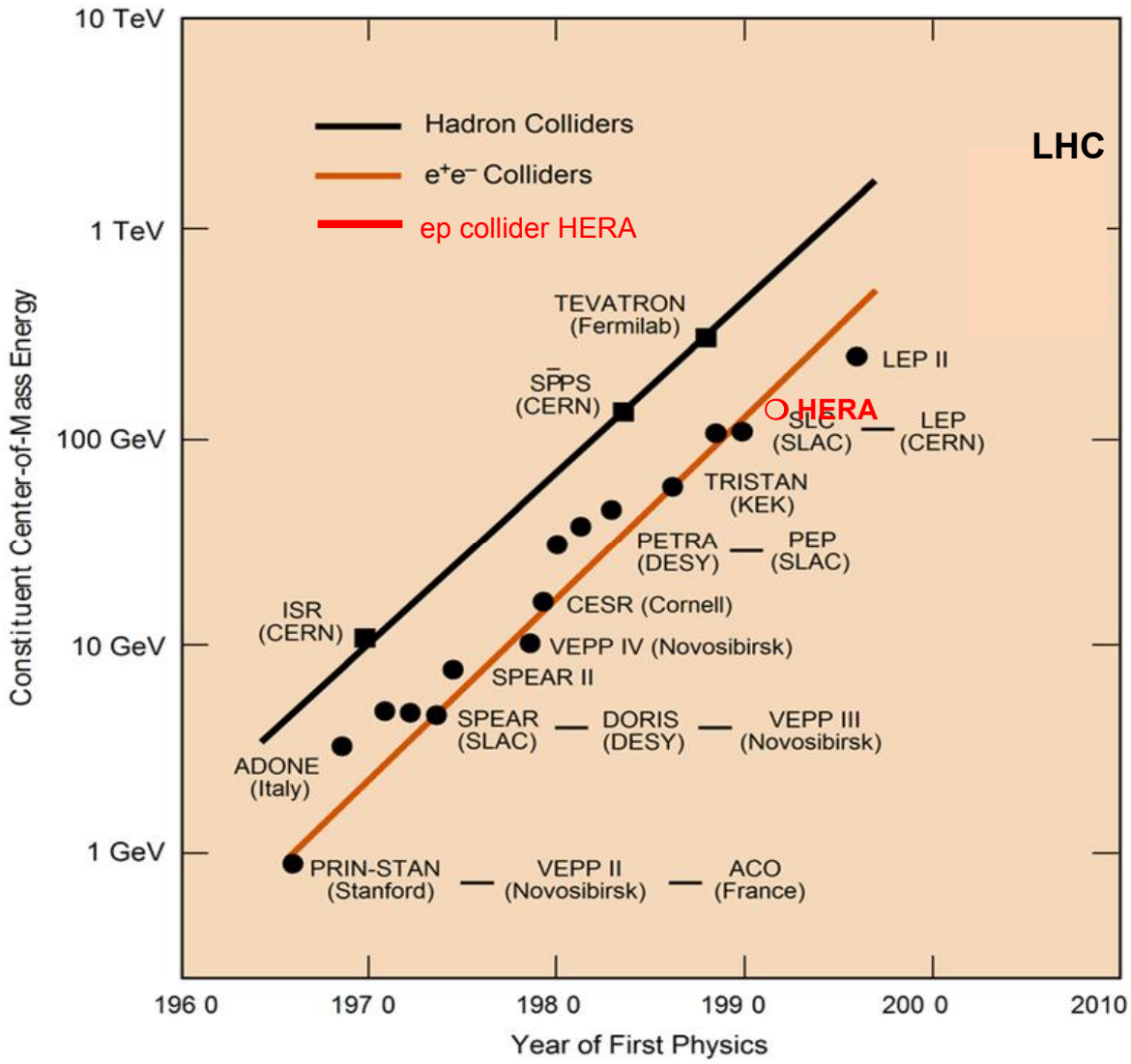
$m_t = 179.3\ \mathrm{GeV}/c^2$



$m_t = 183.0\ \mathrm{GeV}/c^2$



High Energy Physics

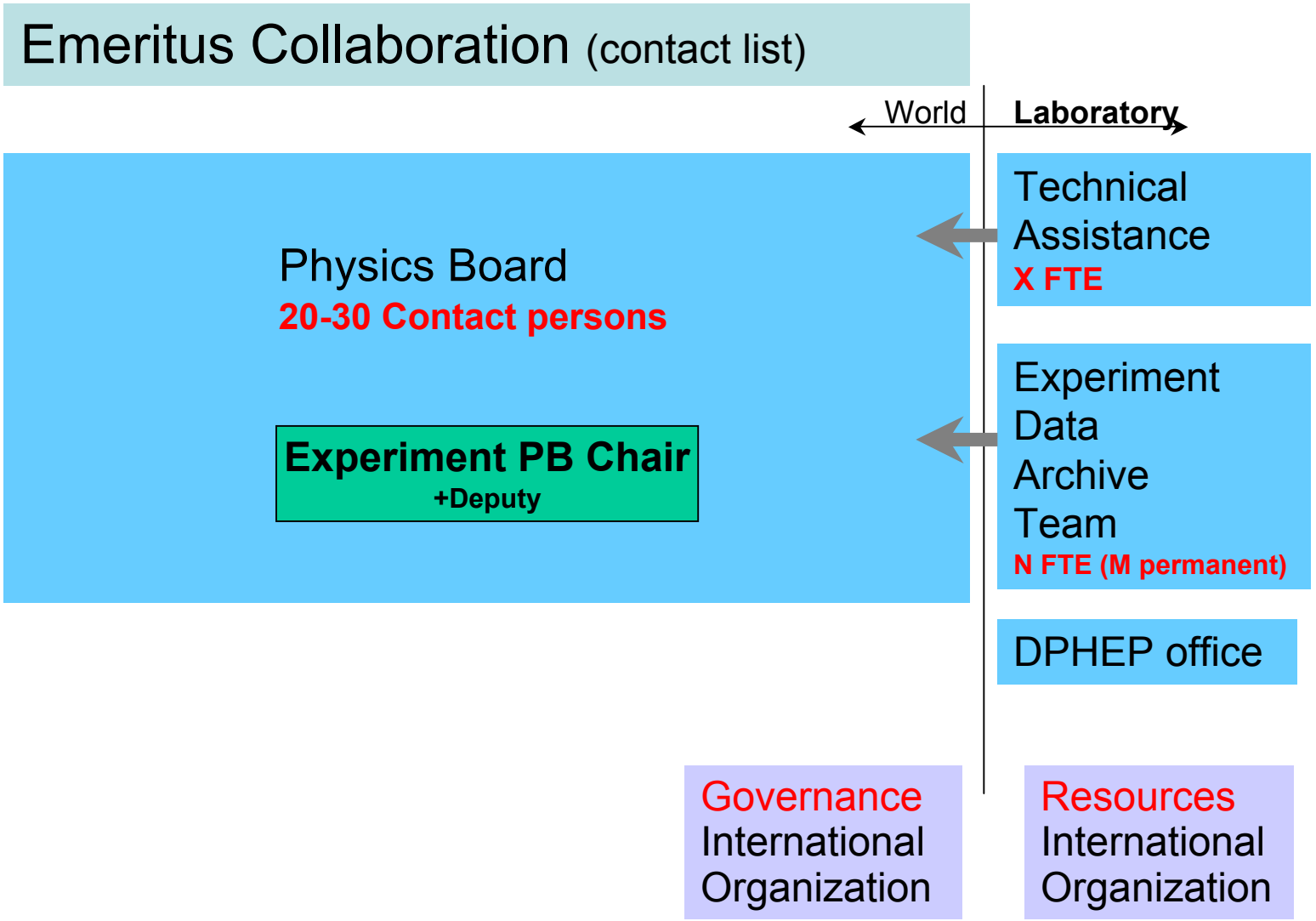


Energy frontier probed with complex experimental Installation

New experiments use to superseed the previous/similar experiments

What is the situation at present?

A long term organization of HEP experiments



The 2010 HEP landscape (colliders)

- LEP 2000

"[…] LEP is scheduled to be dismantled soon so that its 27 km tunnel can become the home for the ambitious LHC proton collider, which is due to come into operation in 2005. "
[CERN Courier, Dec. 1st, 2000]

  - No follow-up decided (ILC?) - after 2020
- HERA: end of collisions in 2007
  - No follow-up decided (LHeC?) - after 2020
- B-factories: Babar 2008, Belle->Belle II
  - Next generation in a few years (2013-2017)
- Tevatron: 2011
  - A majority of the physics program will be taken over at the **LHC**
  - However: p-pbar is unique, no follow-up foreseen

HEP experiments data taking encompass 15-20 years, some are unique

What is the fate of the collected data?

(NB: here "data" = full experimental information)