technische universität
dortmund

Data Preservation and Long Term Analysis in HEP

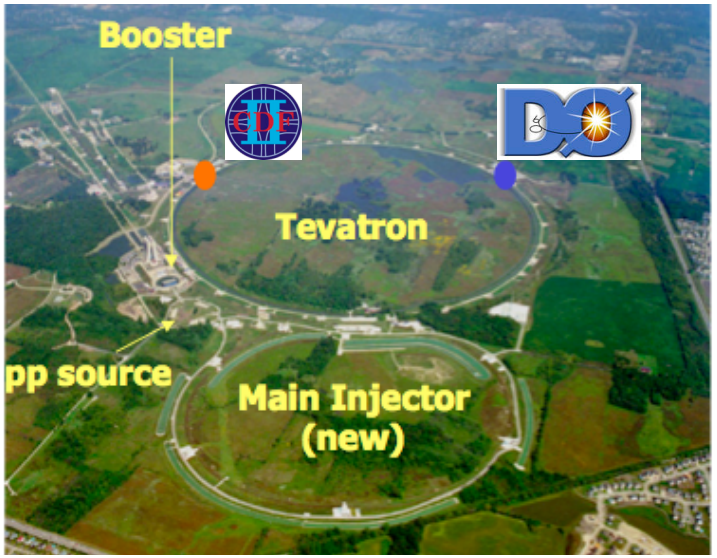
David South (Technische Universität Dortmund)

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

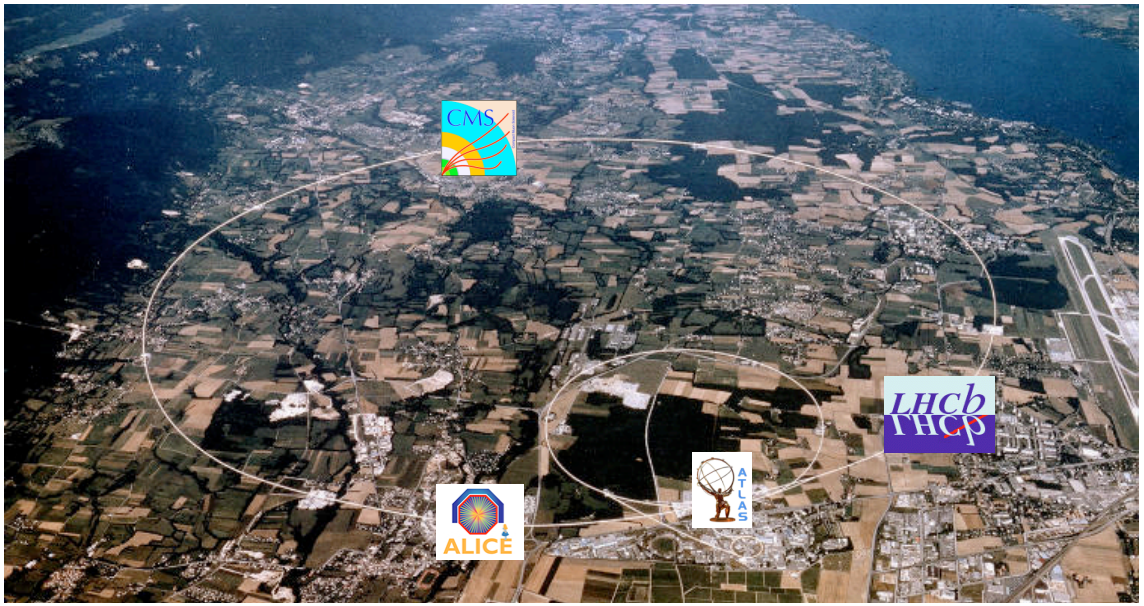
http://www.dphep.org

DESY Zeuthen, Berlin, April 7 2010

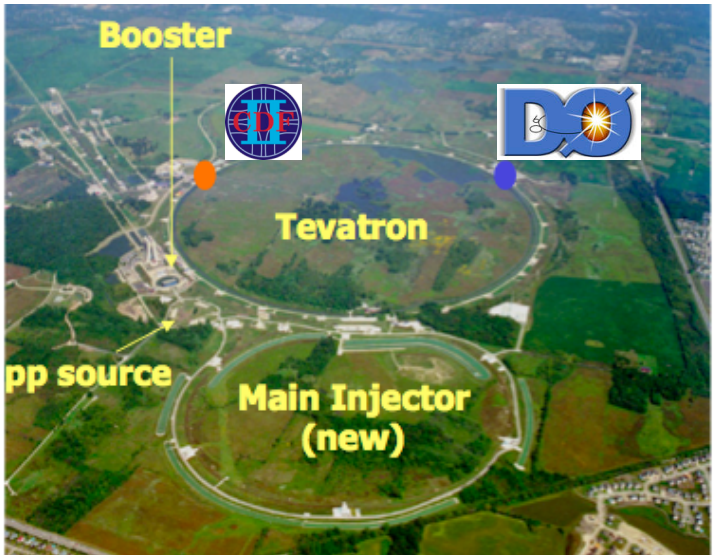
High Energy Physics Data are Unique



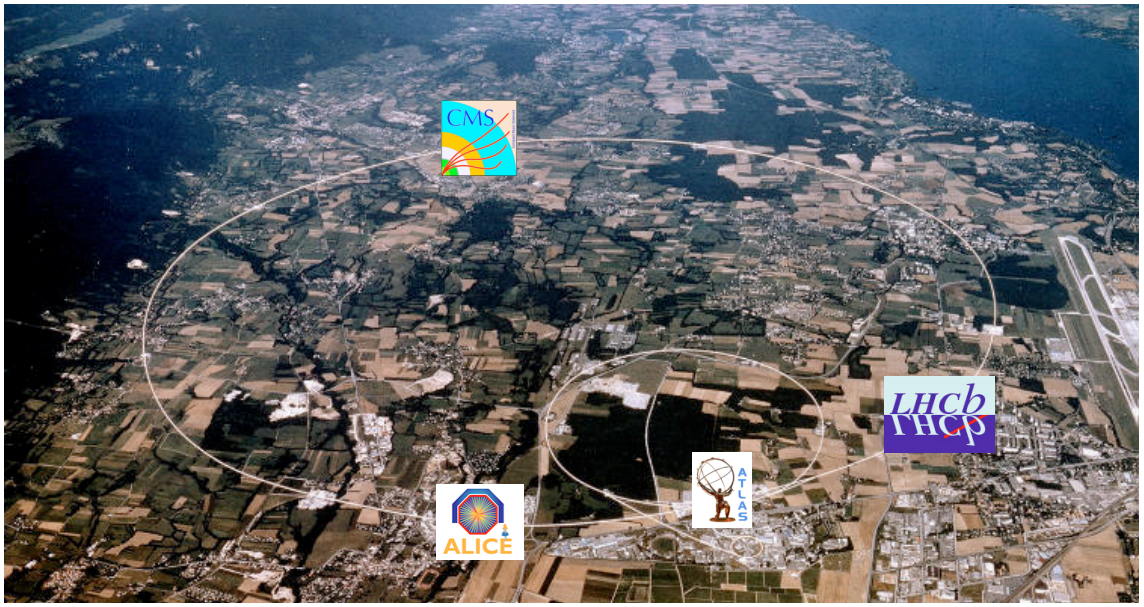
High Energy Physics Data are Unique



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

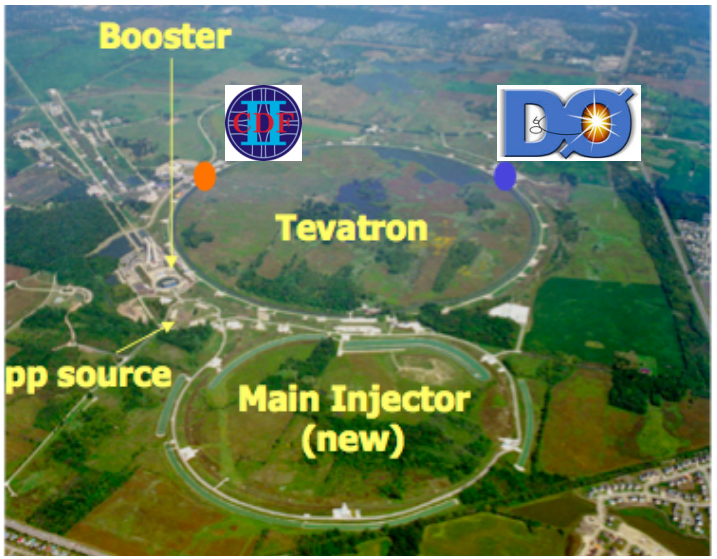


High Energy Physics Data are Unique

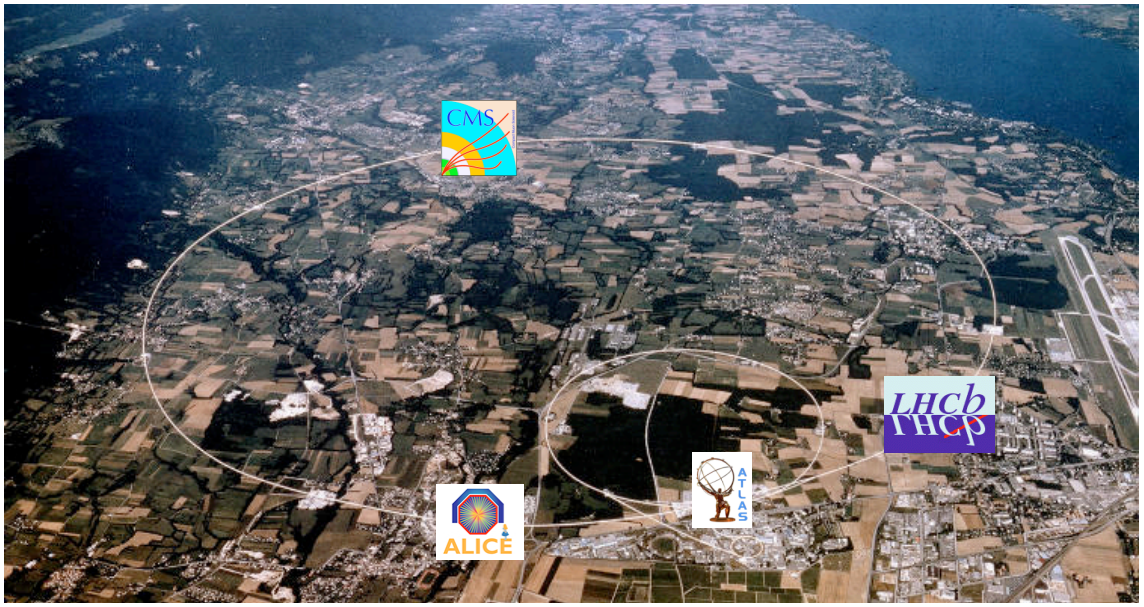


A generation of HEP experiments are concluding their data taking and winding up their physics programmes

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



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

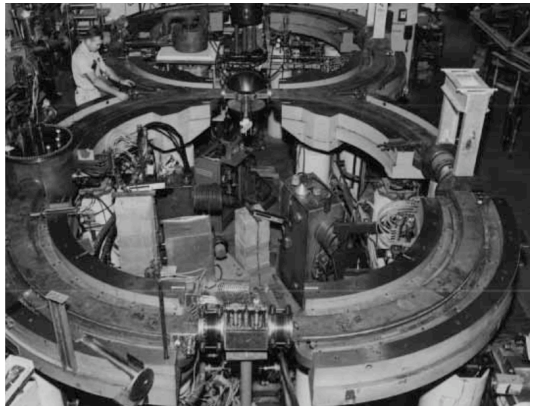
The experimental data from these experiments still has much to tell us, from the existing analyses still to be completed..

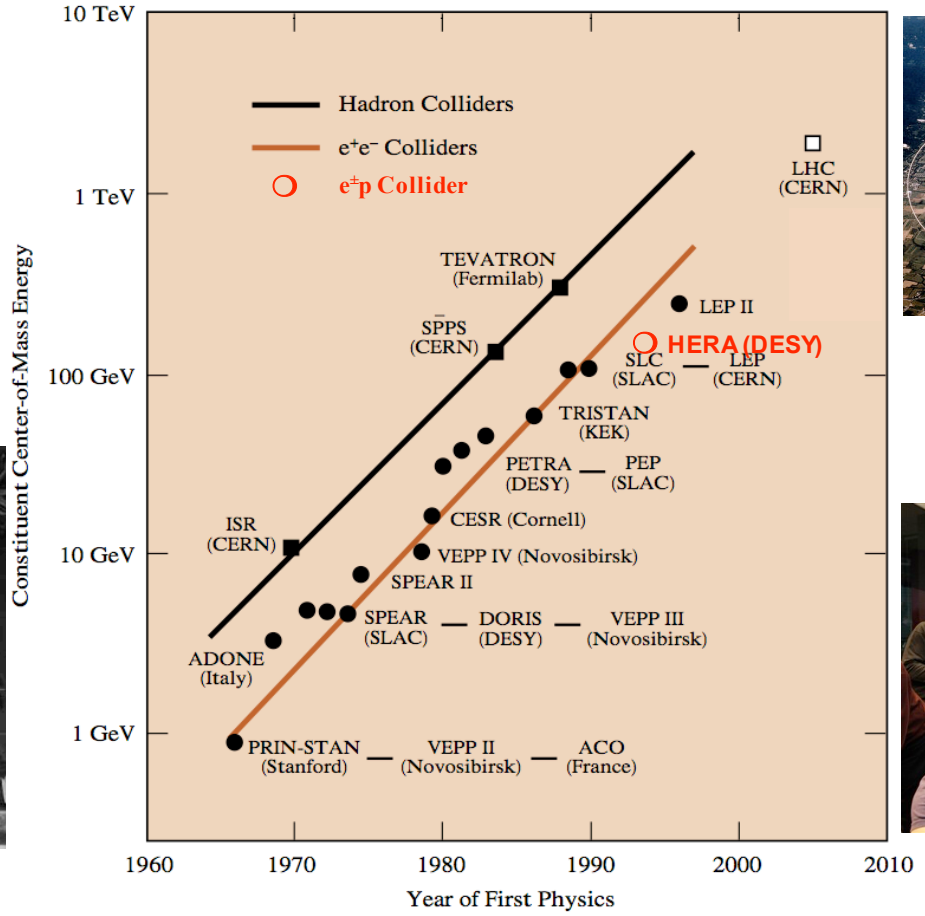
..but they may also contain things we do not yet know, which may only come to light at a later date via LHC data or a new theory

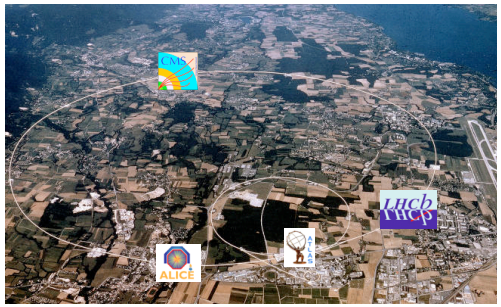
The Last 50 Years of High Energy Physics

PRIN-STAN, built late 1950's

The first colliding-beam machine, a double-ring electron-electron collider, built by a small group of Princeton and Stanford physicists. (Courtesy Stanford University)



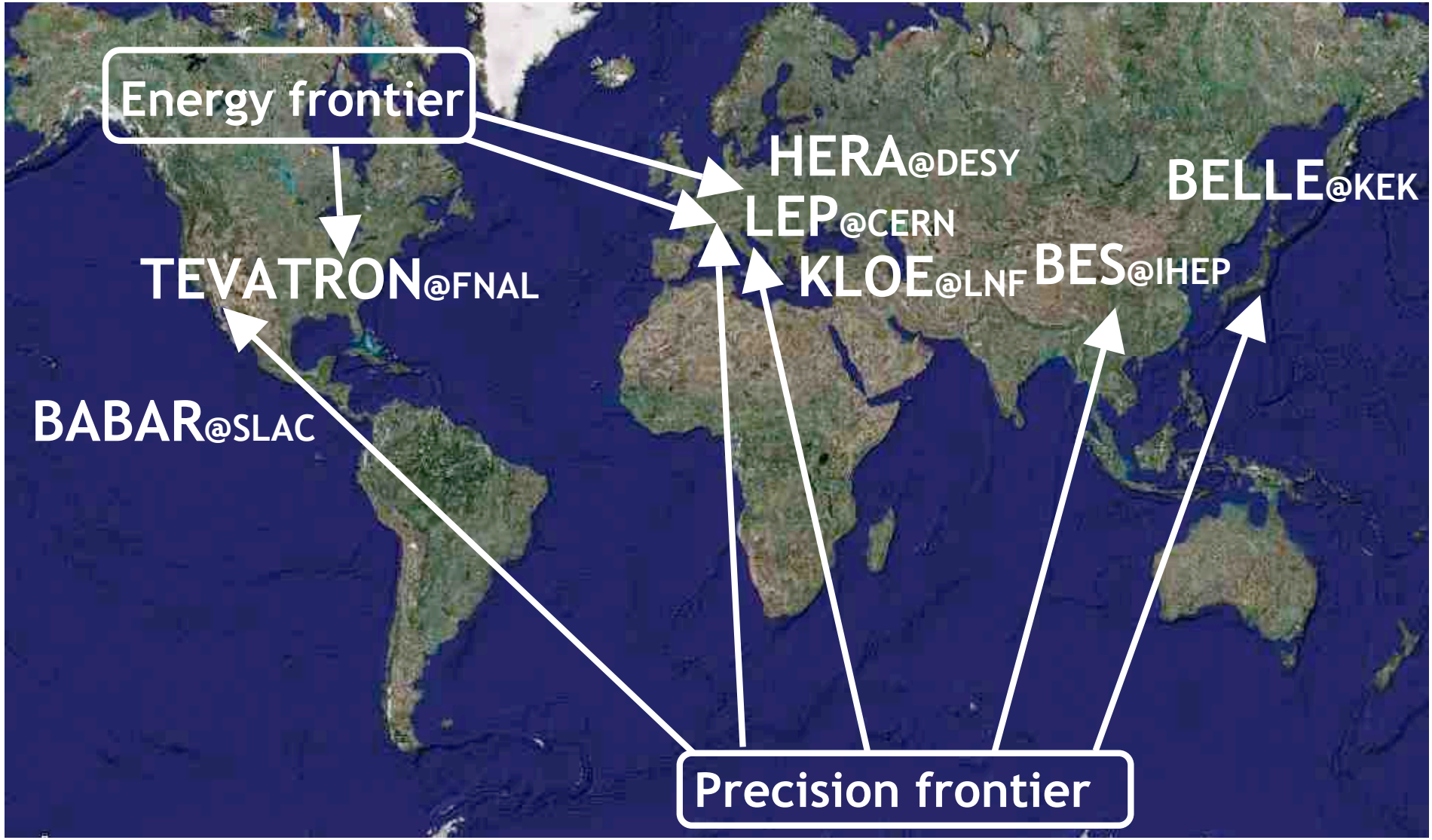




First collisions observed at the LHC in 2008; first data taking at 7 TeV now!

- Energy frontier probed with complex experimental installations
- New experiments normally supercede previous/similar ones
- What is the present situation?

The Pre-LHC Landscape



The 2010 HEP Landscape (Colliders)

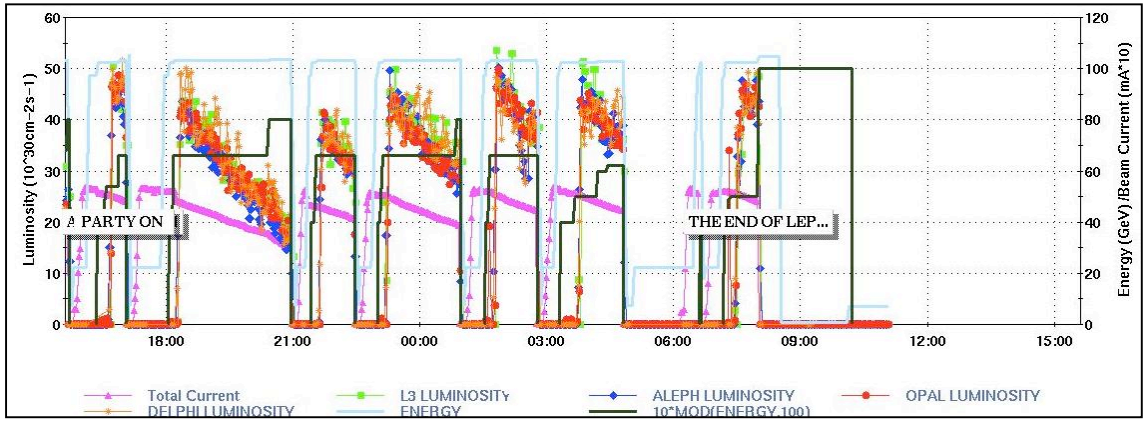
"LEP is scheduled to be dismantled soon so that its 27 km tunnel can become the home for the ambitious LHC proton collider, which is due to come into operation in 2005."
[CERN Courier, Dec. 1st, 2000]

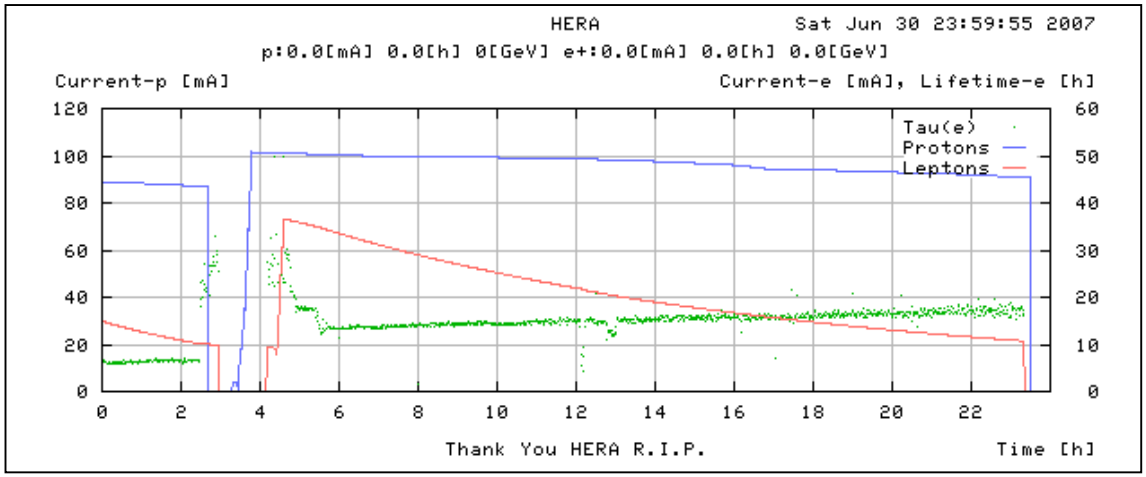
- e^+e^-: LEP ended in 2000
  - No follow-up decided (ILC?) - after 2020
- $e^\pm p$: HERA end of collisions at HERA in 2007
  - No follow-up decided (LHeC?) - after 2020
- B-factories: BaBar ended in 2008, Belle → Belle II
  - Next generation in a few years (2013-2017)
- pp: Tevatron ends soon (in 2011?)
  - The majority of the physics program will be taken over at the LHC
  - However: p-pbar is unique, no follow-up foreseen

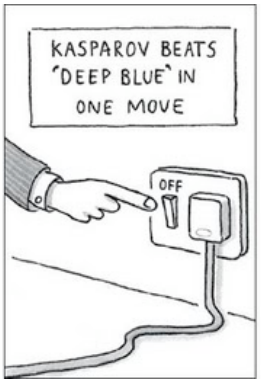
Data taking at HEP experiments takes 15-20 years, and some data are unique

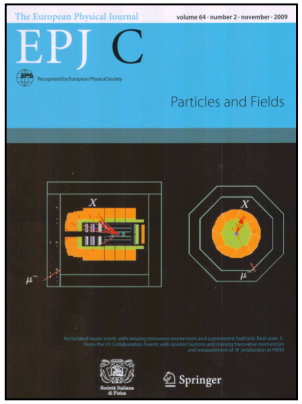
- What is the fate of the collected data? (w here "data" means the full experimental information..)

After the End of Data Taking









- Have a party, dismantle the detector, finalize the analyses, have another party.. *~ 5 years*
- *And then what do you do with the data?*

One Idea …

… and from a recent email

Dear Dr. Diaconu [H1 Spokesperson],

In the tape storage area we still have 4132 tapes of type 3840 containing HERA data.

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?

Yours Sincerely,

Tape administration service

[A large computing centre]

A Common Situation?

"We cannot ensure data is stored in file formats appropriate for long term preservation.

"We cannot ensure those data are still usable - the software for exploiting those data is under the control of the experiments.

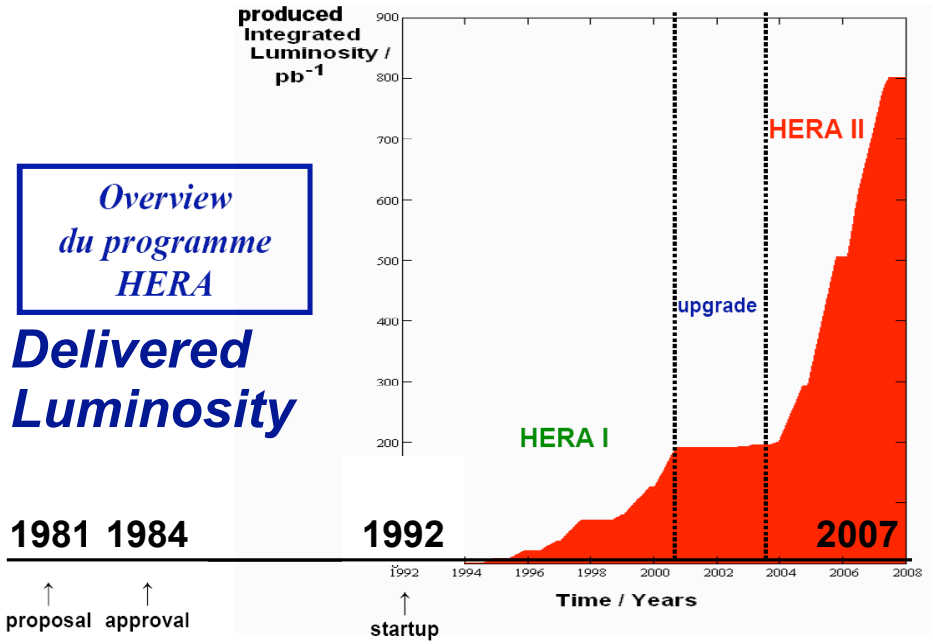
"We are sure most of the data are (not easily) accessible!"

More on this later…

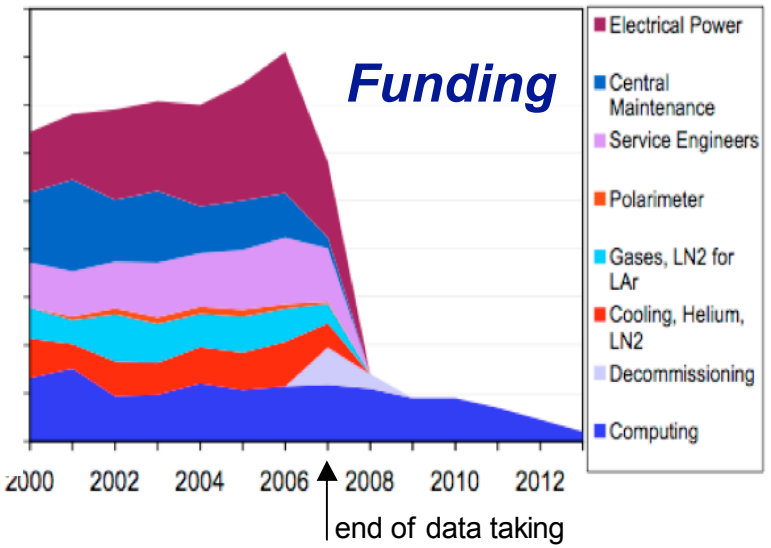
Why should we preserve HEP Data?

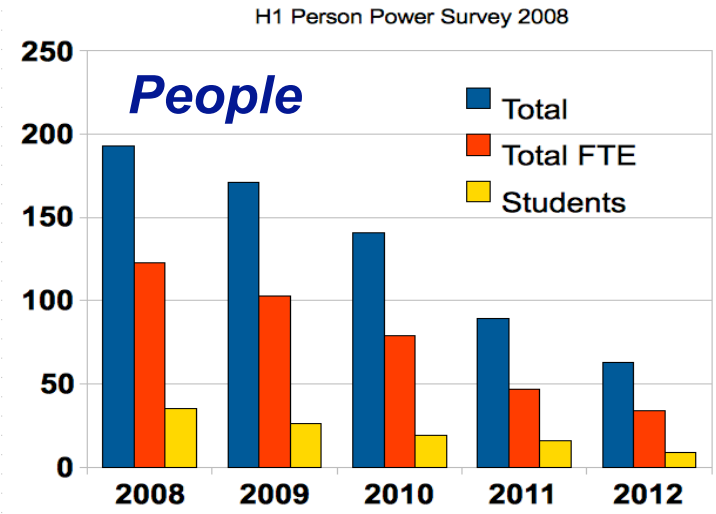
- We may want to re-do previous measurements
  - Increased precision, reduced systematics
  - New and improved theoretical calculations / MC models
  - Newly developed analysis techniques
- We may want to perform new measurements
  - At energies and processes where no other data are available (or will become available in the future)
- Investigate if new phenomena found today
  - Go back and check in the old data

Why is difficult to preserve HEP Data?



- Good data taking period is towards the end of running
- The existing resources (funding and expertise) then decrease when the data taking stops
- Dedicated resources for preservation need to be planned early!





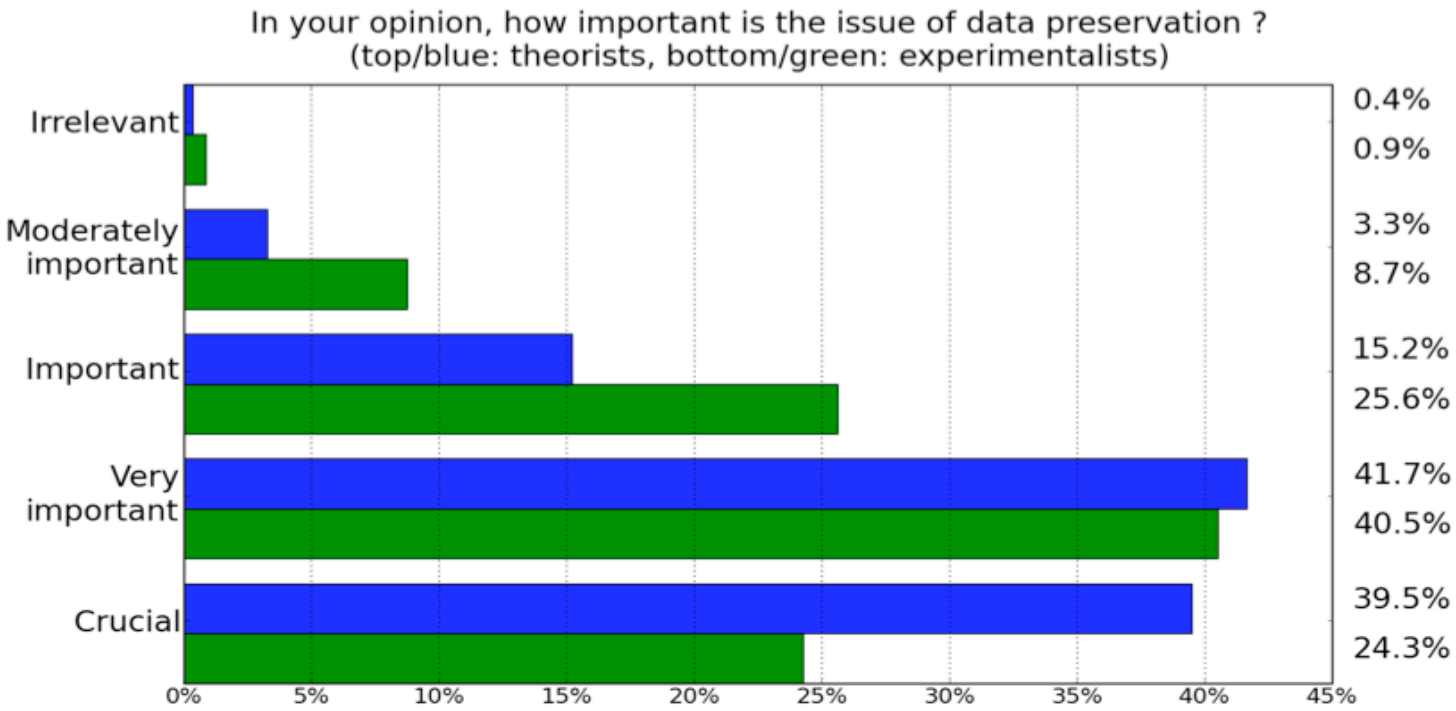
Data Preservation: Support in the HEP Community

PARSE insight
Permanent Access to the Records of Science in Europe

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

arXiv:0906.0485
PARSE.Insight is financed by the European Commission and run at CERN



- 70% of physicists say *very important or crucial!*
- However, no coherent strategy exists: In general, HEP data are lost

UA1, CDF Run 1, LEP?

International Study Group on Data Preservation

- Collider Experiments
  - e^+e^-, ep, $p\bar{p}$
- Computing Centers
- Funding agencies
- About 50 contact persons in total

Coordination
Chair : Cristinel Diaconu (DESY/CPPM)

Working Groups Conveners:
Physics Cases 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)

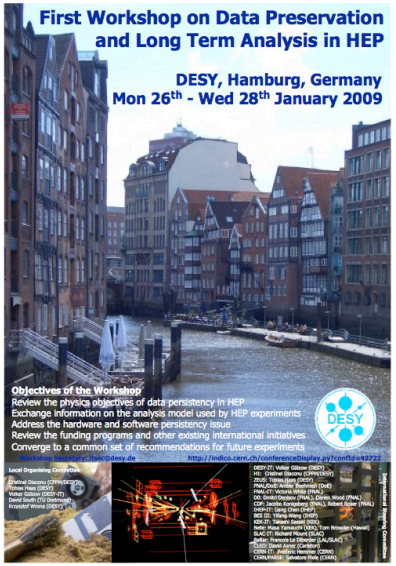
International Advisory Committee
Chairs: Jonathan Dorfan (SLAC), Siegfried Bethke (MPI Munich)

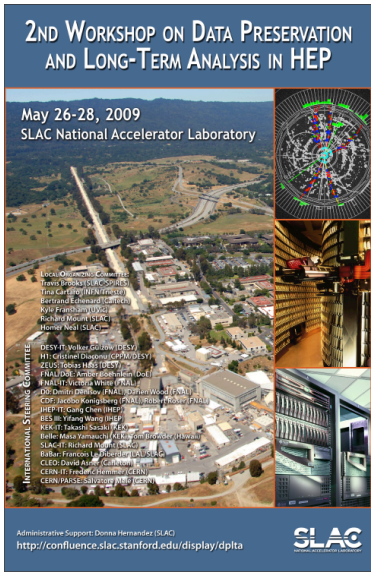
Advisers: Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo)

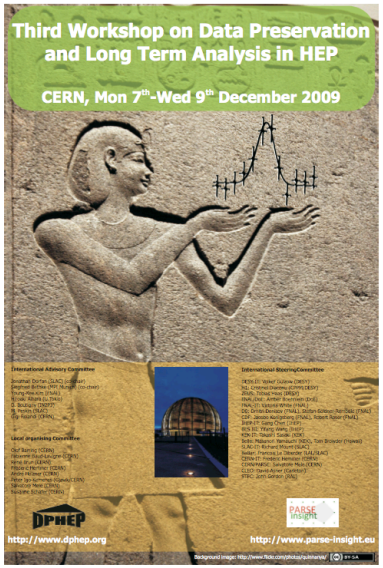
DPHEP Activities

- First contacts in September 2008
- Series of DPHEP workshops held in 2009

@ DESY (Jan) @ SLAC (May) @ CERN (Dec)







- Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields (for example astrophysics)

- Activities endorsed by ICFA summer 2009

DPHEP Visibility

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.

A simulated event in the JADE detector, generated using a refined Monte Carlo program and reconstructed using revitalized software more than 10 years after the end of the experiment. (Courtesy Siggi Bethke.)

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 the need to re-analyse old data. Examples of such analyses already exist and they are likely to become more frequent in the future. As experimental complexity and the associated costs continue to increase, many present-day experiments, especially those based at colliders, will provide unique data sets that are unlikely to be improved upon in the short term. The close of the current decade will see the end of data-taking at several large experiments and scientists are now confronted with the question of how to preserve the scientific heritage of this valuable pool of acquired data.

To address this specific issue in a systematic way, the Study Group on Data Preservation and Long Term Analysis in High Energy Physics formed at the end of 2008. Its aim is to clarify the objectives and the means of preserving data in high-energy physics. The collider experiments BaBar, Belle, BES-III, CLEO, CDF, DØ, H1 and ZEUS, as well as the associated computing centres at SLAC, KEK, the Institute of High Energy Physics in Beijing, Fermilab and DESY, are all represented, together with CERN, in the group's steering committee.

Digital gold mine

The group's inaugural workshop took place on 26–28 January at DESY, Hamburg. To form a quantitative view of the data landscape in high-energy physics, each of the participating experimental collaborations presented their computing models to the workshop, including the applicability and adaptability of the models to long-term analysis. Not surprisingly, the data models are similar – reflecting the nature of colliding-beam experiments.

The data are organized by events, with increasing levels of abstraction from raw detector-level quantities to N-tuple-like data for physics analysis. They are supported by large samples of simulated Monte Carlo events. The software is organized in a similar manner, with a more conservative part for reconstruction to reflect the complexity of the hardware and a more dynamic part closer to the analysis level. Data analysis is in most cases done in C++ using the ROOT analysis environment and is mainly performed on local computing farms. Monte Carlo simulation also uses a farm-based approach but it is striking to see how popular the Grid is for the mass-production of simulated events. The amount of data that should be preserved for analysis varies between 0.5 PB and 10 PB for each experiment, which is not huge by today's standards but nonetheless a large amount. The degree of preparation for long-term data varies between experiments but it is obvious that no preparation was foreseen at an early stage of the programs; any conservation initiatives will take place in parallel with the end of the data analysis.

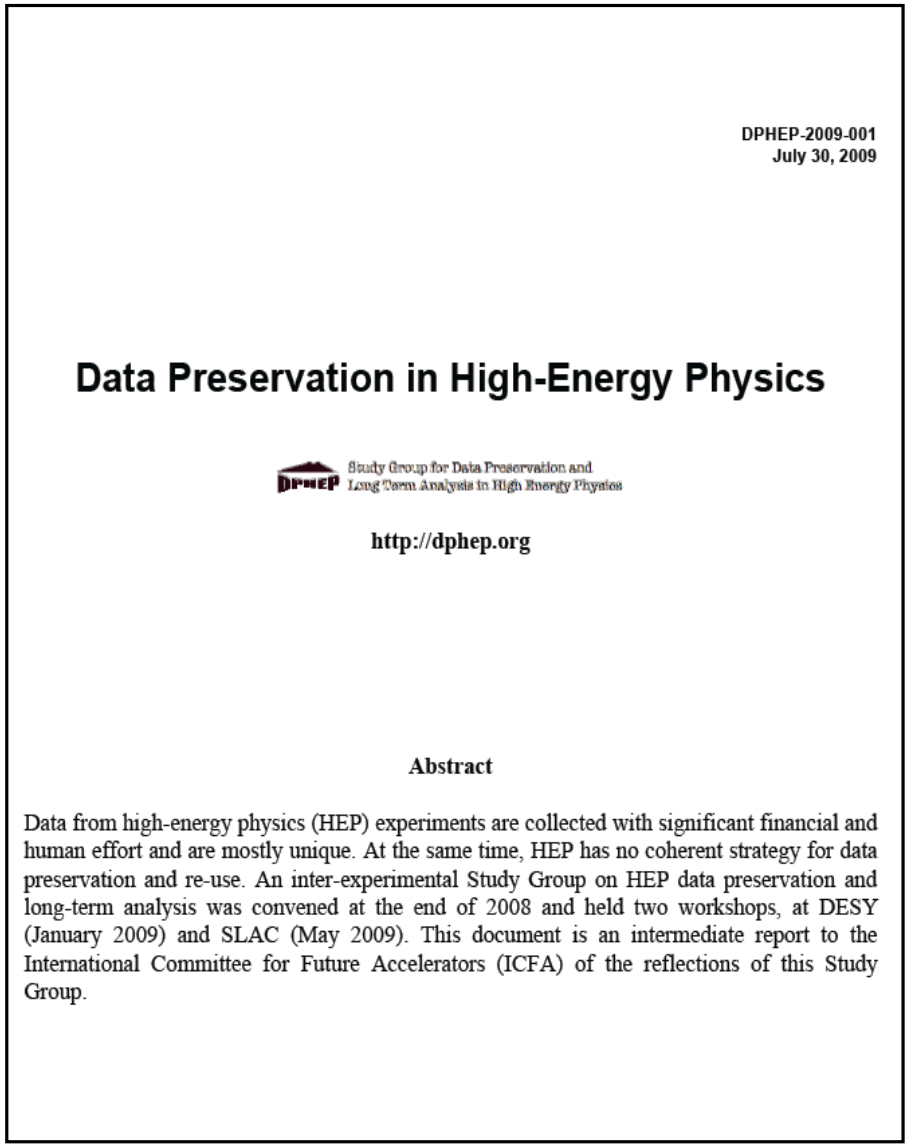
From a long-term perspective, digital data are widely recognized as fragile objects. Speakers from a few notable computing centres – including Fabio Hernandez of the Centre de Calcul de l'Institut, National de Physique Nucléaire et de Physique des Particules, Stephen Wolbers of Fermilab, Martin Gasthuber of DESY and Erik Mattias Wadenstein of the Nordic DataGrid Facility – showed that storage technology should not pose problems with respect to the amount of data under discussion. Instead, the main issue will be the communication between the experimental collaborations and the computing centres after final analyses and/or the collaborations where roles have not been clearly defined in the past. The current preservation model, where the data are simply saved on tapes, runs the risk that the data will disappear into cupboards while the read-out hardware may be lost, become impractical or obsolete. It is important to define a clear protocol for data preservation, the items of which should be transparent enough to ensure that the digital ▷

CERN Courier, May 2009

Symmetry, December 2009

Berliner Zeitung and Frankfurter Rundschau, February 2010

Intermediate DPHEP Report Released Nov 2009



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.

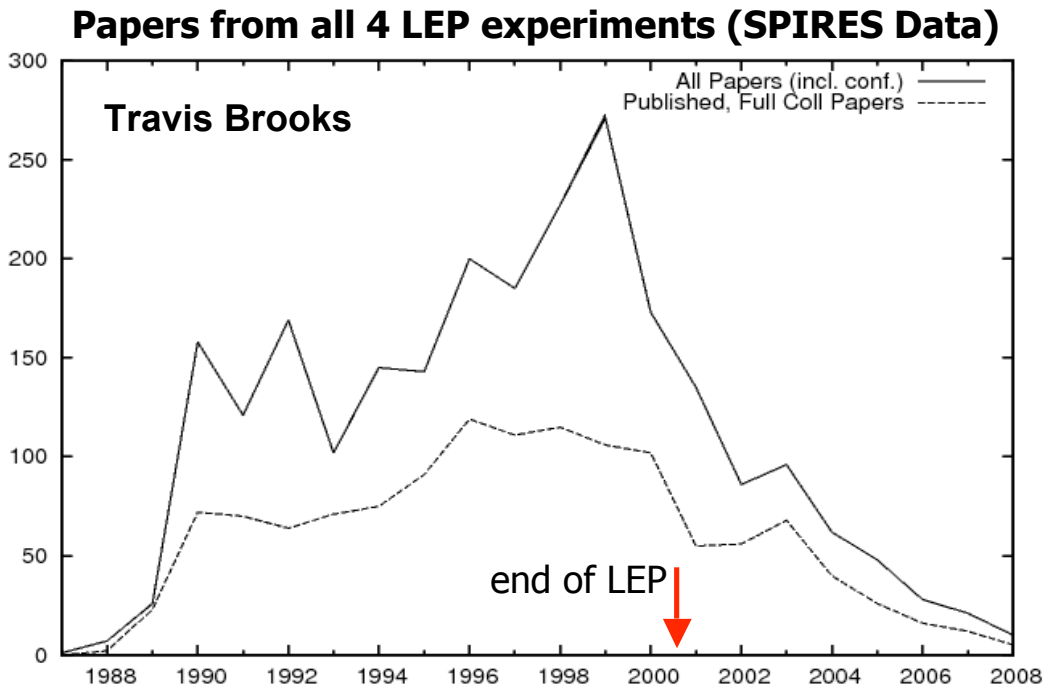
- First recommendations of the group published November 2009 ***arXiv:0912.0255***
- Report covers the four areas
  - 1. Physics Cases
  - 2. Preservation Models
  - 3. Technologies
  - 4. Governance
- In this talk:

  Present the main ideas, project scope and preliminary recommendations of the DPHEP study group

Part 1: Physics Cases for Data Preservation

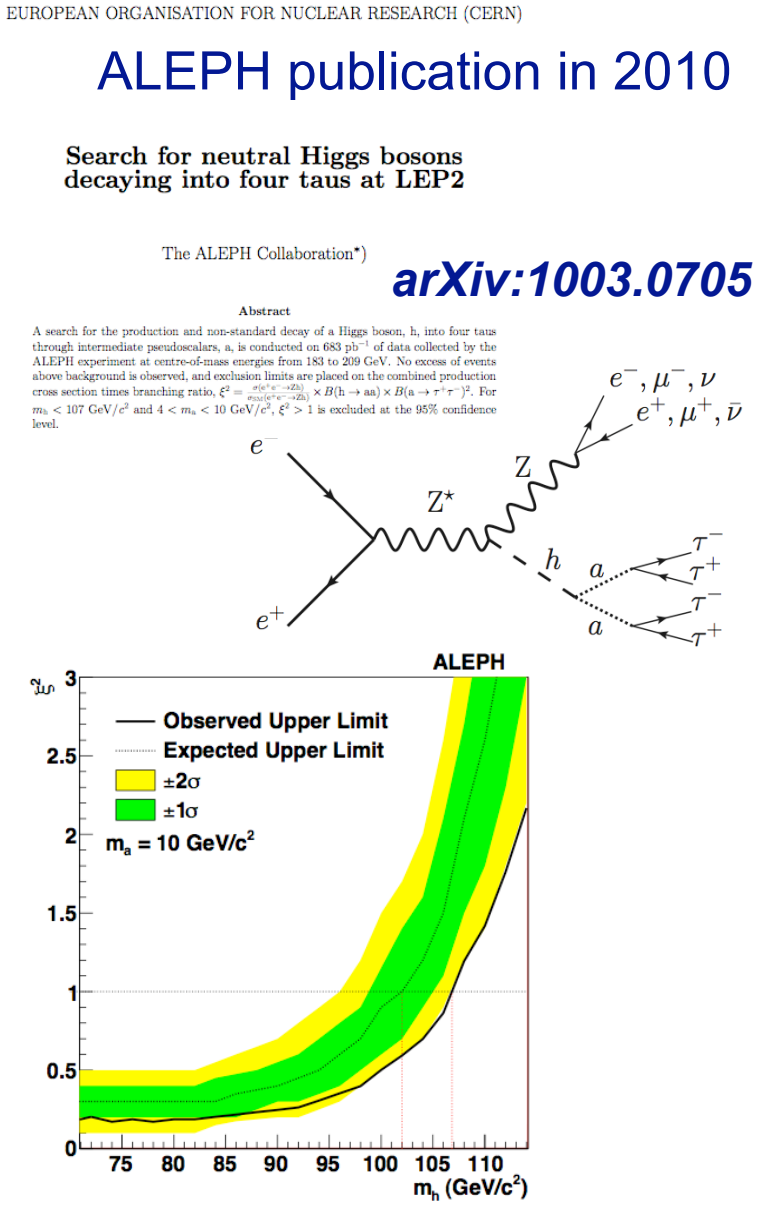
- HEP data are mostly unique and have true scientific potential
- Several physics cases can be presented for preservation
  - Long term completion and extension of the existing physics program: safeguarding the data
  - Cross collaboration between experiments - usually done towards the end of the programmes
  - Re-use of old data: go back and do something new
  - Use in scientific training, education, outreach

Extending the Current Programme



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

ALEPH publication in 2010


EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Search for neutral Higgs bosons decaying into four taus at LEP2

The ALEPH Collaboration*)

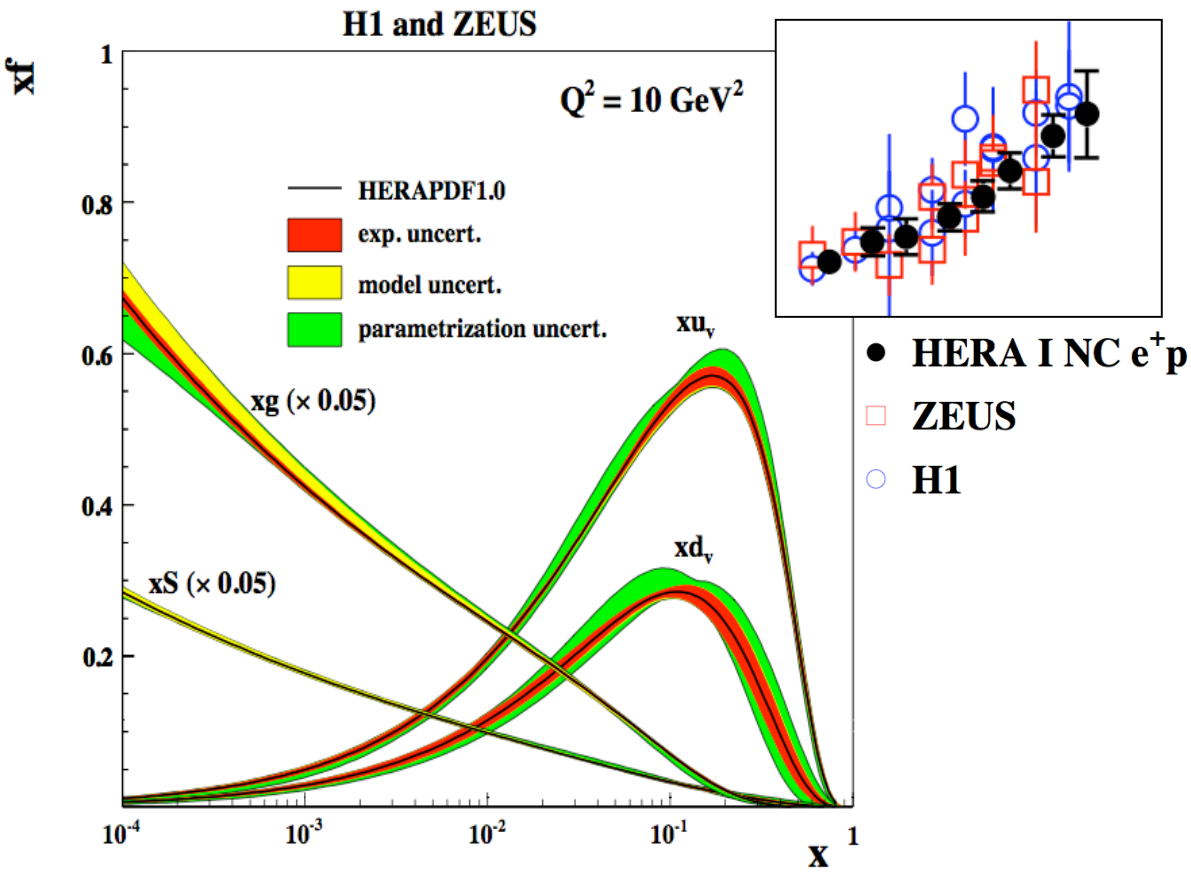
arXiv:1003.0705

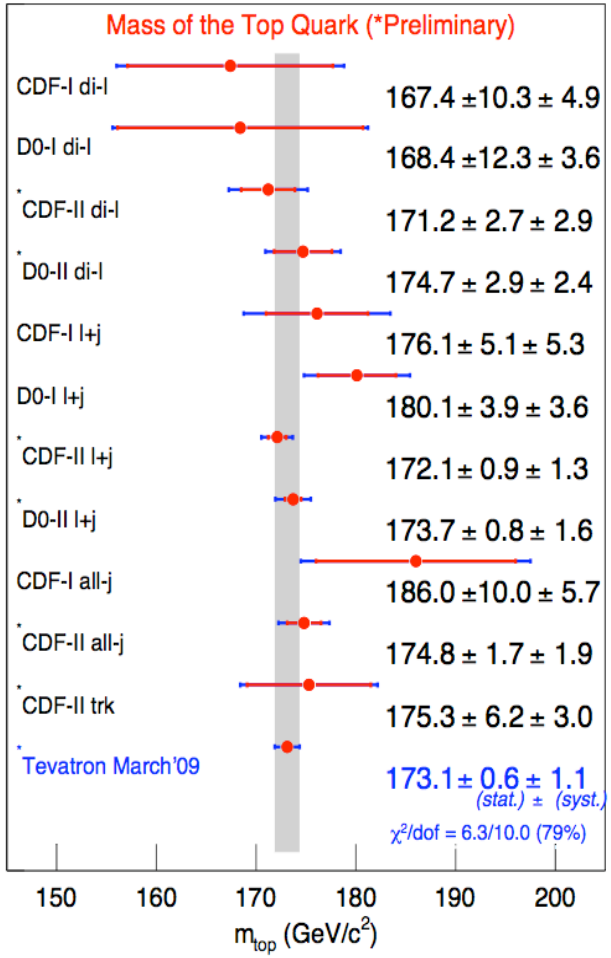
Abstract

A search for the production and non-standard decay of a Higgs boson, h, into four taus through intermediate pseudoscalars, a, is conducted on 683 pb^{-1} of data collected by the ALEPH experiment at centre-of-mass energies from 183 to 209 GeV. No excess of events above background is observed, and exclusion limits are placed on the combined production cross section times branching ratio, $\xi^2 = \frac{\sigma(e^+e^- \to Zh)}{\sigma_{SM}(e^+e^- \to Zh)} \times B(h \to aa) \times B(a \to \tau^+\tau^-)^2$. For $m_h < 107$ GeV/c^2 and $4 < m_a < 10$ GeV/c^2, $\xi^2 > 1$ is excluded at the 95% confidence level.

Cross Collaboration and Combinations

- Combined results already exist from LEP, Tevatron, HERA, BaBar+Belle (in progress)

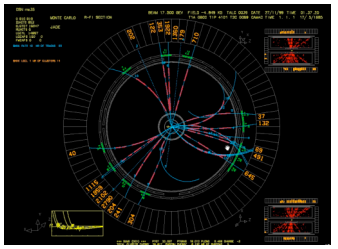




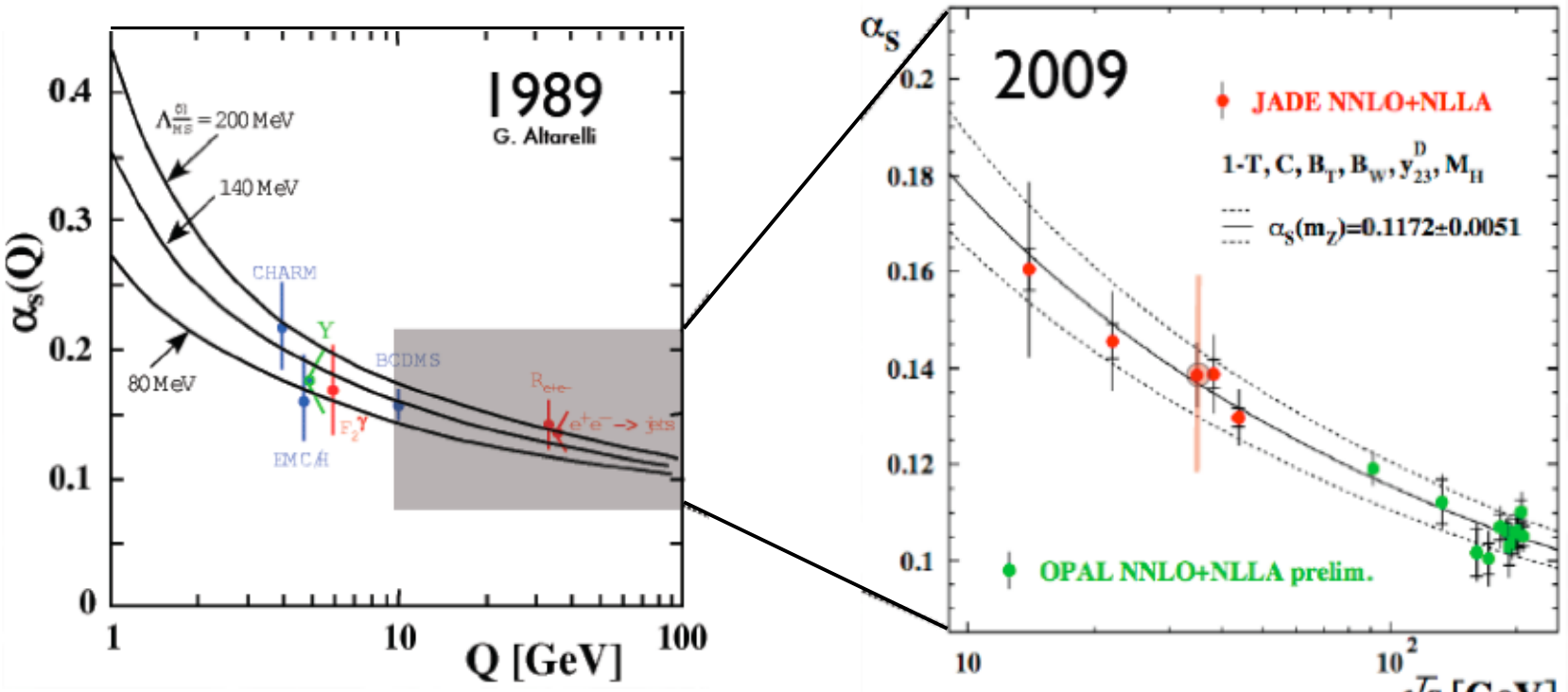
- Preserved data would make possible more combined analyses across experiments

Re-use of Data from Old Experiments

- Improve precision on former measurements
- Apply new and improved theoretical predictions
- Check new physics in the old data samples
- Investigate discrepancies



JADE: full raw data preservation, software revitalisation, needed many individual initiatives

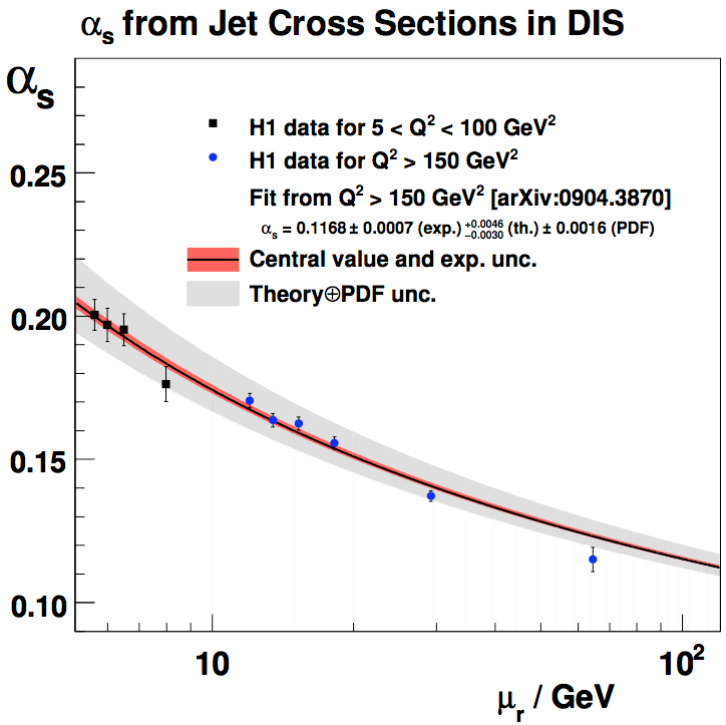


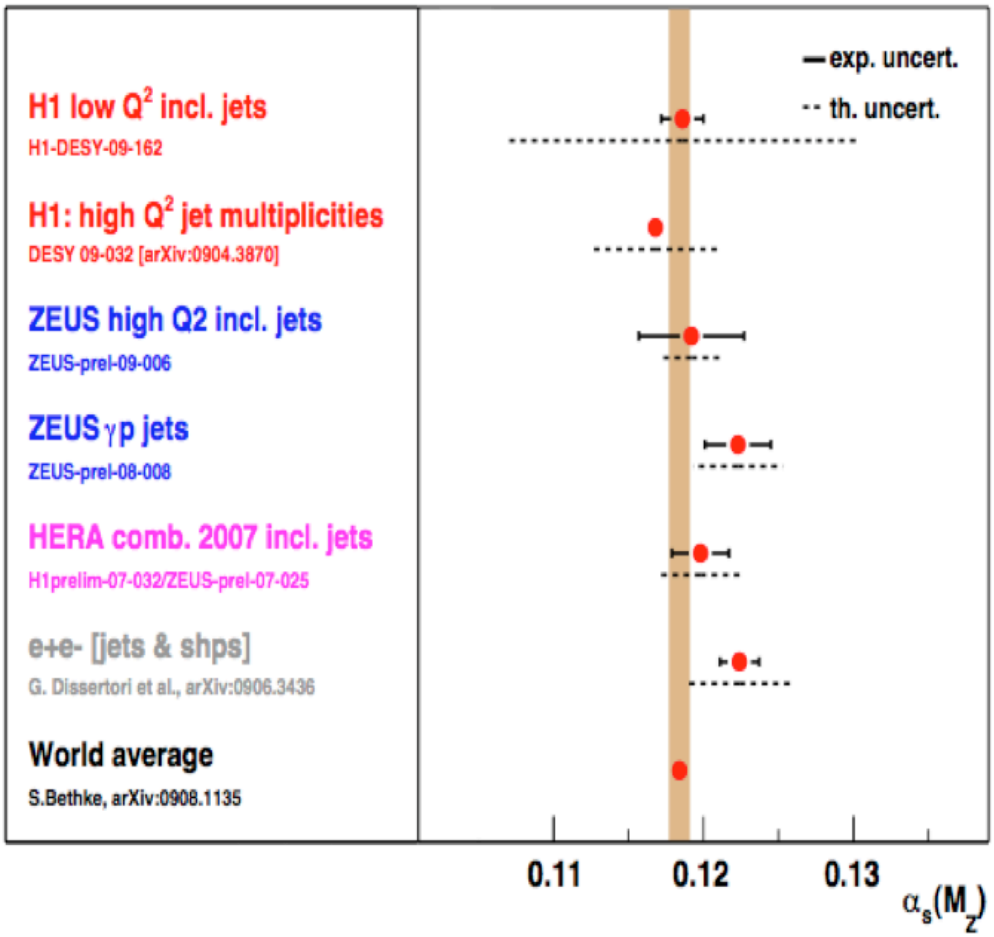
10 recent publications

History may well repeat itself....

- Around 10% of measurements are dominated by non-experimental errors: theory ($N^{n}LO$?) and simulation..

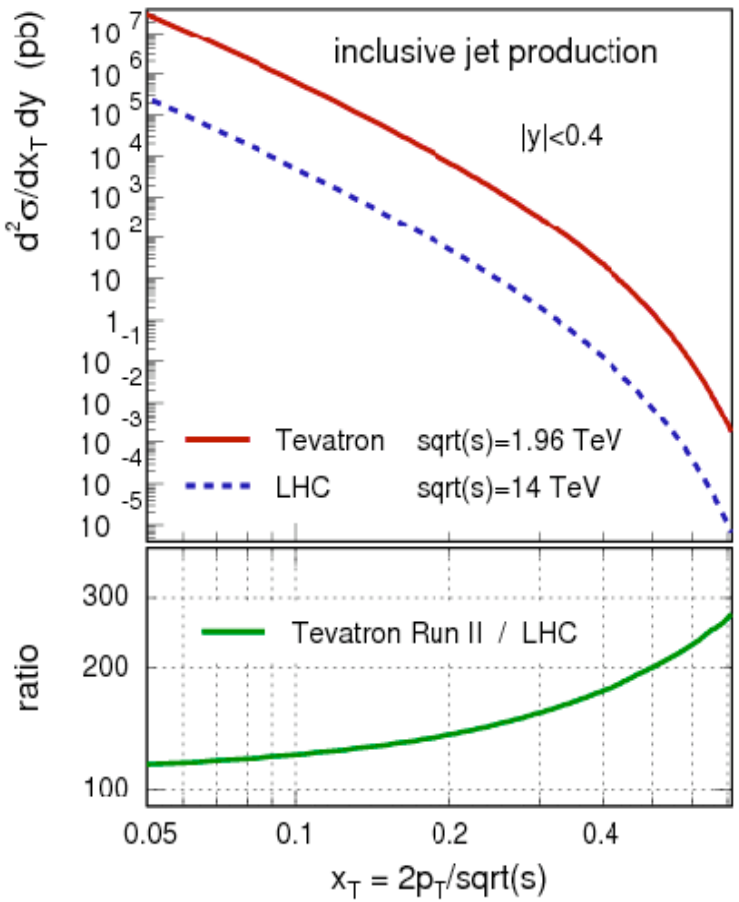
- More recent precision measurements of α_S:





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)

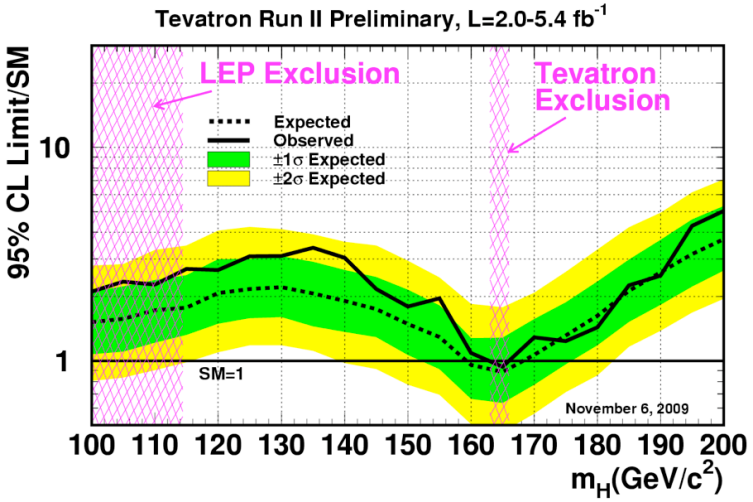
- need more than 1600fb-1 luminosity to compete with Tevatron@8fb-1
- more high-x gluon contributions
- but more steeply falling cross sect. at highest pT (=larger uncertainties)

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

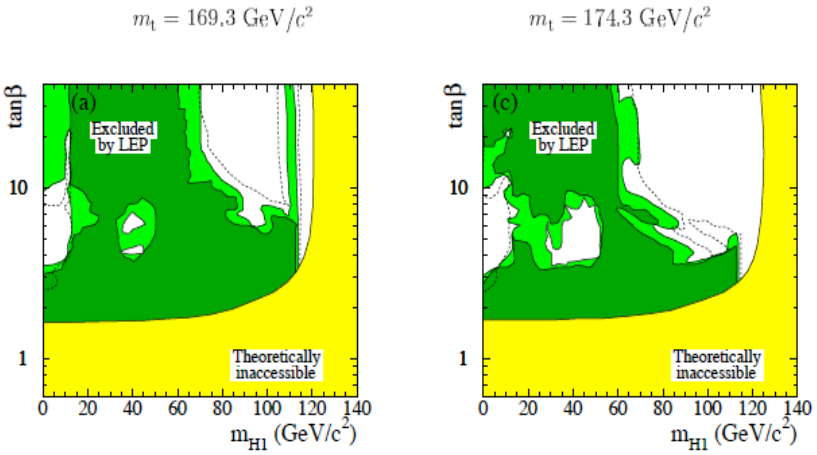
21

M. Wobisch

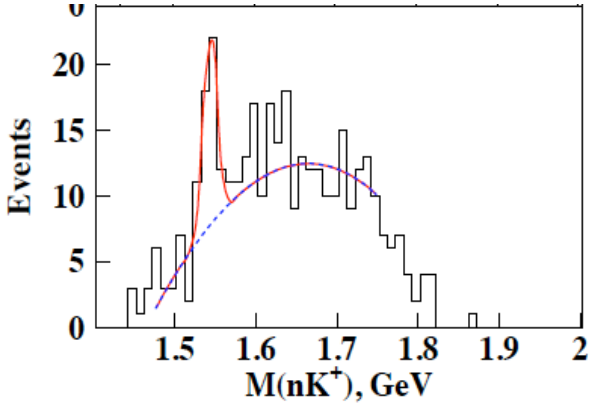
More Examples: Contingency with Future Programmes



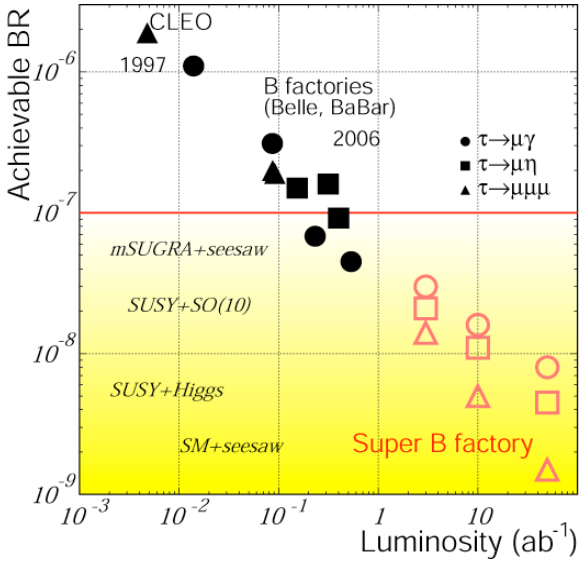
Tevatron/LHC collaboration?



Future LEP analysis optimisation?

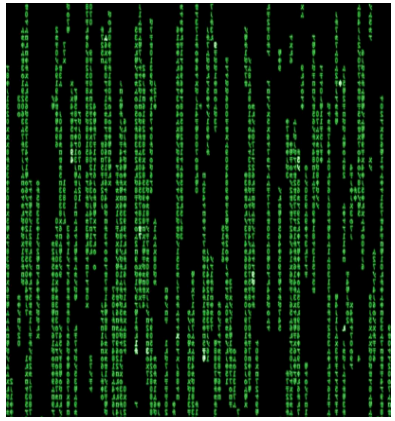


CLAS @ JLAB (low energy)



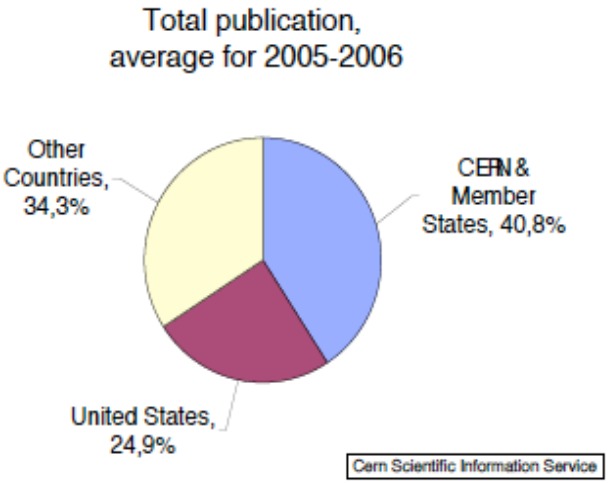
B- and Super-B factories

Scientific Training, Education, Outreach

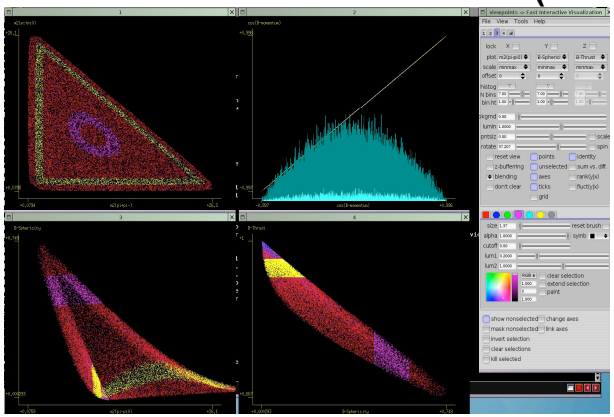


HEP: complicated..

..something for everyone?



Matt Bellis (SLAC)



Outreach projects: *more later*

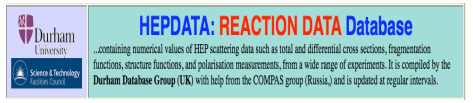
- Improve the overall high level education in HEP
- Improve the connection of HEP-emerging countries to HEP data sets

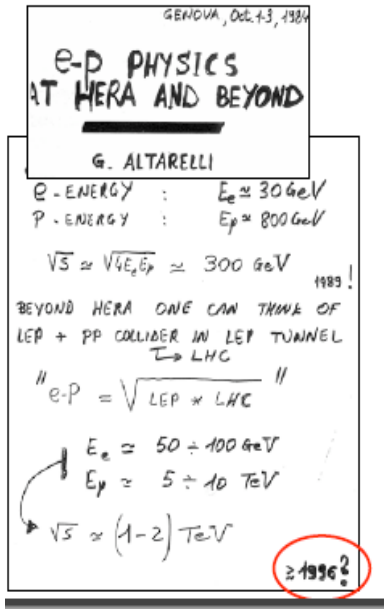
Part 2: Models of Data Preservation

What is "HEP Data" anyway?

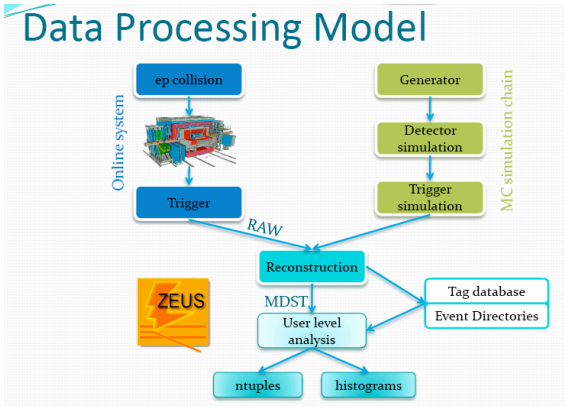
- Digital information
  - Data event files, database
- Software
  - Simulation, reconstruction, analysis, user
- Publications
  - Journals, arXiv, spires, HEP data….
- Documentation
  - Publications, notes, manuals, slides
- Meta information
  - Hyper-news, messages, forums
- Expertise (people)
  - *Often the hardest to secure..*

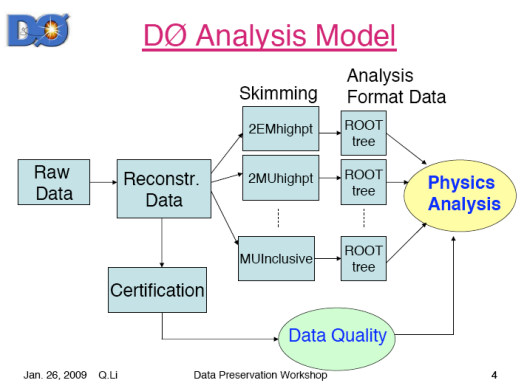
entropy

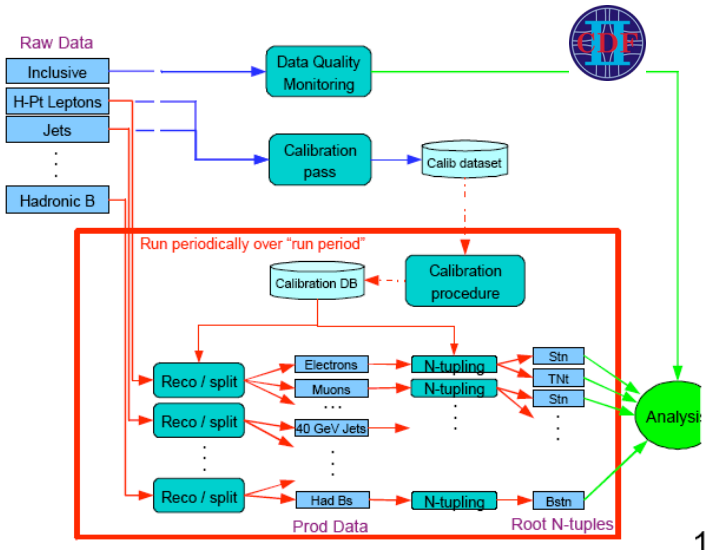


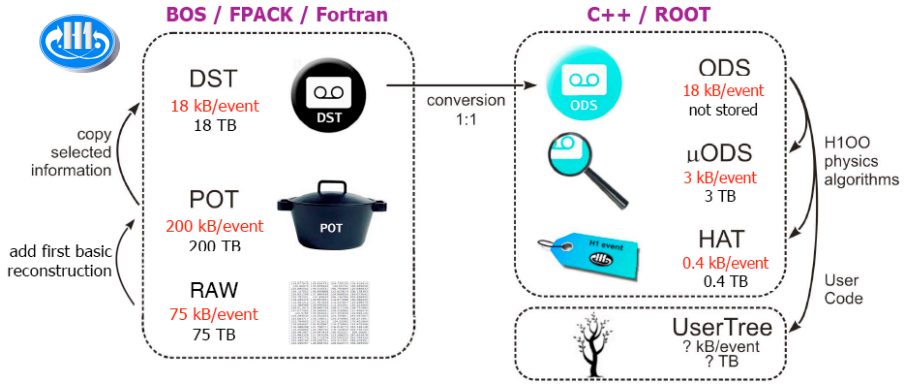


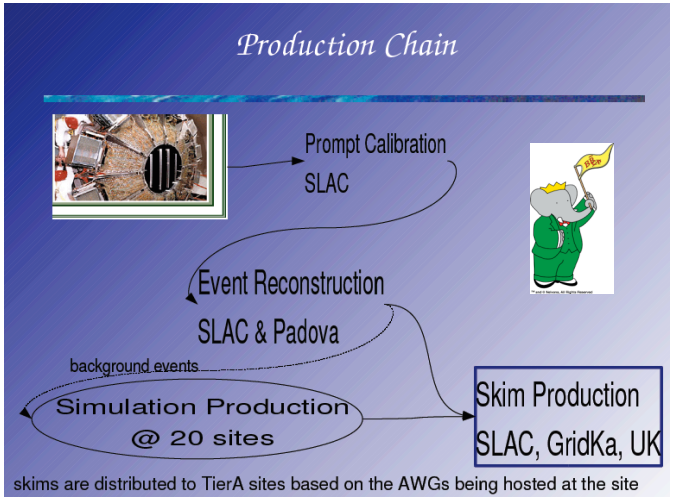
Data Analysis Models in HEP











- Complicated, at first glance different
- Familiar descriptions of data analysis chain, from reconstruction to analysis level
  - RAW → POT → DST → *ntuple* → analysis

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

***Needed by:* JADE, ALEPH**
***Planned by:* BaBar, H1, ZEUS (3-4)**

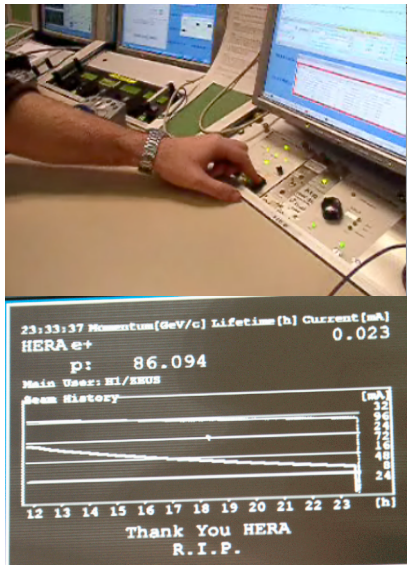
- Each of the higher preservation levels are inclusive, e.g. 3 means "1 to 3"
- Each level implies an R&D project at experiment level
- Simplest levels are 1 and 2, *but this still means some work!*

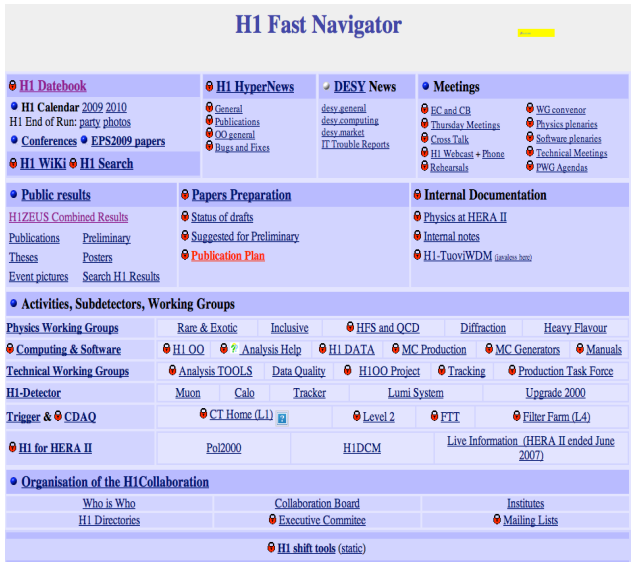
Paper Documentation



- Current location: Where is everything now?
  - Talks from pre-web days, detector schematics, blueprints, logbooks
- Digitisation: Is it a viable / affordable solution?
- Future location: Where can we put everything?
  - Cataloguing will be the biggest headache - can external services help?

Digital Documentation: (Online) (Meta-)Data



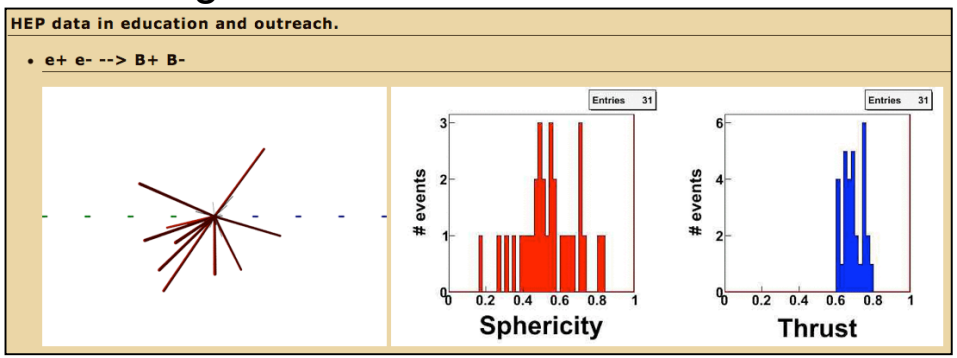


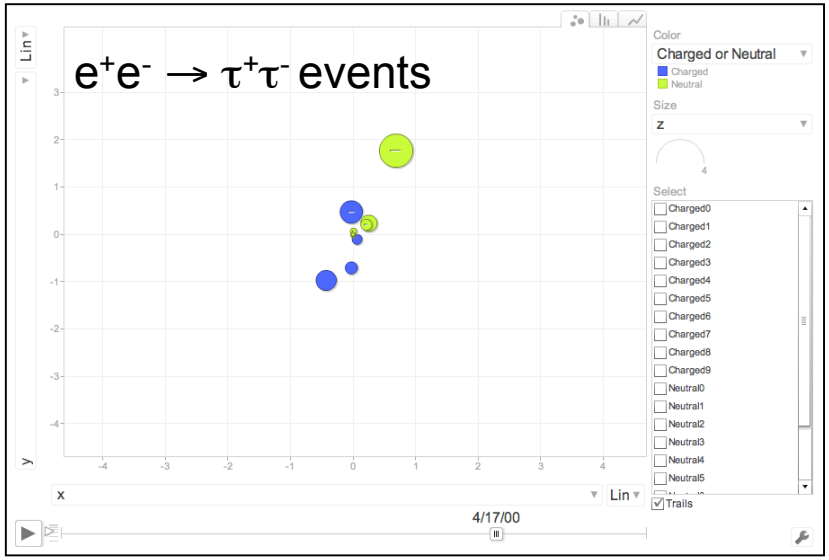
- While we were marking the end of HERA running and data taking a collection of applications were running in the Hall
  - In fact I would guess there were about 20 machines associated with different detector components ticking away, with online monitoring

- Ensuring the "standard" web-based documentation is up to date and complete is also challenging
  - In particular when there are many parts / responsible people

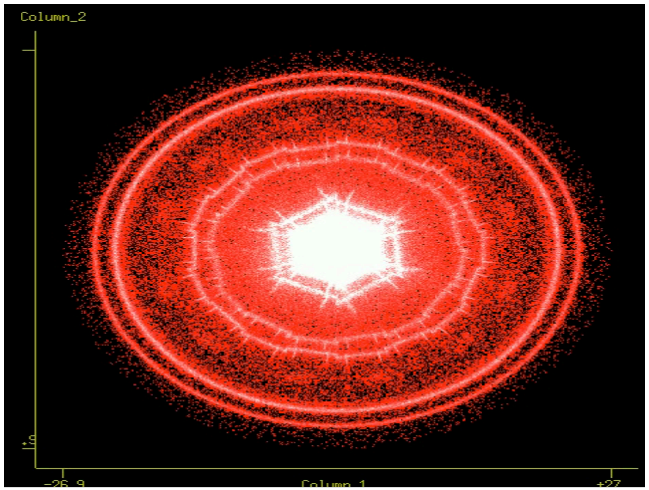
A Simplified Format? Outreach Data and Tools

Movie of generic $e^+e^- \rightarrow B^+B^-$ events

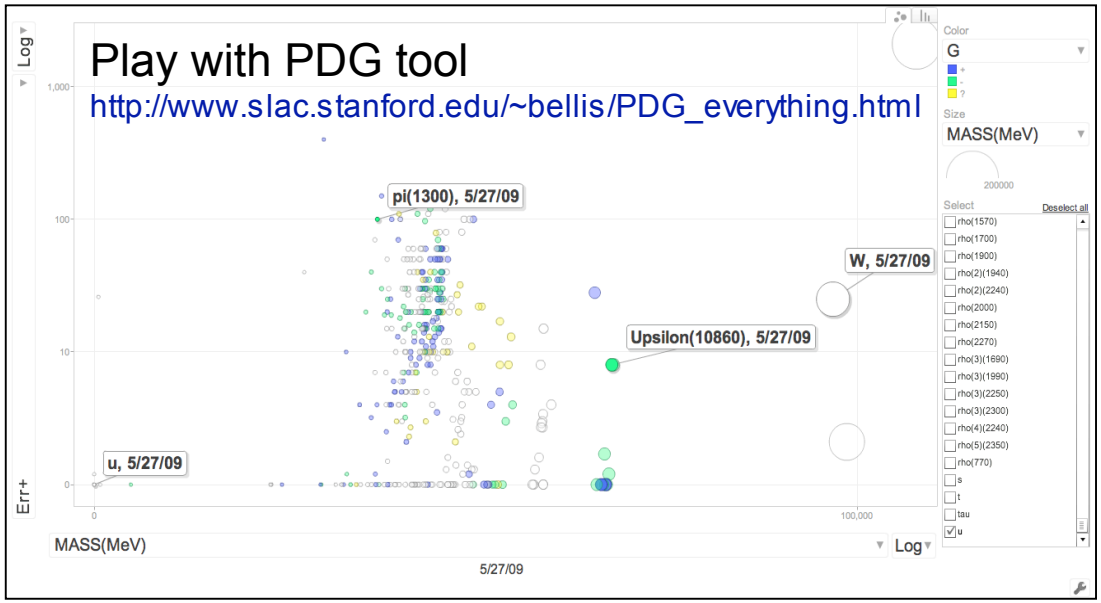




Several outreach tools already being used in classrooms



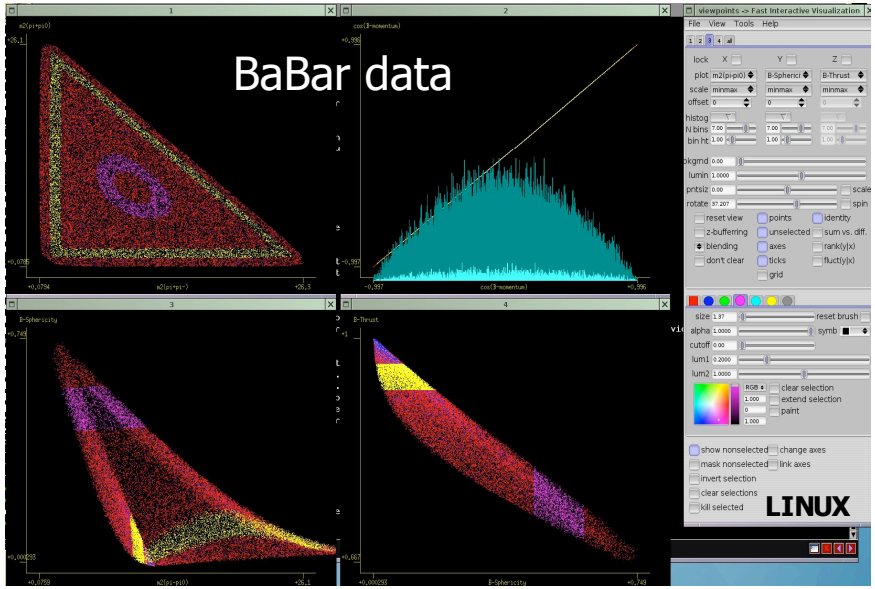
Tomography studies using converted photons in BaBar Silicon Vertex-Tracker

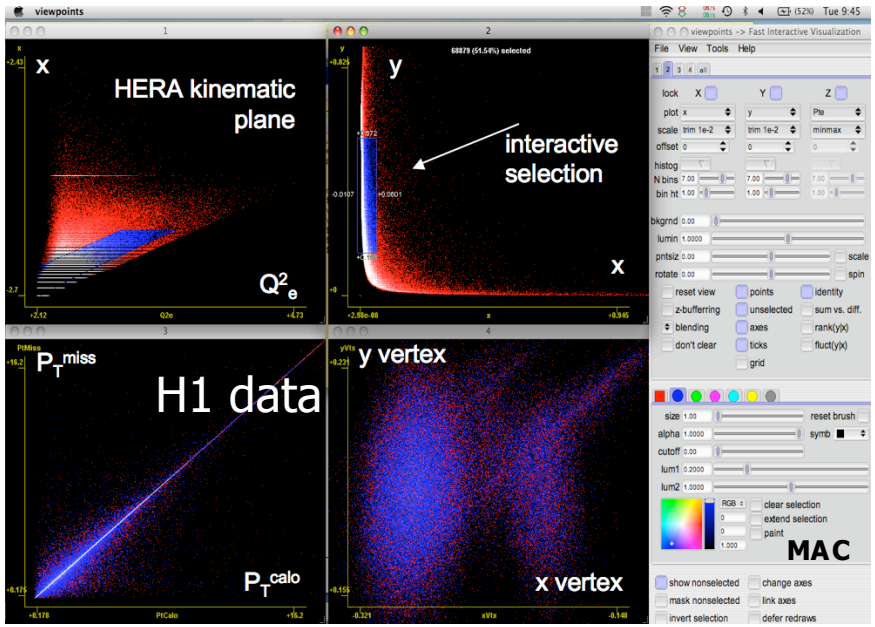


http://www.slac.stanford.edu/~bellis/HEP_data.html

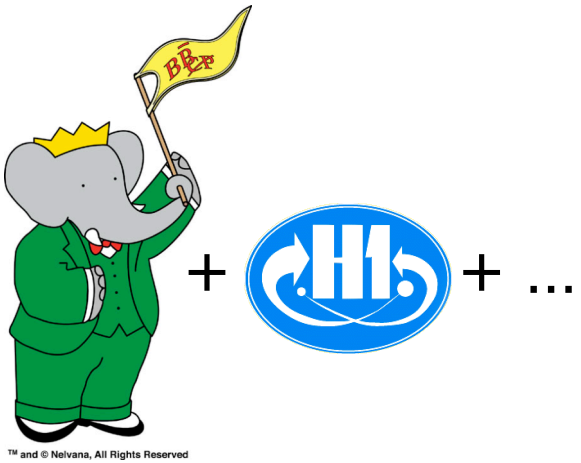
Outreach Example: Using *Viewpoints* from NASA

- Many attractive outreach tools available, like *Viewpoints* from NASA
- Simple data format: input using text file of kinematics of HEP events
- Nice example of new collaborative work between BaBar and H1 via DPHEP
- LASS data has also been used

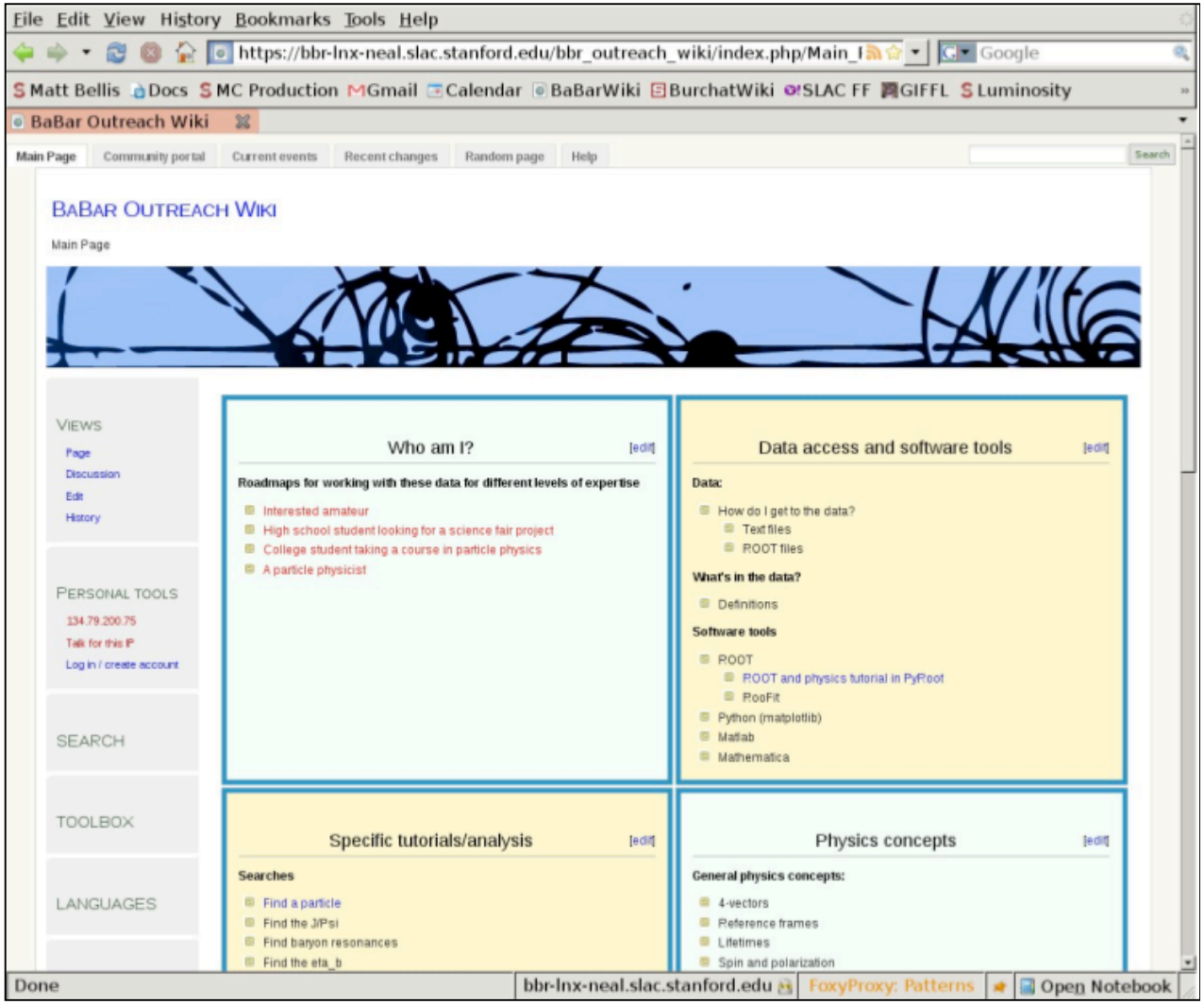




A HEP Outreach wiki?



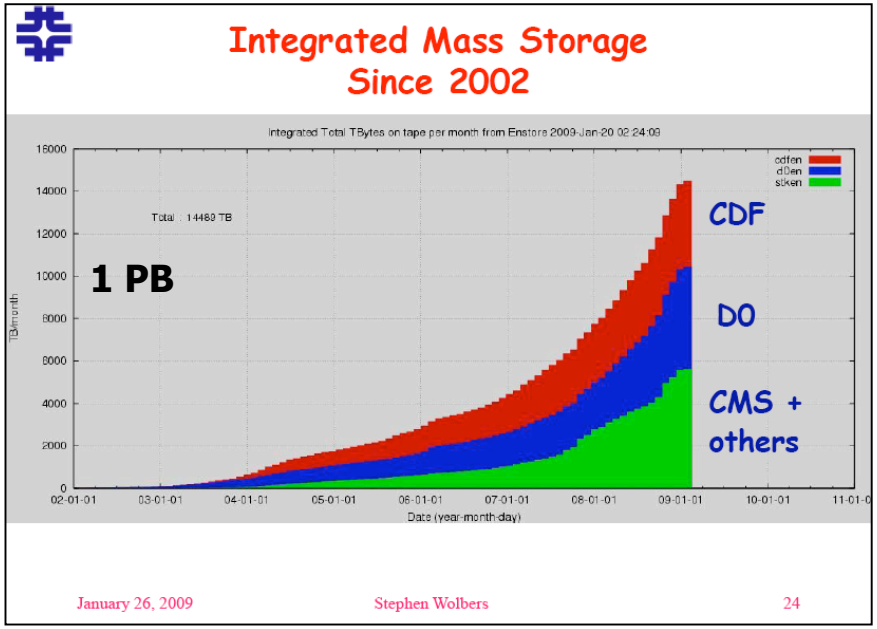
- Recent beginnings of a HEP wide effort, based on first ideas from BaBar
  - A true HEP data portal for outreach
  - Exercises using real HEP data aimed at a variety of levels
  - **Important:** Runs *in parallel* to main preservation efforts (level 4)



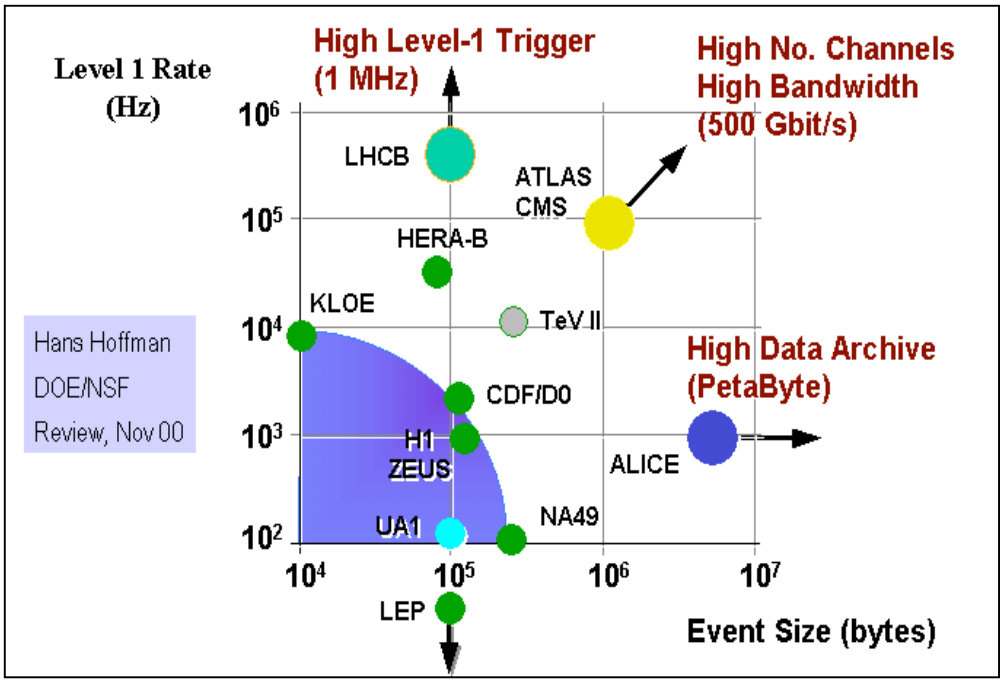
Matt Bellis (SLAC)

Part 3: Technologies

Size of Data Samples



Stephen Wolbers



- Storage technology should be comfortable by the end of the experiment: *Data preservation is not about the data!*
- However, regular migration of the data to latest technologies should be considered and carefully planned

Technological Issues

- Computing centres are (in principle) able to store the data
  - Discussions in DPHEP lead to a number of 0.5 to 10 Pb / exp

Total cost of data migration = 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: virtualisation and cloud computing

- Interface with experiments needs to be defined
  - Procedures, agreements, resources
  - Supervision and custodianship of data sets, archival expertise

Generic (Technological) Models for Data Preservation

- HEP data models could follow one of three directions

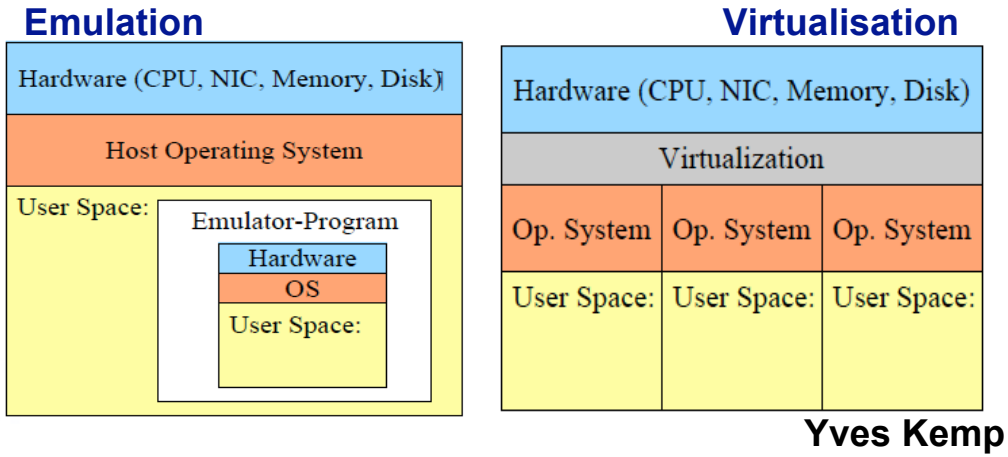
- Technology preservation
  - Freeze the hardware
  - Limited capability, one day it will however fall apart

- Technology emulation
  - Prepare it once (?) and migrate the "middleware"

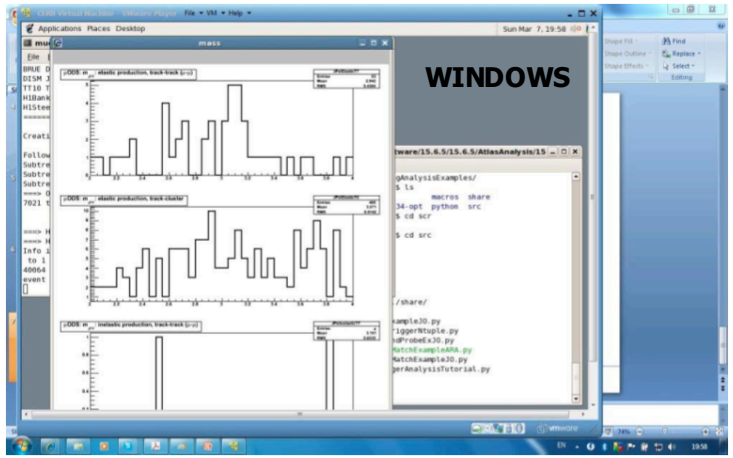
- Continuous migration
  - Follow technology changes
  - Adjust, redesign, recompile etc
  - Requires the most manpower, but has the most benefits…

Scientific Linux

Emulation and Virtualisation Techniques



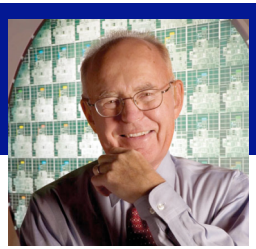
Yves Kemp



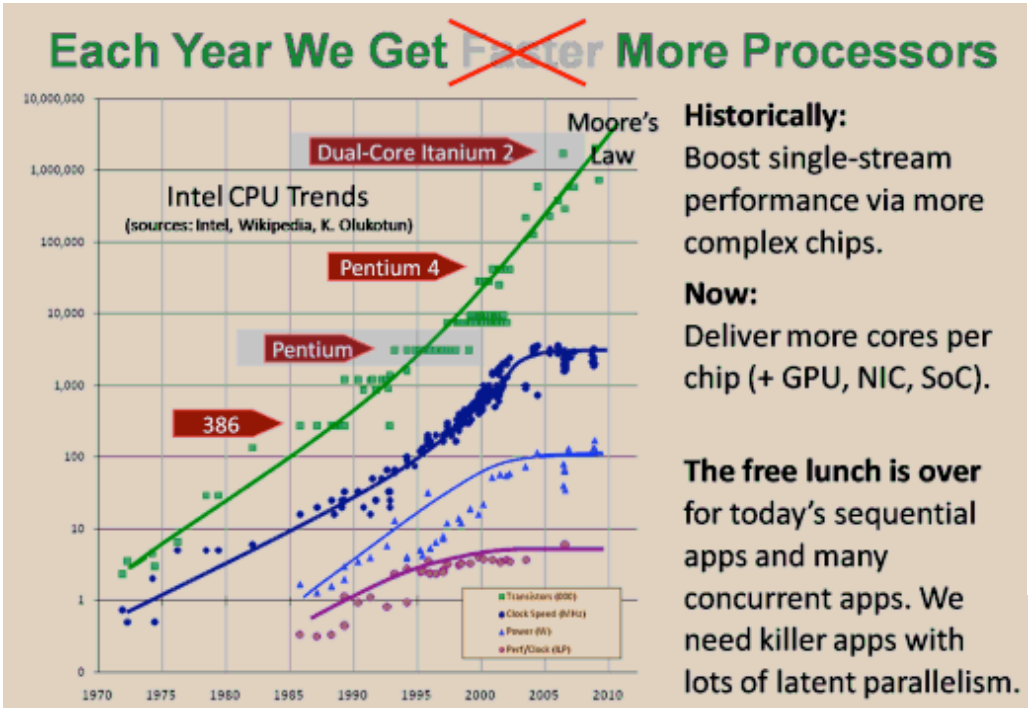
Mihajlo Mudrinic

- Different operating systems can be "preserved", using virtualisation techniques
  - Virtual environments also very useful, with enough validation, to aid potential OS transitions
- For HERA (and Tevatron?) experiments relatively new idea
  - But already included in the standard modus operandi for the LHC
  - Several HEP projects try out such strategies, using e.g. CERN-VM

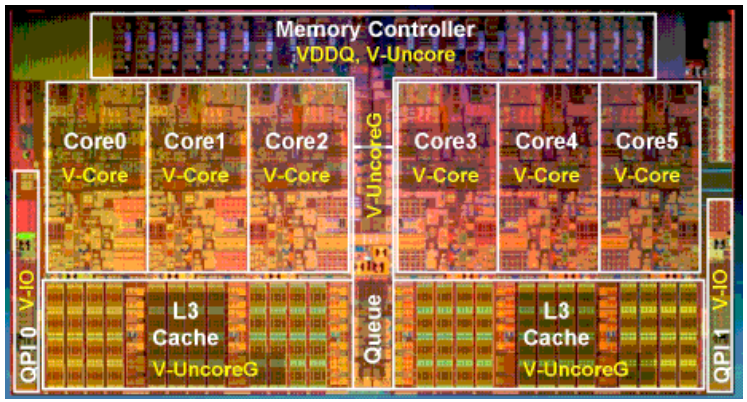
Computing Power



Moore's Law continues..

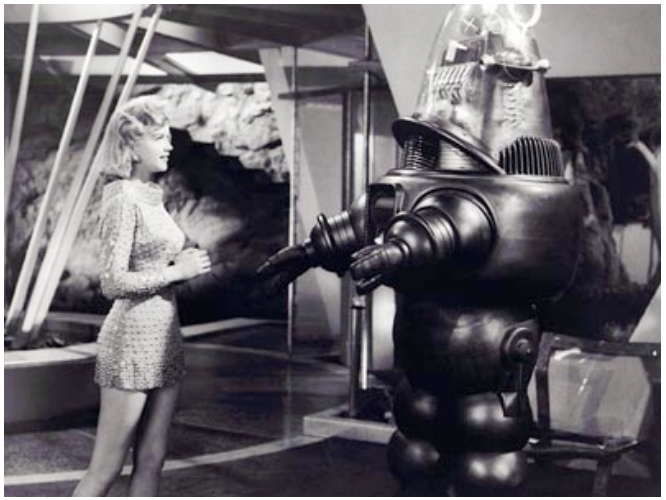


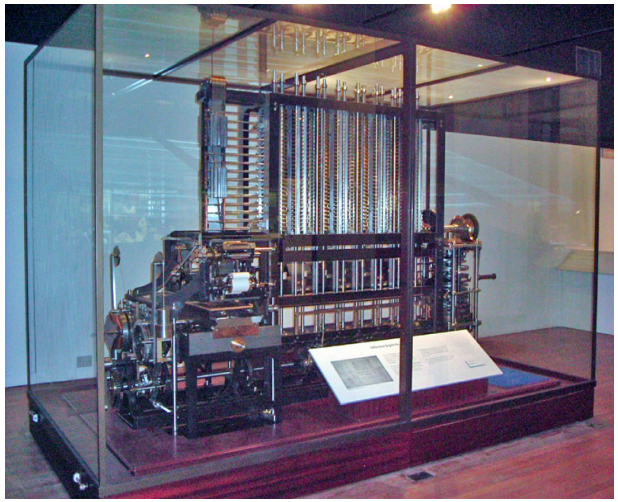
- Any archival system should be able to absorb future technological evolutions..
  - ***Parallelism*** crucial for future applications

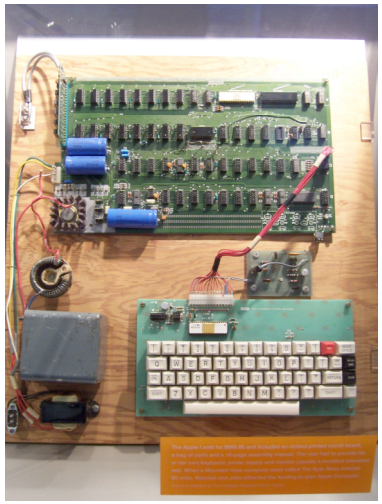


New Westmere 6-core from intel

Hardware Persistency

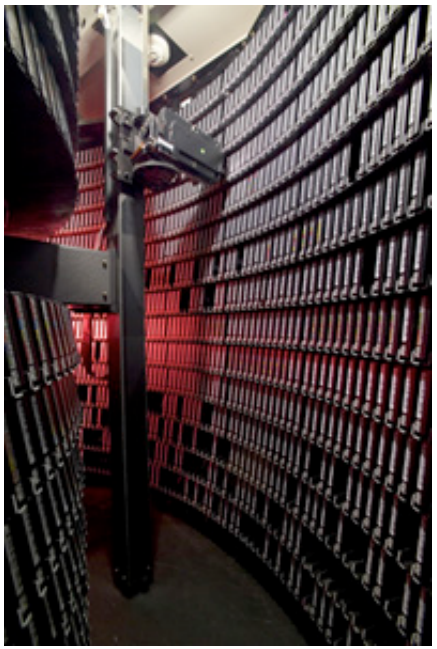


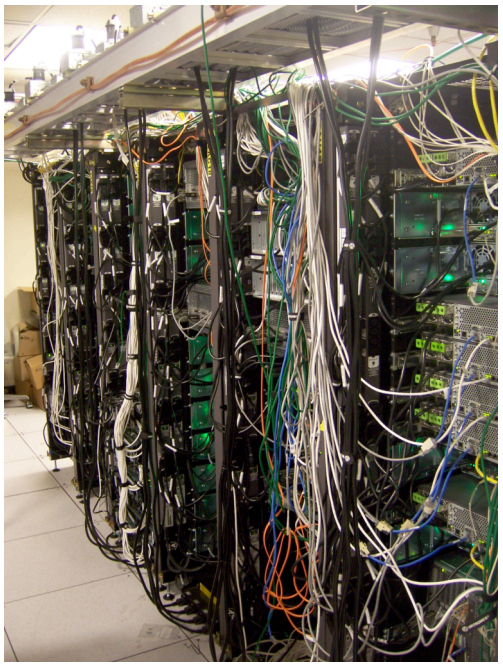




- Hardware now "old" after about 6 years, often no budget left to replace it

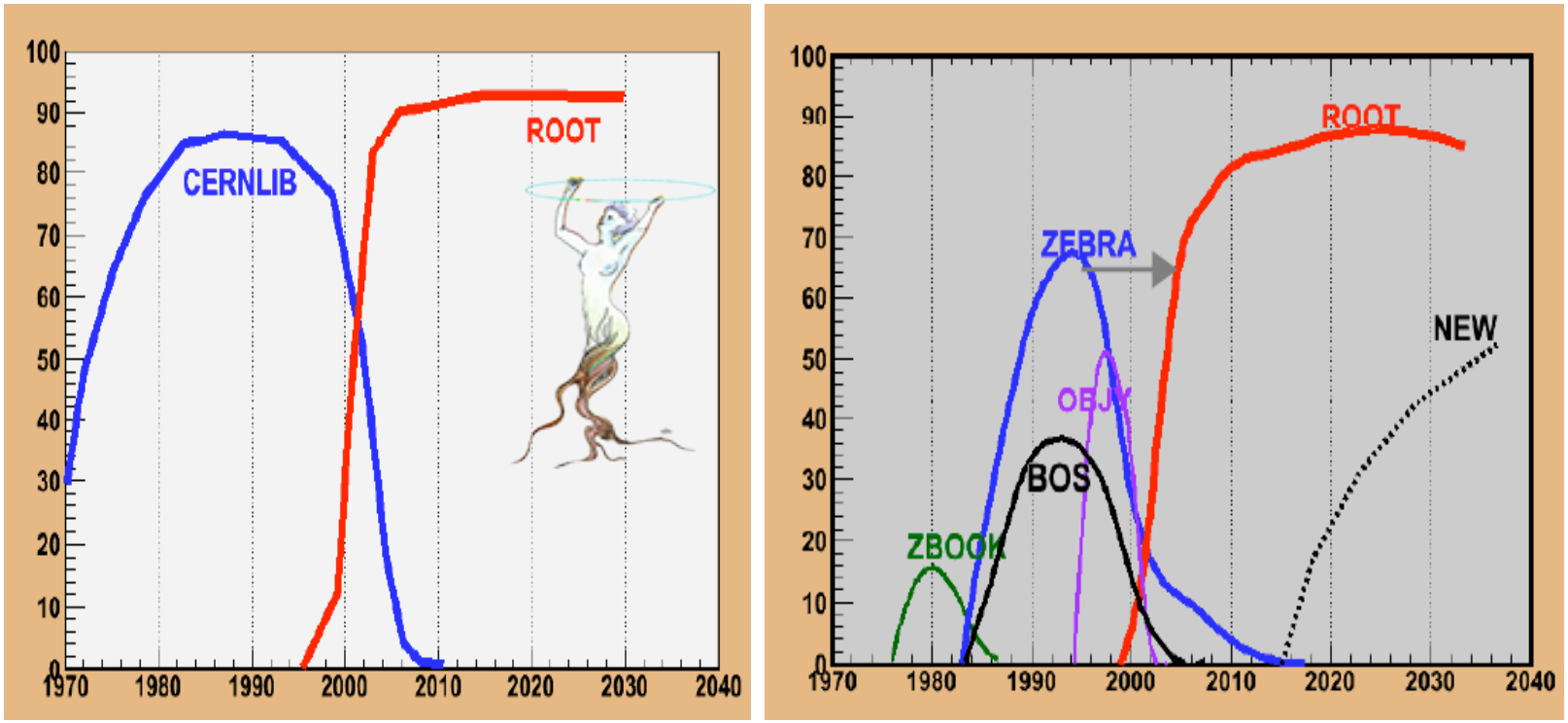
- New hardware has increased storage capacity (big tapes) and 64 bit (and beyond)





Software Persistency

Libraries **I/O**

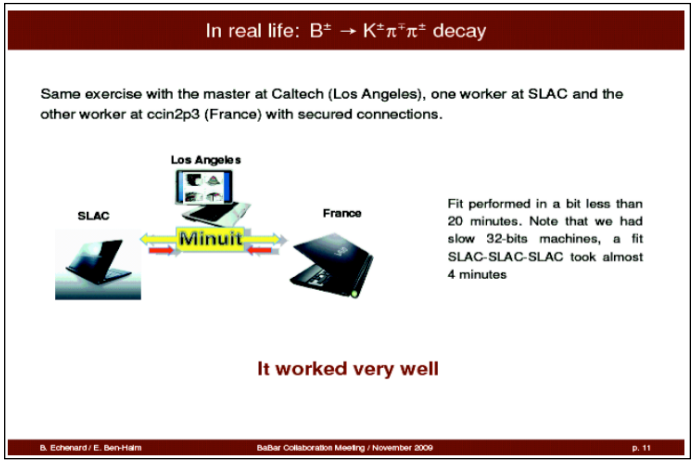


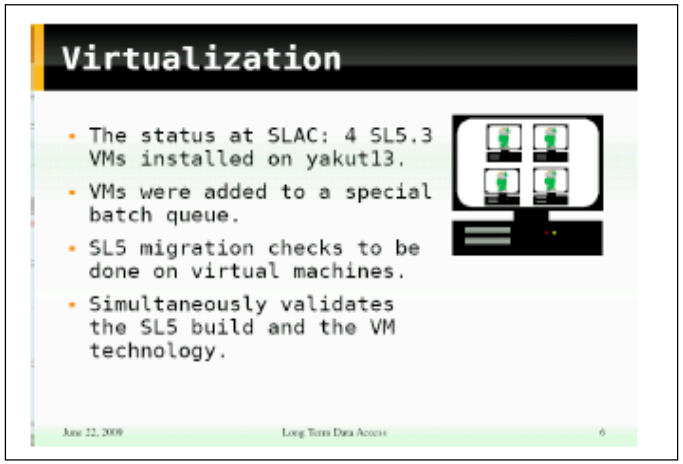
Rene Brun

- Software is a source of concern: maintenance, migration, validation
  - ROOT offers the needed coherence in next few decades (and documentation)
  - Fortran not always trivial: *gcc* goes to *gfortran* in SL4 to SL5 transition
  - Other dangers: commercial software (e.g. databases) may cause problems..

Another Example: BaBar Archival project

Collaboration between Belle and BaBar





Homer Neal

- BaBar analysis model is moving to an "Archival Mode"
- Use of virtualisation for validation of OS transition and for performing combined analyses with Belle
- Also major advances in use of Cloud Computing, running analyses on Amazon™ CPU resource

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

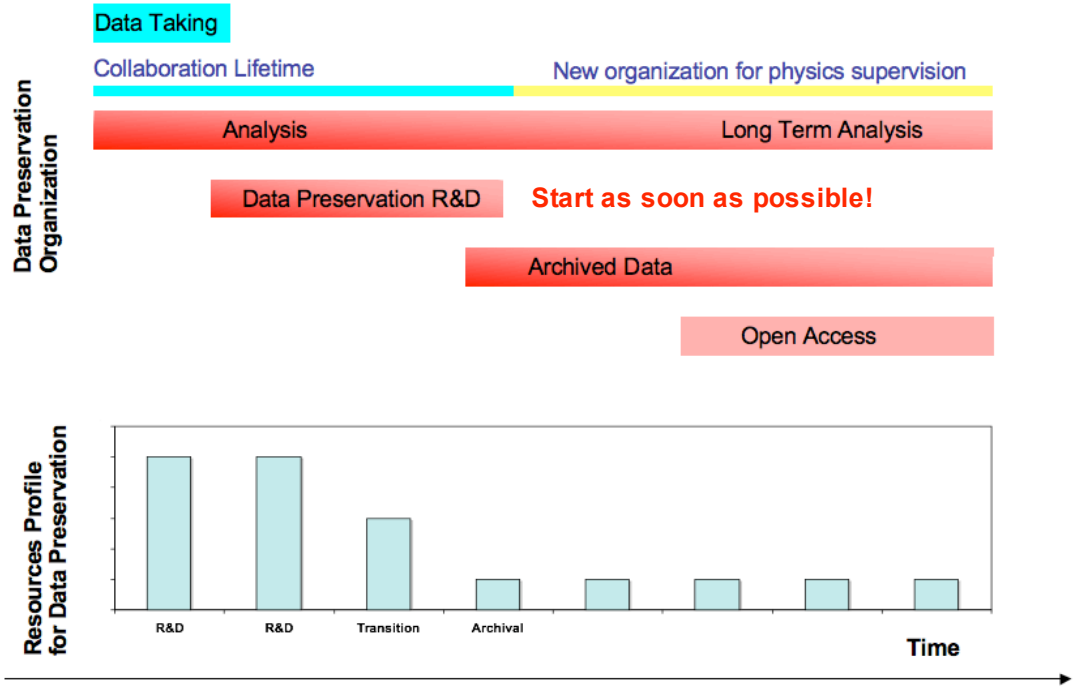
Part 4: Governance

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

Transition Scenario and Resources at the Experimental Level

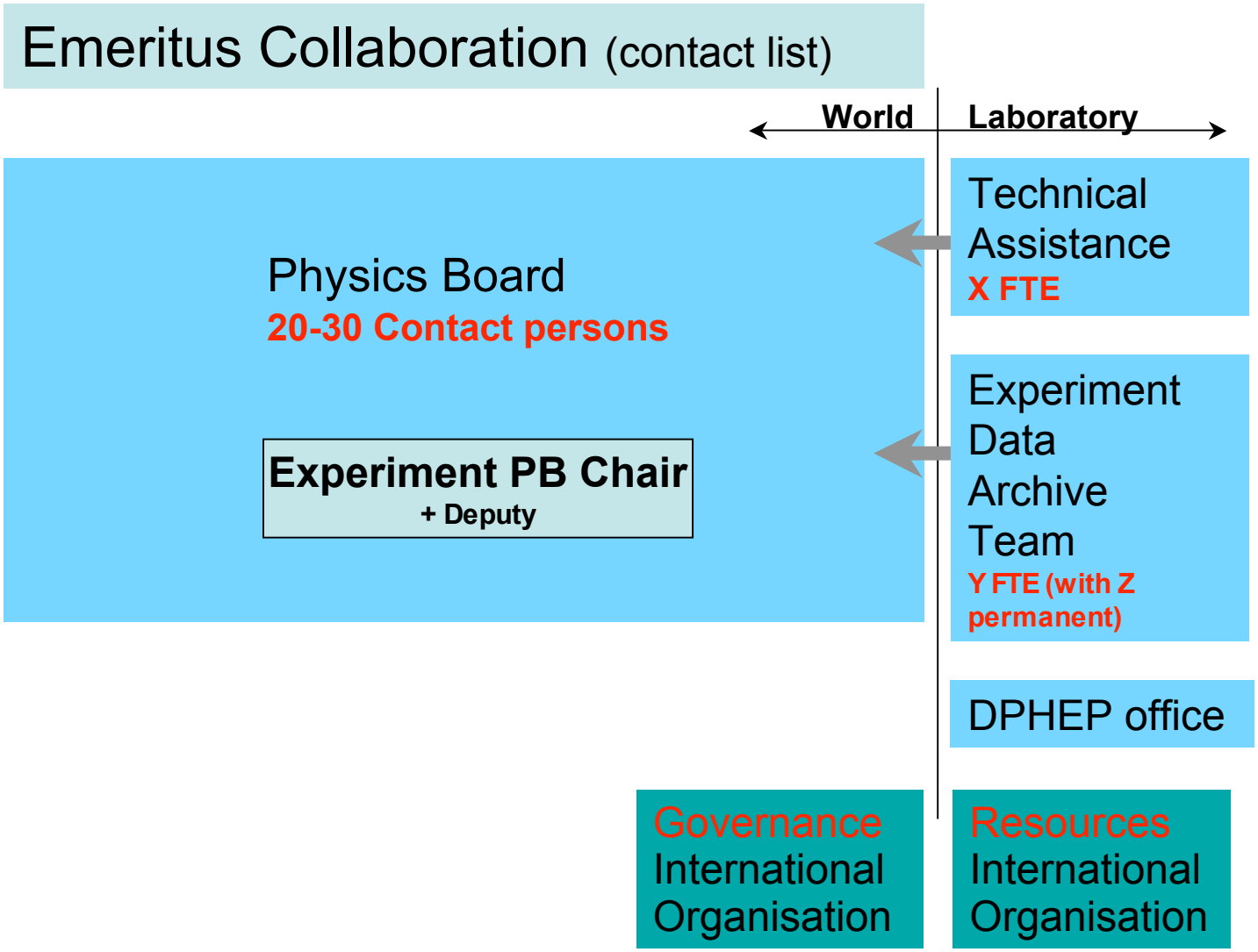
- Planning the transition to a long term analysis model
- R&D needed
  - Data migration models
- Data Archivist position
  - Long term custodianship
  - Similar to other fields
- Resources / experiment
  - Surge of 2-3 FTEs (2-3 yrs)
  - To be compared to

    300-500 FTEs (many years)
  - Archival position: 1 FTE per experiment/lab



Costs are less than 1% of the original investment

Example of Long Term Organisation of a HEP Experiment



Towards an International Organisation

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

Interaction with Other Fields

- Input is very valuable for HEP
  - Little experience in the field of data preservation and open large scale access
  - Connections with Data Archivists
  - General projects on digital preservation
  - Astrophysics

A Word from the Archivists

Jean Deken

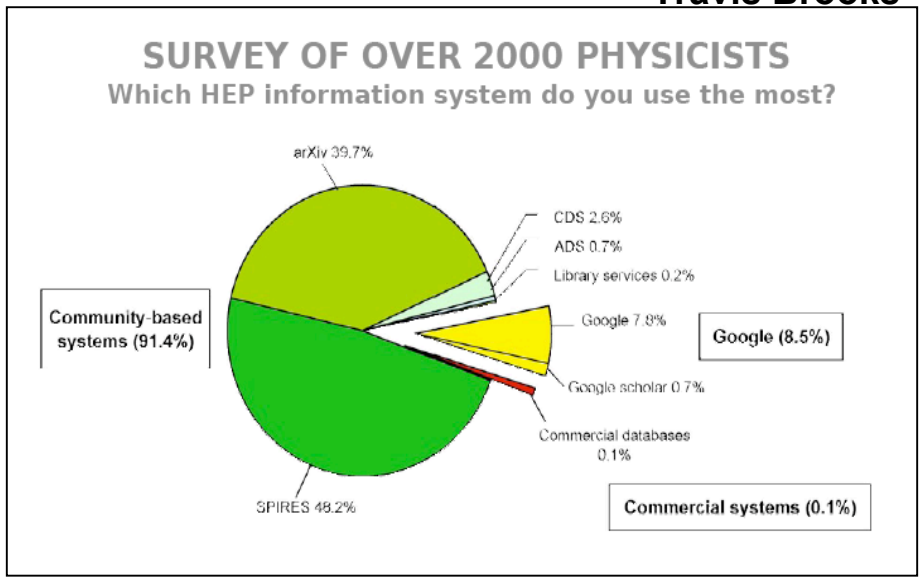
DOE Research and Development Records Schedule

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

Travis Brooks



- Input from the experts as to potential future schemes

SPIRES → INSPIRE

http://inspirebeta.net/

- Attractive opportunity from *INSPIRE* to host further documentation, meta-data, and even the data themselves associated with the publications

Other Fields

- 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 report and references)

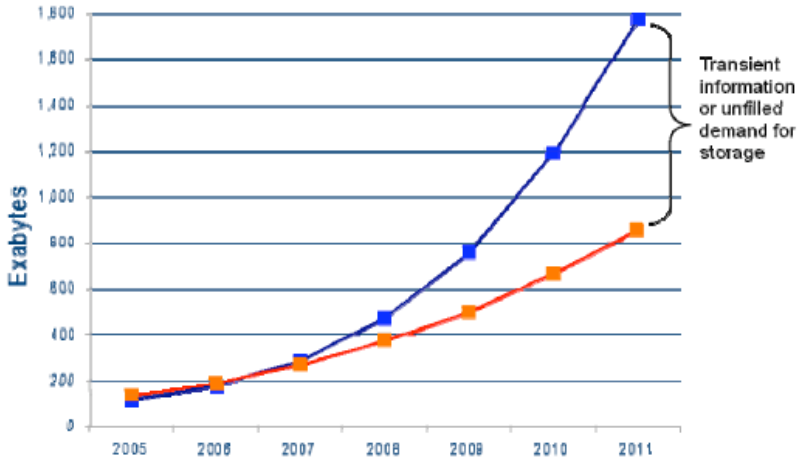


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

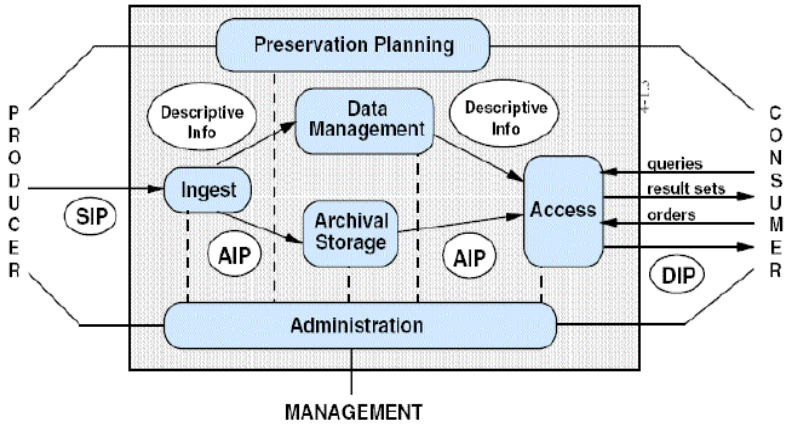
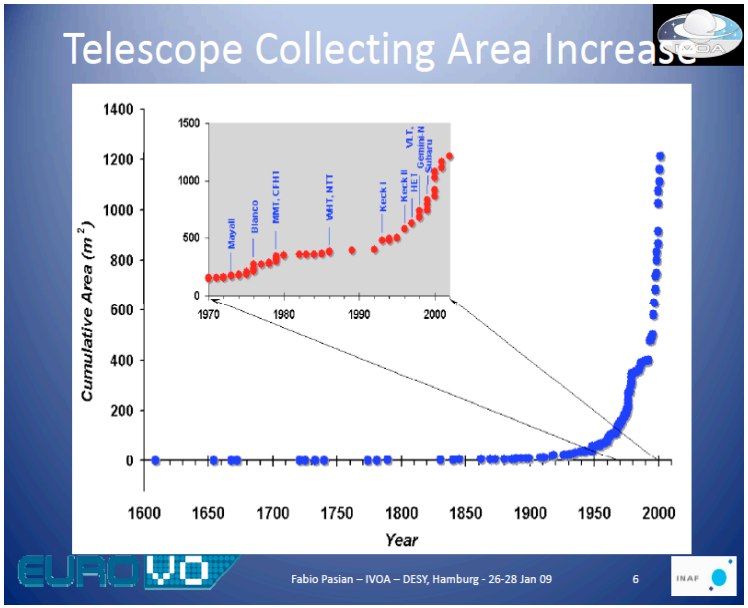
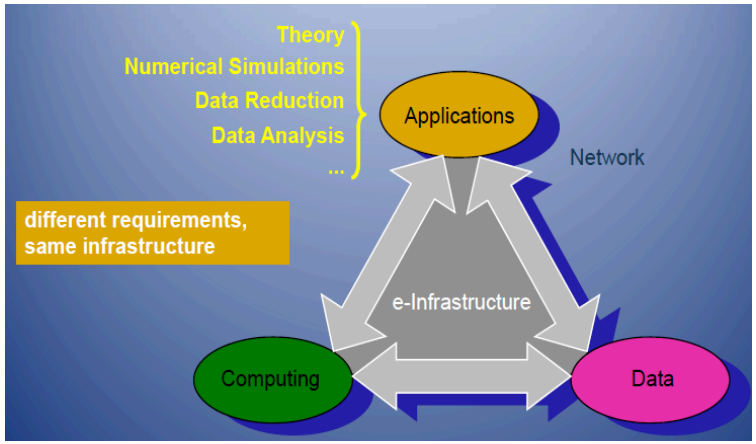


FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

- Scientific Data is a major component of the ongoing efforts (helps because of its complexity)
- Some scientific fields are well advanced, e.g. Astrophysics

Virtual Observatories in Astrophysics

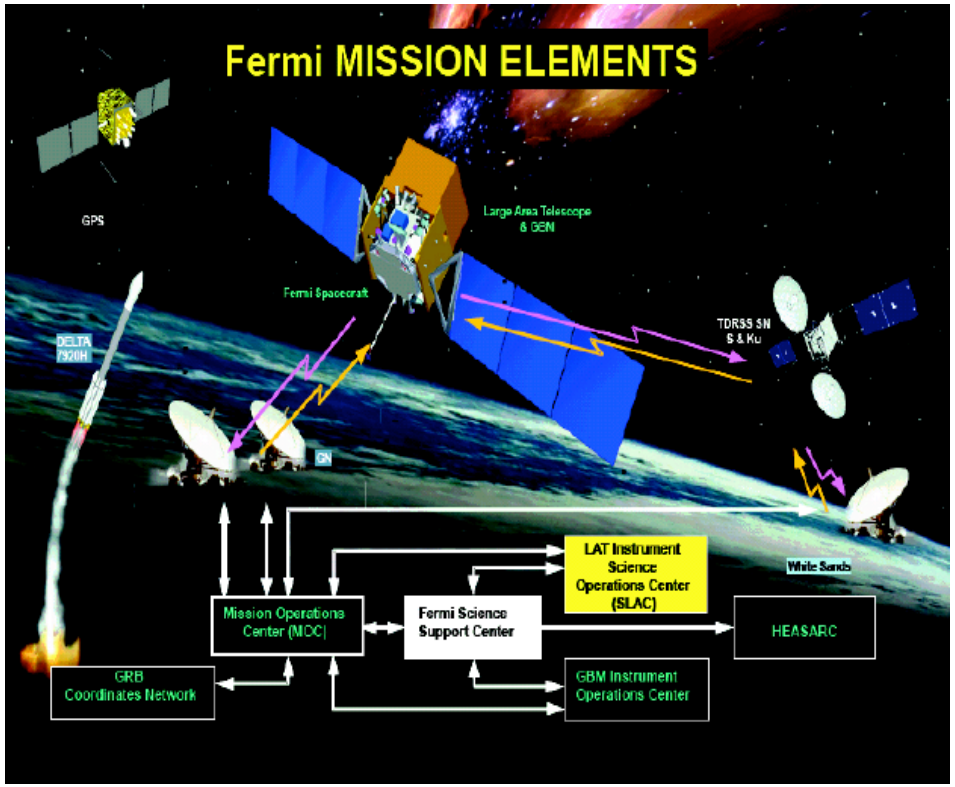


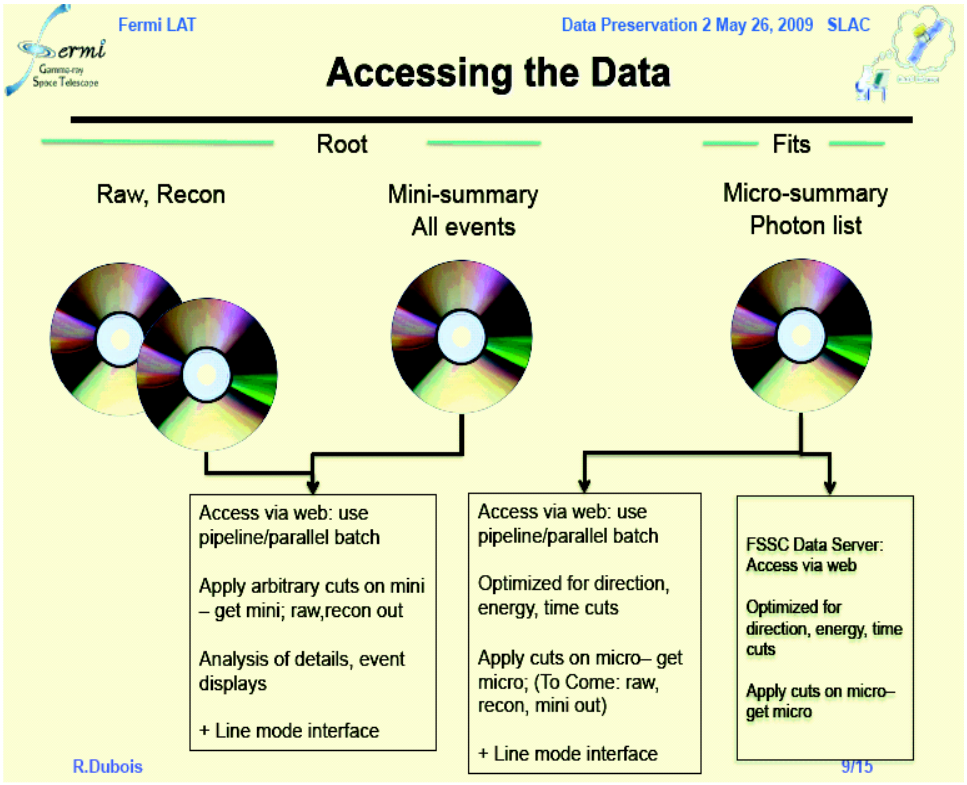


Fabio Pasian

- Data archives operable by many
- Work on standards and access to:
  - Data, simulation, mining techniques
- International, multi-experiment

Open Access in Astrophysics





R. Dubois

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

DPHEP 2009: Intermediate Recommendations

- Document presented to ICFA: 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 centres*
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*

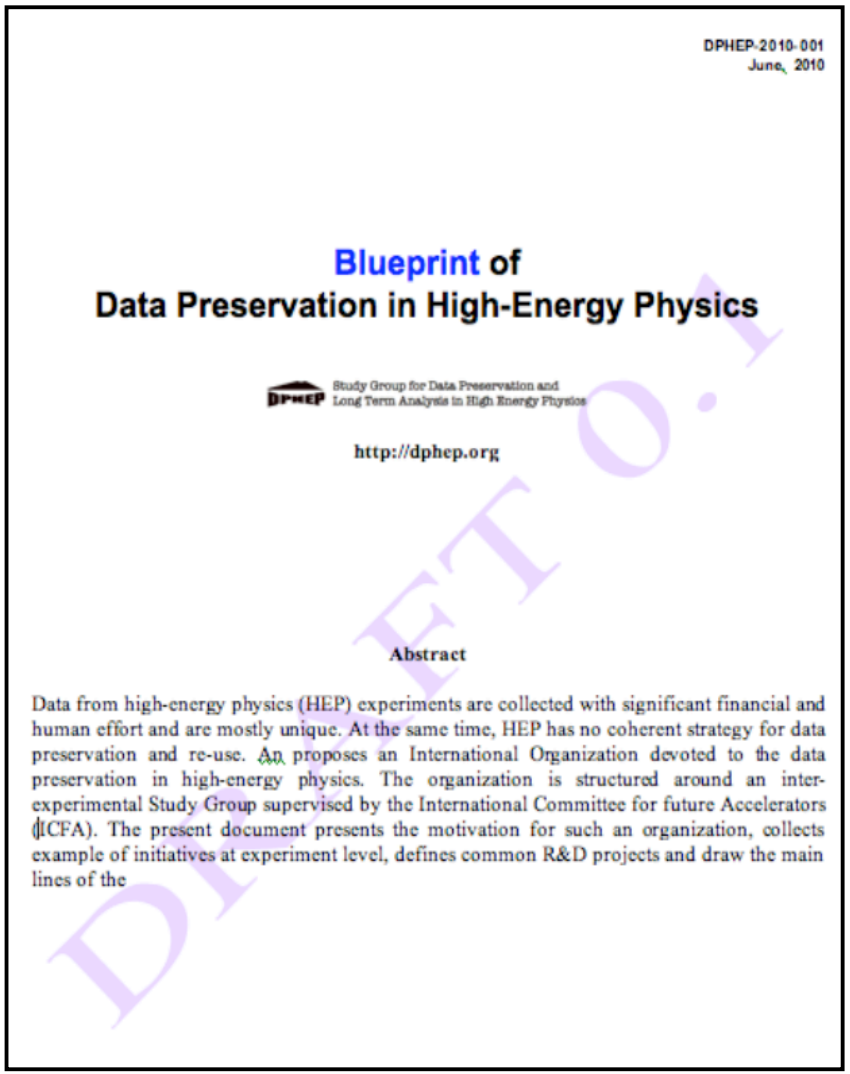
Feedback from the HEP Community

- Support from major labs expressed:
  - DESY, CERN, Fermilab, SLAC, IHEP
- ICFA: August 2009 (update presented to ICFA in Feb. 2010)
  - 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 Blueprint
  - 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**

Next DPHEP Document: Blueprint Plans



DPHEP-2010-001
June, 2010

Blueprint of
Data Preservation in High-Energy Physics

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

http://dphep.org

DRAFT 0.1

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 proposes an International Organization devoted to the data preservation in high-energy physics. The organization is structured around an inter-experimental Study Group supervised by the International Committee for future Accelerators (ICFA). The present document presents the motivation for such an organization, collects example of initiatives at experiment level, defines common R&D projects and draw the main lines of the

- Contents
  - Concrete R&D projects to enable data preservation with cost estimates
  - Experiment specific and across several initiatives
  - Skeleton for local, regional, lab, national and international proposals
  - Gathering expertise in preparatory phase
- Much of the text was written at the CERN workshop in January
- To be published in 2010

Conclusion and Outlook

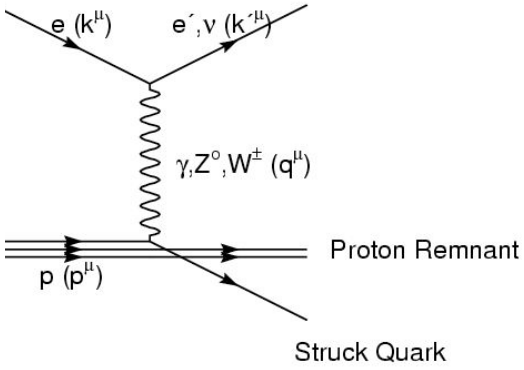
- Data preservation in HEP is important because:
  - Relevant physics cases for future use can be made
  - It is timely, given the current experimental situation and plans
  - It may enhance the return on the initial investment in the experimental facilities
  - It provides additional research at particularly low cost
- But it 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
  - Providing recommendations for future HEP experiments

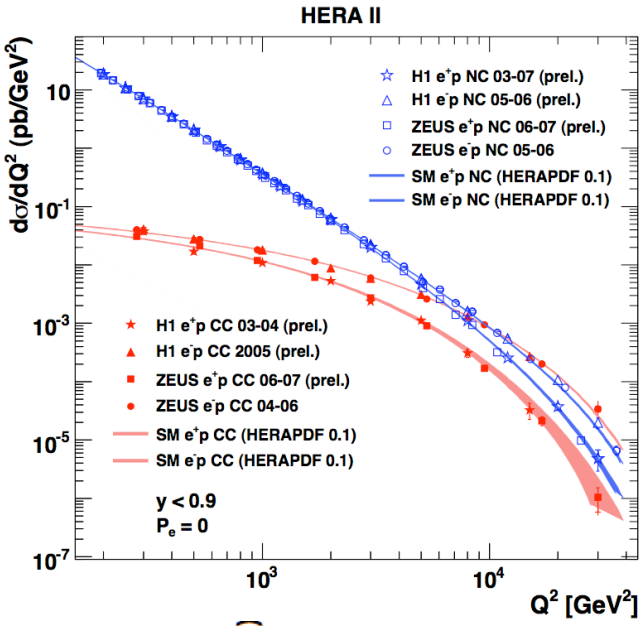
Extra Slides

An H1 Data Format for Outreach

- Producing H1 data in a format suitable for outreach purposes is an attractive proposal
  - To run in parallel with the main preservation effort
- The physics content of such a format is essentially defined by the outreach plans
  - What can the user learn by studying ep collision data?
- This then starts to define the variables, quantities and even the outreach projects themselves

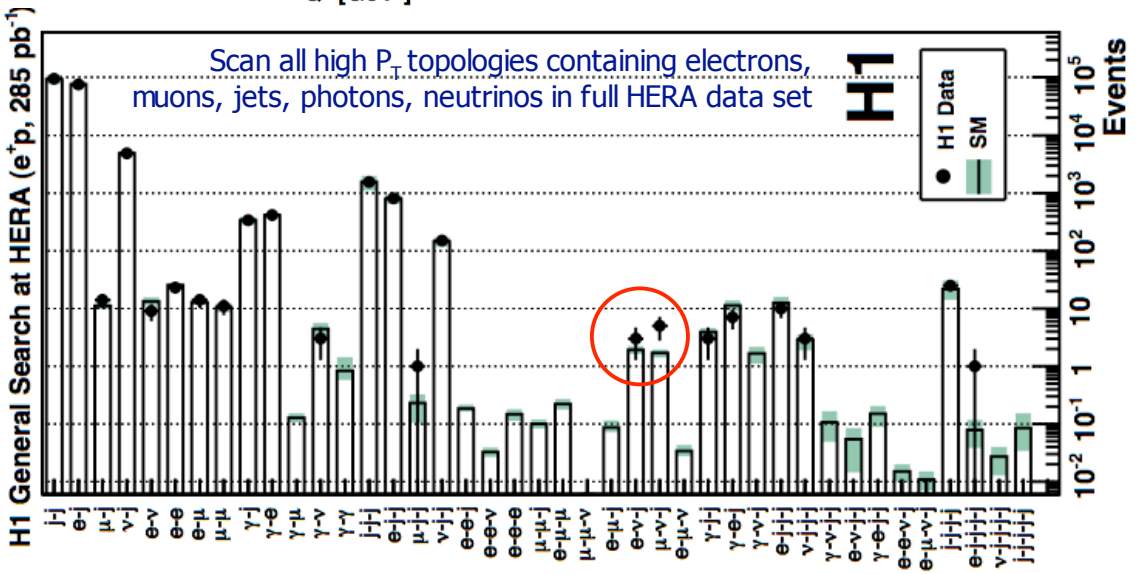
Outreach Potential of HERA Data



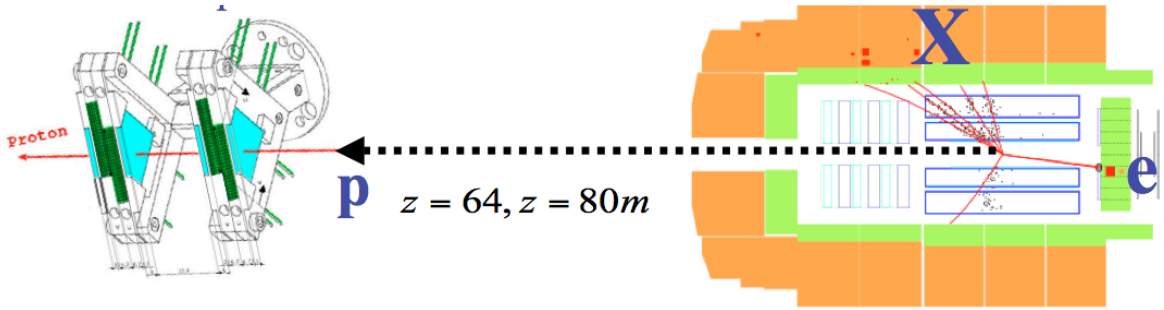


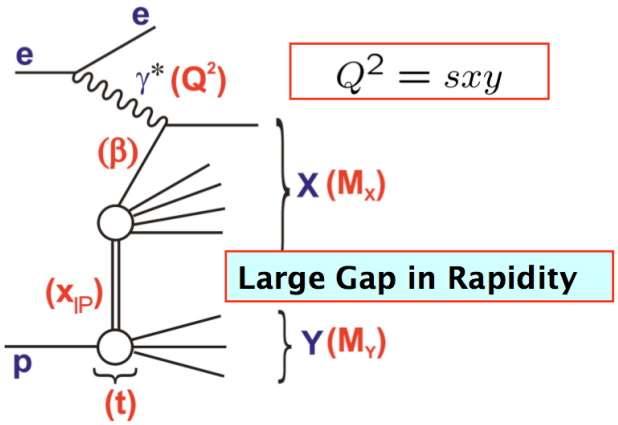
1) Basics of deep-inelastic ep scattering, and understanding the differences between NC and CC events, electroweak unification

2) Looking at high P_T event topologies, and in regions where the SM expectation is low: look for deviations in the data

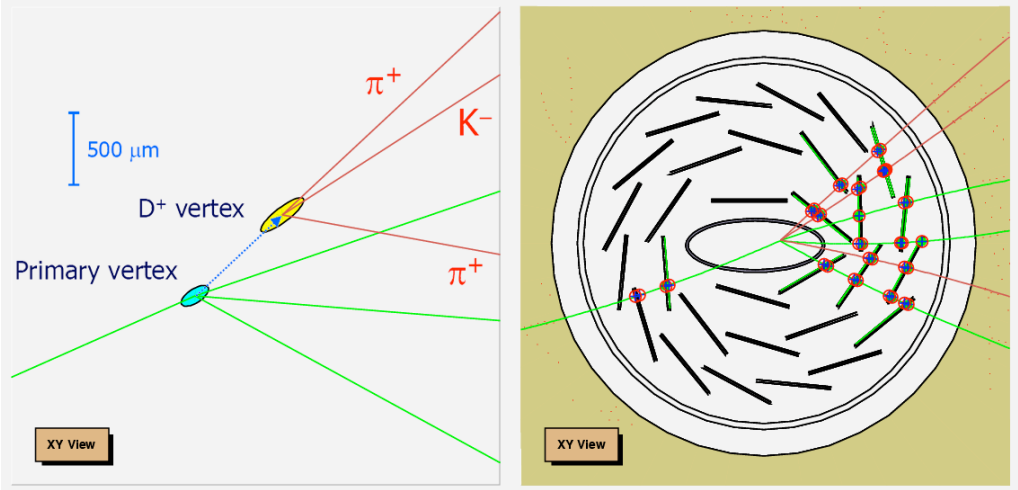


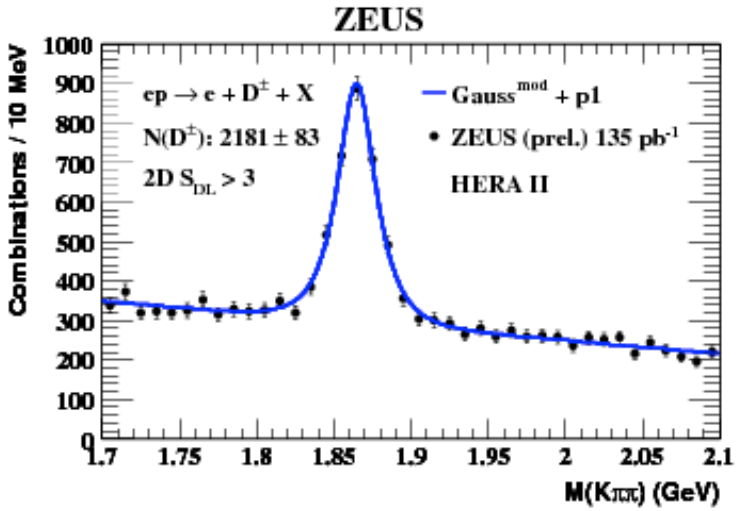
Outreach Potential of HERA Data





3) Diffractive events and their topologies





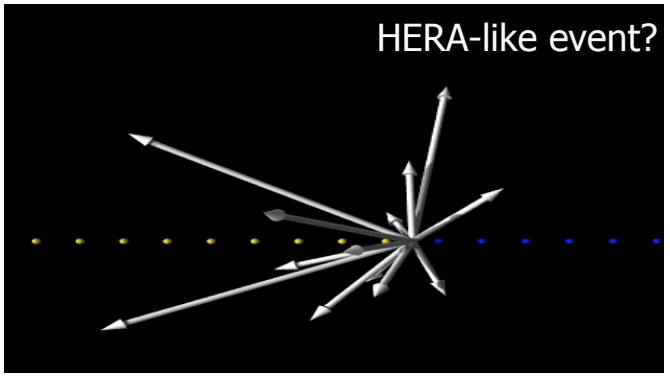
4) Fraction of total DIS cross section from heavy flavours (charm and beauty): particle spectroscopy, inclusive and maybe even lifetime methods (ambitious!)

Outreach Format: Technical Issues

- An outreach format seems reachable from the current software, and would come somewhere in content between the existing HAT and μODS formats

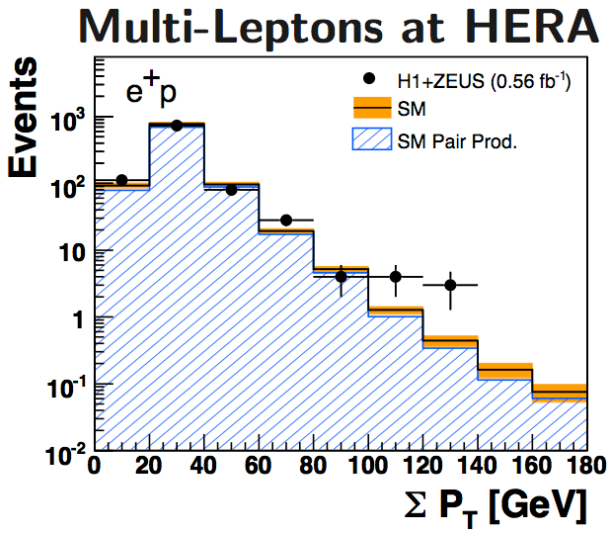
- What about the actual data format?
  - Should consist of simple data types: floats, ints, and arrays..
  - Independent of H1OO, but based on ROOT types (TClonesArrays etc)
  - A single format to cover all outreach projects would be preferable
  - If one wants to include comparison to MC, a universal event weighting scheme which takes into account all efficiencies from triggers, vertex finding and so on, may be prohibitively complex
  - If we only deal with data, then the situation is much simpler

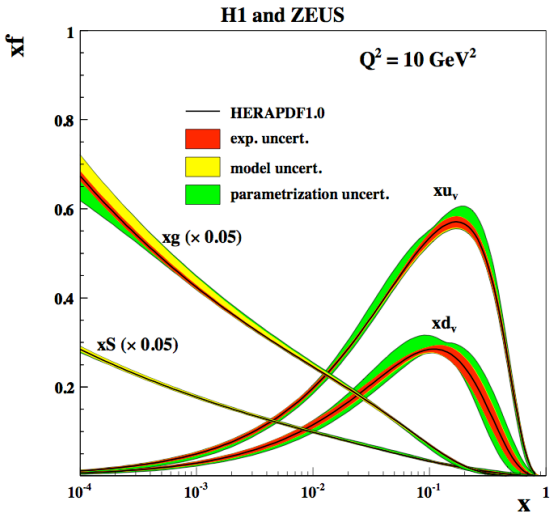
- Would be nice to have something that can interface to Matt Bellis' work in terms of user applications
  - Will certainly be followed up

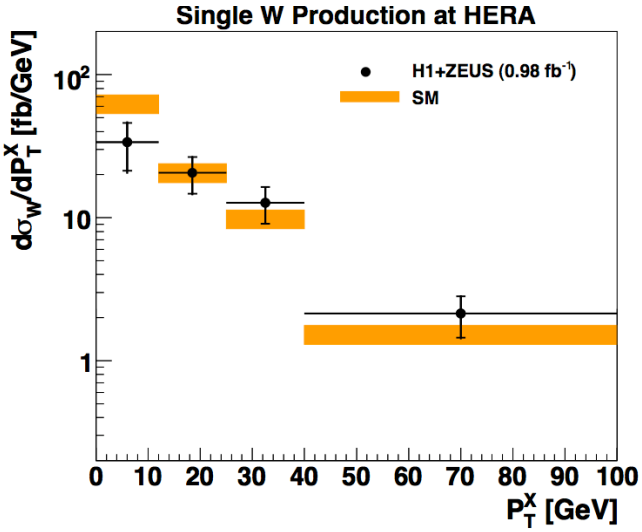


Outreach Format: HERA Format?

- Such a format would be a candidate for combining $e^{\pm}p$ data from the H1 and ZEUS (and even HERMES?) experiments
- 2009 saw the first combined H1+ZEUS publications:







- Some ideas came out of the first HERA data preservation meeting
  - Different strategies in some areas: learn from shared experiences
  - Joint HERA financial proposal would give better chance of support?

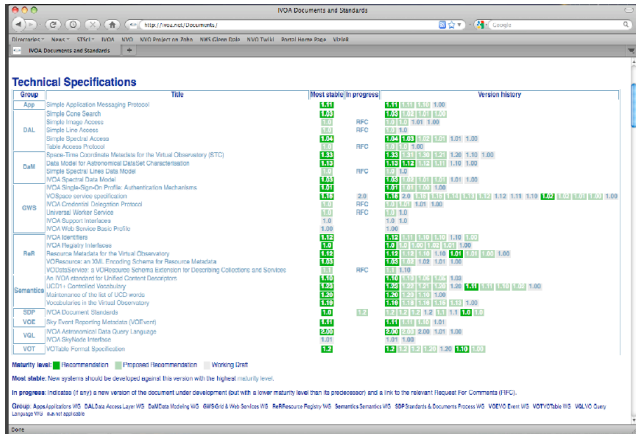
International Virtual Observatory for Astrophysics

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

Virtual Observatory 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)

Blueprint Content

Chapter 1: Executive Summary and General remarks

Chapter 2: The Scientific Potential of the Data Preservation in High Energy Physics

Chapter 3: Experiments Data Preservation projects
A: project/hardware/resources
B: governance, international scene

Chapter 4: Inter-experiment R&D survey

Chapter 5: DPHEP