Data Preservation in High Energy Physics.

Cristinel Diaconu
CPPM/DESY

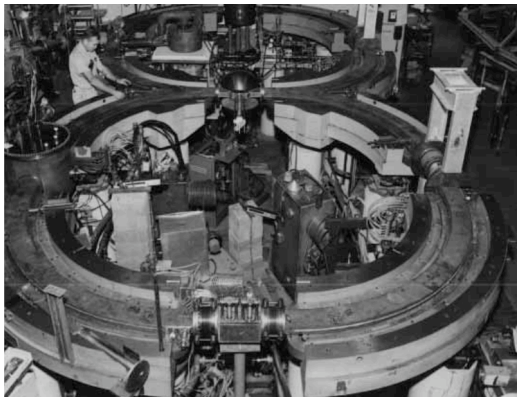
Study Group for Data Preservation and Long Term Analysis in High Energy Physics

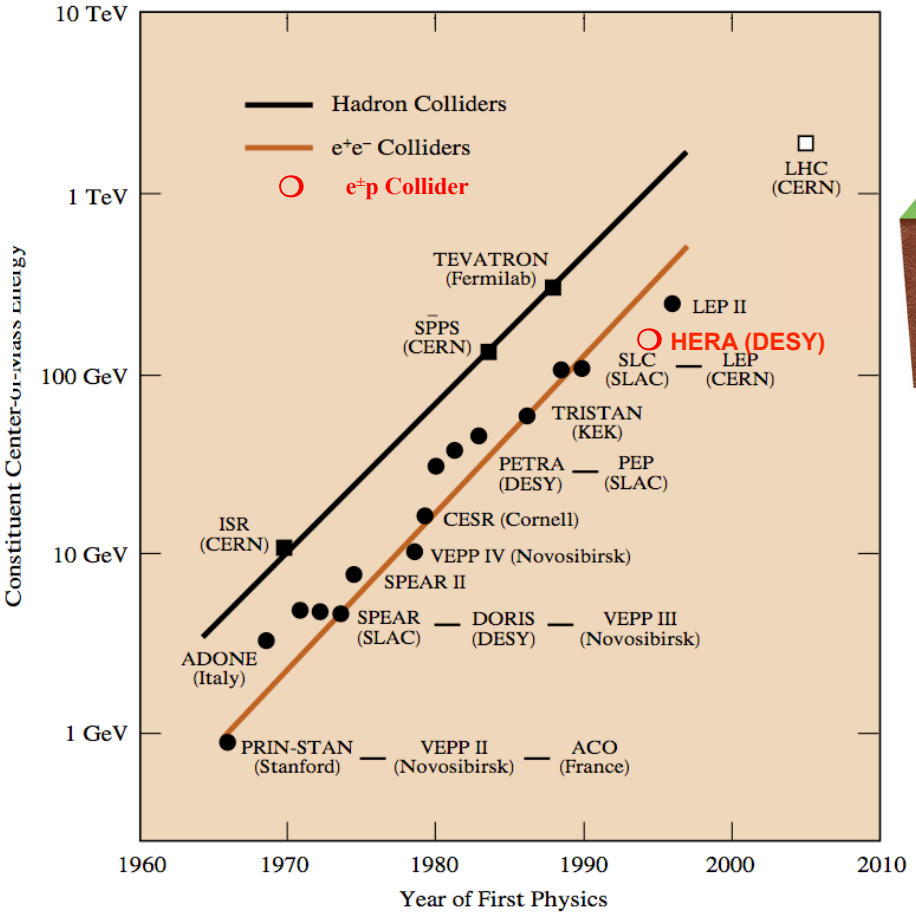
http://www.dphep.org

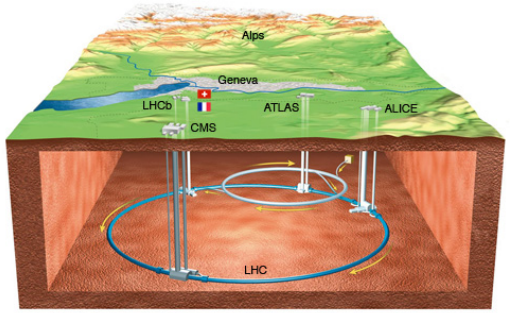
The Last 50 Years of High Energy Physics

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)



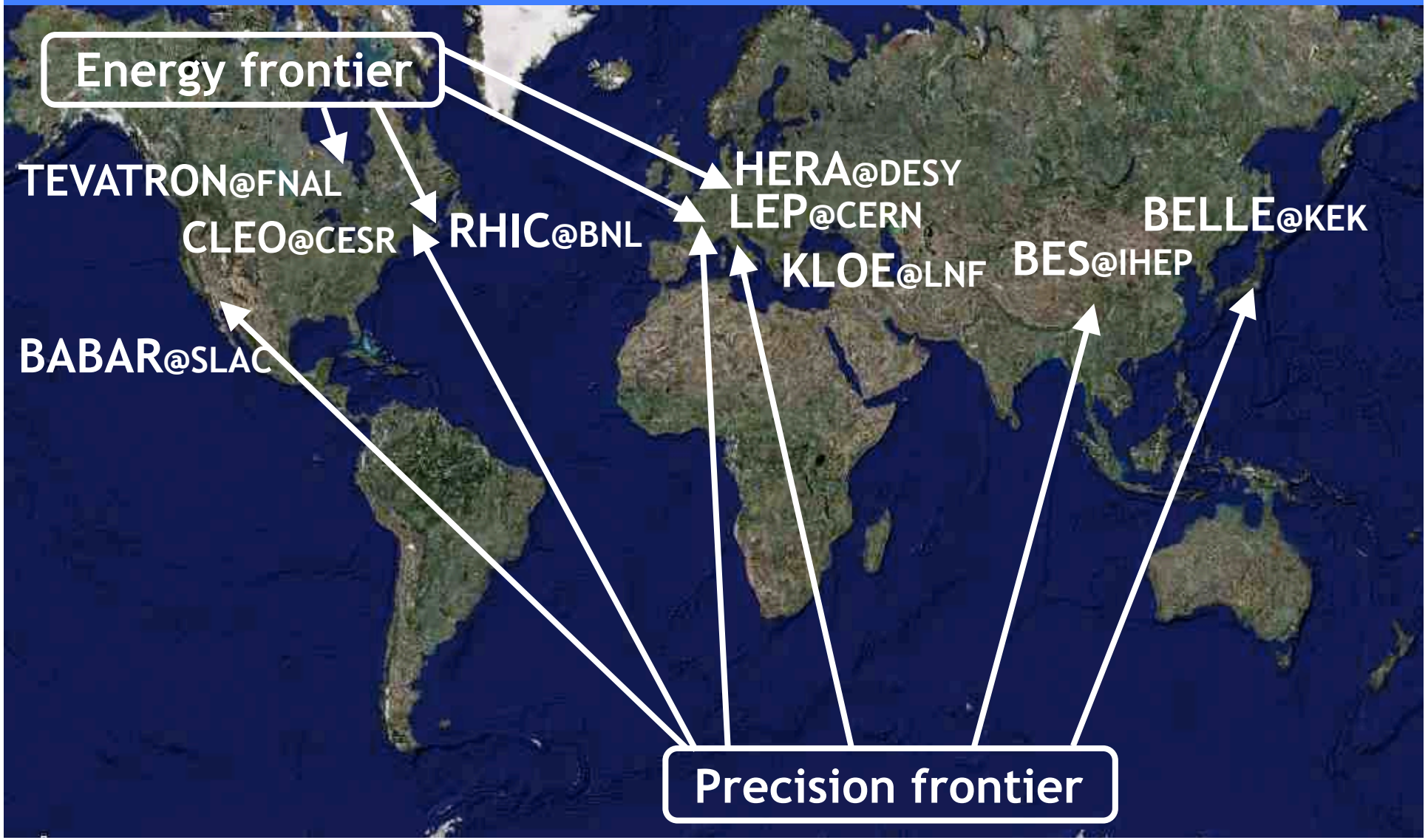




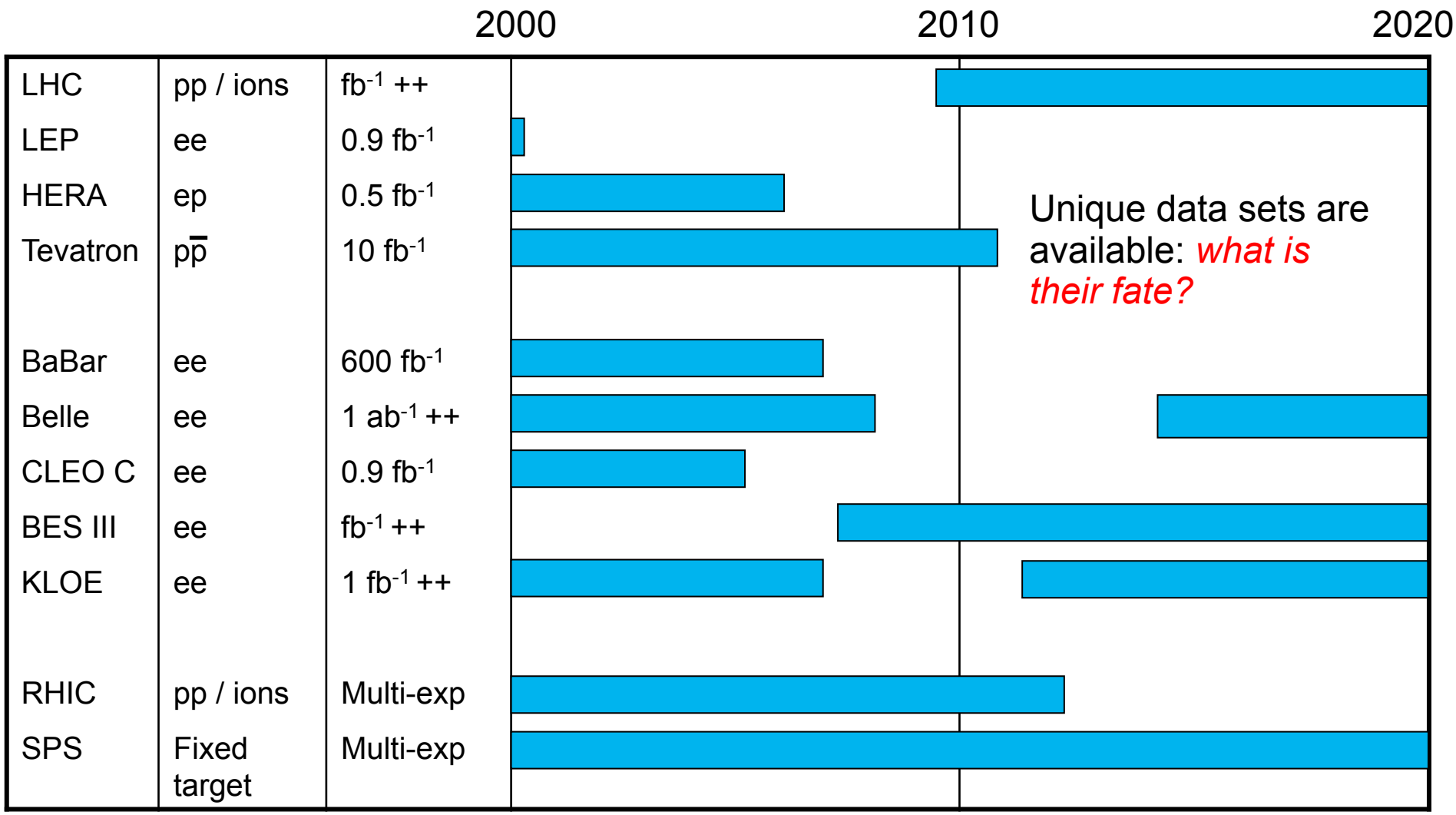
First collisions observed at the LHC in 2008; first data taking at 7 TeV now!

- Energy frontier probed with complex experimental installations
- New experiments normally supercede previous/similar ones - but not always..
- What is the present situation?

Active Experiments in the Pre-LHC Landscape

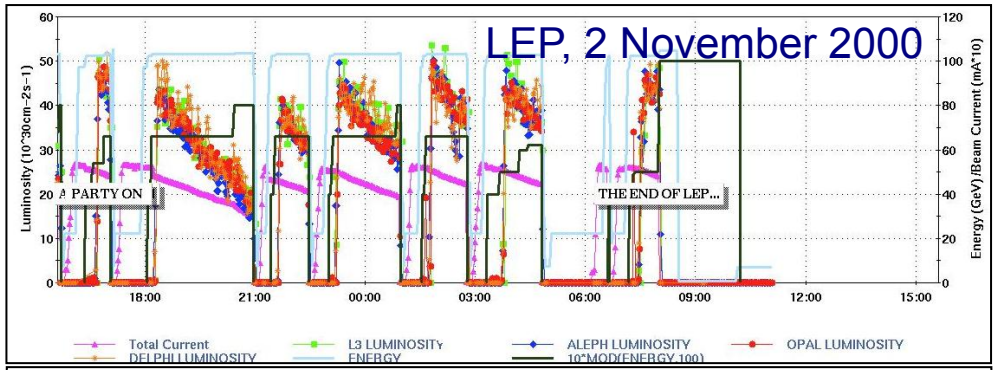


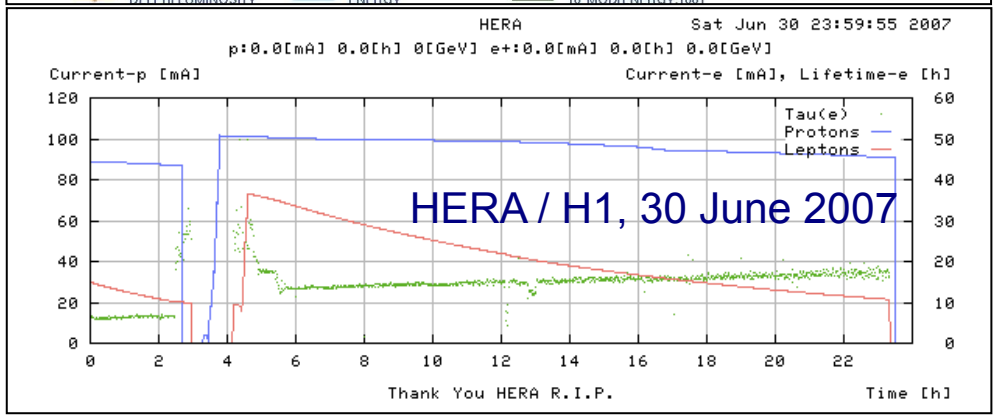
HEP Experimental Programmes in ± 10 Years

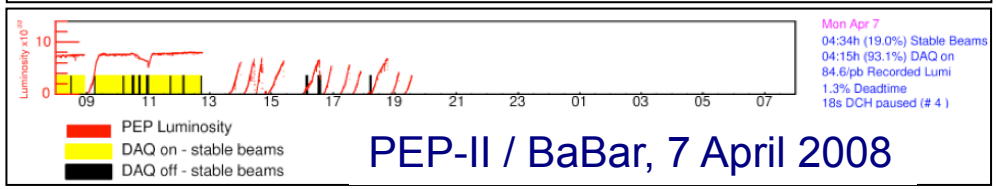


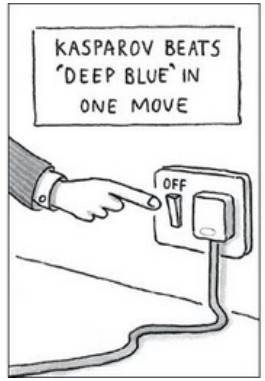
[not all programmes, dates are approximate, just to give the picture]

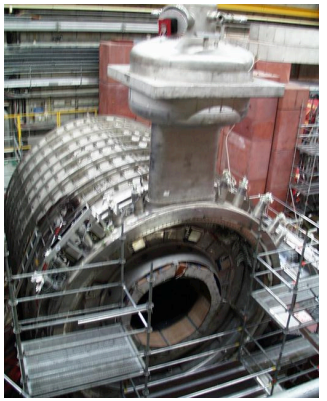
After the End of Data Taking

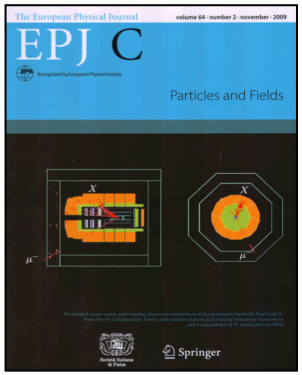












- Have an end of run party, dismantle the detector, finalize the analyses,.. *all in all about 5 years*
- *And then what do you do with the data?*

A Few Communiqués Suggest a Common Problem…

To Whom it may concern,

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 admin. service [a large computing centre]

> Some other choice quotes:

"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!"

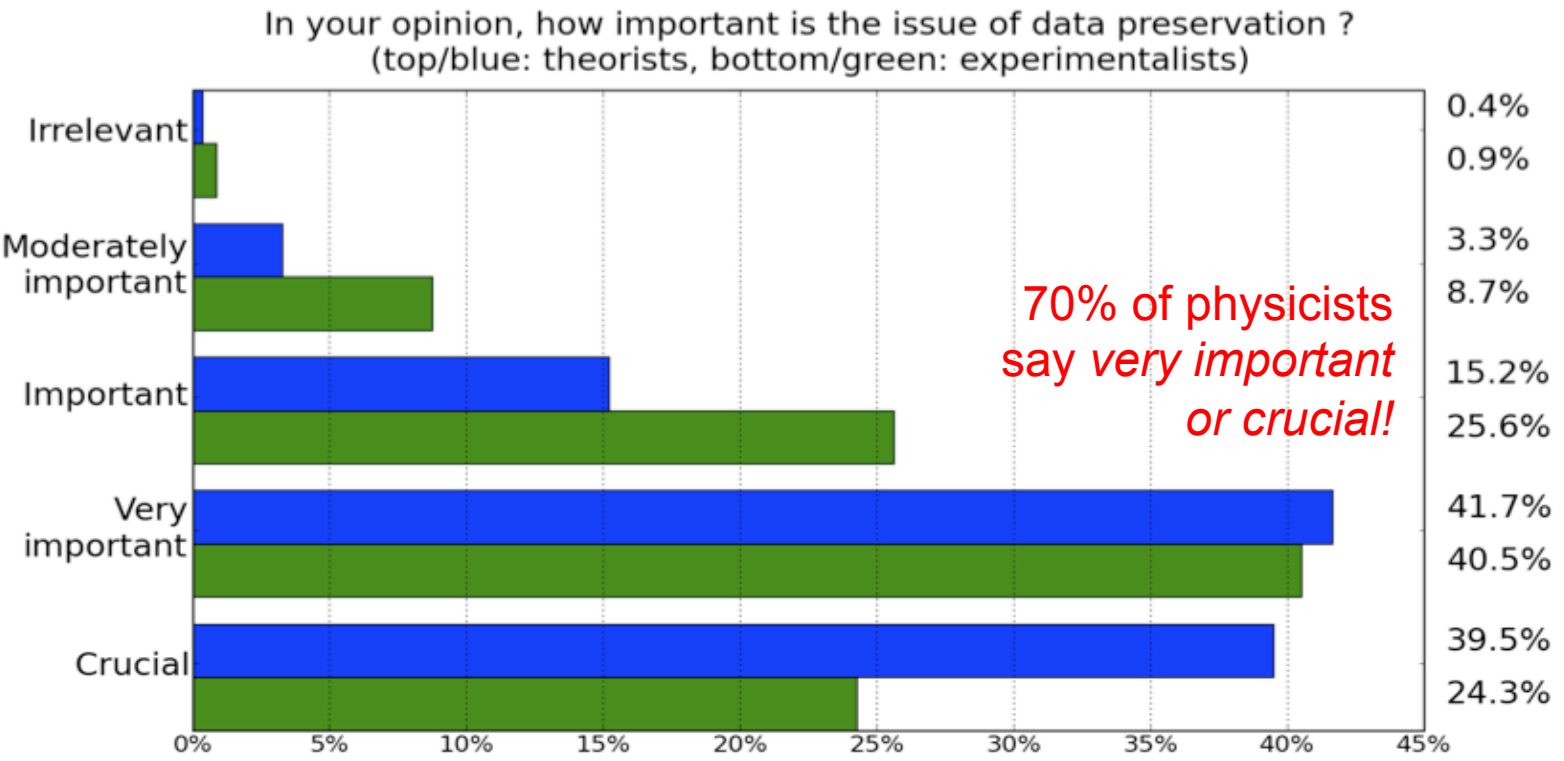
Past Experiences of Data Preservation in HEP

- No tradition, no model
- Data is lost or practically unavailable after a few years
- DP Not part of the planning, software design or budget of a HEP experiment
- Preservation examples are so far individual initiatives

PARSE.Insight: Support in the HEP Community

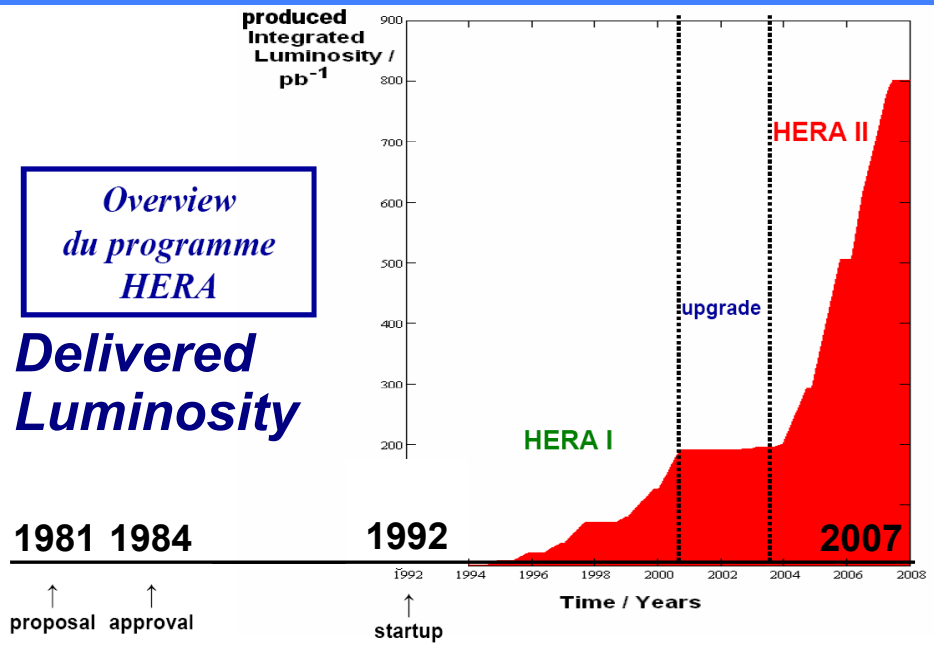
PARSE.Insight is financed by the European Commission and run at CERN
arXiv:0906.0485

S. Mele

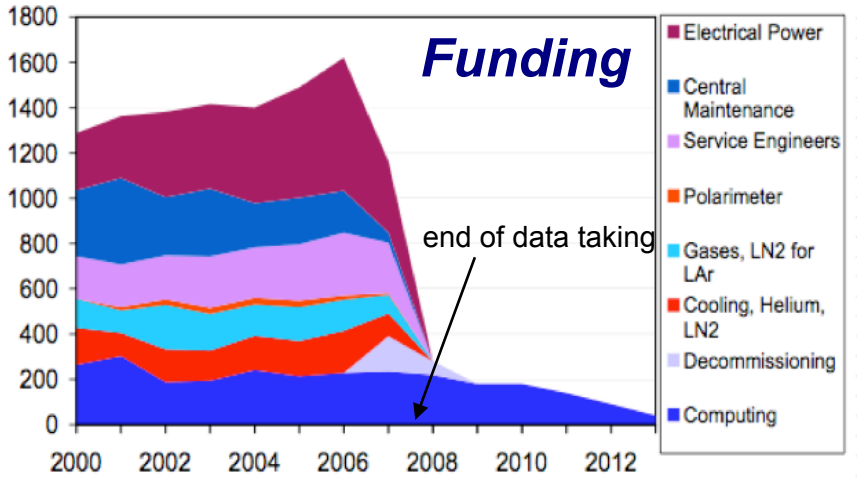


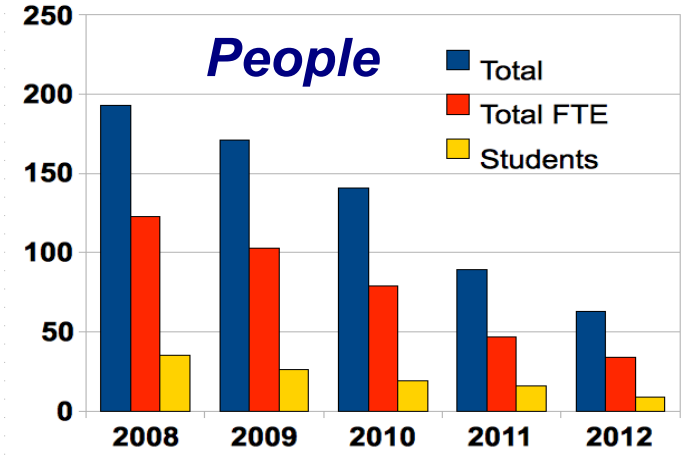
- However, no coherent strategy exists: in general, HEP data are lost
- The task in hand is to provide a coherent set of guidelines for future experiments to ensure the longevity of our data

Why is it Difficult to Preserve HEP Data?

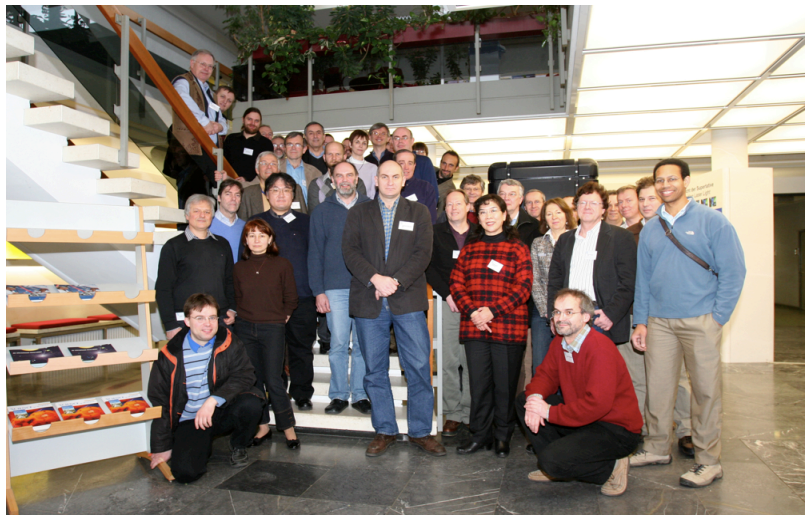


- Good data taking period is towards the end of running
- The existing resources (funding and expertise) then decrease when the data taking stops





DPHEP: International Study Group on Data Preservation



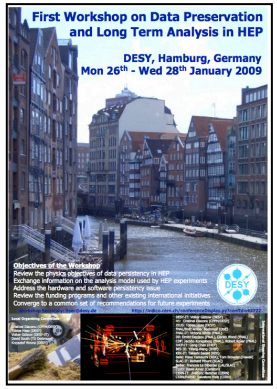
DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

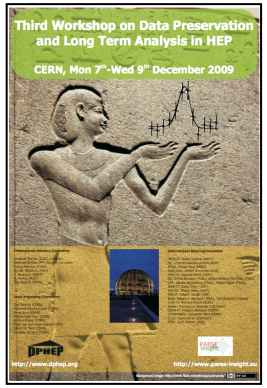
- Group has grown since 2008 to over 100 contact persons
- Endorsed by ICFA summer 2009
- **LHC** experiments joined in 2011

- Chair: Cristinel Diaconu (DESY/CPPM)
- Working Groups
  - Physics Cases: François Le Diberder (SLAC/LAL)
  - Preservation Models: D. South (DESY), Homer Neal (SLAC)
  - Technologies: Stephen Wolbers (FNAL), Yves Kemp (DESY)
  - Governance: Salvatore Mele (CERN)
- International Steering Committee
  - Participants from ee, ep and pp collider experiments
  - Associated computing centres at the labs
  - Some funding agencies
- International Advisory Committee
  - Chairs: Jonathan Dorfan (SLAC), Siegfried Bethke (MPIM)
  - Advisers: Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo), Alex Szalay (JHU)

DPHEP Activities

- First contacts established in September 2008
- Series of DPHEP workshops held since 2009
  - Jan2009: DESY May 2009: SLAC Dec 2009: CERN Jul 2010: KEK May 2011: Fermilab





- Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields such as astrophysics and others handling large data sets
- With the ultimate aim of providing a set of recommendations concerning data preservation for past, present and future HEP experiments

DPHEP Visibility

CERN Courier, May 2009

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.

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

Science

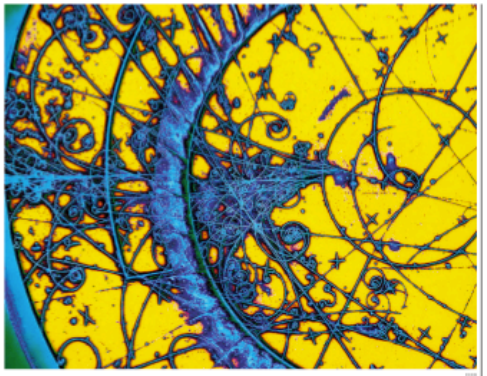
Rescue of Old Data Offers Lesson for Particle Physicists

Old data tends to get forgotten as physicists move on to new and better machines.

February 2011

Symmetry, December 2009

Wissenschaft



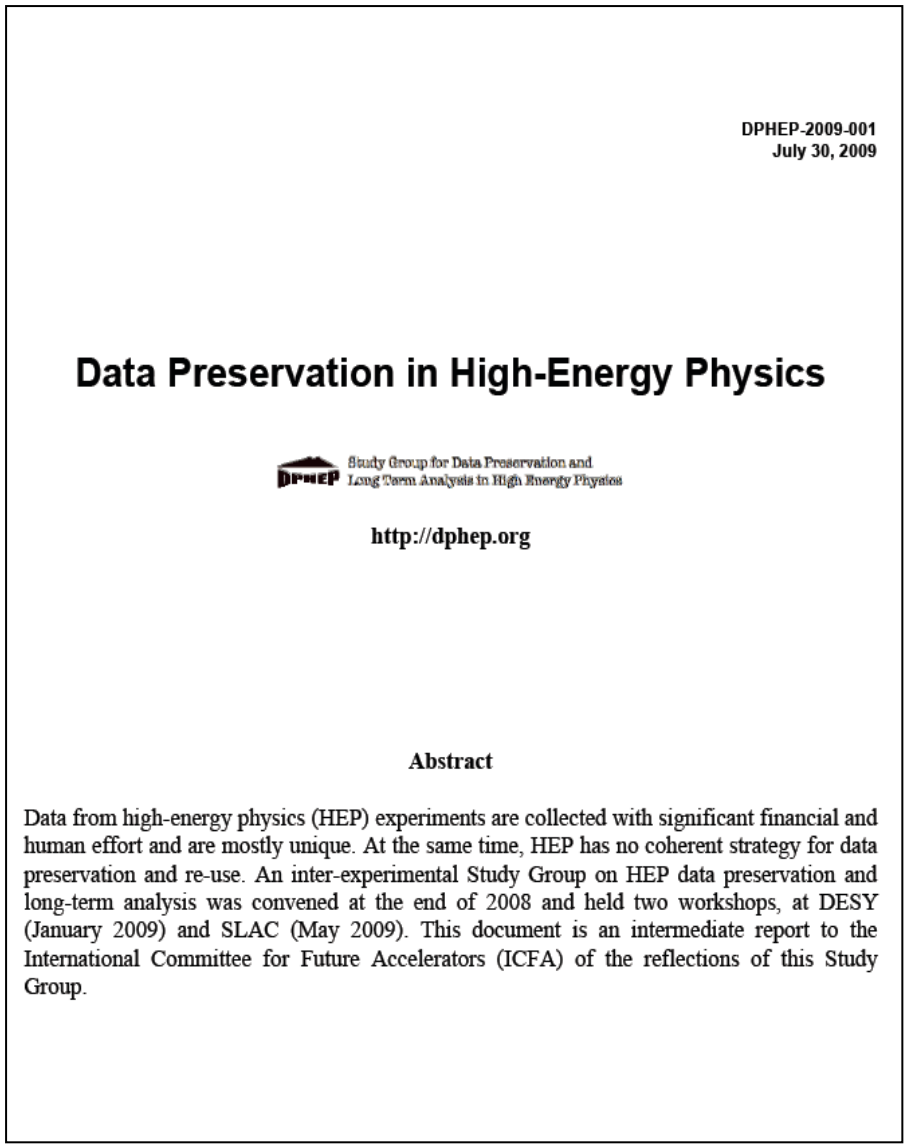
Die Hieroglyphen von morgen

An Beschleunigern sind immense Datenmengen entstanden – die Archivierung beginnt erst jetzt

Der Teilchenzoo

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.

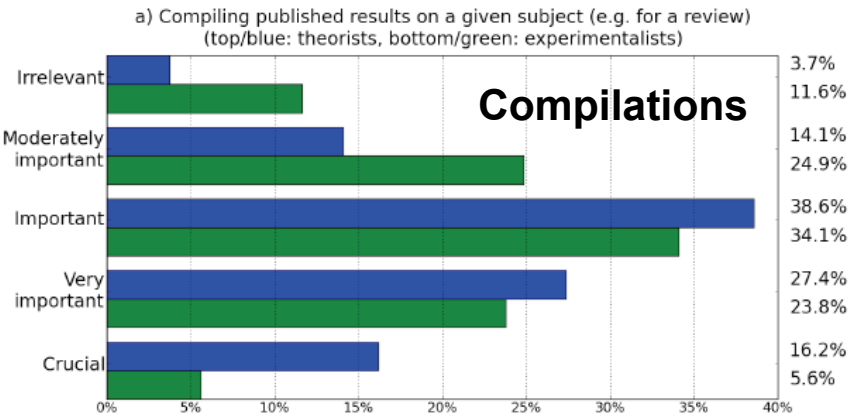
- First recommendations of the group published November 2009 ***arXiv: 0912.0255***
- The report covers the four key areas
  - Physics Case for Data Preservation
  - Preservation Models
  - Technologies
  - Governance

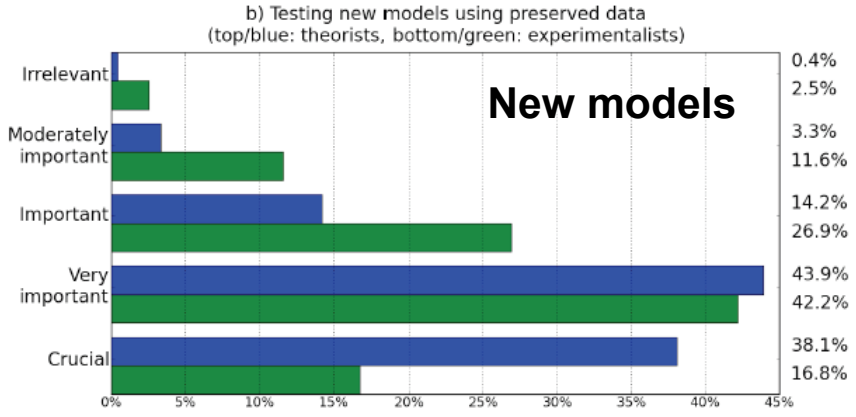
Physics case: Why would we want to re-use old HEP Data?

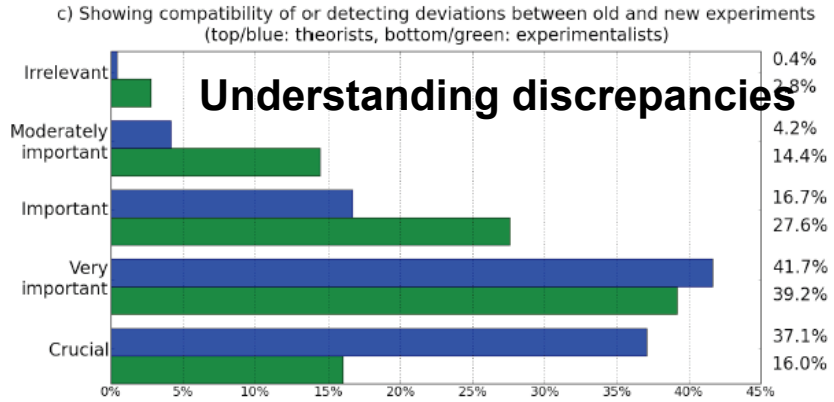
- We may want to re-do previous measurements
  - Increased precision, reduced systematics
  - New and improved theoretical calculations / MC models
  - Newly developed analysis techniques
- We may want to perform new measurements
  - At energies and processes where no other data are available (or will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

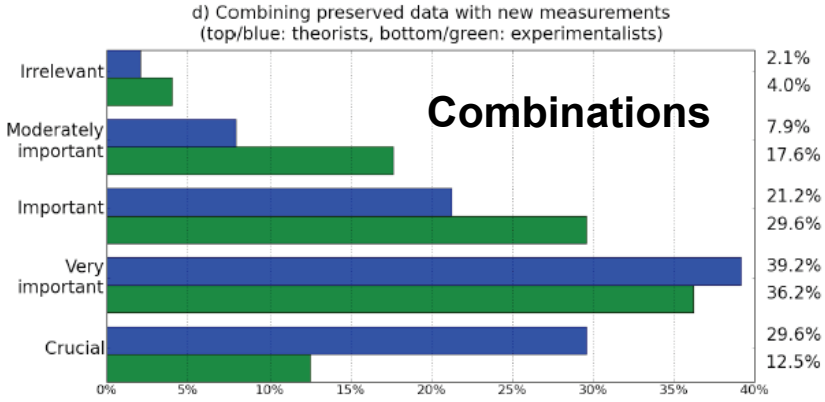
Support from the HEP community

Preserving HEP data is important for:



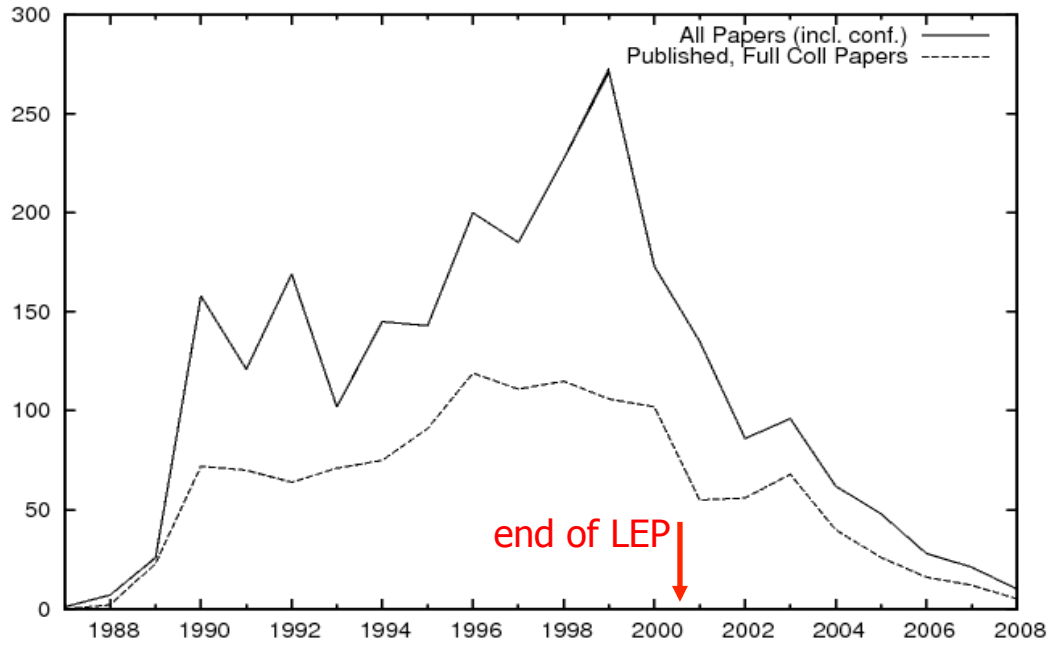






The Long Tail of LEP

Papers from all 4 LEP experiments (SPIRES Data)

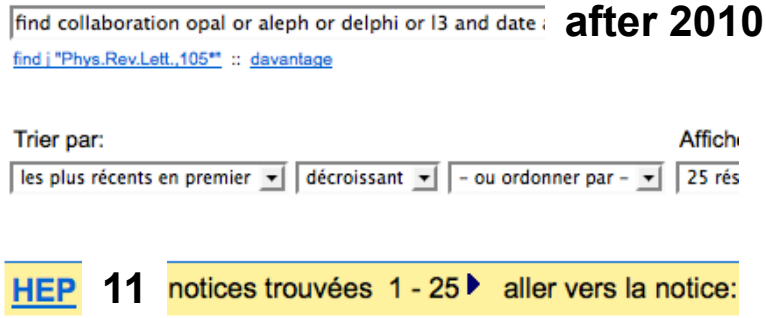


	All	ALEPH	DELPHI	L3	Opal
All physics	**345**	**65**	**114**	**85**	**81**
Electroweak	**89**	**17**	**26**	**22**	**24**
QCD	**85**	**19**	**25**	**19**	**22**
Higgs searches	**37**	**6**	**14**	**8**	**9**
SUSY searches	**25**	**4**	**7**	**5**	**9**
Exotica search	**34**	**5**	**12**	**10**	**7**
Flavor physics	**30**	**6**	**15**	**4**	**5**
Exclusive channels	**21**	**3**	**8**	**8**	**2**
Cosmo-LEP	**12**	**3**	**3**	**6**	**-**
Other	**13**	**2**	**4**	**3**	**3**

LEP Publications after 2004

S.Mele, P.Igo-Kemens

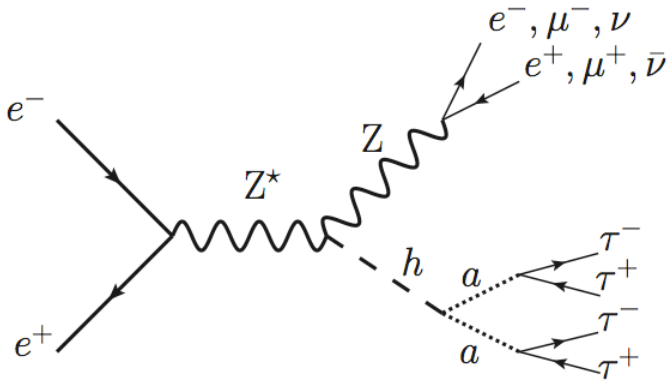
- Physics subjects are published after the end of collisions and/or collaborations
- 5-10% of the papers are finalized in the "archival mode"
  - Large number of publications well after data taking stopped
  - Large variety of topics
  - Legacy publications (full data, combined results) came later

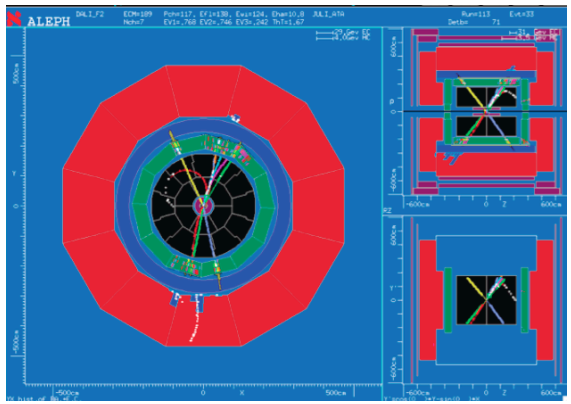


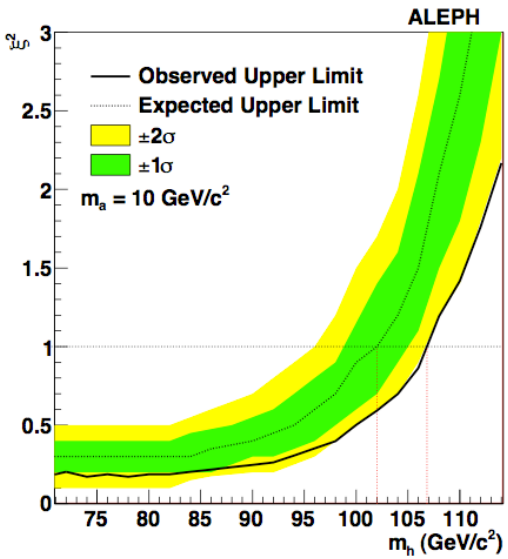
after 2010

Searches still possible

- Theory and “common sense” evolve
- Unique physics case analysed 10 years after the end of collisions (and 5 years after the official end of the collaboration)







EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Search for neutral Higgs bosons decaying into four taus at LEP2

The ALEPH Collaboration*)

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.

Better theory, better methods

OPAL [2011] arXiv:1101.1470 [hep-ex]

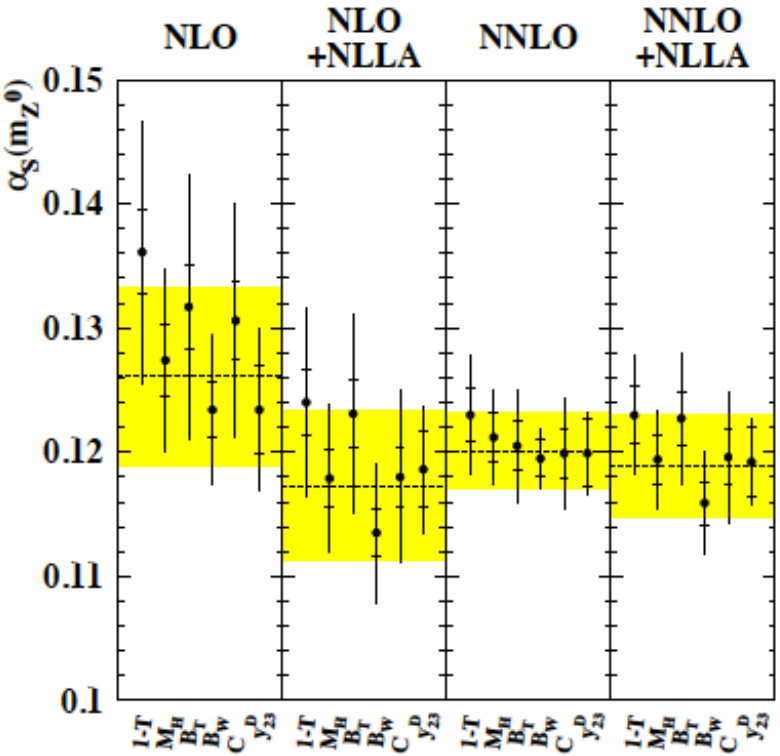


Figure 6: α_S results combined over all OPAL c.m. energies for different event shape variables and different QCD calculations as indicated on the figure. The shaded bands and dashed lines show the values of $\alpha_S(m_{Z^0})$ combined from these values with total uncertainties. The inner and outer uncertainty bars show the combined statistical and experimental and total uncertainties, respectively.

DELPHI, Eur.Phys.J. C71 (2011) 1557

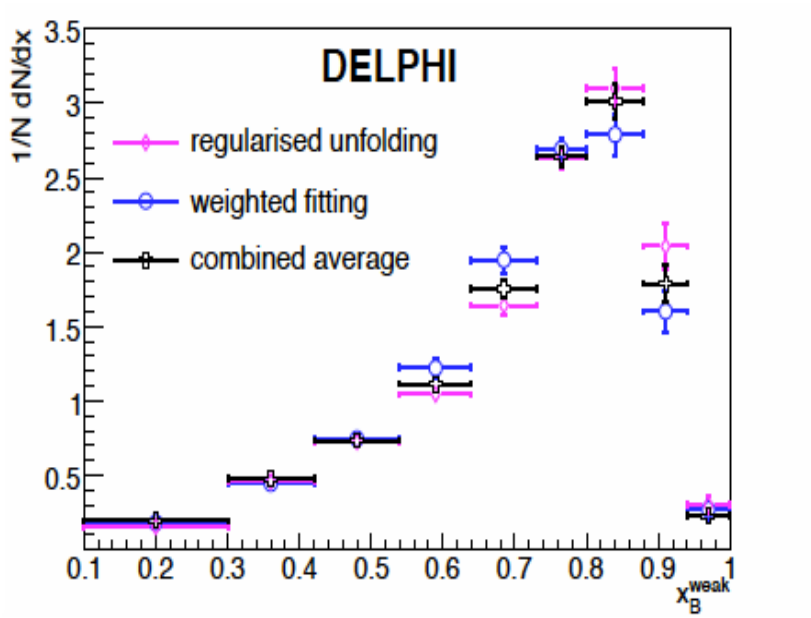
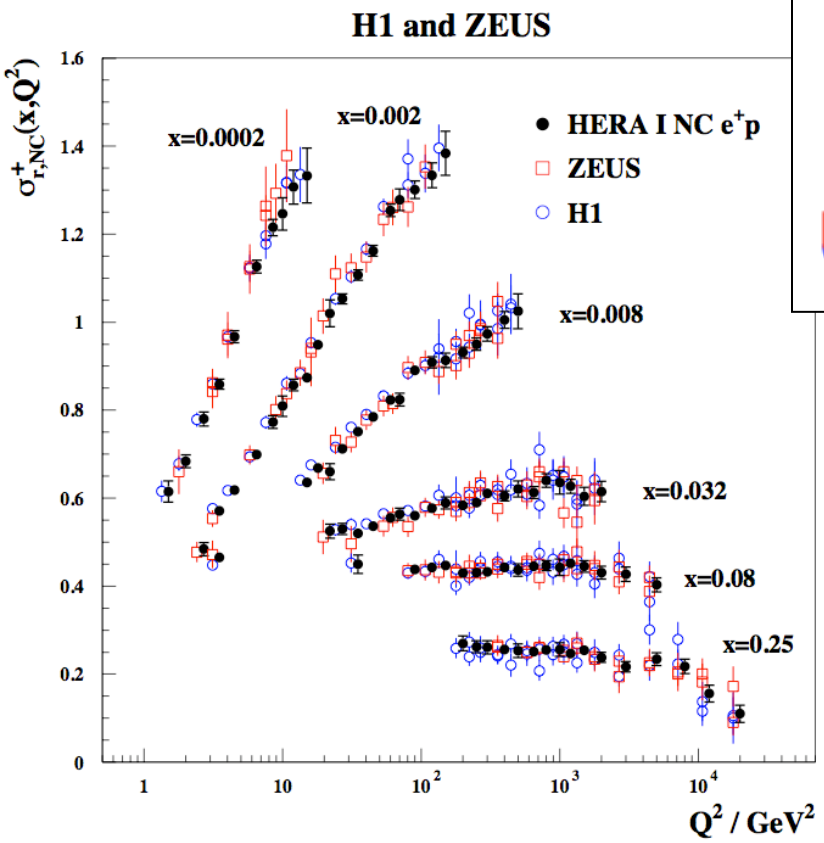
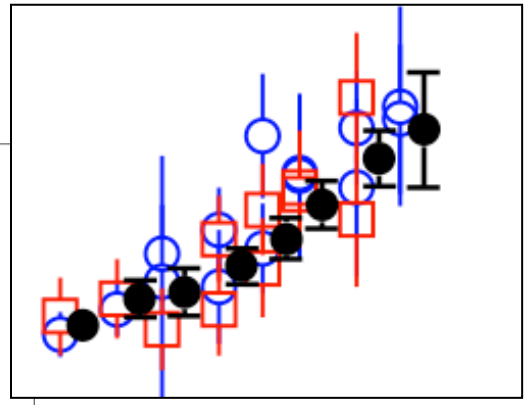


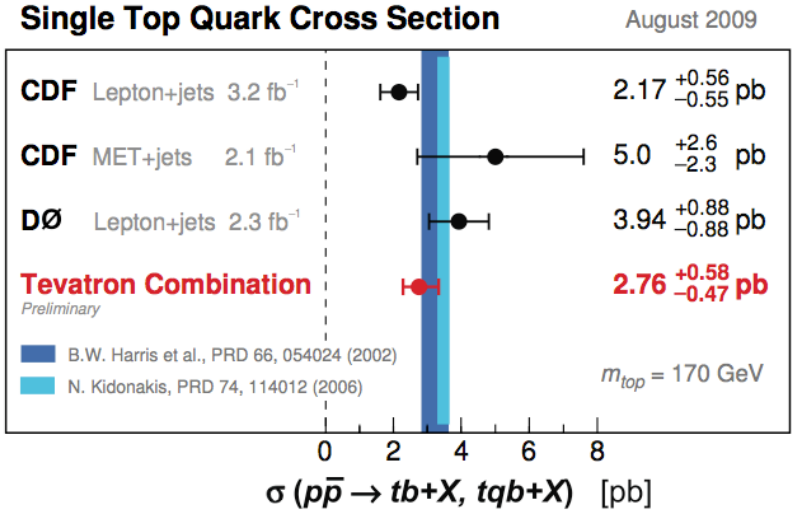
Figure 6: Measured fragmentation distributions in the two analyses and their combined average. Uncer-

Cross Collaboration and Combinations

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







- Preserved data would make possible more combined analyses across experiments

Dark photons: subject is new, data is old

Figure from Bluemlein, Brunner (2011)

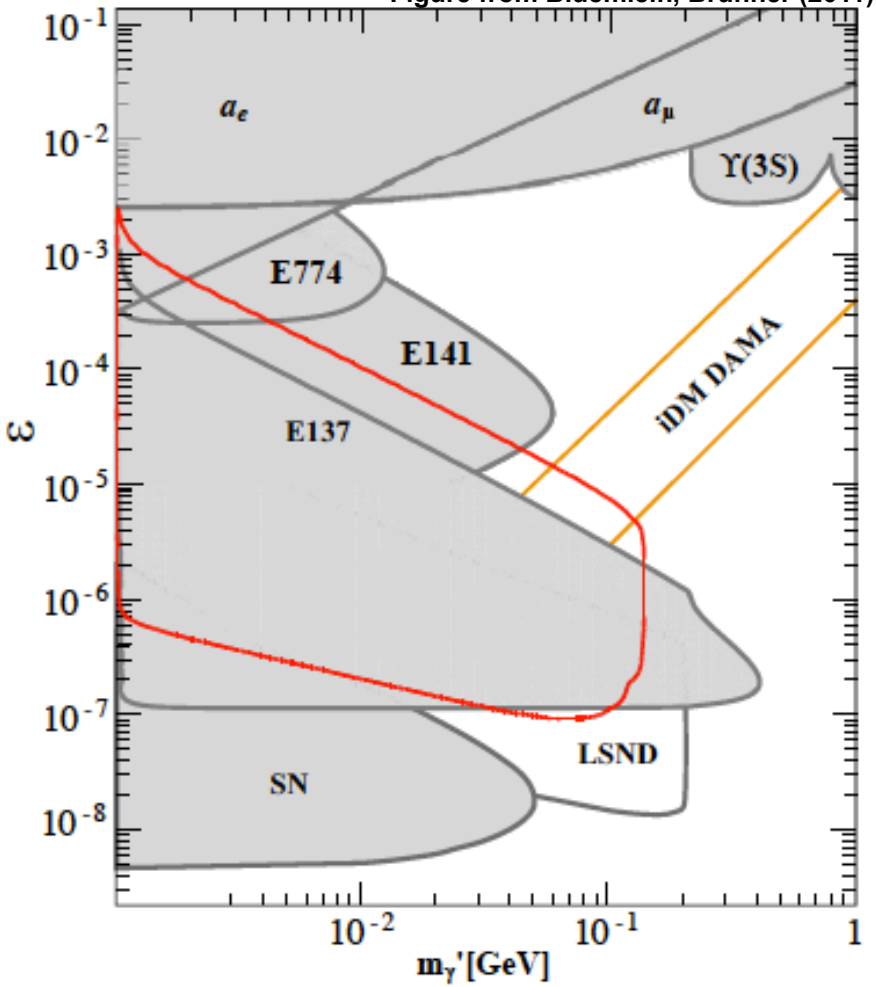
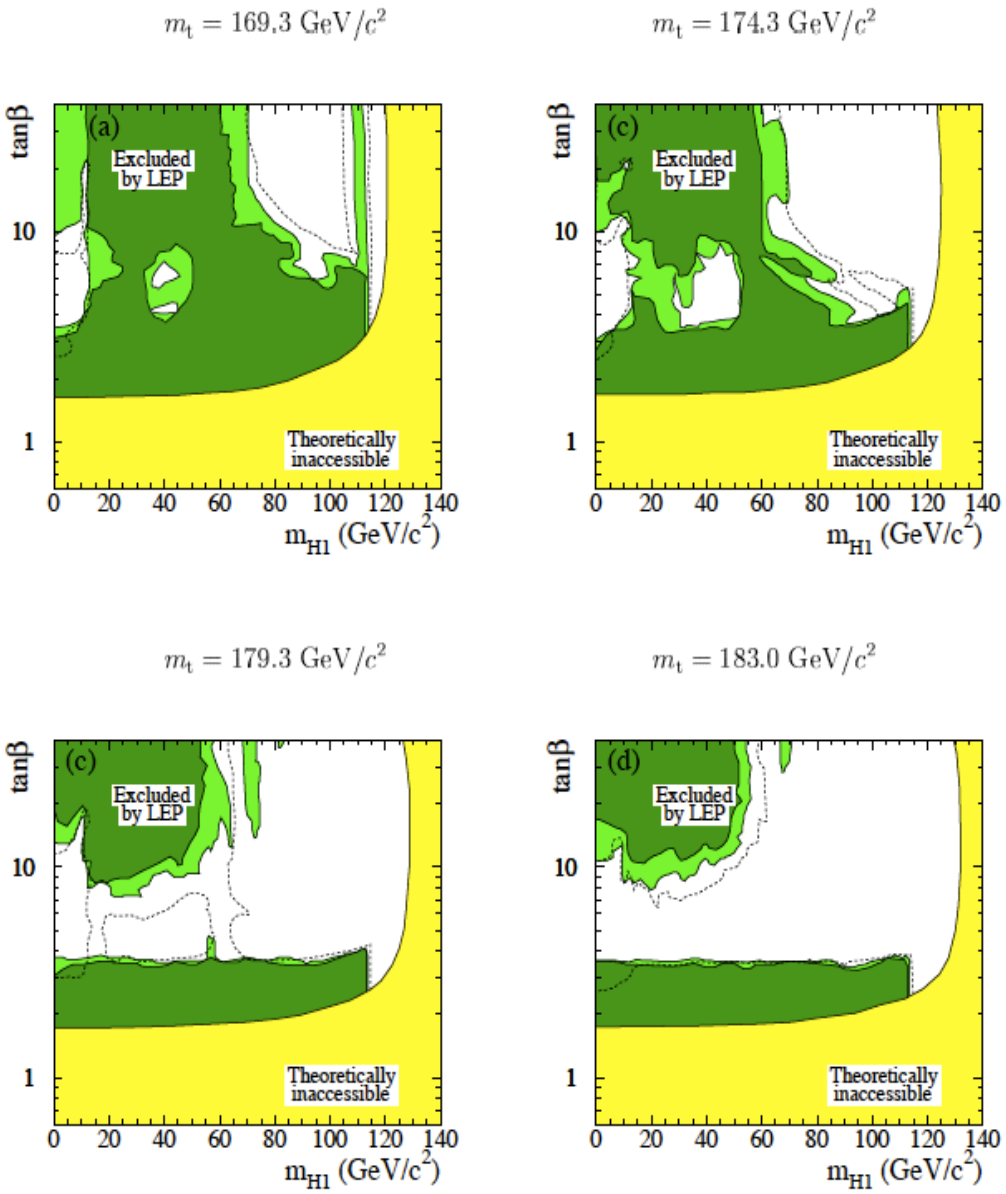


Figure 5: Comparison of the present exclusion bounds (red line) with other limits from the measurement of the anomalous magnetic moments a_e and a_μ [19], $\Upsilon(3S)$ decay [20], the beam dump experiments E137, E141, E774 [21–23], and supernovae cooling [4,24]. We indicate the prospects for LSND [7,25] (open grey-bounded area), and the DAMA/LIBRA region (open orange bounded area) [26]. The limits for $\epsilon > 10^{-7}$ have been taken from Ref. [6].

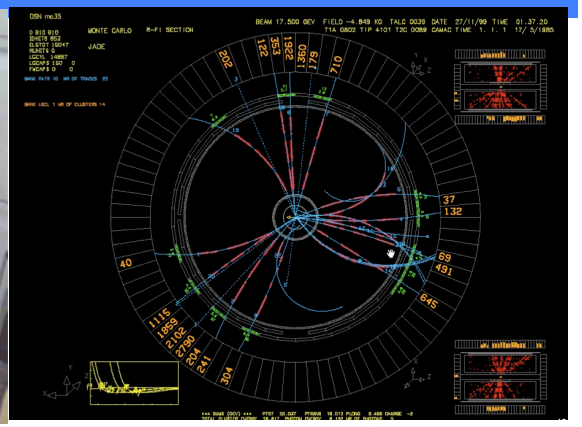
Excluded?

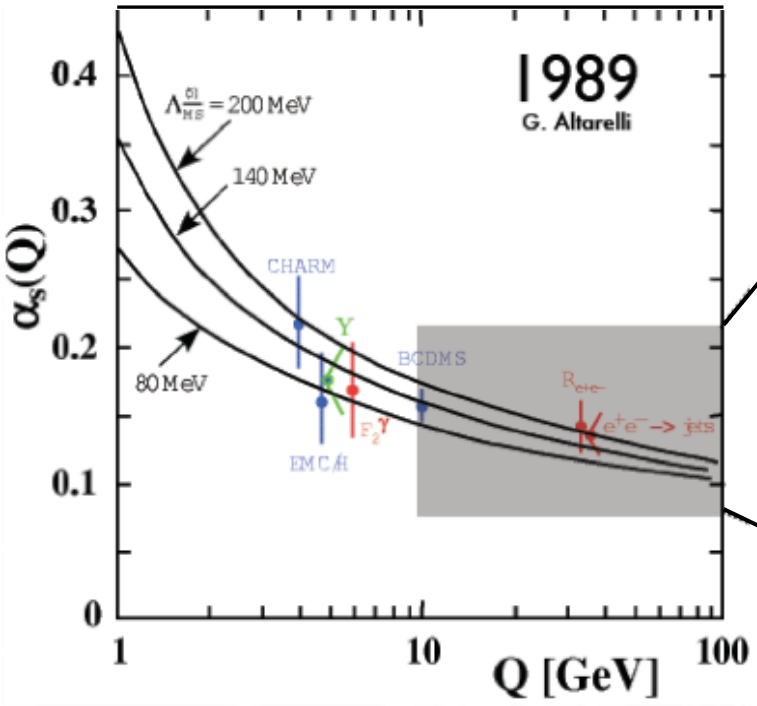
- Some external parameters may be not well known
- Re-optimisation may be a case for re-analysis

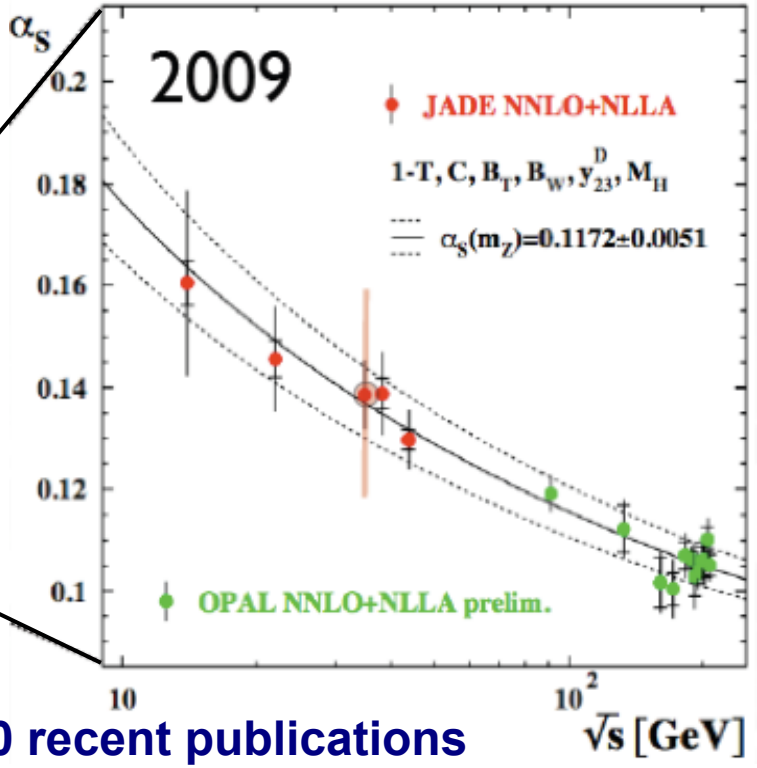


Successful Resurrection of JADE Data Analysis

- Required full raw data preservation, software revitalisation, needed many individual initiatives…



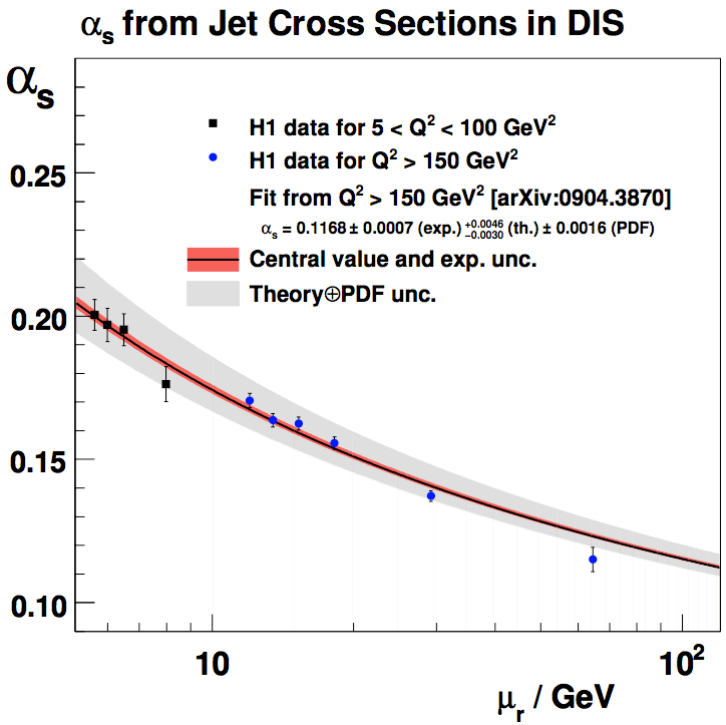


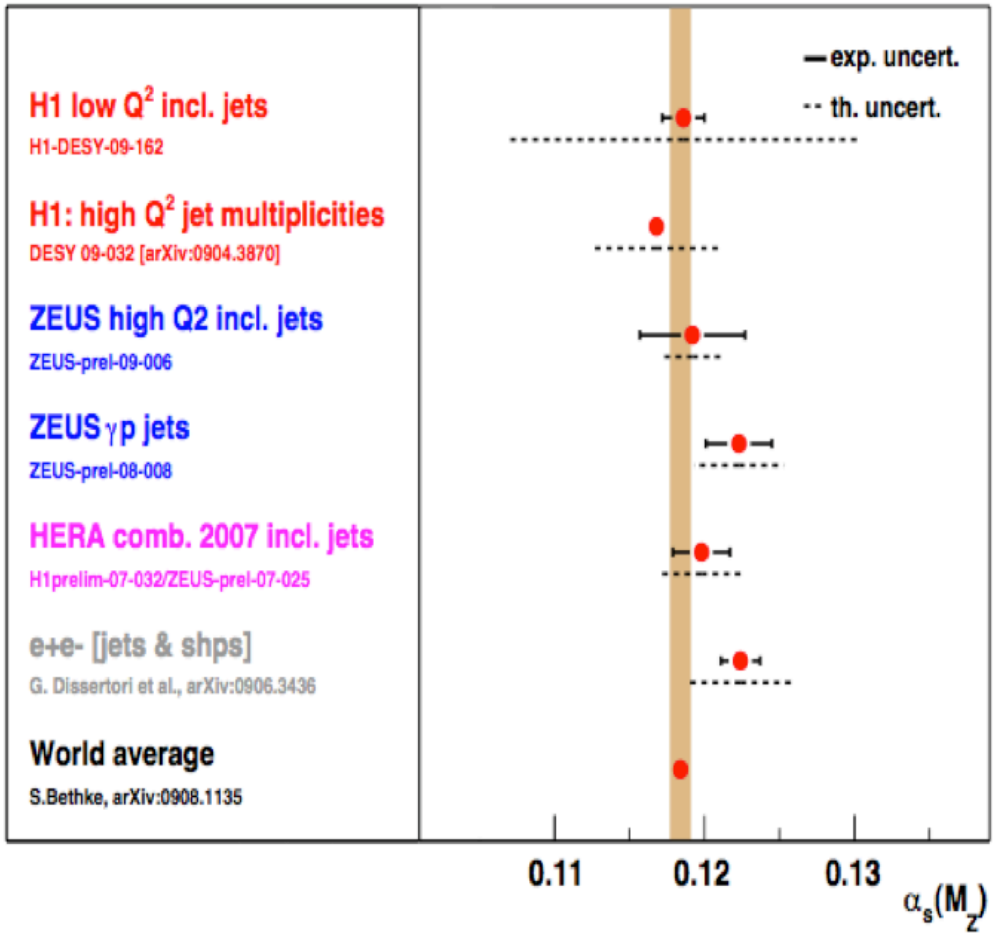


10 recent publications

History may well repeat itself….

- Around 10% of measurements are dominated by non-experimental errors: theory (N^nLO?) and simulation..
- More recent precision measurements of α_S:

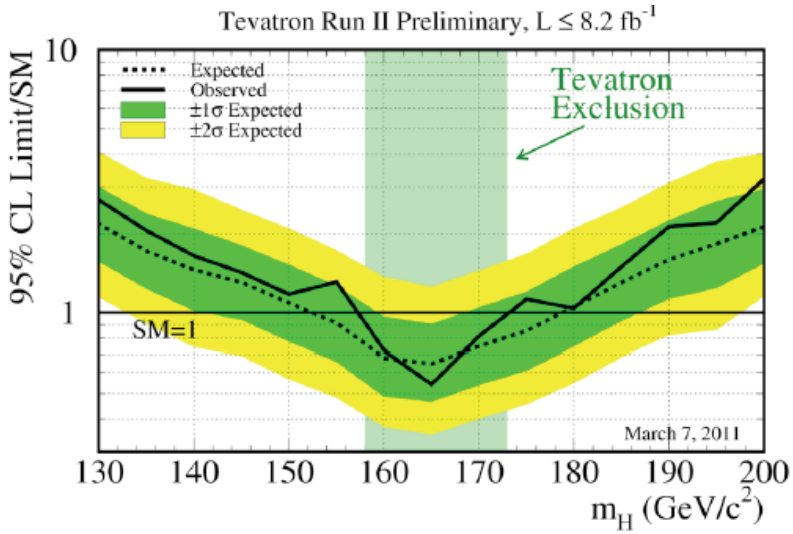


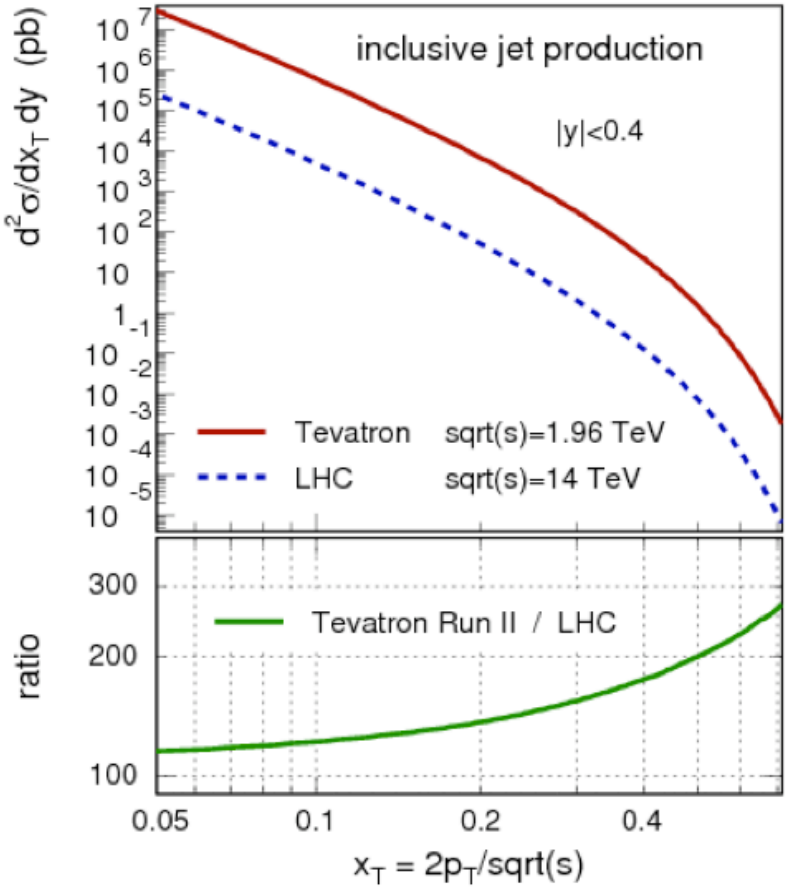


LHC will not completely take over Tevatron physics

On January 10, 2011, it was announced that the Tevatron will cease operations at the end of September, 2011,[4] as it has been made obsolete by the LHC,[5] which began operations in early 2010. The main ring of the Tevatron will probably be reused in future experiments, and its components may be transferred to other particle accelerators.[6] http://en.wikipedia.org/wiki/Tevatron

- Tevatron proton-antiproton collisions are in fact unique
- There is a physics case for Tevatron-LHC combinations

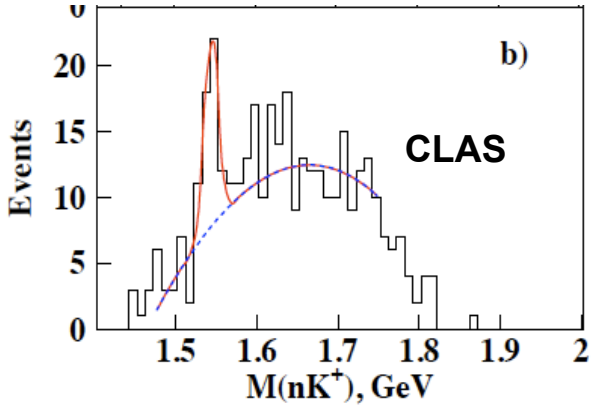


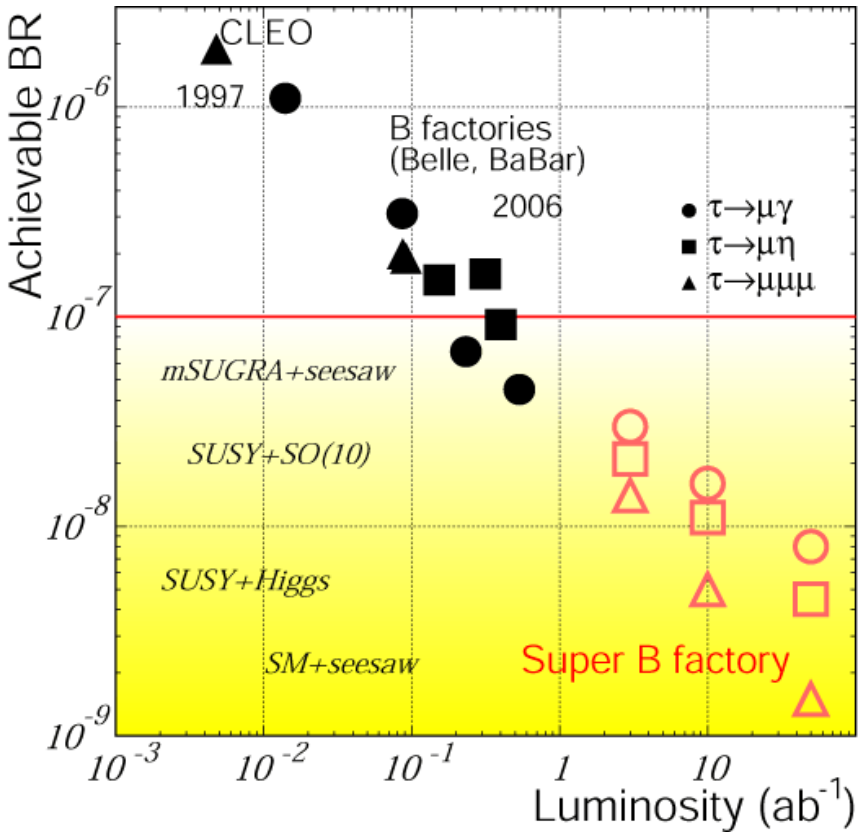


More examples…

- B- and SuperB-factories
- Low energy
- …and many others
  - your favourite?

...surprises can occur at lower energies too



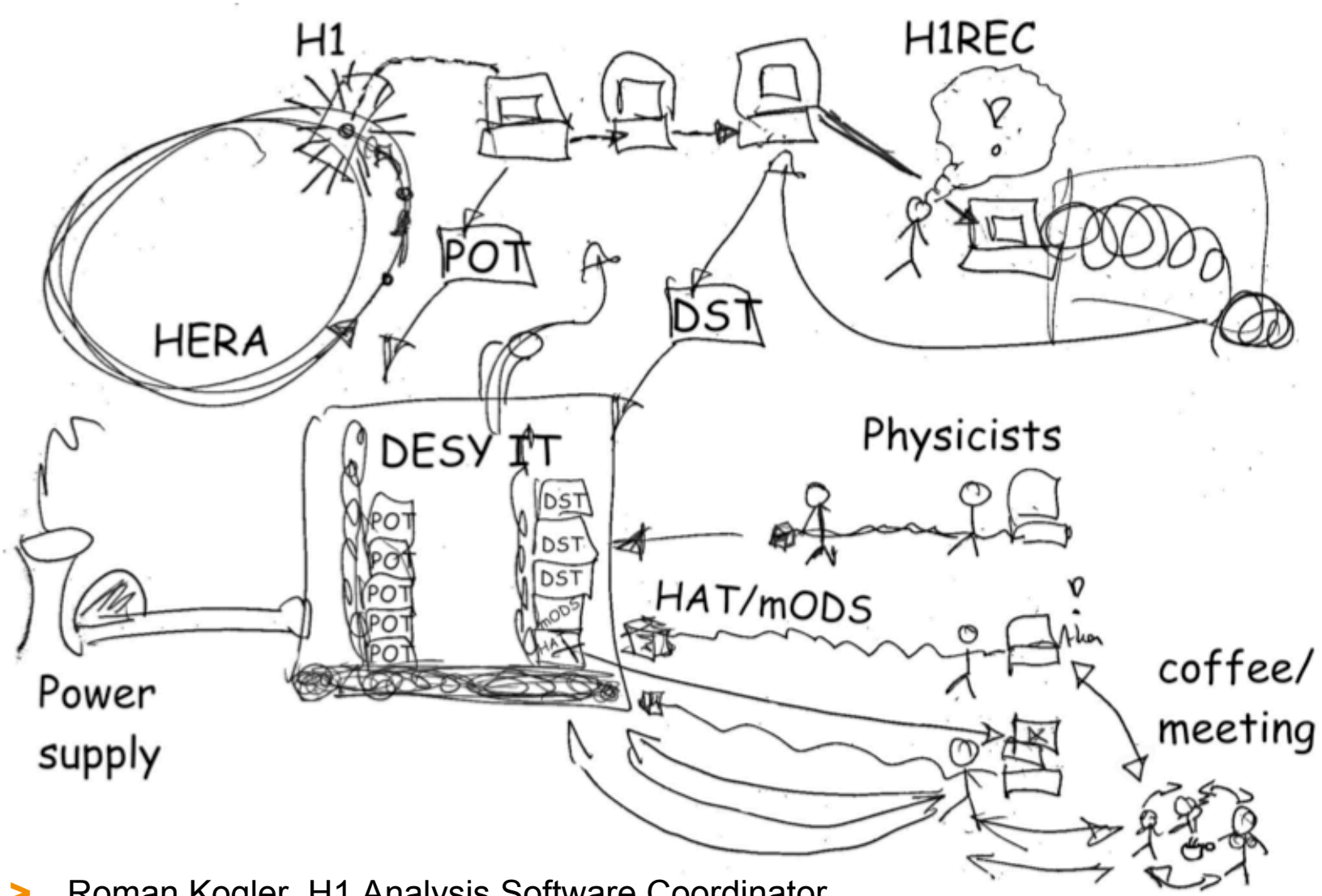


Part 1: Physics Cases for Data Preservation

- HEP data are mostly unique and have true scientific potential
- Several physics cases can be presented for preservation
  - Long term completion and extension of 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 (see Part 2)

Part 2: Models of Data Preservation

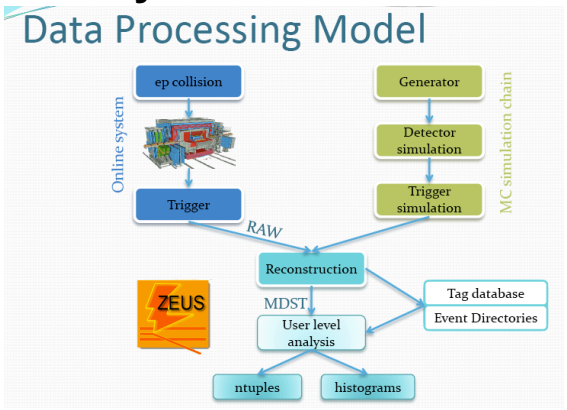
What we do

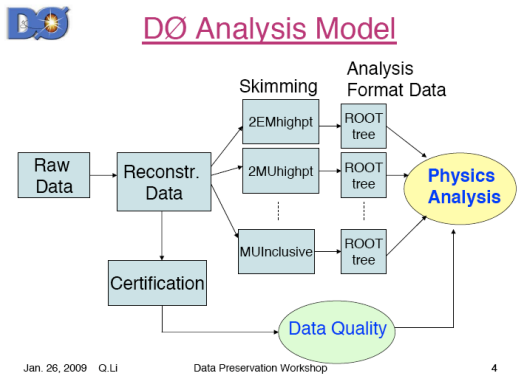


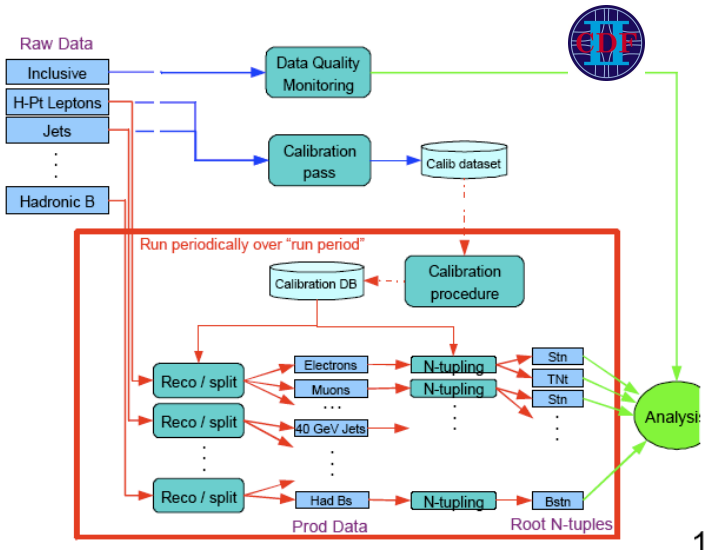
> Roman Kogler, H1 Analysis Software Coordinator

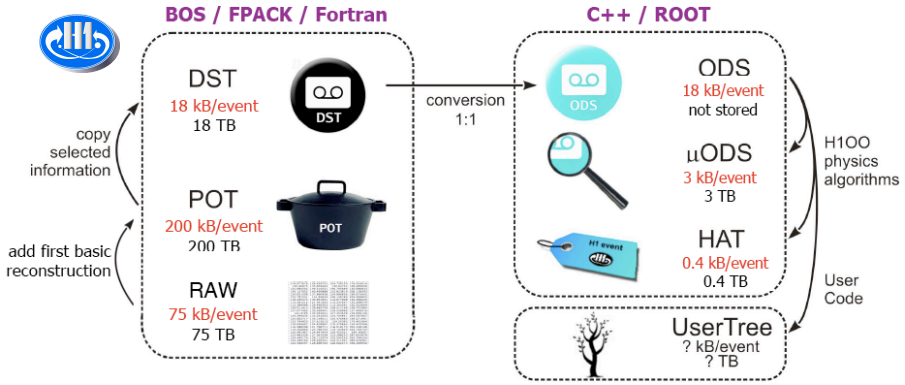
Data Analysis Models in HEP

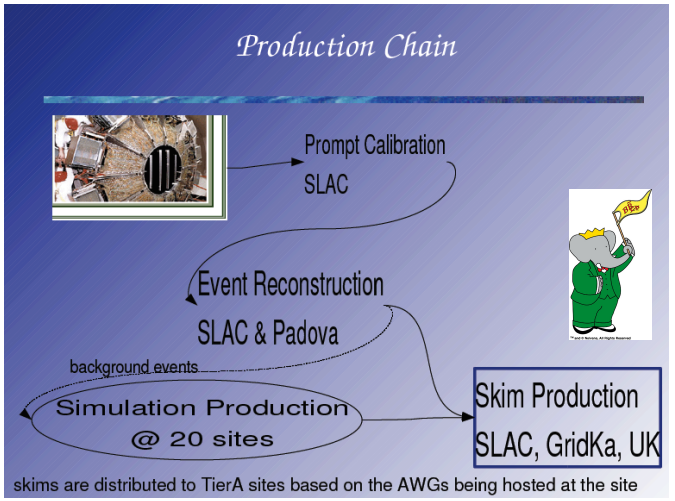
Survey of 2009







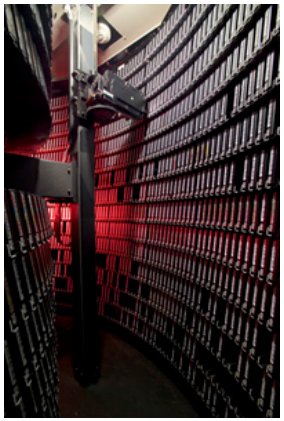


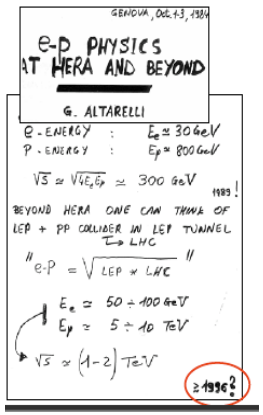


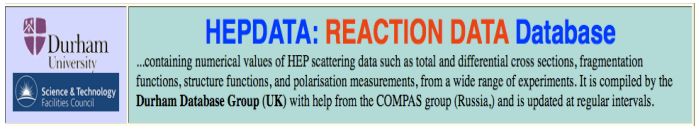
- Complicated, at first glance different
- Familiar descriptions of data analysis chain, from reconstruction to analysis level
  - RAW → POT → DST → *ntuple* → analysis

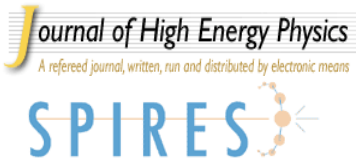
+ Belle, BES3, CLEO, Hermes,…

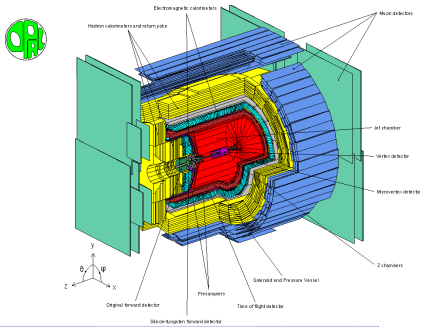
What is “HEP Data” anyway?











- Digital information: Data event files, database
- Software: Simulation, reconstruction, analysis, user
- Publications: Journals, arXiv, Spires/INSPIRE, HEPDATA
- Documentation: Publications, notes, manuals, slides
- Meta information: Hyper-news, messages, wikis, forums
- Expertise (people): Often the hardest to secure

Data Preservation Models identified by DPHEP

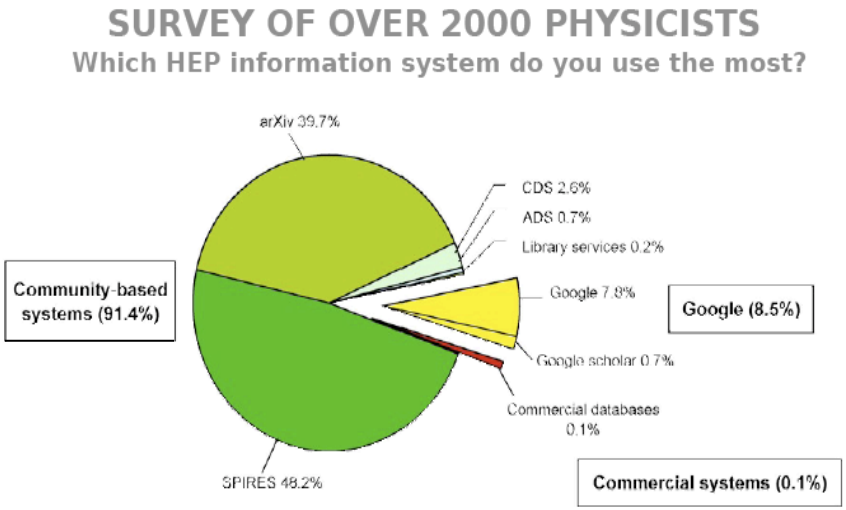
Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve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

Cost, complexity, benefits

- Only with the full flexibility does the full potential of the data remain
  - Level 4 type programme was required by the JADE and ALEPH re-analyses
- BaBar, H1, HERMES aim for DPHEP level 4, ZEUS between levels 3 and 4
  - Still some different approaches, can benefit from each other's experiences
- Levels 1 and 2 still require some work!

Documentation

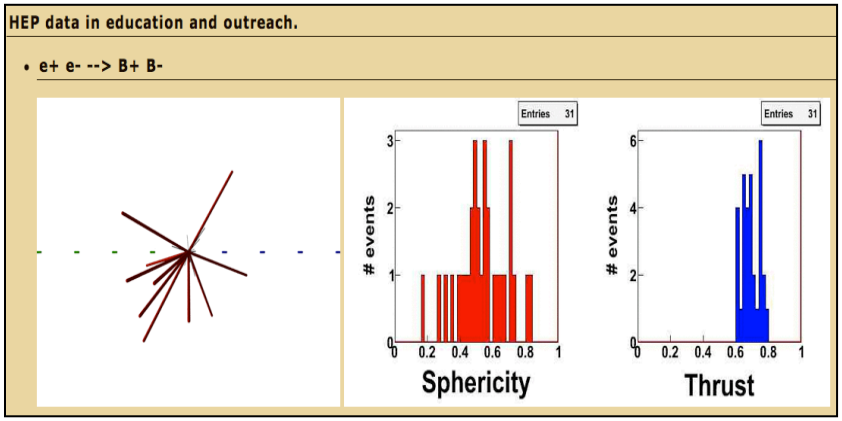
- Requirements: complete, clear, up-to-date
- A long list of things to consider:
  - Paper documents and drawings, on-line data and tools, detector configuration files, electronic logbooks, detailed run information, web content on outdated servers with dead links, wikis, meetings, talks, ... the list goes on..
  - Machine-related data
- Opportunity provided by *INSPIRE*
  - Project started (see later)
- Summary publications are also important
  - Project: B-Factories Legacy Book



SPIRES → INSPIRE

DPHEP Level 2: Scientific Training, Education, Outreach

- Improve the overall high level education in HEP
- Improve the connection of HEP-emerging countries to HEP data sets
- Provide access to HEP data to more people
- Within the preservation effort: a chance to have a coherent approach across experiments

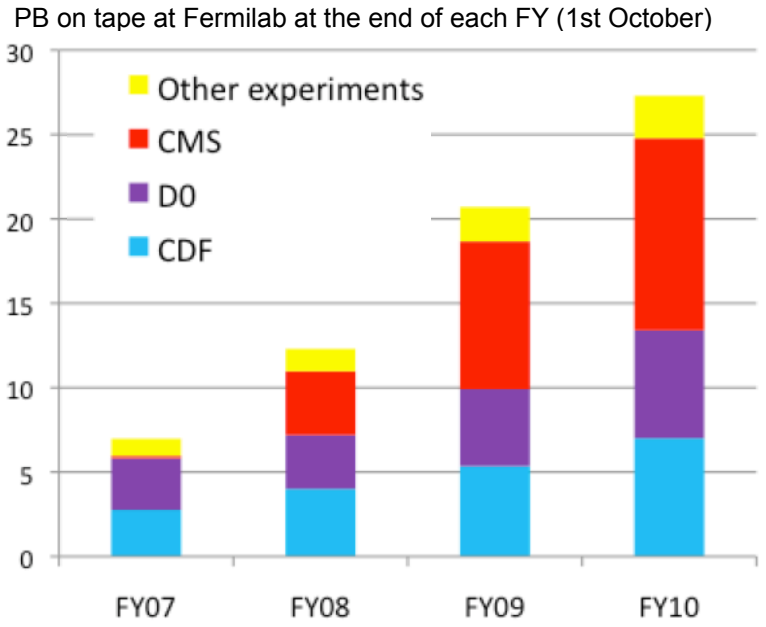


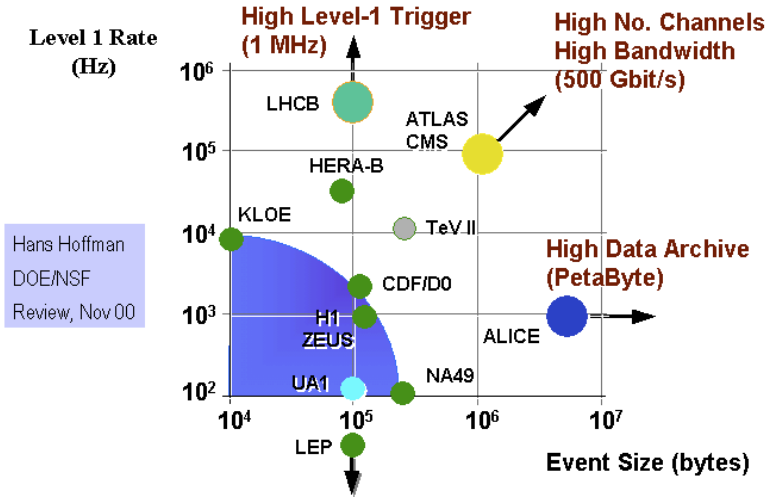
Part 3: Technologies

“Digital information lasts forever - or five years, whichever comes first.”
Jeff Rothenberg, RAND Corp.

How much Data are we talking about?

- Discussions in DPHEP lead to a number of around 0.5 to a few PB
  - Depending on preservation model (coming up!)
- Computing centres are, at least by volume arguments, able to store the data
  - Data preservation is not about the data!
- Regular migration of the data to latest technologies should be considered and carefully planned
- However, currently employed storage systems may not be suited for archival storage
  - Regular integrity checks of the full sample
- Any **archival system** should be able to absorb future technological evolutions

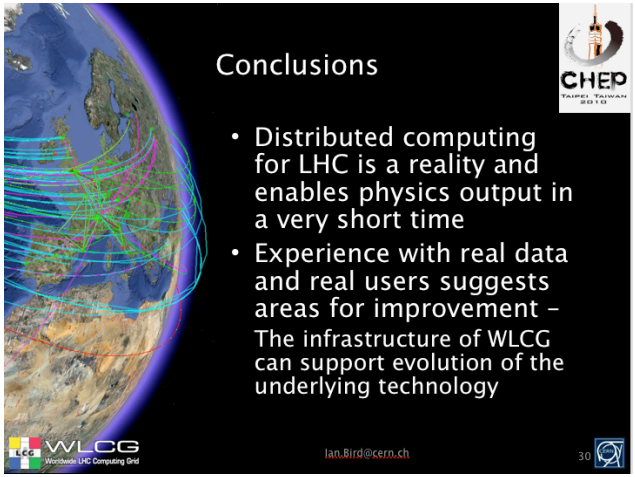


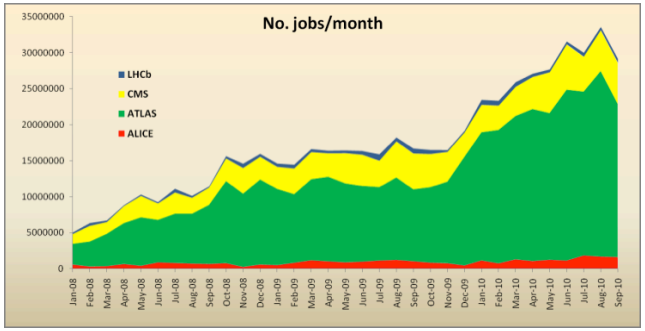


The Technology Continues to Evolve

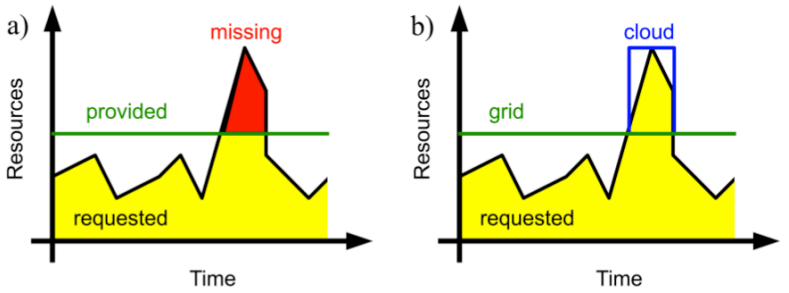
- Newer hardware has increased performance CPU/storage capacity
  - Parallelism crucial for future applications
  - Only runs on more modern new OS only?
  - 64 bit only and beyond?
- Hardware now "old" after a few years, transition and budget need careful planning

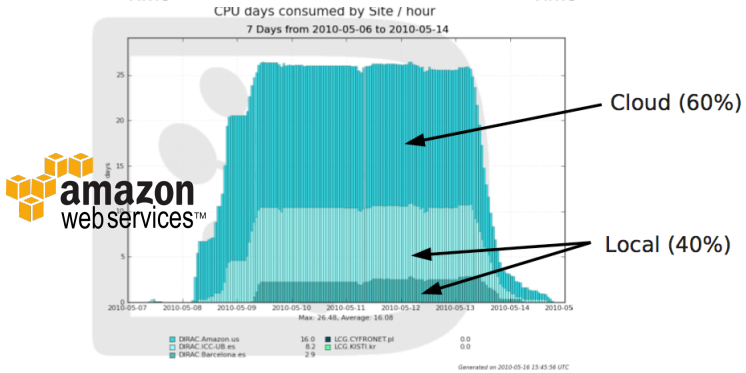
The Data Analysis Model is also Evolving





- Cloud computing allows to buy resources on demand
  - Well suited to absorb peaks in varying resource demand





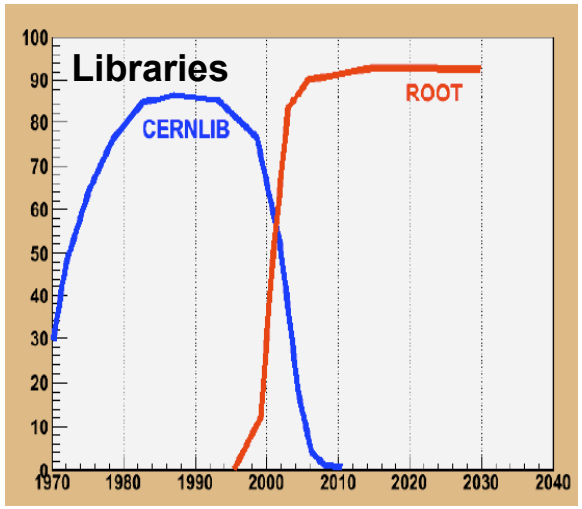
- 170M events (3.6 TB) produced in 6 days
- Amazon Spot Instances → 0.20 USD / 10k events

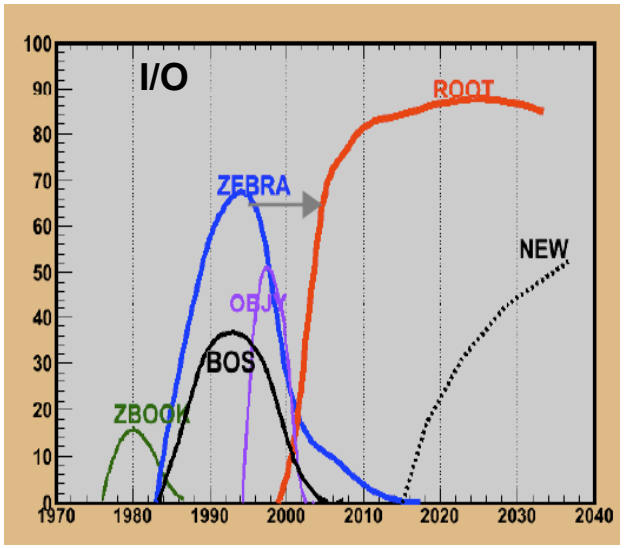
- Success of the GRID during first year of LHC; Belle buy time on the Cloud
- Where does analysis of preserved data take place and under which protocols?
- Is a global solution possible? (analysis grid-ification, large scale multi-center?)

A serious issue: the software maintenance

- Freezing: Technology preservation
  - Virtualisation techniques provide the software environment, freeze the hardware
  - Preparation step is not saved, lifetime limited as well
- Better: Continuous migration
  - Follow technology changes, external software, new OS, redesign, recompile etc
  - Virtualisation can help here too
- Preparation is not trivial
  - New operational model
  - Dependencies etc.
- Supervision is needed for both data and software
  - Data archivist position

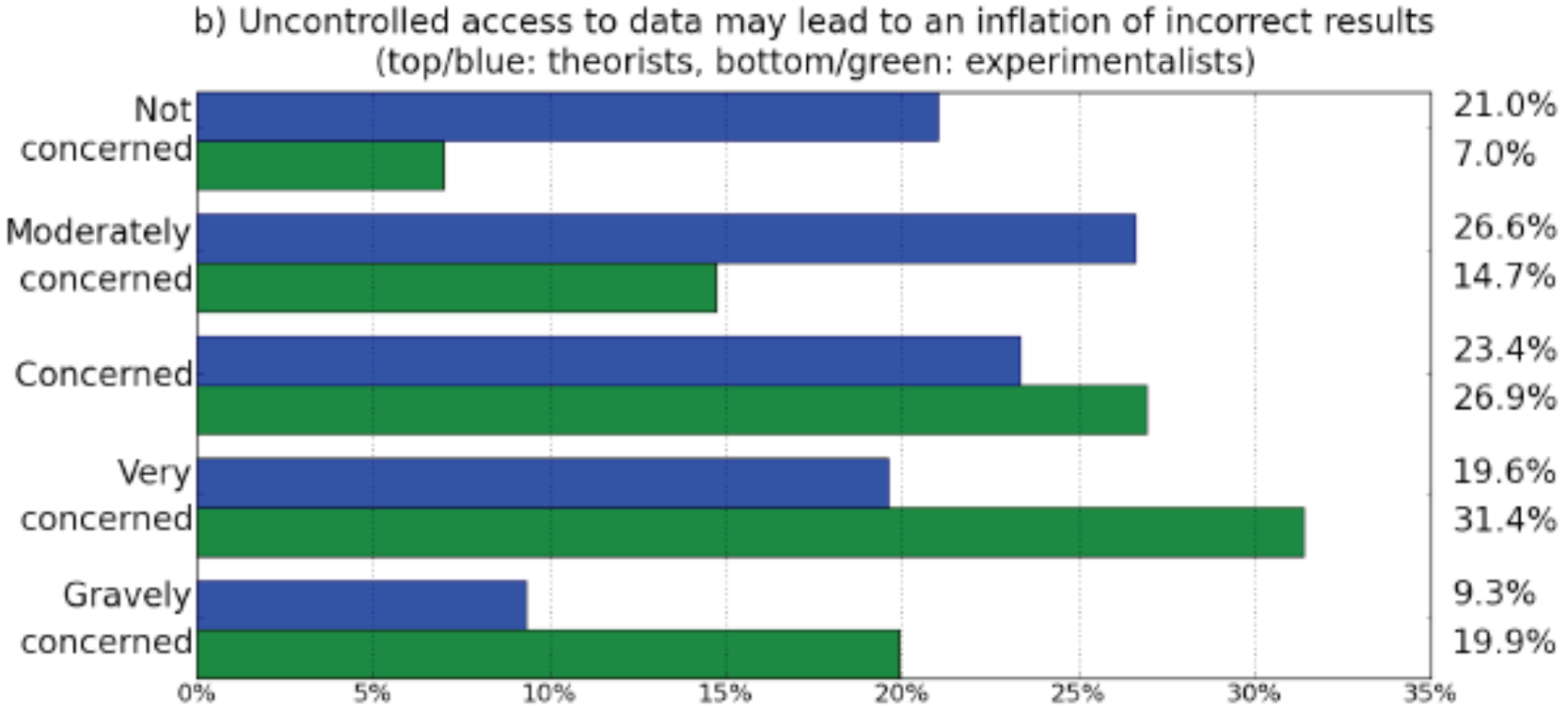
R. Brun





Risks of re-use?

Parse.insight



Governance issues are very important

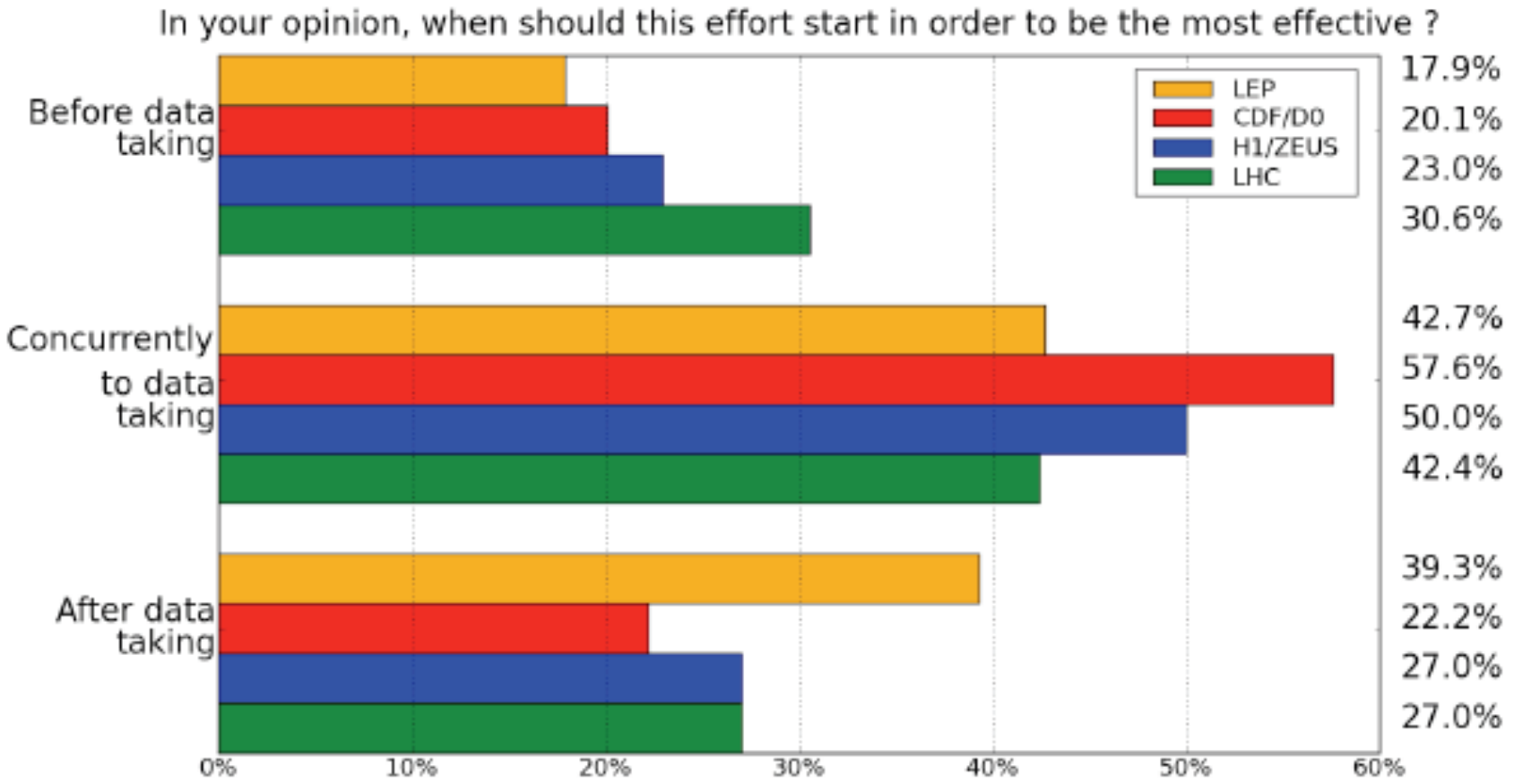
"Errors using inadequate data are much less than those using no data at all." Charles Babbage

Part 4: Governance

- HEP Collaborations function as international bodies with well defined policies over a few decades
  - A long term data management plan must include a solid governance solution
- 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

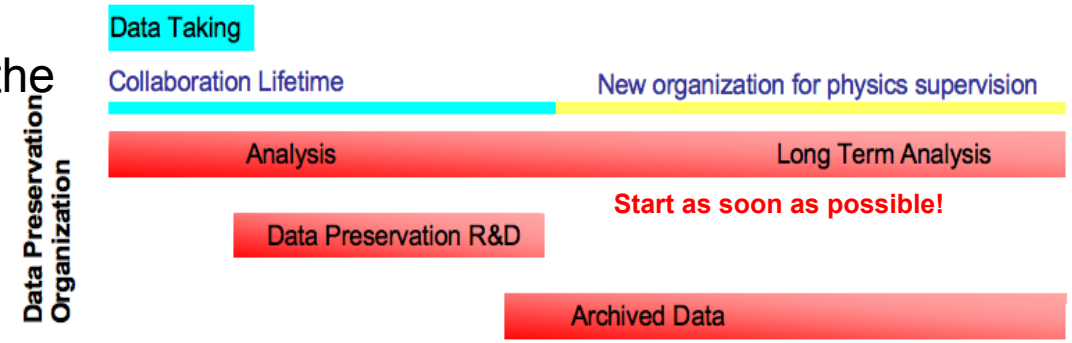
When is the best to start Data Preservation in HEP?

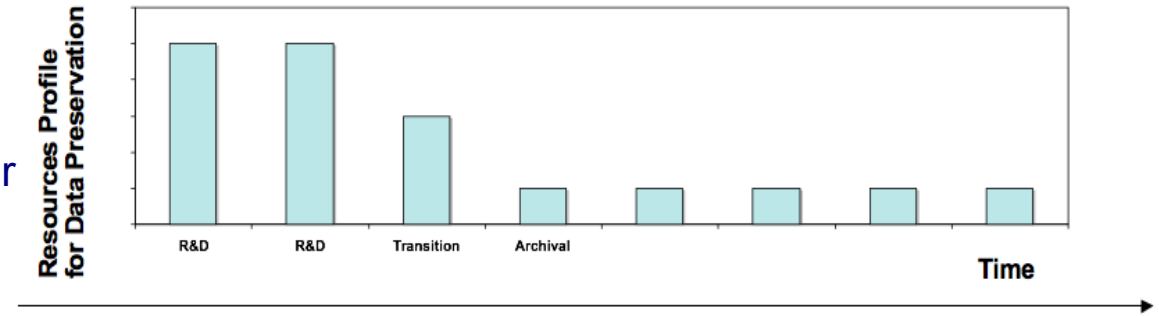
Parse.insight



Transition Scenario and Resources at Experimental Level

- Planning the transition to a long term analysis model
- R&D phase needed to develop the projects for the transition
- Data Archivist position
  - Long term custodianship of the physics data
- Resources / experiment
  - Typically a surge of 2-3 FTEs for 2-3 years, followed by steady 0.5-1.0 FTE per experiment/lab
  - This should be compared to 300-500 FTEs for many years / experiment!

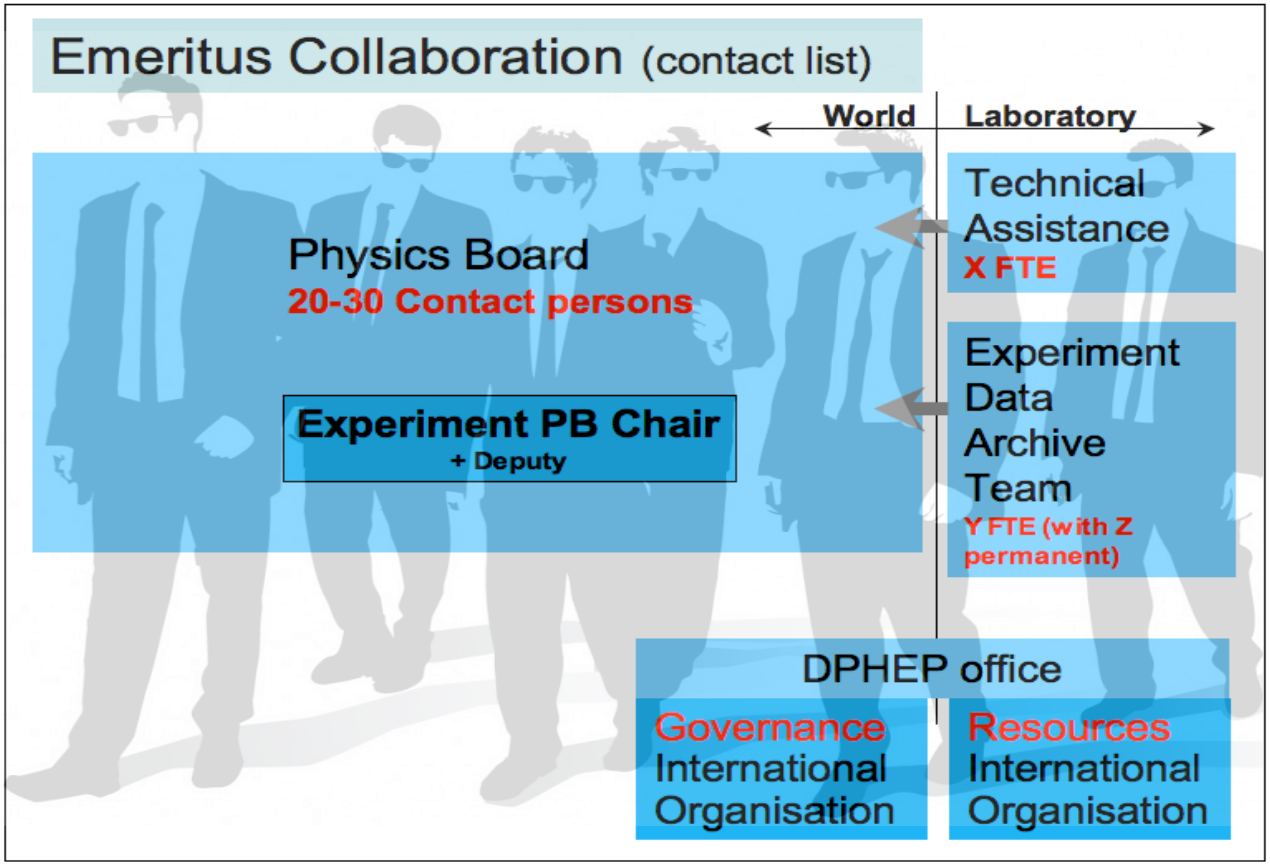




Cost estimates represent typically much less than 1% of the original investment

Long Term Governance

- Future structure of the collaboration should also be considered by HEP experiments
  - Experimental organisation risks being left in an undefined state
  - Transition should also be planned in advance of the projected end date



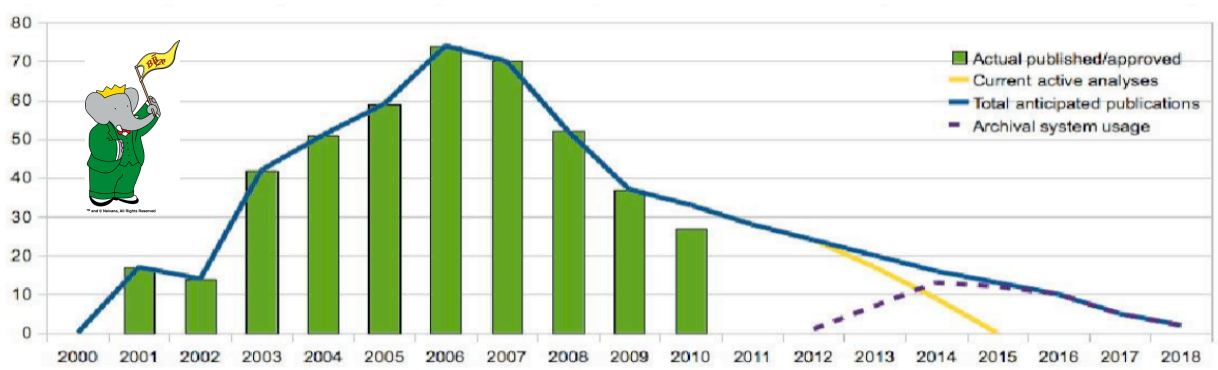
DPHEP Recommendations (end 2009)

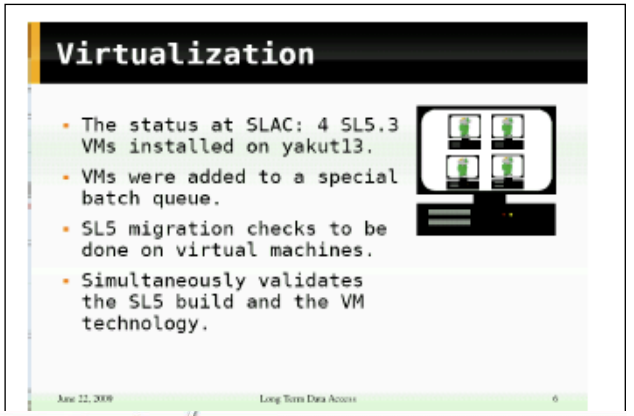
- An urgent and vigorous action is needed to ensure data preservation in HEP
- The preservation of the full analysis capability of experiments is recommended, including the preservation of reconstruction and simulation software
- An interface to the experiment know-how should be introduced: **data archivist**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

Progress in the last two years

- Data preservation plans/projects started by experiments
  - Technology and Organisation within the experiments
  - 2 dedicated projects funded (DESY, SLAC)
  - Expanding the capabilities of the collected data starts to be recognised as an cost-effective research
- Common multi-experiment projects start to be defined
  - Preservation technologies
  - Documentation
  - Outreach
- Communication to ICFA, evaluation bodies, community
  - Blueprint in preparation
- March 2011: LHC experiments joined DPHEP: **a new dynamics**

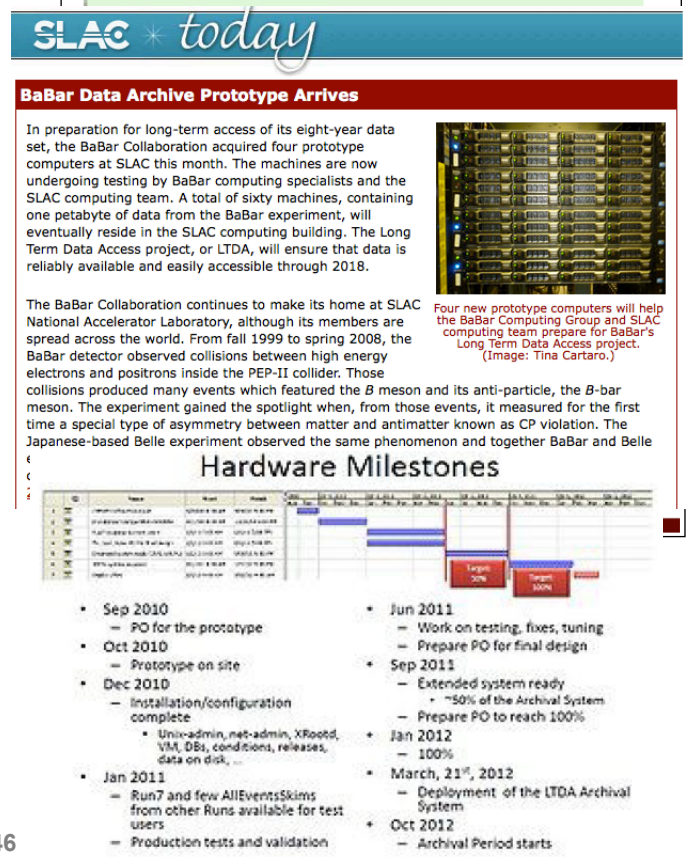
Data Preservation at BaBar





- BaBar moving to an "Archival Mode", preserving analysis ability beyond 2012
  - In a very advanced state
- Use of virtualisation and cloud computing

Resources for projects at BaBar taken into account in funding model during analysis phase !


SLAC today

BaBar Data Archive Prototype Arrives

In preparation for long-term access of its eight-year data set, the BaBar Collaboration acquired four prototype computers at SLAC this month. The machines are now undergoing testing by BaBar computing specialists and the SLAC computing team. A total of sixty machines, containing one petabyte of data from the BaBar experiment, will eventually reside in the SLAC computing building. The Long Term Data Access project, or LTDA, will ensure that data is reliably available and easily accessible through 2018.

Four new prototype computers will help the BaBar Computing Group and SLAC computing team prepare for BaBar's Long Term Data Access project. (Image: Tina Cartaro.)

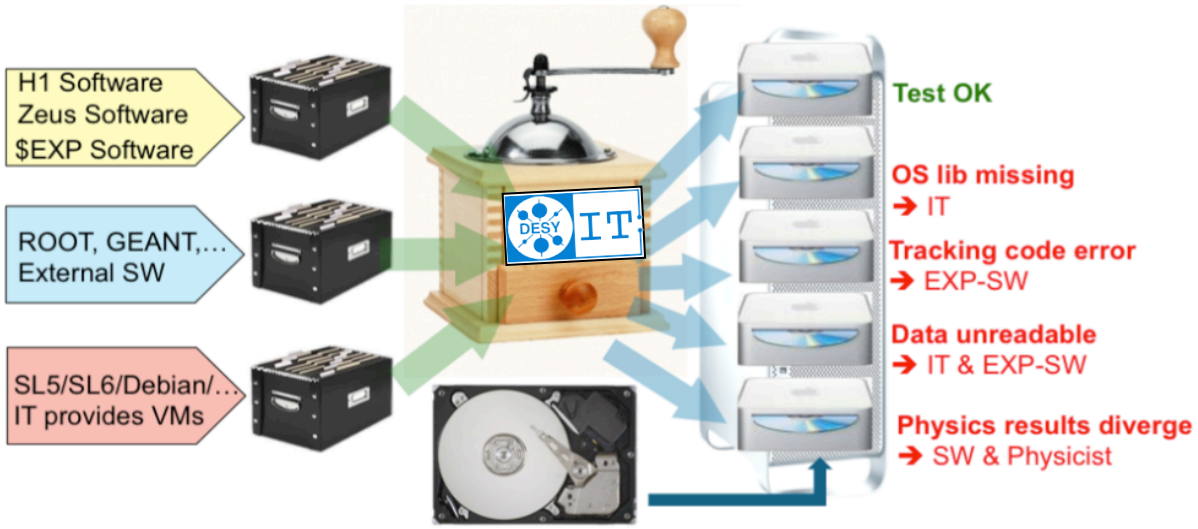
The BaBar Collaboration continues to make its home at SLAC National Accelerator Laboratory, although its members are spread across the world. From fall 1999 to spring 2008, the BaBar detector observed collisions between high energy electrons and positrons inside the PEP-II collider. Those collisions produced many events which featured the *B* meson and its anti-particle, the *B*-bar meson. The experiment gained the spotlight when, from those events, it measured for the first time a special type of asymmetry between matter and antimatter known as CP violation. The Japanese-based Belle experiment observed the same phenomenon and together BaBar and Belle

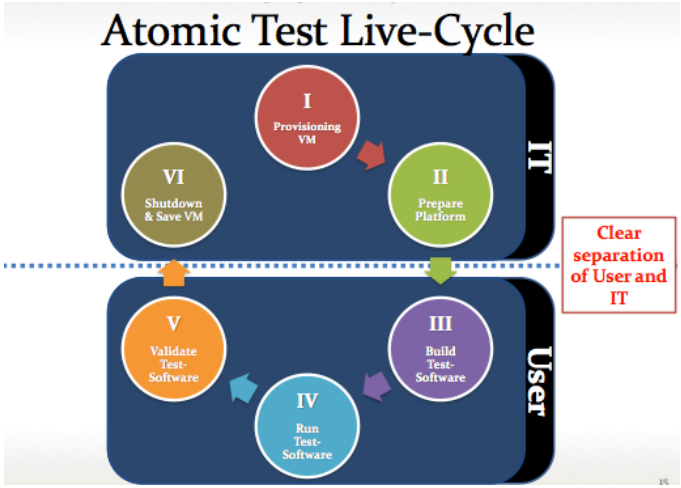
Hardware Milestones

- Sep 2010
  - PO for the prototype
- Oct 2010
  - Prototype on site
- Dec 2010
  - Installation/configuration complete
    - Unix-admin, net-admin, XRootd, VM, DBs, conditions, releases, data on disk, ...
- Jan 2011
  - Run7 and few AllEventsSkims from other Runs available for test users
  - Production tests and validation
- Jun 2011
  - Work on testing, fixes, tuning
  - Prepare PO for final design
- Sep 2011
  - Extended system ready
    - ~50% of the Archival System
  - Prepare PO to reach 100%
- Jan 2012
  - 100%
- March, 21st, 2012
  - Deployment of the LTDA Archival System
- Oct 2012
  - Archival Period starts

Towards a Generic Solution at DESY-IT

- Validation of experimental software using a virtual environment





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.

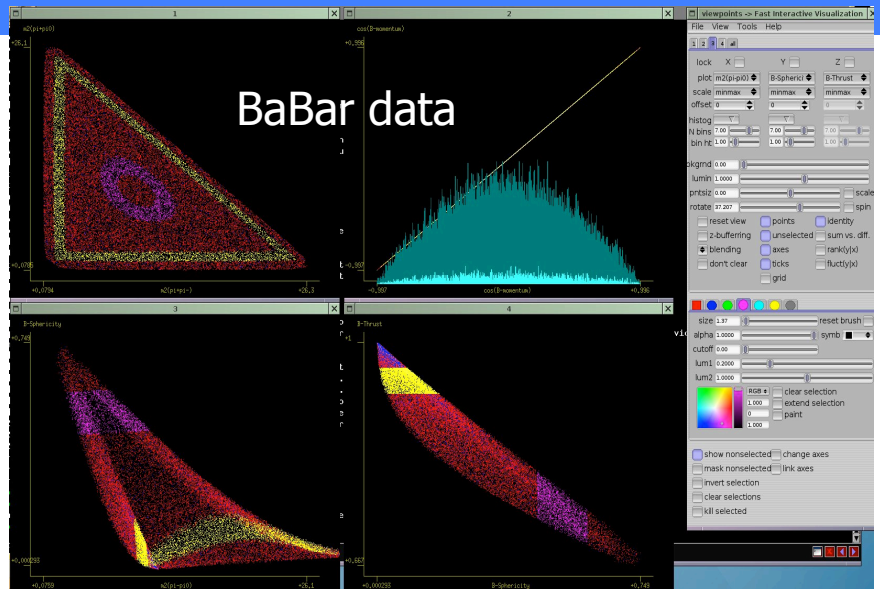
- Generic solution, for all HERA experiments: **validate the whole analysis chain**
  - Useful collaboration for future OS, external software transitions
  - Successful pilot project implemented, full project (people and $) now secured
  - Multi-center cooperation should be envisaged

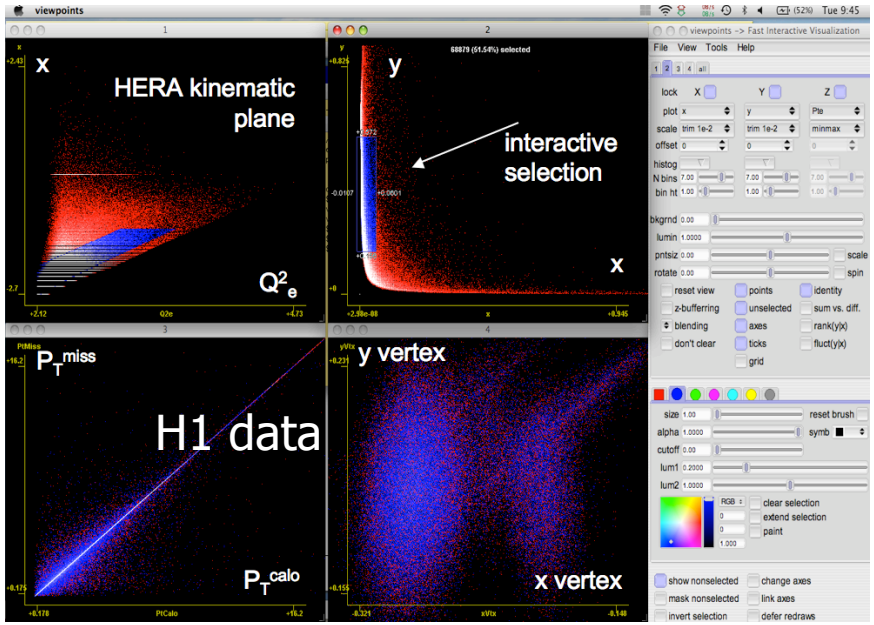
Outreach

- Many attractive outreach tools available, like *Viewpoints* from NASA
- Simple data format: input using text file of kinematics of HEP events

# Q2e	y	x	PtCalo	PtMiss	Ex	Ptx	Phix	Thx	Ee	Pte	Phie	The	Empz	xVtx
377.673	0.174	0.021	2.769	2.769	109.685	15.153	-11.780	8.231	26.226	17.665	164.138	137.656	58.542	-0.237
185.111	0.399	0.005	2.133	2.133	41.933	12.652	87.669	36.327	18.252	10.544	-93.948	144.713	57.878	-0.248
187.320	0.211	0.009	2.584	2.584	51.742	9.773	78.869	13.682	23.482	12.160	-106.349		148.813	55.164
264.266	0.508	0.005	0.238	0.360	35.343	11.738	-138.270		64.975	15.984	11.407	41.034	134.465	57.043
229.056	0.043	0.052	4.204	5.067	65.601	19.196	72.870	17.842	28.485	14.885	-98.351	148.685	58.941	-0.237
275.596	0.121	0.022	4.277	4.282	78.331	18.413	51.596	14.380	26.750	15.562	-139.235		144.425	55.018
240.102	0.183	0.013	3.513	3.434	67.134	17.402	85.049	17.201	24.719	14.004	-92.840	145.491	56.060	-0.266
451.996	0.209	0.021	1.723	1.723	49.126	17.196	66.018	24.927	25.936	18.913	-114.452		133.180	55.810
524.251	0.572	0.009	2.170	2.170	43.738	17.555	171.073	61.182	16.573	14.987	-11.341	115.274	58.543	-0.249
391.944	0.000	0.000	2.107	2.107	183.513	21.270	75.875	6.693	31.602	19.959	-108.713		140.834	58.375
201.600	0.212	0.009	4.441	4.441	44.890	17.098	-92.989	27.261	23.578	12.605	86.968	147.683	55.361	-0.243
335.881	0.052	0.064	16.769	16.769	29.256	1.142	-90.021	2.250	29.219	17.848	83.461	142.349	52.723	-0.242
286.039	0.009	0.315	2.514	2.514	194.560	18.922	-83.365	5.616	29.944	16.837	92.126	145.787	56.826	-0.254
207.703	0.137	0.015	8.095	8.095	84.993	21.487	82.549	15.129	25.701	13.389	-95.886	148.605	53.237	-0.258
387.371	0.358	0.011	1.272	1.272	70.266	15.460	-91.456	17.173	21.232	15.772	93.071	132.027	55.684	-0.236
855.333	0.589	0.017	2.588	2.588	88.511	23.066	70.622	21.191	21.306	20.499	-110.759		105.828	56.196
154.527	0.667	0.002	3.589	3.589	72.273	8.810	174.450	101.604	10.598	7.176	-28.067	137.379	92.478	-0.240
304.756	0.025	0.121	1.622	1.622	128.020	17.765	-145.756		8.522	29.678	17.240	39.272	144.486	55.298
278.950	0.627	0.004	3.813	0.726	37.163	9.588	124.342	60.056	12.831	10.205	-53.435	127.311	52.247	-0.243
456.769	0.204	0.022	1.621	1.568	48.542	18.231	-134.045		26.029	26.095	19.063	50.043	133.070	55.126
275.593	0.050	0.055	0.296	0.296	39.621	16.300	-71.596	25.608	28.728	16.184	109.363	145.712	58.077	-0.249
890.585	0.853	0.010	7.314	4.893	193.920	16.386	173.467	6.104	12.119	11.435	-5.872	70.653	48.847	-0.248
328.267	0.018	0.185	2.294	2.294	200.323	16.451	-16.314	4.712	30.090	17.959	169.425	143.356	54.999	-0.250
156.729	0.466	0.003	2.340	1.529	64.710	10.664	-88.089	17.271	16.150	9.146	93.350	145.507	59.919	-0.257
270.064	0.025	0.105	8.909	8.909	304.993	24.646	-177.466		4.653	29.349	16.225	10.869	146.439	56.009

- Nice example of new collaborative work between BaBar and H1 via DPHEP
- Discussions about common formats ongoing
  - B-lab (KEK) example considered

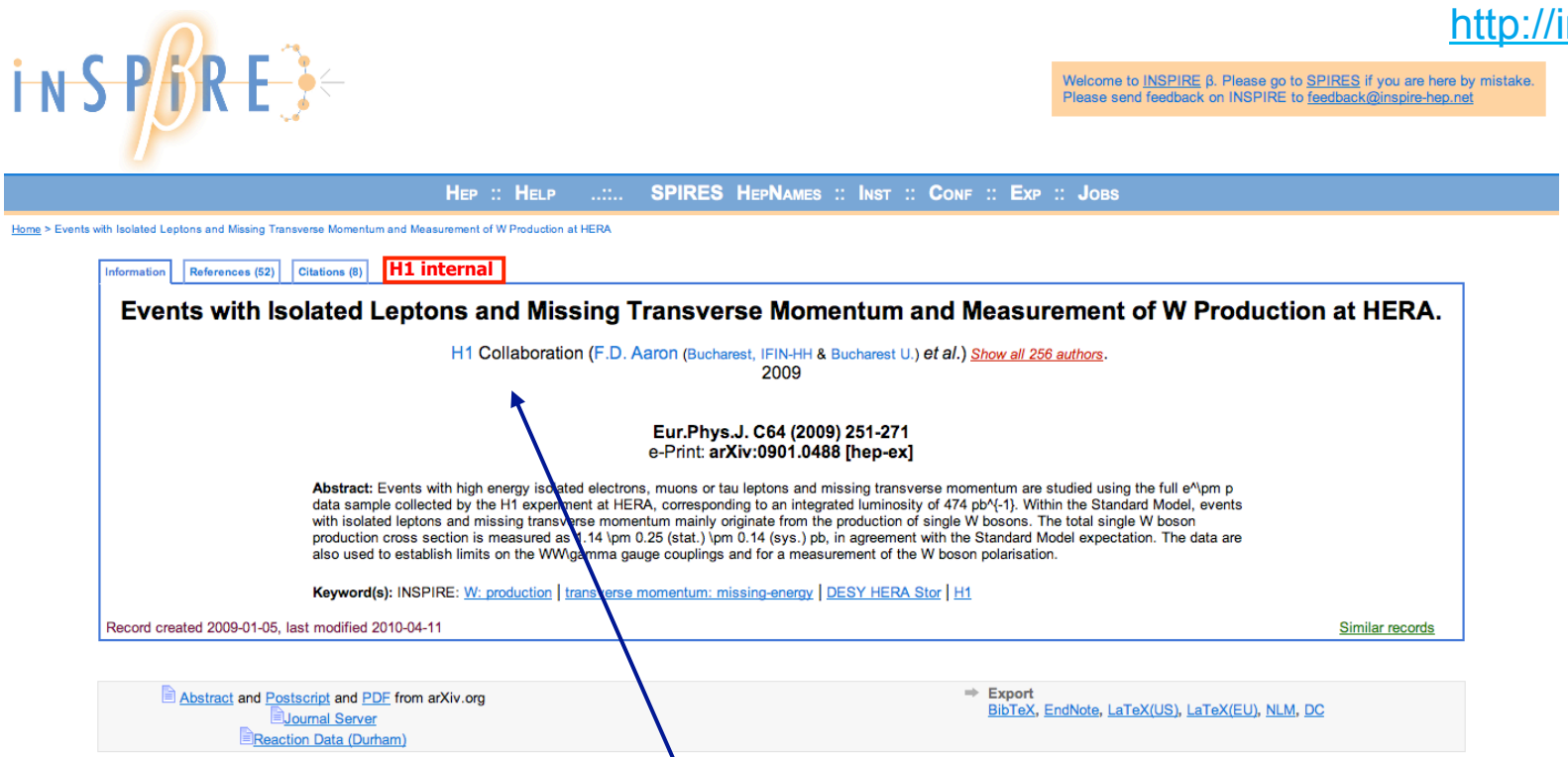




http://www.stanford.edu/group/burchat/cgi-bin/bellis_mediawiki/index.php/Viewpoints_NASA

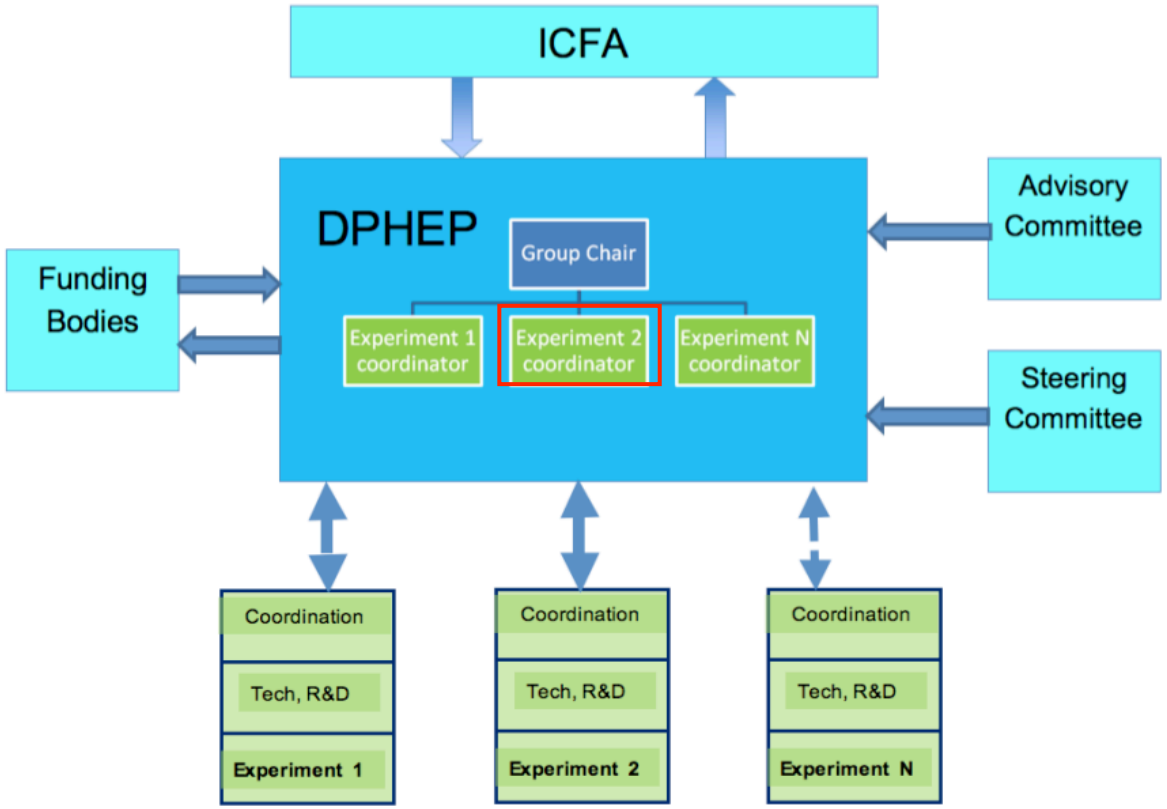
Common project on documentation: INSPIRE

http://inspirebeta.net/



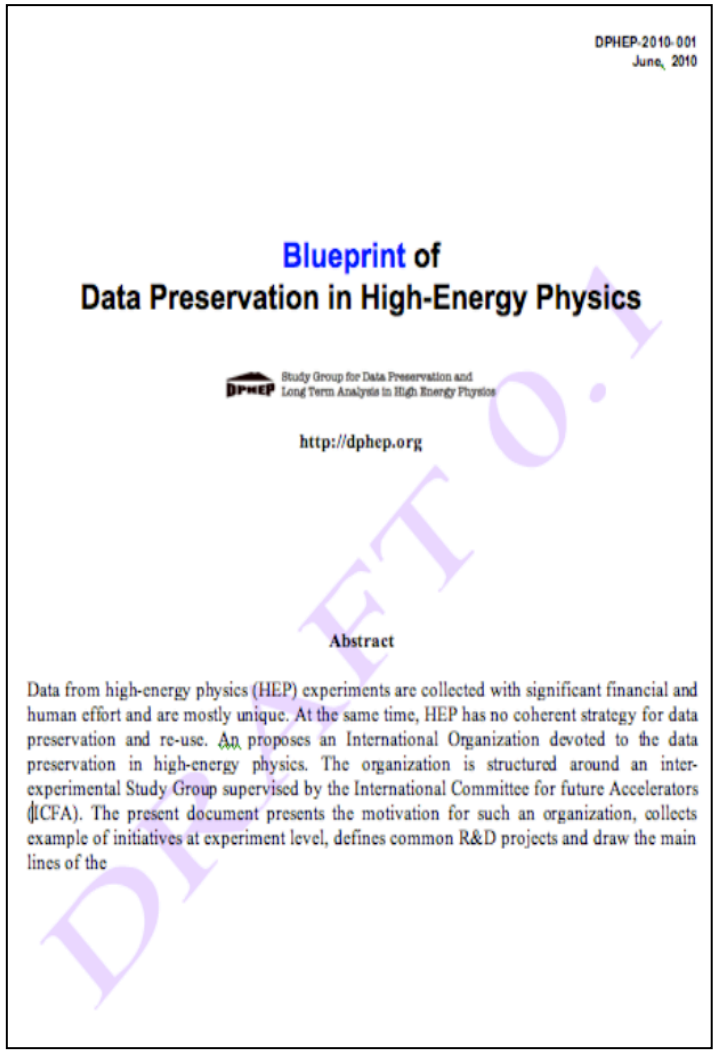
- Envisage an additional link for the collaboration members only
- Provides additional information (notes, slides etc.)
- Reduced data and macros also possible

DPHEP Organisation



- Support expressed by major laboratories and committies: ICFA, HEPAP, FALC
- Funding plan in preparation: a managing director to be installed, contributions from laboratories expected

Blueprint for Data Preservation in HEP in Preparation



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

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

- More details and specifics, focussing on:
  - An expanded physics case for data preservation
  - A description of the existing experiment-level and lab-level strategies for data analysis and preservation
  - Global transverse data preservation projects, such as outreach, validation, documentation
  - Future international coordination of activities
- Cost estimates also included
- Should provide a skeleton for local, national and international proposals, for past an future experiments
  - LEP!
  - Other branches of HEP (HI, neutrinos etc.)

From the executive summary

- In summary, the DPHEP Study Group identified the following priorities, in order of urgency:

- **Priority 1: Experiment Level Projects in Data Preservation.** Large laboratories should define and install data preservation projects in order to avoid catastrophic loss of data once major collaborations come to an end. The recent expertise gained during the last 18 months indicate that an extension of the computing effort within experiments with **a person power of the order of 2-3 FTEs** leads to a significant improvement in the ability to move into a long-term data preservation phase. Such initiatives exist already or are being defined in the participating laboratories and are followed attentively by the Study Group.

- **Priority 2: International Organisation DPHEP.** The efforts are best exploited by a common organisation at the international level. The installation of this body, already prefigured by the ICFA Study Group, requires **a Project Manager (1 FTE**) to be employed as soon as possible. The effort is a joint request of the Study Group and could be assumed by rotation among the participating laboratories.

- **Priority 3: Common R&D projects.** Common requirements on data preservation are likely to evolve into inter-experimental R&D projects (three concrete examples are given above**, each involving 1-2 dedicated FTE, across several laboratories**). The projects will optimise the development effort and have the potential to improve the degree of standardisation in HEP computing in the longer term. Concrete requests will be formulated and the activity of these projects will be steered by the DPHEP organisation.

  - These priorities could be enacted with **a funding model** implying contributions from the three regions (Europe, America, Asia) and strong connections with laboratories hosting the data samples.

Conclusion and Outlook

- HEP data are mostly unique and have true scientific potential
- 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
- 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: DPHEP
- Blueprint for Data Preservation is on the way
  - **This workshop is an important milestone**

The reflection on Data Preservation in HEP has only started

Daniel Burnham

- But: « Make no little plans »