Data Preservation and Long Term Analysis in HEP

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

http://www.dphep.org

The HEP landscape (colliders)

- HERA: end of collisions in 2007
  - No follow-up before at least two decades
- 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

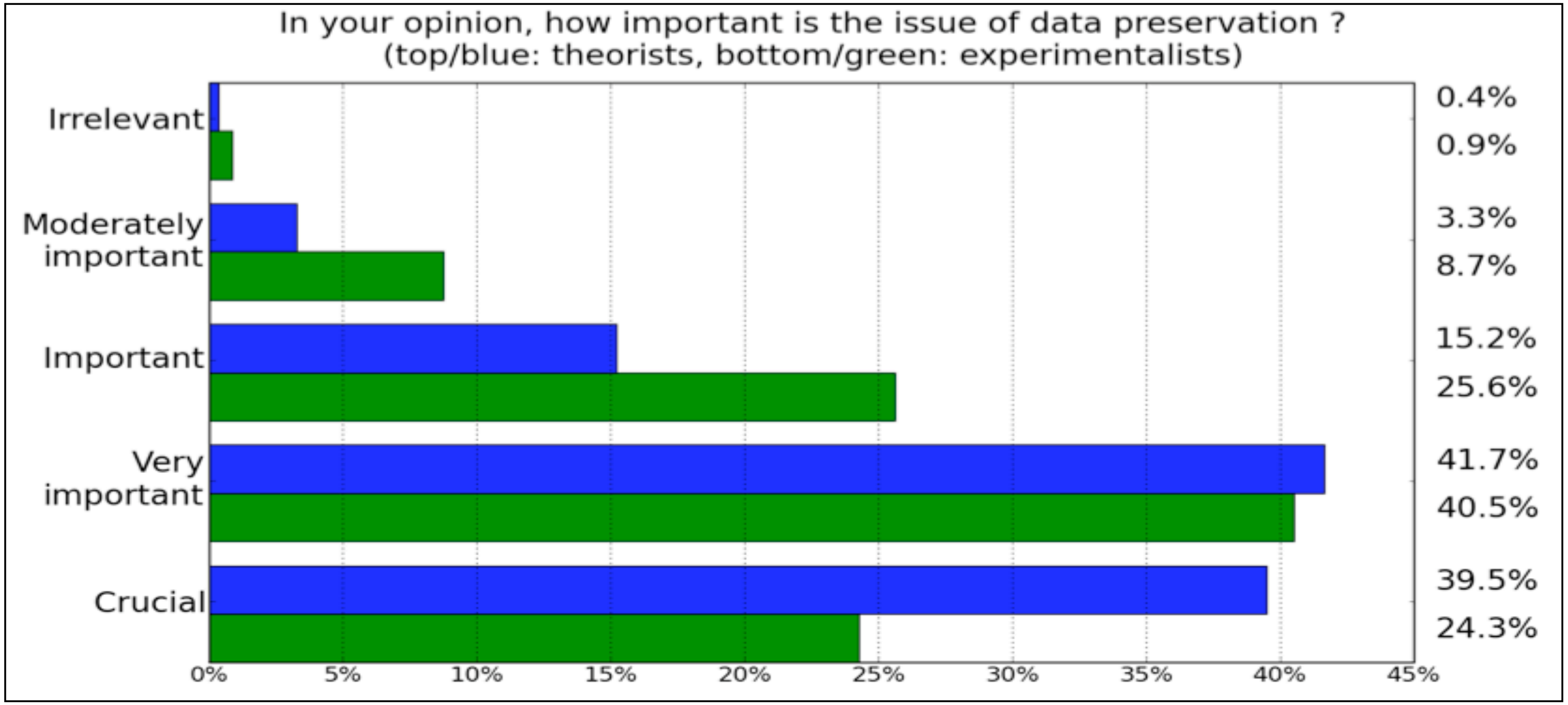
HEP experiments data taking encompass 10-15 years, some are unique

What is the fate of the collected data?

(NB: here “data” = full experimental information)

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

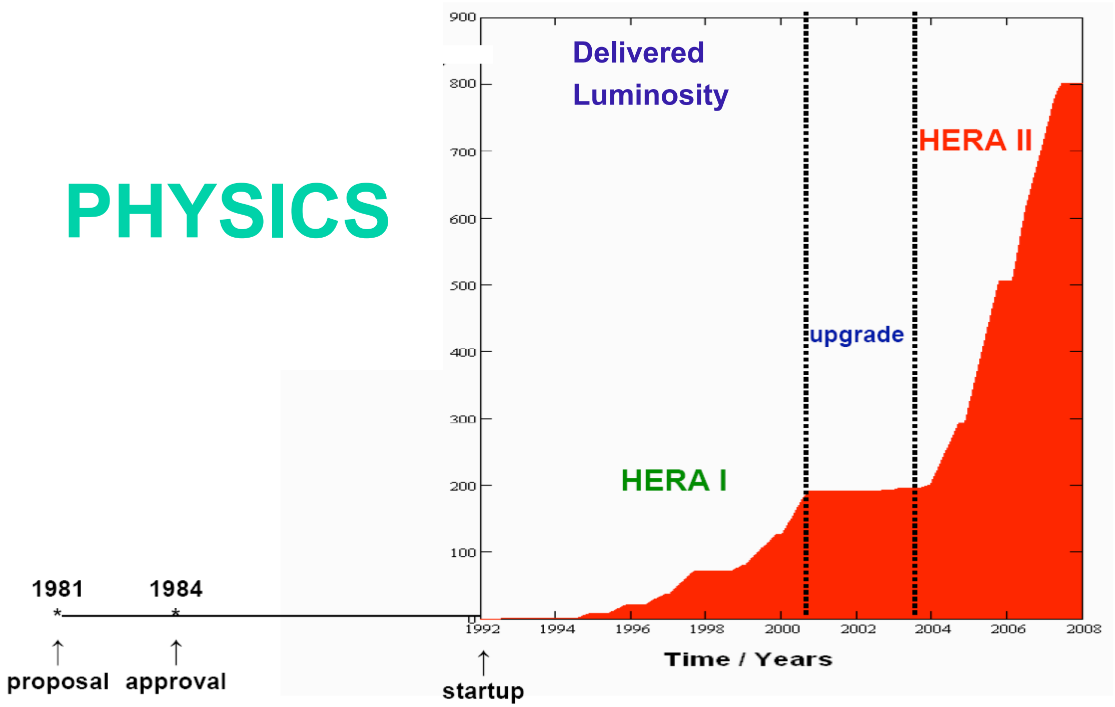


70%: very important or crucial
However, no coherent strategy exists: in general, HEP data is lost

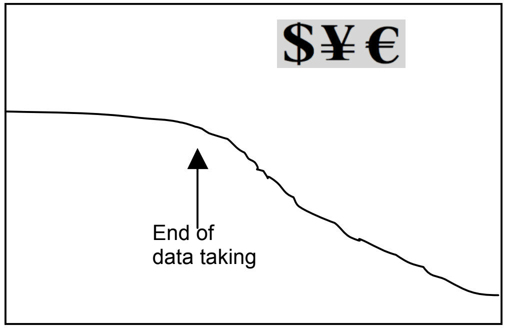
Why is difficult to preserve HEP data?

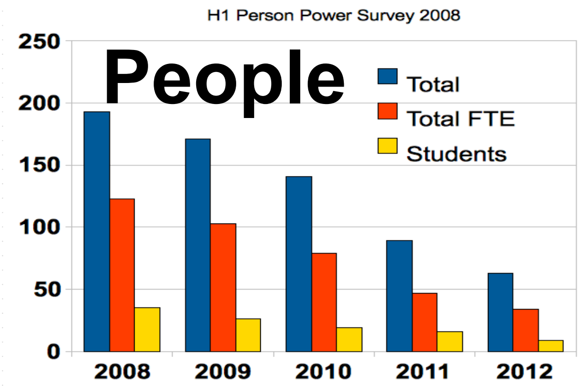
- Good physics is collected at the end, but:
- The resources decrease after the end of data taking
  - Dedicated resources need to be planned

PHYSICS



Funding





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)

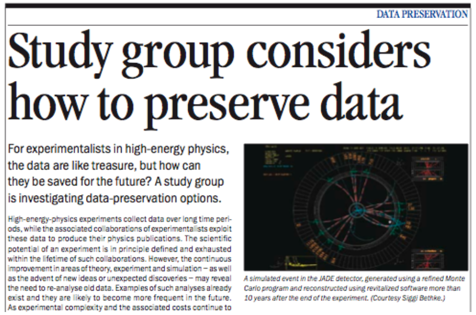
International Advisory Committee

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
- Two workshops in 2009: DESY and SLAC
  - 40-50 participants, 40 talks, proceedings
  - Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields (astrophysics)
- Objectives 2009:
  - Draft report for ICFA
  - Make the report available for debate in the HEP community



DATA PRESERVATION

Study group considers how to preserve data

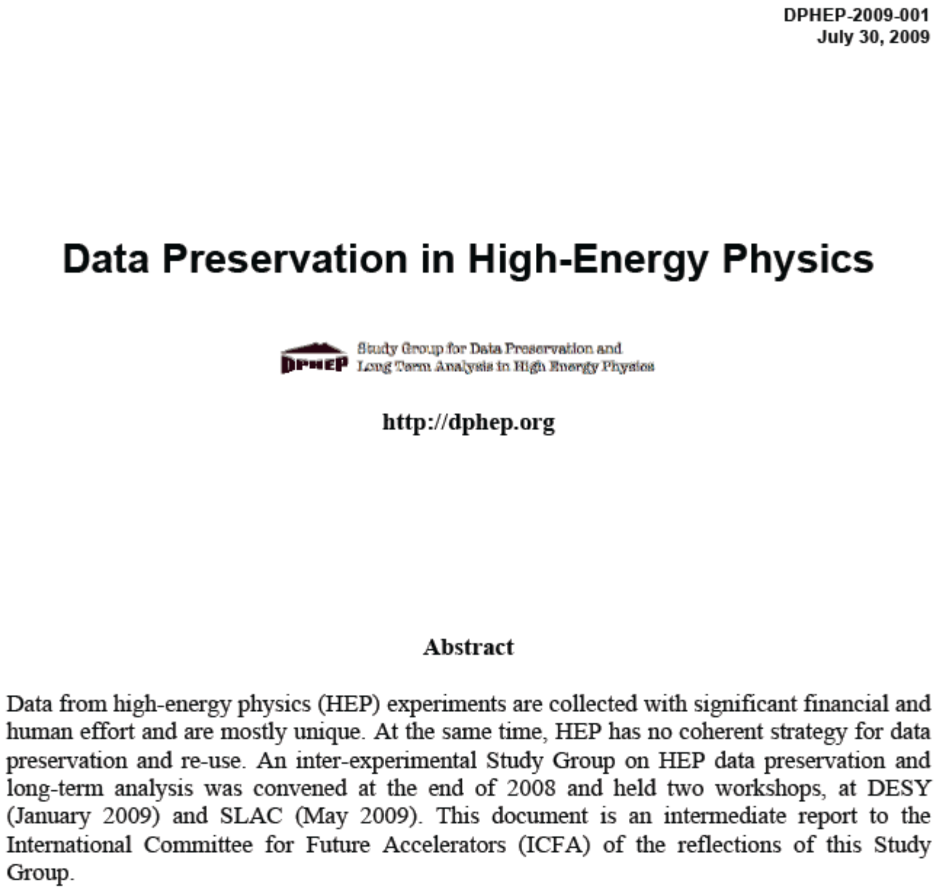
For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

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 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

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.)

CERN Courier, May 2009

Preliminary document submitted to ICFA



DPHEP-2009-001
July 30, 2009

Data Preservation in High-Energy Physics

DPHEP 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.

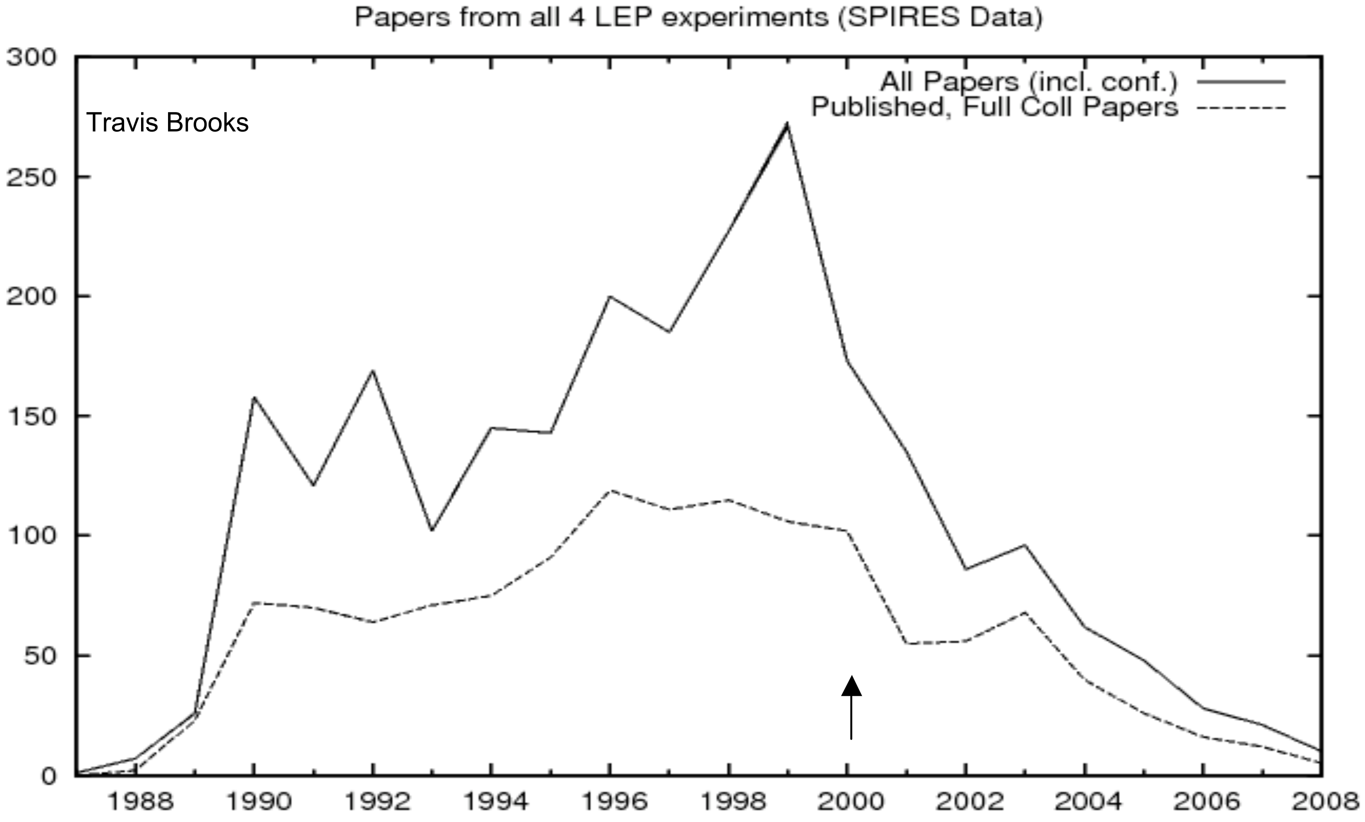
In this this talk: present the main ideas, preliminary recommendations, plans

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

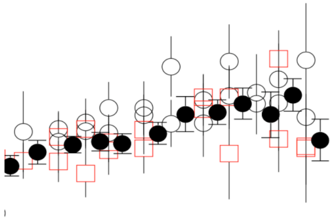
Physics Case I

- Long term completion and extension of the physics program



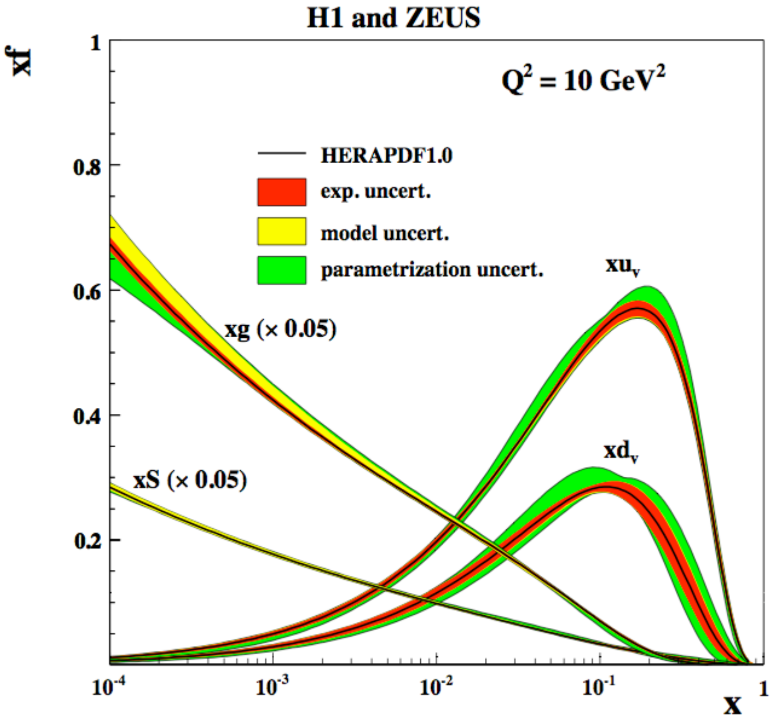
Physics subjects are published after the end of collisions/collaborations
5-10% of the papers are finalized in the “archival mode”

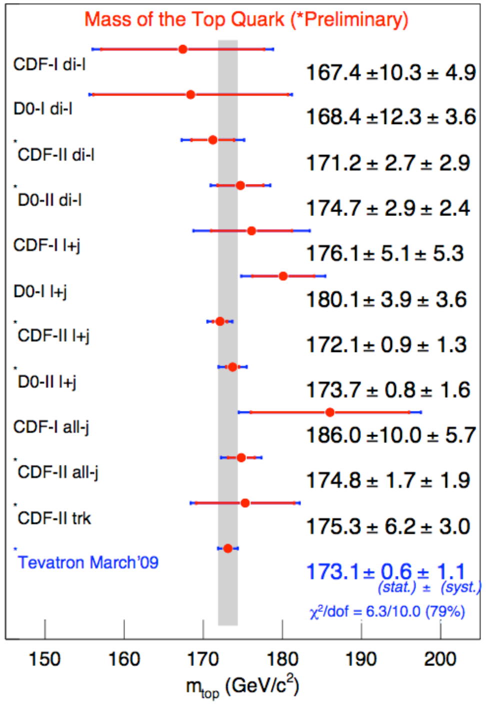
Physics Case II



- Cross collaborations

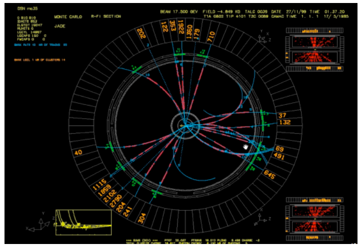
Already exist at LEP, Tevatron, HERA, Babar+Belle (in progress)





Preserved data would make possible more combined analyses across experiments

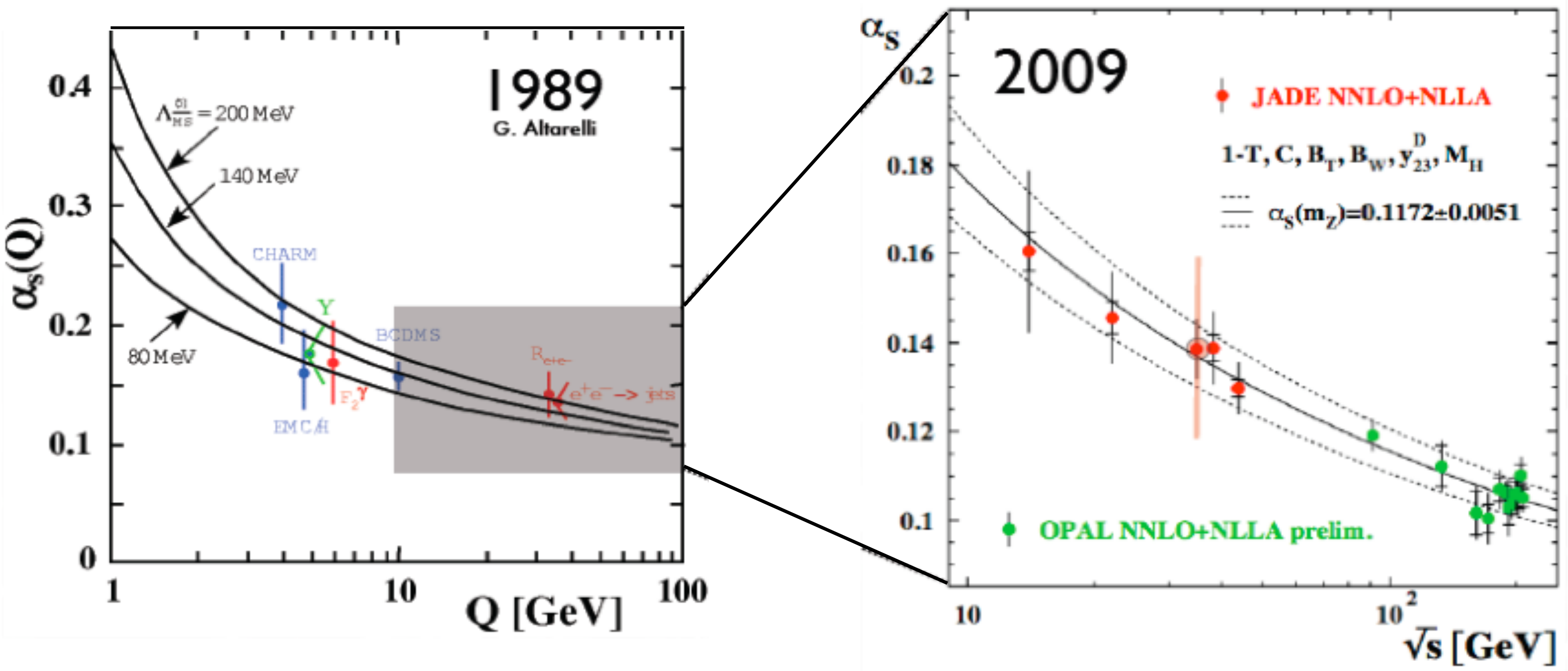
Physics Case III



JADE: raw data preservation,
software revitalisation
individual initiative

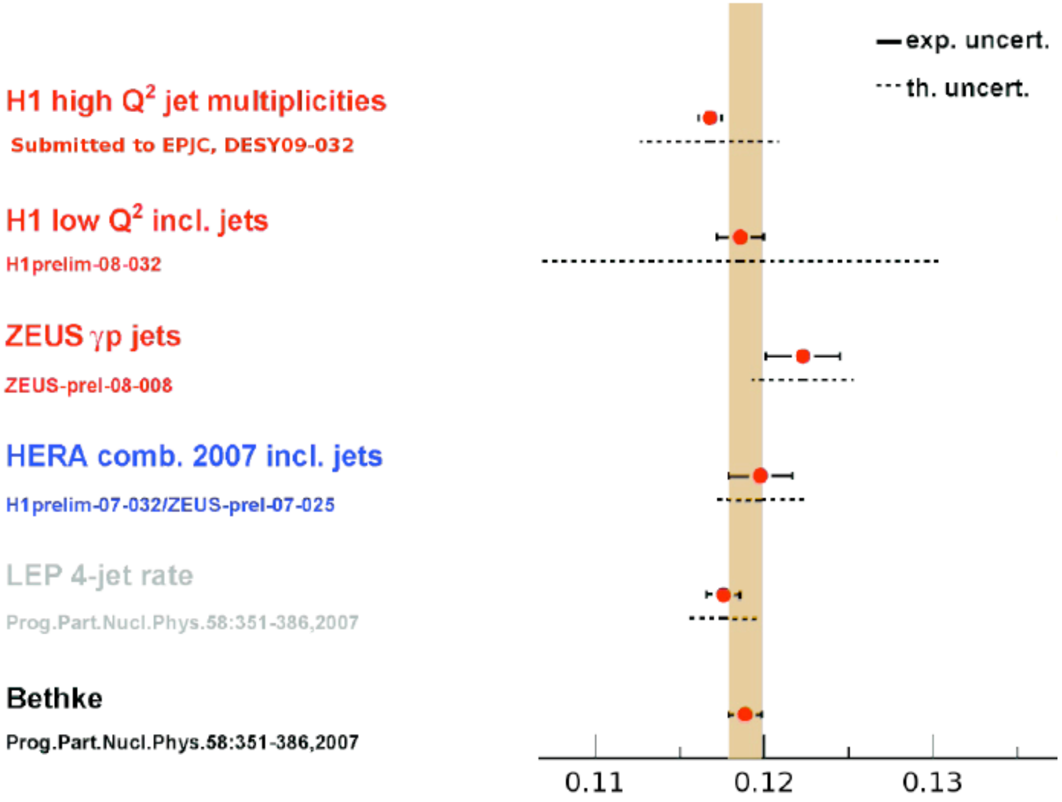
10 publications

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

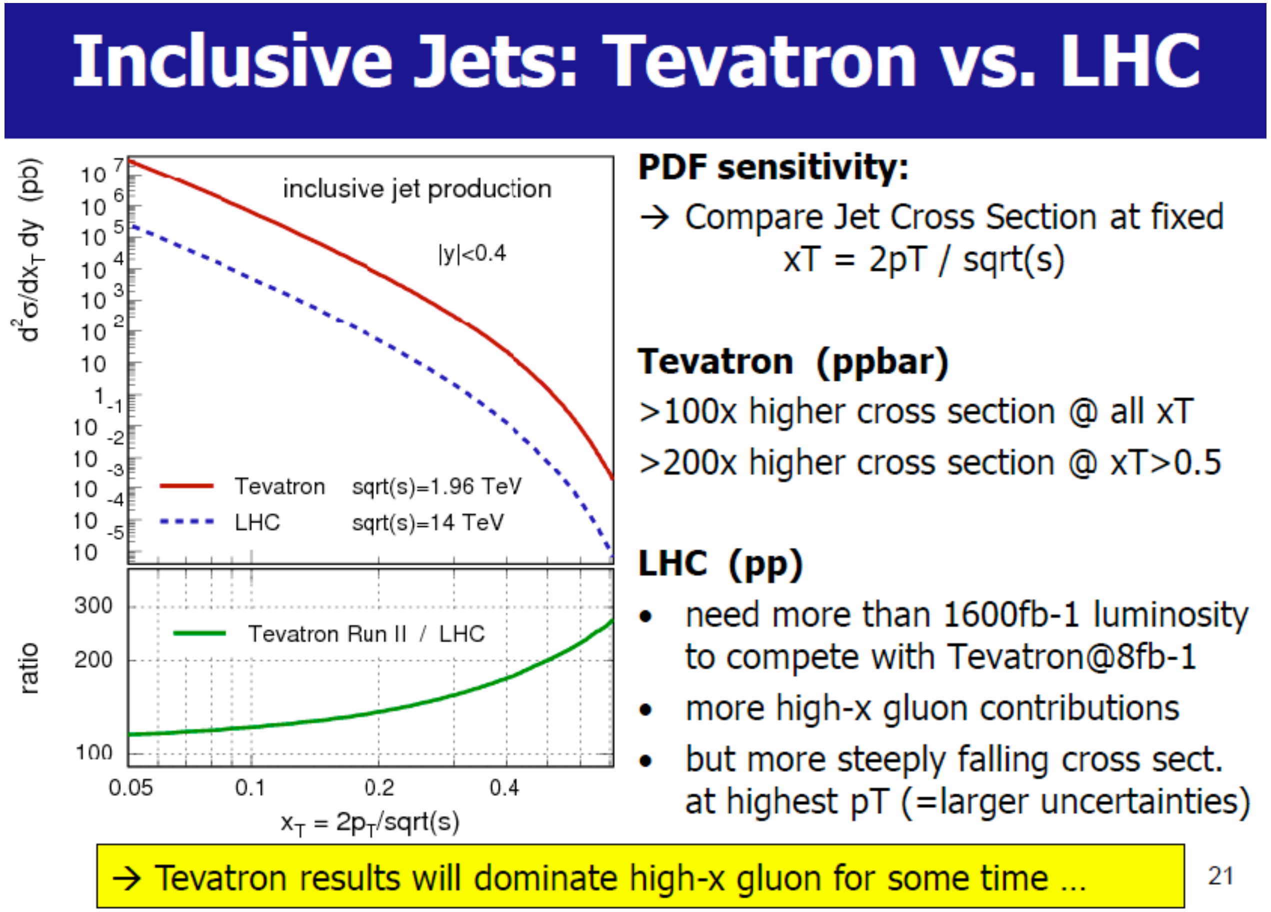


The history may well repeat itself….

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



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)

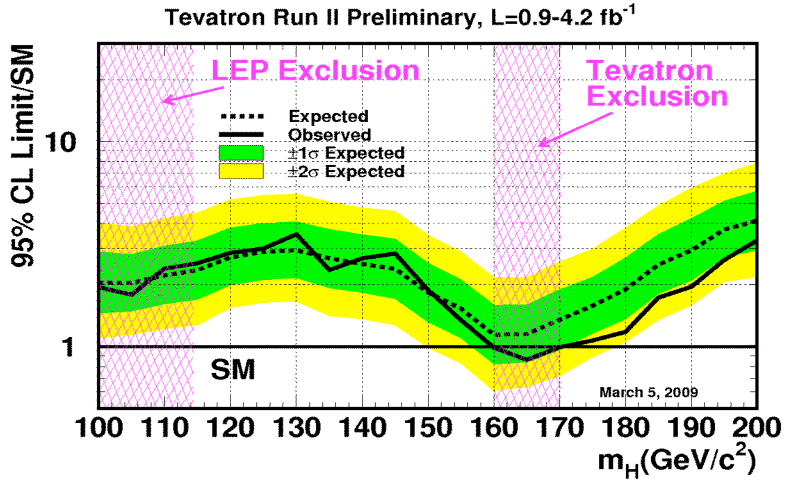
- need more than 1600fb-1 luminosity to compete with Tevatron@8fb-1
- more high-x gluon contributions
- but more steeply falling cross sect. at highest pT (=larger uncertainties)

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

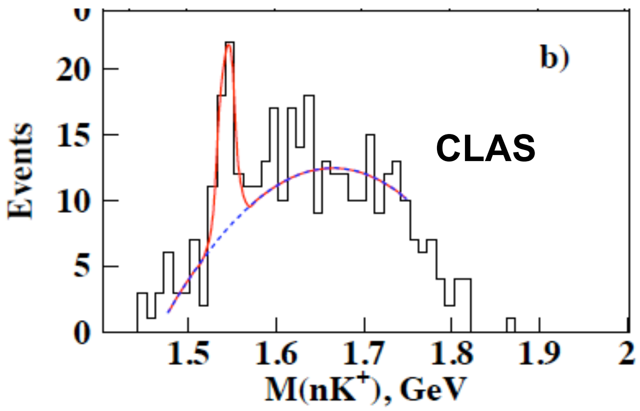
M. Wobisch

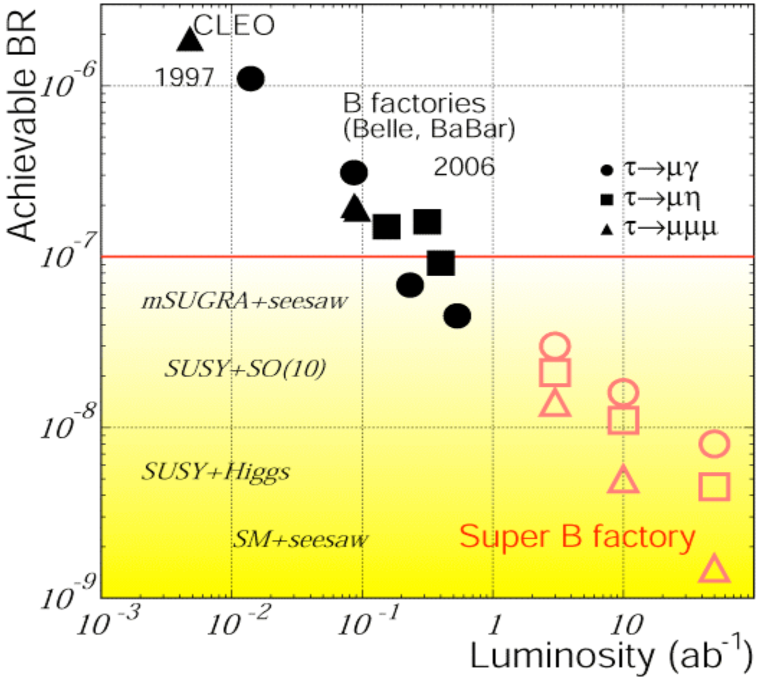
More examples: contingency with future programs

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



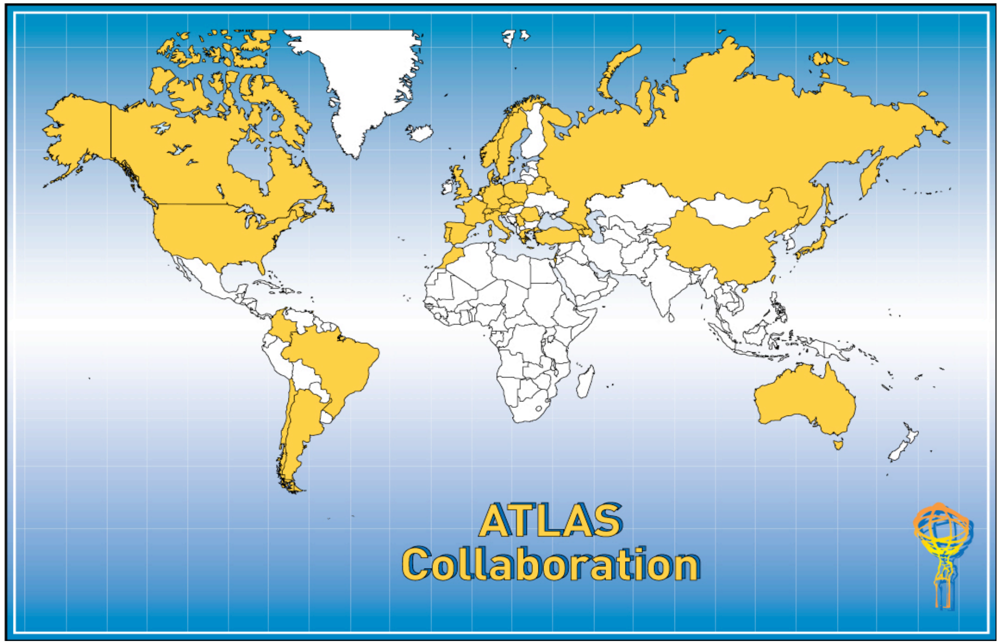
...surprises can occur
at lower energies too

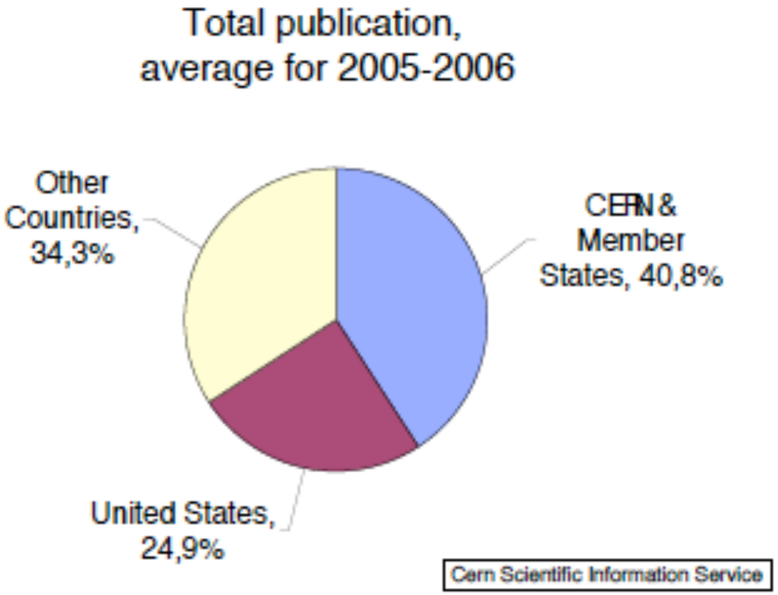




Physics Case IV

- Scientific training, education, outreach





Improve the overall high level education in HEP

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

What is “HEP data”?

- Digital information: event files, database
- Software: simulation, reconstruction, analysis, user
- Documentation: publications, notes, manuals
- “Meta” information: news, messages
- Expertise (people)

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

Each level implies an R&D project at experiment level

Technological issues

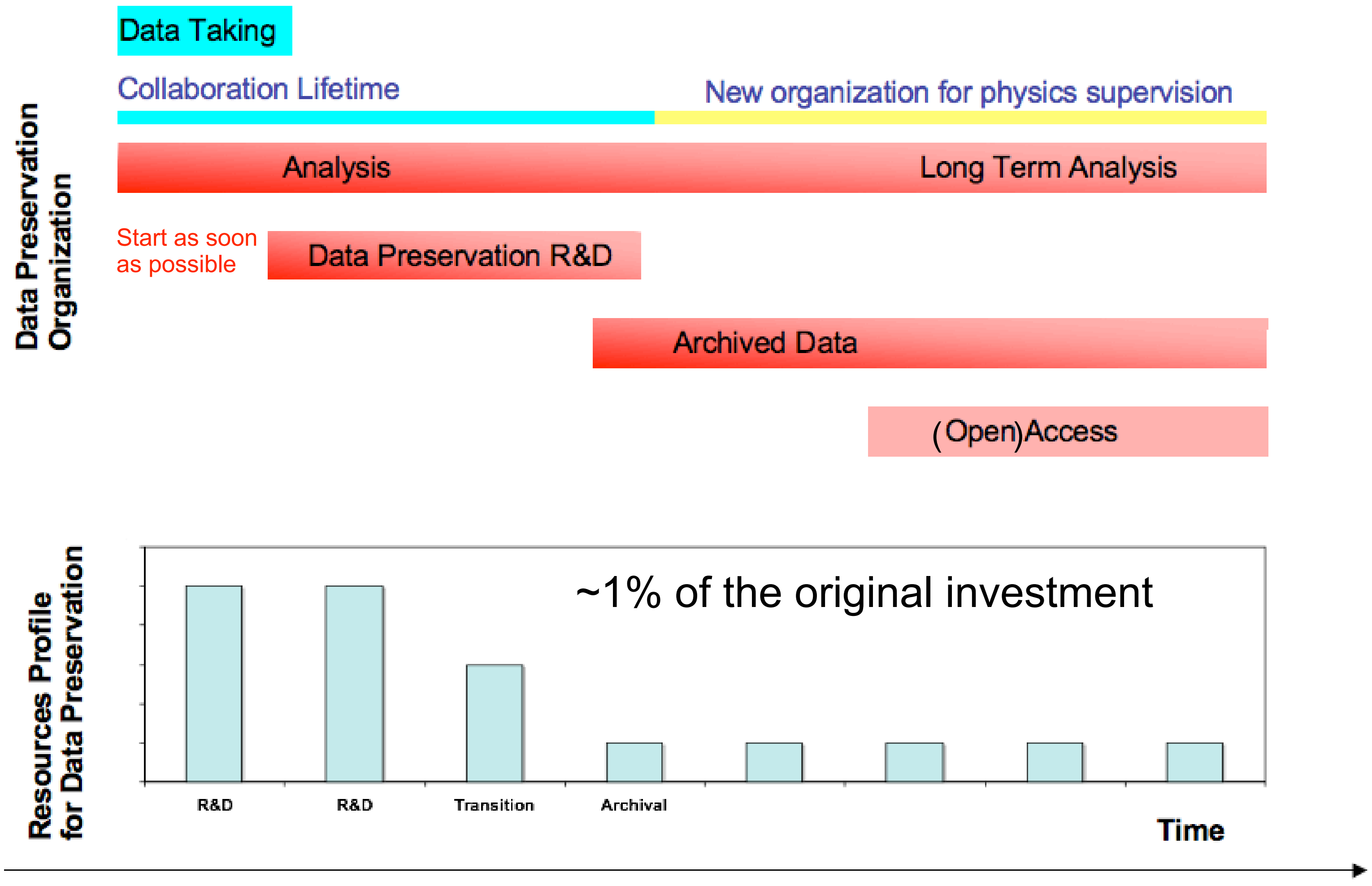
- Computing centers are (in principle) able to store the data
  - 0.5 to 10 Pb /exp.
  - Total cost of data storage double current costs: 1 + 1/2 + 1/4 + 1/8 .. = 2
- Technological evolution and data migration
  - Software maintenance is the real issue
  - Preservation, emulation, migration
  - New possibilities: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

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)



Data Preservation Organization

Data Taking

Collaboration Lifetime

New organization for physics supervision

Analysis

Long Term Analysis

Start as soon as possible

Data Preservation R&D

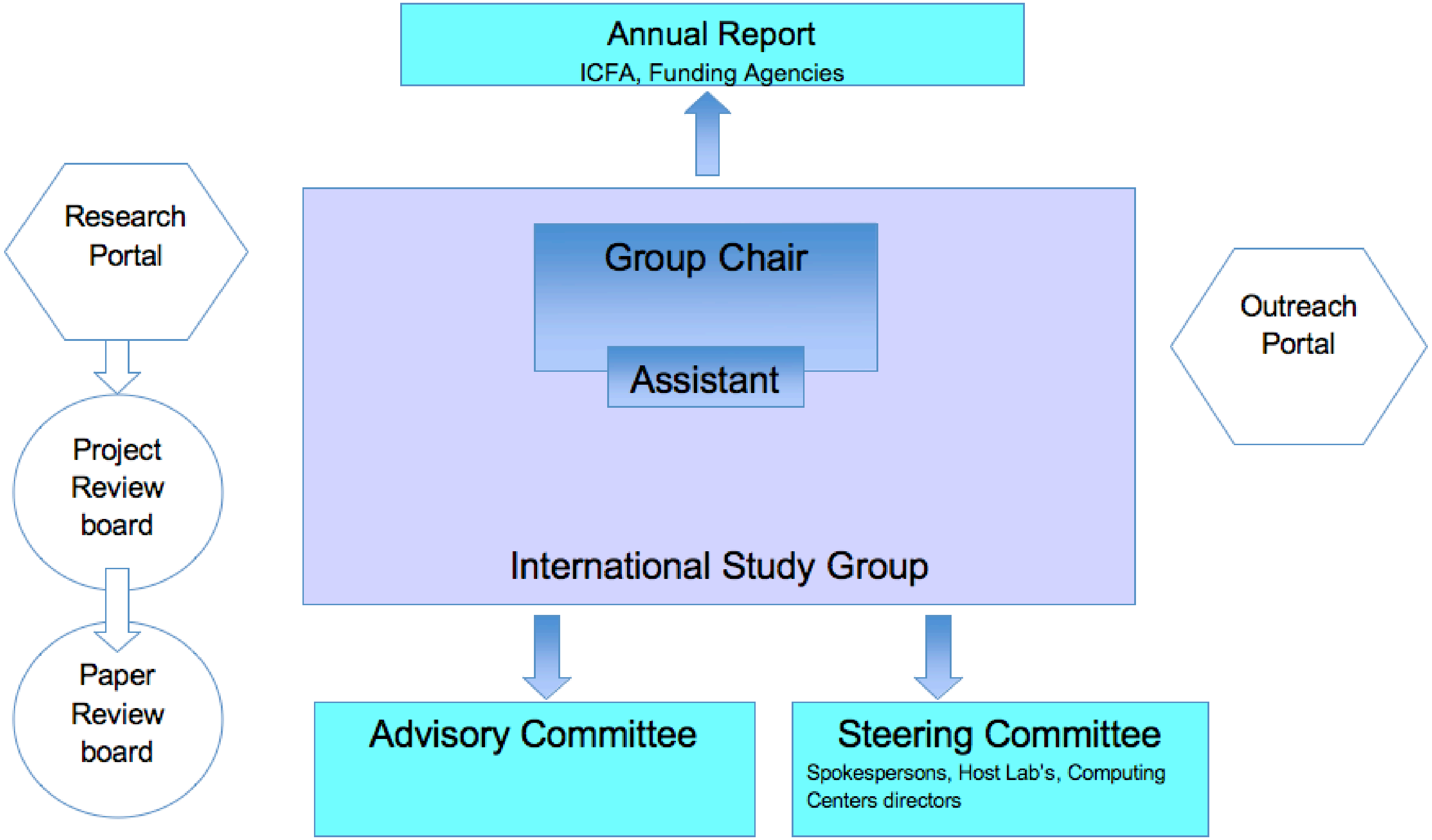
Archived Data

(Open)Access

Resources Profile for Data Preservation

~1% of the original investment

Towards an International Organization



Preliminary Recommandations

- 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. Data preservation beyond the end-date of experiments opens up future scientific opportunities. Given the present status of experimental programs at most facilities, an urgent and vigorous action is needed to ensure data preservation in HEP.
2. Different levels of data preservation and usability are possible. The preservation of the full analysis capability of experiments is recommended, including the preservation of reconstruction and simulation software. A dedicated project in each experiment is needed to assess the corresponding technological requirements.
3. The technological aspects of data preservation are well within the reach of large computing centres in HEP. Nevertheless, an interface to the experiment know-how should be introduced. The most efficient solution would be the creation of a data archivist position, in charge with the preservation of the data analysis capabilities.
4. The preservation of HEP data requires a synergic action of all stakeholders: experimental collaborations, laboratories and funding agencies. A clear and internationally coherent policy should be defined and implemented.
5. An International Data Preservation Forum is proposed as a reference organisation, with the mandate to organise and overview HEP data preservation initiatives; to discuss and propose solutions to technological or policy issues; to evolve into a clearing house for policies for access and re-use of preserved data. The Forum should represent experimental collaborations, laboratories and computing centres.

Feedback from the Advisors

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

- Very positive feed back to the initiative
- Document more examples of physics case
- Be more quantitative on preservation models
- Clarify the physics supervision and the relation with the open access philosophy
- Encourage full preservation model (level 4) for full physics capabilities and publications
- Associate time scales with the preservation models
- Explore models used in astrophysics
- Strong support to follow a global approach

ICFA decisions

- 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)

Milestones for 2009/2010

- Document made public by the end of the year
  - Including advisory committee and ICFA recommendations
- Two workshops
  - 7-9 December 2009 (CERN)
    - Review proposals for preservation models, more quantitative estimations
    - Steps towards global organization
    - Include public (HEP) discussion chaired by CERN DG.
  - Mid 2010
    - Prepare blueprint for concrete proposals, including costs estimates
- Prepare for Data Preservation funding programs (EU/DOE/NSF)

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**