

# Innovation in scholarly communication: Vision and projects from High-Energy Physics

**Rolf-Dieter Heuer**

DESY - Research Director HEP

CERN - Director-General Elect



APE2008

Berlin - January 22-23 2008

# Outline

- Introduction
- High Energy Physics as a case study
  - The Publishing Landscape in HEP
- Open Access
  - SCOAP<sup>3</sup>: A New Publishing Model
- What's on a scientist's mind?
  - Future HEP information systems
- The next frontier
  - (Open) Access to (usable) data

# Introduction



Progress in information technology and evolving needs within the scientific community drive changes in scholarly communication

# Introduction



- We need
- access to (comprehensive) information
- quality assurance
- reasonable costs
- state-of-the-art information tools

High Energy Physics ideal testbed for innovations

- driving force in information management
- long history in Open Access

# High-Energy Physics (or Particle Physics)

"What is the world made of?" & "What holds it together?"

HEP aims to understand how our Universe works:

- discover the constituents of matter and energy
- probe their interactions
- explore the basic nature of space and time

Experimental HEP

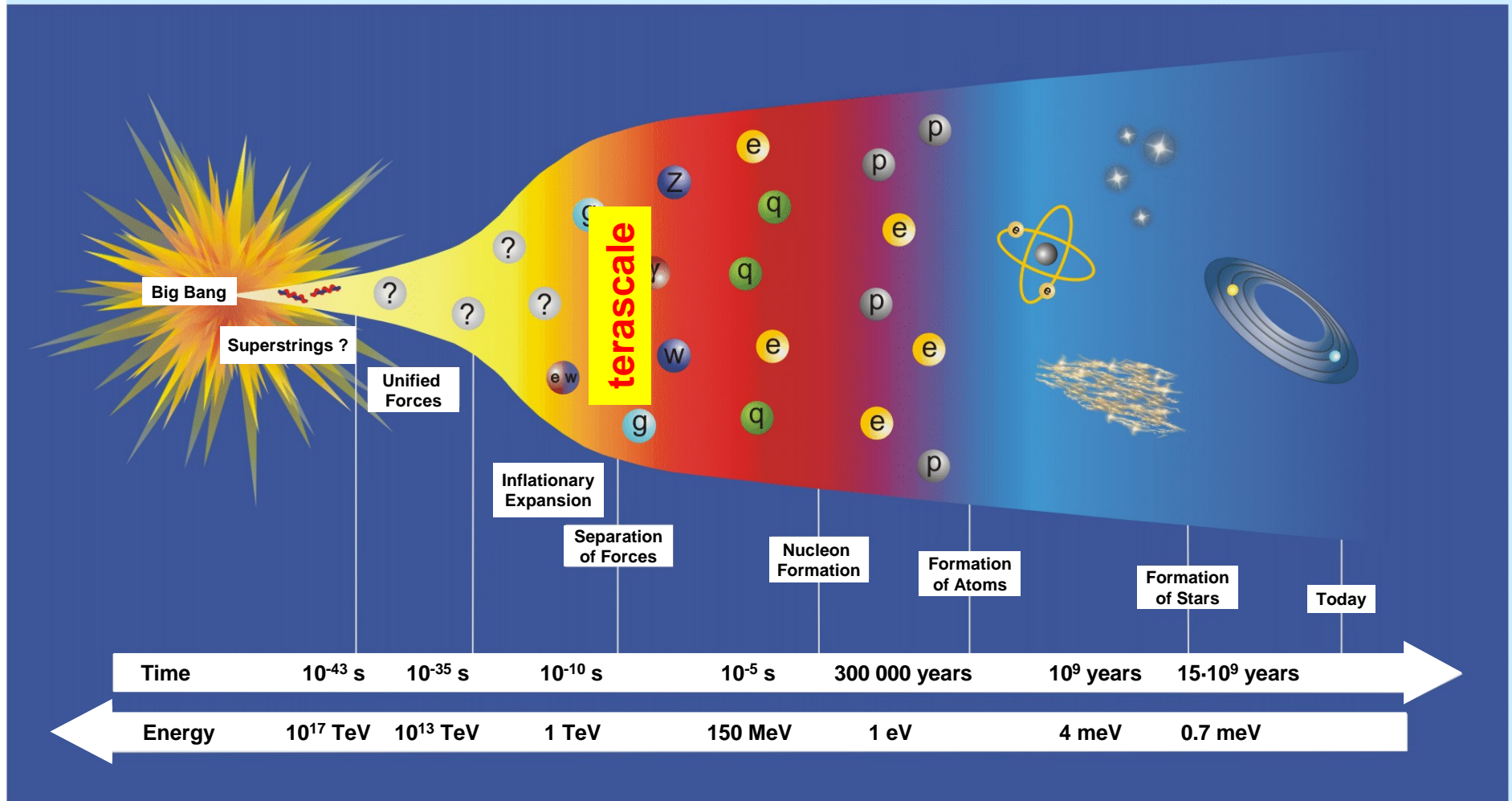
builds the largest scientific instruments ever to reach energy densities close to the Big Bang  
(Half of the community, 20% of literature)

Theoretical HEP

predicts and interprets the observed phenomena  
(Half of the community, 80% of literature)

# Vision

- Revolutionary advances in understanding the microcosm
- Connect microcosm with early Universe



Particle Physics at the **Energy Frontier** with highest collision energies ever will change our view of the universe

# DESY: Deutsches Elektronen-Synchrotron (since 1959)

- one of the leading accelerator centers worldwide
- development of large accelerator facilities for both particle physics and research with photons
- 1800 staff
- 3000 guests per year from 45 countries
- discovery of the gluon (carrier of the strong force) in 1979  
→ EPS price
- development of superconducting (TESLA) technology for European XFEL and International Linear Collider ILC
- leading member of the Germany-wide Helmholtz Alliance 'Physics at the Terascale'
- SPIRES literature database (SLAC / DESY/ Fermilab)



# CERN: European Organization for Nuclear Research (since 1954)

- The world leading HEP laboratory, Geneva (CH)
- 2500 staff (mostly engineers)
- 9000 users from all over the world (mostly physicists)
- 3 Nobel prizes (Accelerators, Detectors, Discoveries)

- Invented the web



- Commissioning the 27-km (6000 M€) LHC accelerator
- Runs a 1-million objects Digital Library

The CERN Convention (1953) contains what is effectively an early Open Access manifesto:

“... the results of its experimental and theoretical work shall be published or otherwise made generally available”



# High Energy Physics as a case study

The Publishing Landscape in HEP

# The HEP “preprint culture”

- In the '60s HEP scientists not willing to wait ~1 year for their articles to reach their peers through journals
- *Preprints* became main vehicle of information in HEP
- Mass mailing of hard-copies:  
*Ante-litteram Open Access* paid by big Institutes  
(DESY costs: ~close to 1MDM/year)
- HEP libraries classify preprints received worldwide
  - HEP Index published biweekly by DESY 1963 - 1996

L. Goldschmidt-Clermont, 1965,

[http://eprints.rclis.org/archive/00000445/02/communication\\_patterns.pdf](http://eprints.rclis.org/archive/00000445/02/communication_patterns.pdf)

L. Addis, 2002,

<http://www.slac.stanford.edu/spires/papers/history.html>

# The HEP “preprint culture”

## Revolution #1: '70s

IT starts to meet libraries

SPIRES (1974): e-catalogue of preprint and publications

## Revolution #2: '90s

HEP preprints and Internet indissolubly linked

arXiv (full-text server) by Paul Ginsparg at LANL in 1991

## Revolution #3: '91

the web by Tim Berners-Lee at CERN

First U.S. WWW server at SLAC in '91 to access SPIRES

Summer 1992, SPIRES links to the arXiv for full-texts

SPIRES now contains metadata for >750 000 HEP articles,  
adding ~4500 records every month

arXiv has about 450 000 full-texts,  
adding ~5000 new articles every month

# In the era of electronic journals, the “preprint-culture” lives on

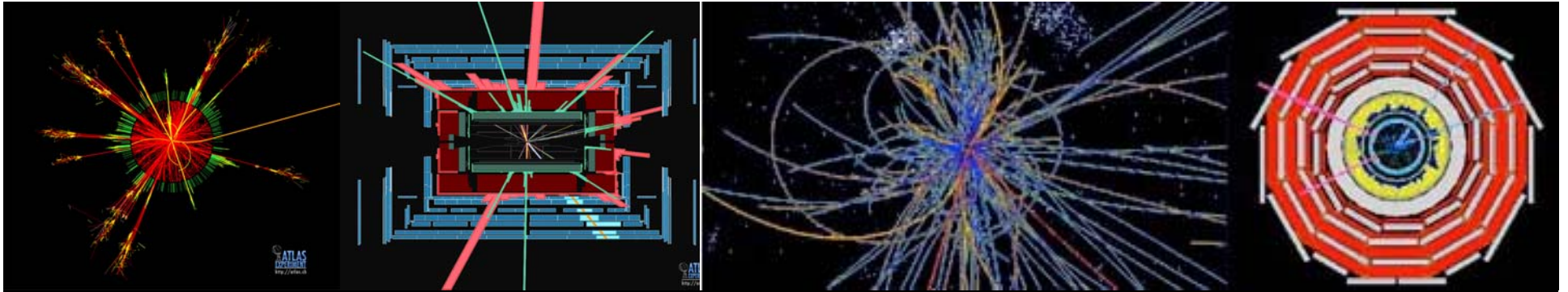


CERN circa 2005

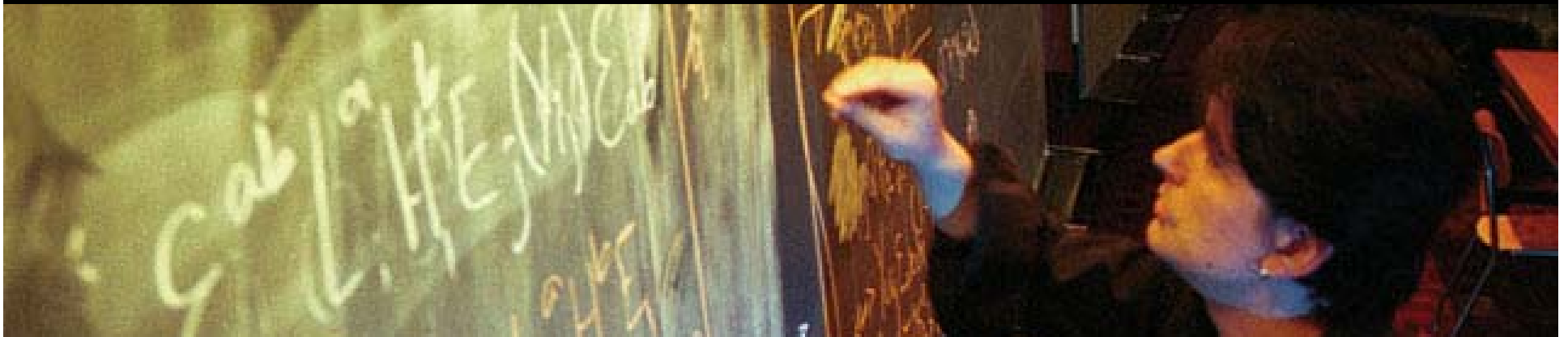
## But can journals survive?

# HEP and its journals

- Journals are losing their century-old role as vehicles of scholarly communication.
  - Still, evaluation of institutes and (young) researchers is based on prestigious peer-reviewed journals.
  - The main role of journals is to assure high-quality peer-review and act as keepers-of-the-records
  - The HEP community needs high-quality journals, our “interface with officialdom”
  - As an “all-arXiv discipline” HEP is at risk to see its libraries cancel important journals due to spiraling subscription costs.
  - Prestigious HEP journals are in danger of losing their sustainability.
- new business model combining OA and sustainability



**Open Access:**  
Grant anybody, anywhere and anytime  
access to the (peer-reviewed)  
results of (publicly-funded) research

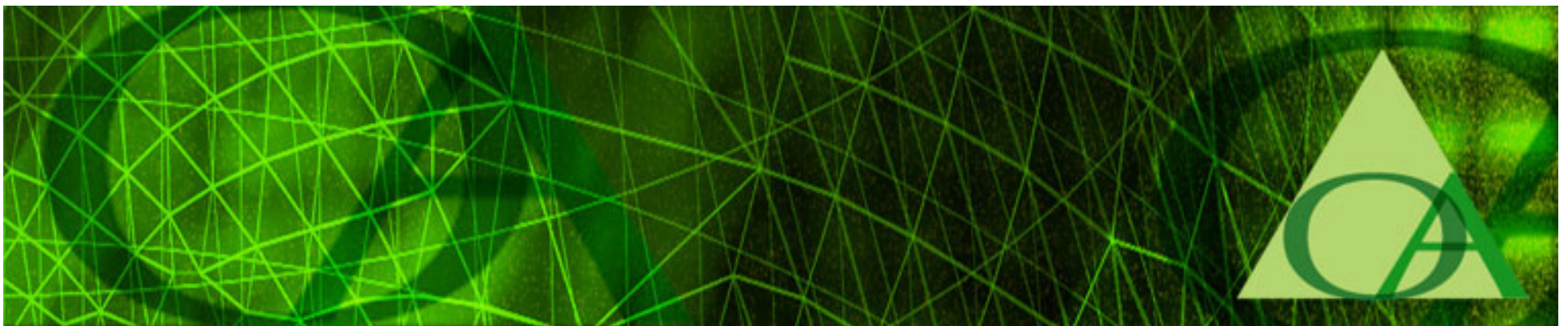


# HEP and Open Access: a synergy

- HEP is decades ahead in thinking Open Access:
  - Mountains of paper preprints shipped all over the world by HEP institutes for 40 years (at author/institute expenses!)
  - HEP launched arXiv (1991), the archetypal Open Archive
  - The first free peer-reviewed electronic HEP journals:
    - *Journal of High Energy Physics* (1997)
    - *Physical Review Special Topics Accelerators and Beams* (1998)
- Small and connected community (<20000 scientists)
- Small number of articles (<10000)
- Small publishing landscape (< 10 journals)
- Reader and author communities largely overlap
- Open Access, second nature: posting on arXiv before even submitting to a journal is common practice.
  - No mandate, no debate. Author-driven. Evident benefits
  - Revised version post peer-review routinely uploaded

# HEP and Open Access

After preprints, arXiv and the web,  
Open Access journals  
are the natural evolution of  
HEP scholarly communication





# Is it all about vocal librarians? Strong support from the LHC collaborations

*"We, the   \* Collaboration, strongly encourage the usage of electronic publishing methods for   \* publications and support the principles of Open Access Publishing, which includes granting free access of our   \* publications to all. Furthermore, we encourage all   \* members to publish papers in easily accessible journals, following the principles of the Open Access Paradigm."*

5400 scientists  
building the largest  
scientific instruments ever

  \*  
} ATLAS; approved on 23rd February 2007  
} CMS; approved on 2nd March 2007  
} ALICE; approved on 9th March 2007  
} LHCb; approved on 12th March 2007

*"The Strategic Helmholtz Alliance 'Physics at the Terascale' fully supports the goal of SCOAP3 of free and unrestricted electronic access to peer-reviewed journal literature in particle physics . . . Will benefit scientists, authors, funding agencies and publishers alike. Unrestricted access to published scientific results is essential for wide dissemination and efficient usage of scientific knowledge, . . . raising awareness on open-access publishing in their communities and encourage their authors to publish in open-access journals."*

The Alliance is a German network comprising 17 universities, 2 Helmholtz institutes and 1 Max Planck institute. *Theorists, experimentalists, computing and accelerator scientists*

# The 2832<sup>nd</sup> EU Competitiveness Council

[http://www.consilium.europa.eu/ueDocs/cms\\_Data/docs/pressData/en/intm/97236.pdf](http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf)

"[The EU Council] recognizes the strategic importance for Europe's scientific development of current initiatives to develop sustainable models for open access [...]" and "underlines the importance of effective collaboration between different actors, including funding agencies, researchers, research institutions and scientific publishers, in relation to access [... to], scientific publications [...]". It "invites Member States to enhance the co-ordination between Member States, large research institutions and funding bodies on access [...] policies and practices"

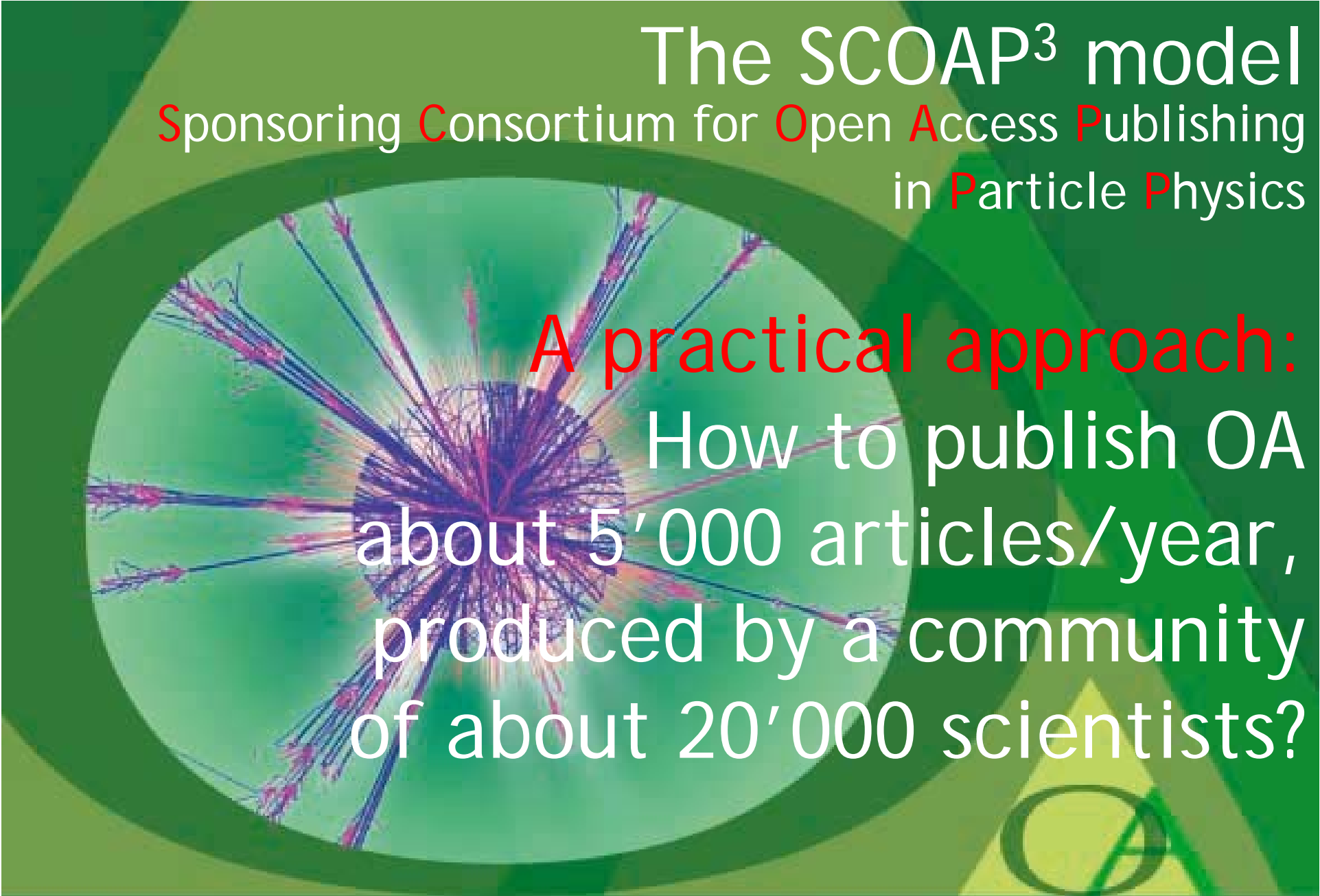
**These principles are precisely the pillars of the SCOAP<sup>3</sup> model**



# SCOAP<sup>3</sup>

## The next step for Open Access

- goals
- organization
- funding



# The SCOAP<sup>3</sup> model

Sponsoring Consortium for Open Access Publishing  
in Particle Physics

**A practical approach:**  
How to publish OA  
about 5'000 articles/year,  
produced by a community  
of about 20'000 scientists?

<http://scoap3.org/files/Scoap3ExecutiveSummary.pdf>  
<http://scoap3.org/files/Scoap3WPReport.pdf>

# SCOAP<sup>3</sup> in one sentence

A consortium sponsors HEP publications and makes them OA by re-directing subscription money.

Today: (funding bodies through) libraries buy journal subscriptions to support the peer-review service and to allow their patrons to read articles.

Tomorrow: funding bodies and libraries contribute to the consortium, which pays centrally for the peer-review service. Articles free to read for everyone.

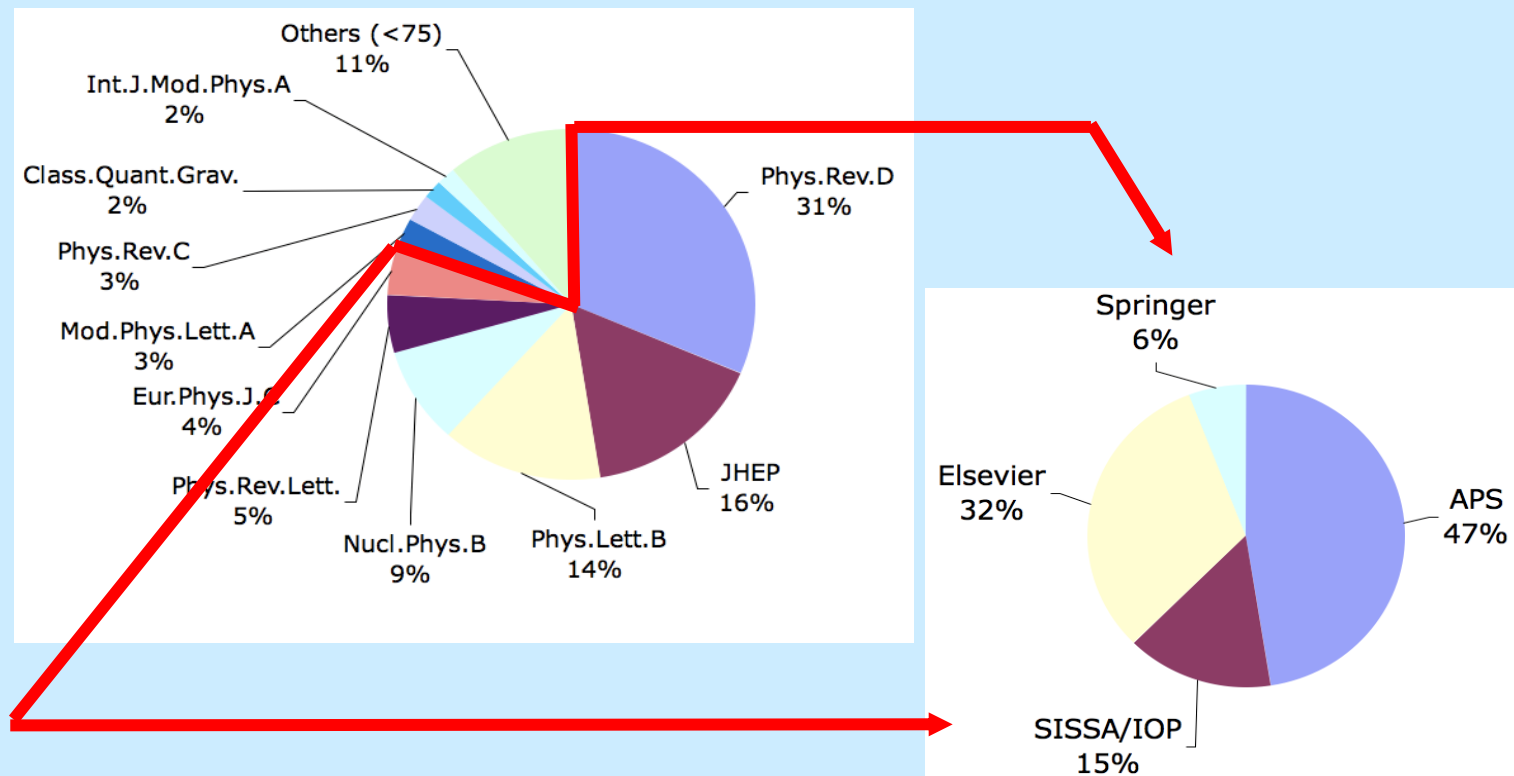
*Visit [scoap3.org](http://scoap3.org)*



# Potential initial partners of SCOAP<sup>3</sup>

Journals where HEP researchers mostly publish today

6 journals with mainly HEP content  
+ 2 important mixed journals (PRL, NIMA)  
from 4 publishers: APS, Elsevier, SISSA/IOP, Springer  
cover ~80% of HEP literature



# Guesstimating the budget envelope

- *Physical Review D* (APS) operates with **2.7M€/year** (31% of arXiv:hep)
- *Journal of High Energy Physics* (SISSA/IOP) needs **~1M€/year** (19% of arXiv:hep)

**HEP Open Access price tag: 10M€/year**

- A published PRD article costs APS **~1500€**
- 6-8 leading journals publish 5000-7000 articles a year





# How to organize this?

40 funding agencies

ATLAS

400 M€  
(Excluding person-power)

1000 contracts

O(50) funding bodies

10 M€/a

SCOAP<sup>3</sup>

O(10) contracts  
with publishers

HEP is used to large collaborations

It works already on a much bigger scale

Establish OA with the same structure

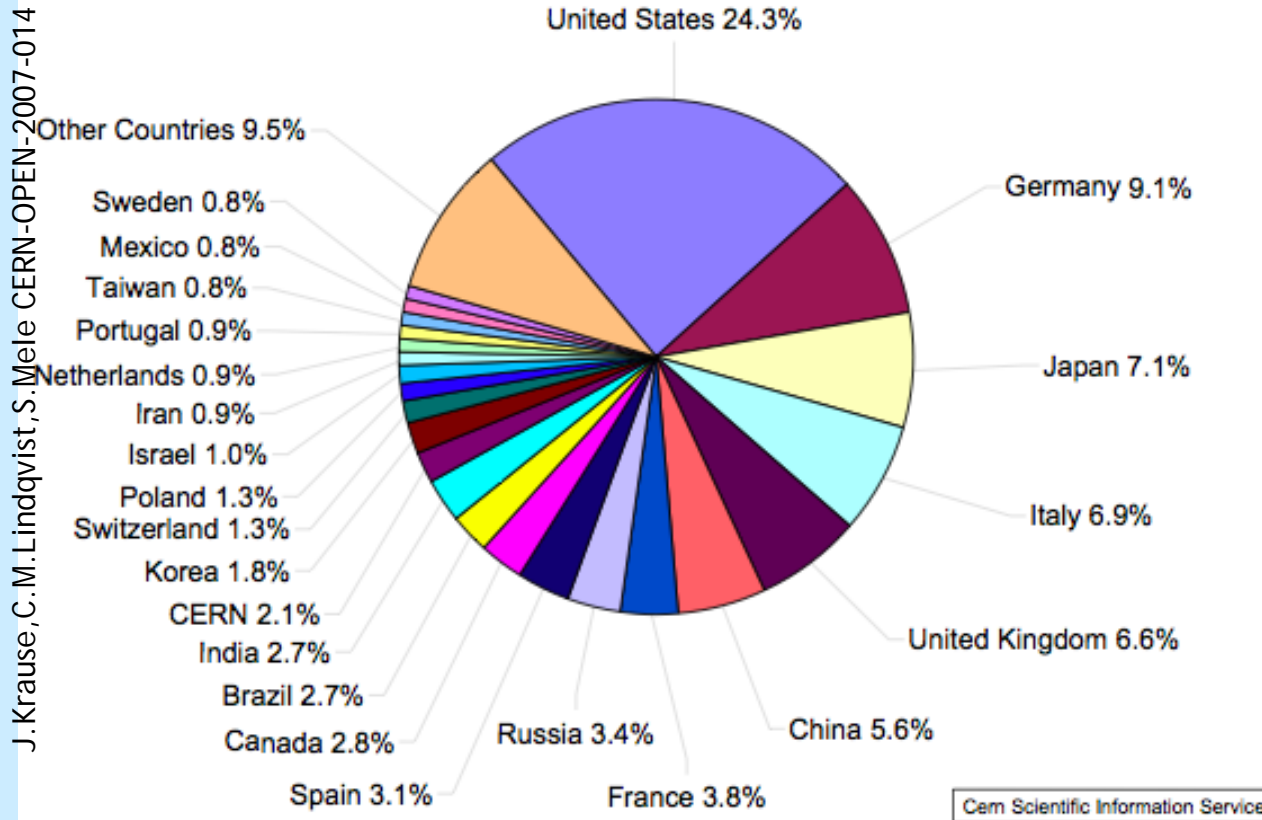
# SCOAP<sup>3</sup> fund-raising

- SCOAP<sup>3</sup> financing to be distributed according to a “fair-share” model based on the distribution of HEP articles per country, accounting for co-authorship.
- Make a 10% allowance for developing countries who at the beginning might not contribute to the scheme.
- Once a sizeable fraction of budget is pledged send a tender to publishers and determine final budget
- **The model is viable only if every country is on board!** Allowing only SCOAP<sup>3</sup> partners to publish Open Access simply replicates the subscription scheme.
- **Goal: SCOAP<sup>3</sup> operational for the first LHC articles!**



# SCOAP<sup>3</sup> fund-raising

Distribution of HEP articles by country, average 2005-2006

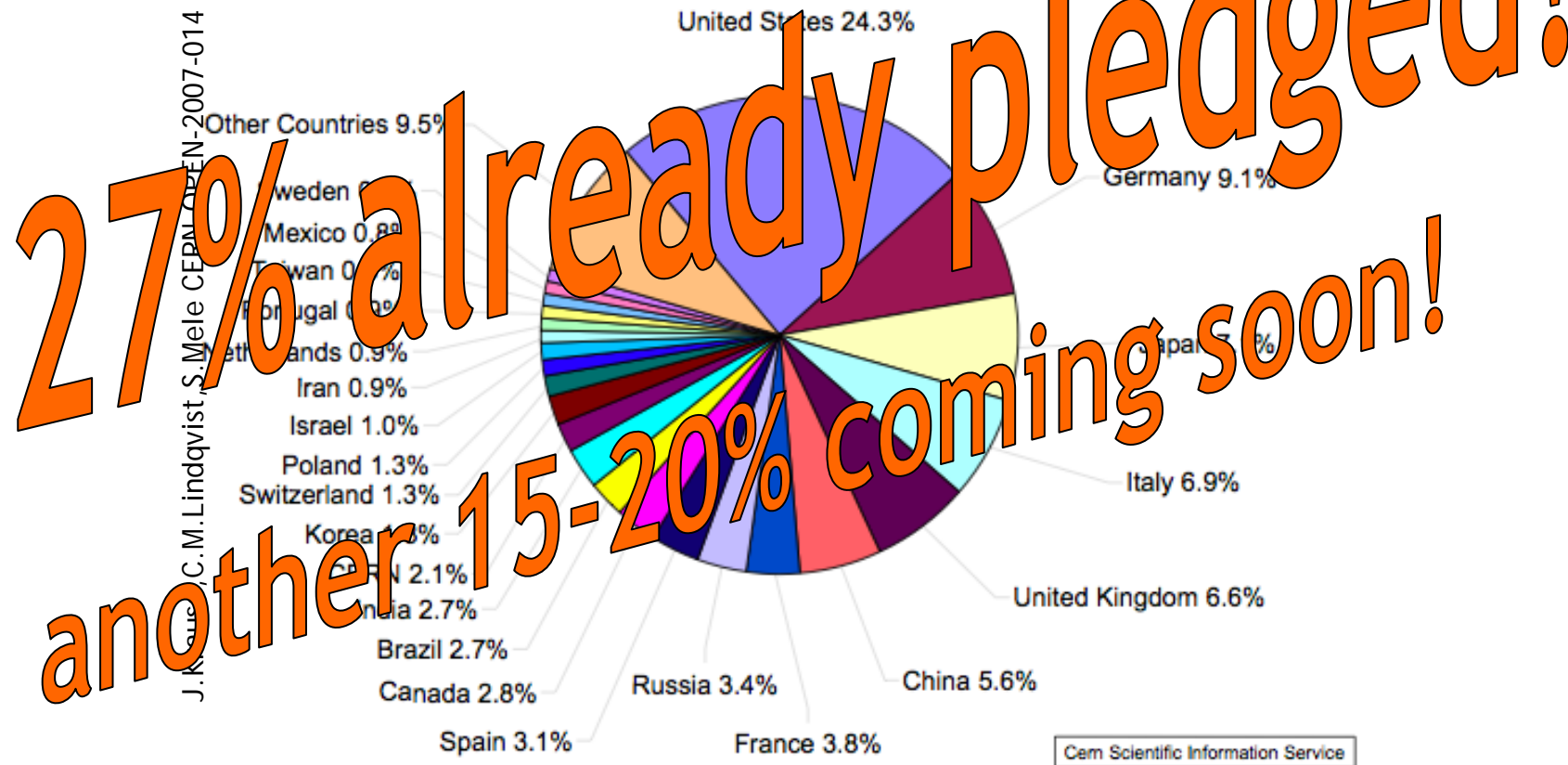


J. Krause, C. M. Lindqvist, S. Mele CERN-OPEN-2007-014

Germany, France, Italy, Greece, CERN, Sweden, Slovakia, Denmark, Norway, Austria have already joined. Most European countries expected to join soon. Intense discussions in Asia and the Americas. Leading US libraries signing up.

# SCOAP<sup>3</sup> fund-raising

Distribution of HEP articles by country, average 2005-2006



Germany, France, Italy, Greece, CERN, Sweden, Slovakia, Denmark, Norway, Austria have already joined. Most European countries expected to join soon. Intense discussions in Asia and the Americas. Leading US libraries signing up.

# SCOAP<sup>3</sup> in a nutshell

- Establish Open Access in HEP publishing in a transparent way for authors.
- Convert existing high-quality peer-reviewed journals to Open Access, in a sustainable way.
- Operate along the blueprint of large scientific collaborations.
- Price tag of 10M€/year to be shared according to the distribution of HEP articles per country.
- 27% of the budget has been pledged in a few months! Another 20% coming soon.
- The model has high potential but is only viable if every country contributing to HEP is on board!
- Our model could be rapidly generalized to fields with similarly tightly-knit communities.

# What's on a scientist's mind?

## Future HEP information systems

- needs
- wishes
- possibilities

# Time for a modern e-infrastructure

Preprints stay main HEP communication channel, "*just*" submission and search have evolved  
Still primitive text-mining

today

tomorrow

But what about

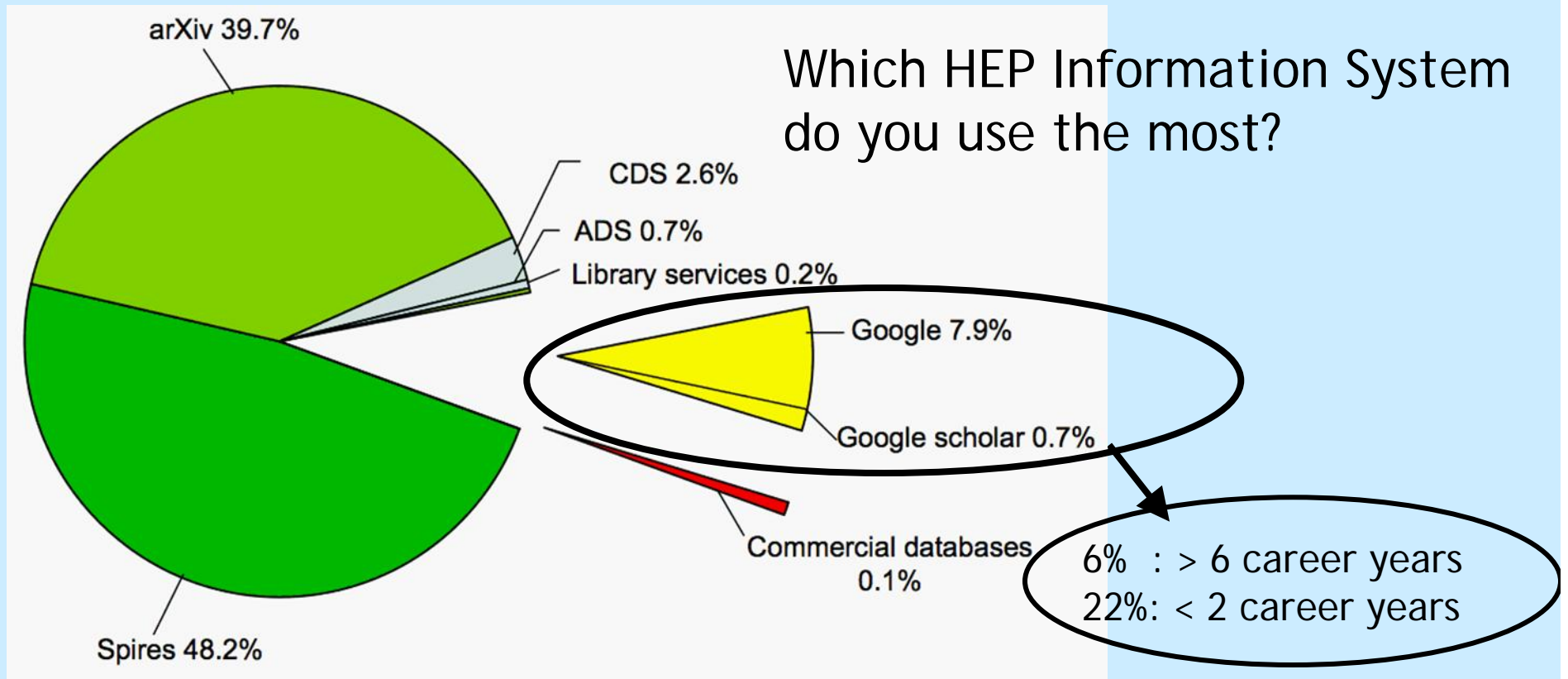
- conference slides ?
- searching tables and plots ?
- aggregating all instances (slides, proceedings, preprint, article, data) ?

Complex needs → modern e-infrastructure

# Information search in HEP

A poll of the HEP community

>2000 answers (10% of the community!)



## 91 % Community services

- 40 % Subject repositories
- 51 % Lab-supported databases

## 9% Google

## <0.1% Commercial services



# SPIRES & arXiv

SPIRES database @ SLAC  
since 1974 (ftp-server)  
1991 first US-www server

HEP-Content:

- bibliographic information
- standardized keywords
- links to full-text
- match journals/preprints
- citation analysis

Input from SLAC, Fermilab  
and DESY (former HEP-Index)

arXiv @ LANL now  
@ Cornell University  
since 1991

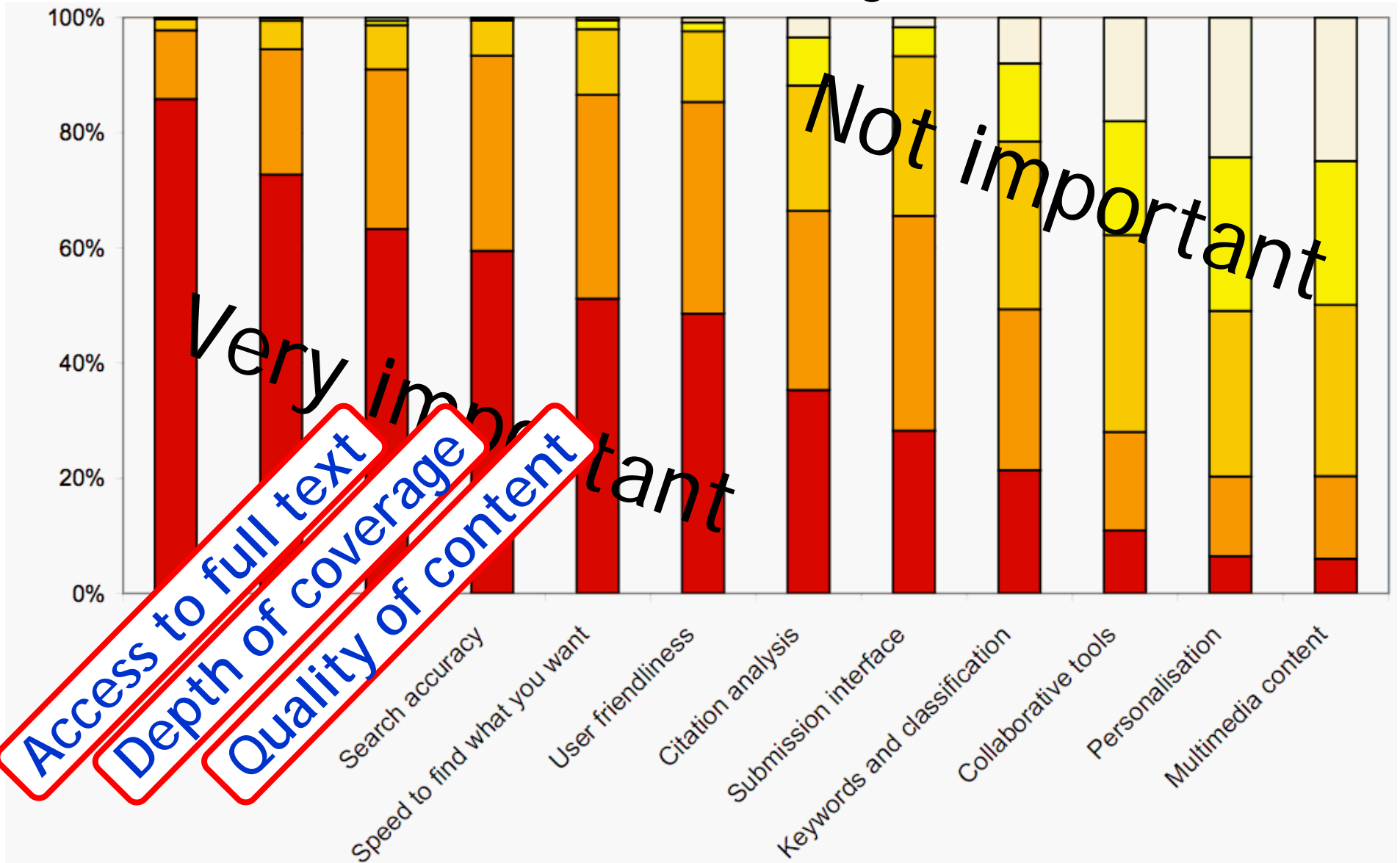
full-text preprint server

input by authors

automated submission  
and indexing

Maintained by hosting Institution,  
free of charge for users worldwide.

# How important are these features of an information system?

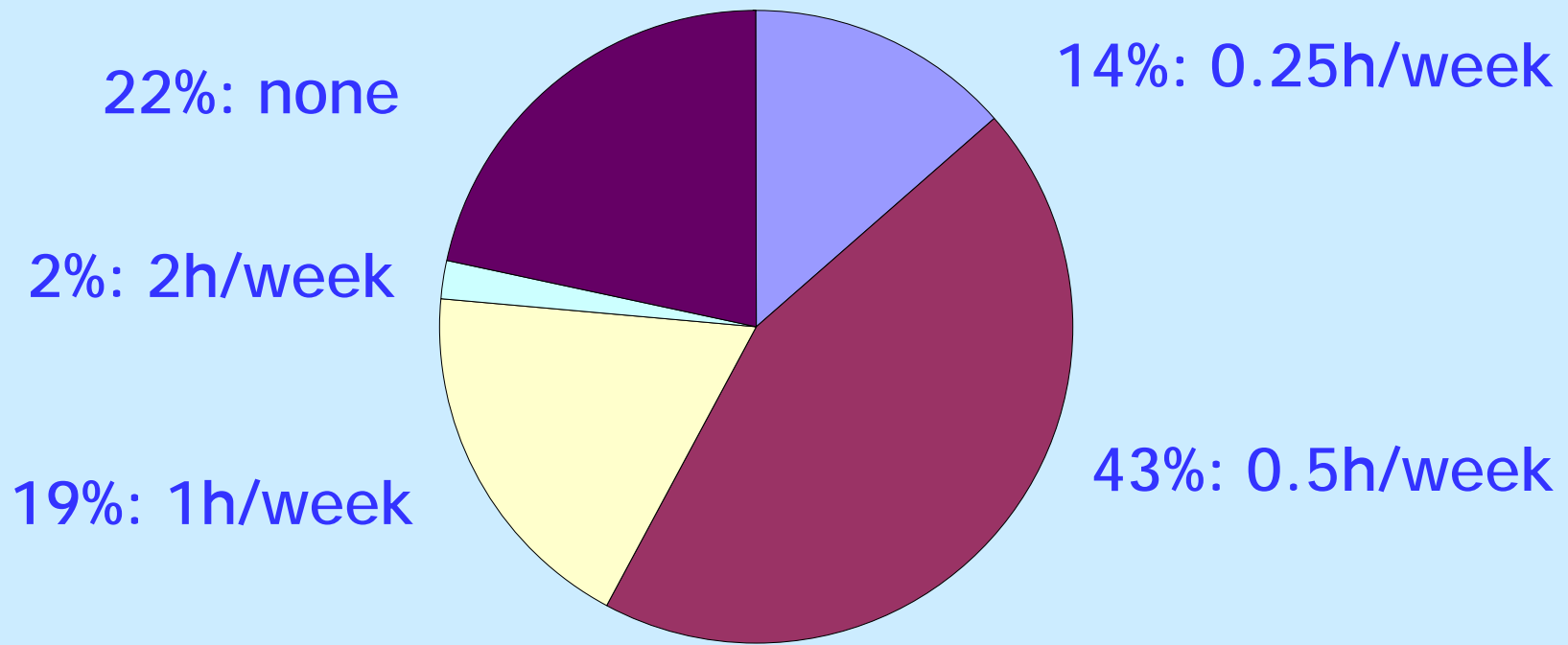


# Which changes do you expect?

## Summary of recurrent and inspiring answers

- Seamless (open) access to older articles
- Improved (full-text search and) access to public experiment notes (grey literature)
- Indexing of conference .ppt slides (interlinked with the corresponding article)
- “Publication” of “ancillary” material:
  - Data in tables & figures; correlation matrices
  - Data (high-level objects)
- (A new kind of) Peer-reviewing overlaid on arXiv
- “Smarter” search tools (related papers)
- Fragments of computer code accompanying equations

# Would users invest time in online community service (here content tagging)?



On average 30 min/week

Immense potential to be harnessed

# Vision for an e-Infrastructure for HEP scientific communication

May'07: HEP Information Summit @ SLAC

May'08: next Summit @ DESY

kick-off and brain-storming of all concerned parties to

1. Build a complete HEP information platform
2. Enable text- and data-mining applications
3. Demonstrate and deploy Web2.0 applications
4. Preservation and re-use of research data



## 1. Build a complete HEP information platform

- Integrate the content of present repositories and databases to host the entire body of metadata and the full-text of all OA publications, past and future
- Create the one-stop shop 30-million hits/year platform where all HEP researchers go for their information needs
- Integrate conference material (pre-grey literature)

## 2. Enable text- and data-mining applications

- Detect relations between documents carrying similar information
- Create datasets to exercise new hybrid metrics to measure the impact of articles, authors and groups
- Extract numerical information from figures and tables within published articles.

## 1. Build a complete HEP information platform

- Integrate the content of present repositories and databases to host the entire body of metadata and the full-text of all OA publications, past and future
- Create the one-stop shop 30-million hits/year platform where all HEP researchers go for their information needs
- Integrate conference material (pre-grey literature)

## 2. Enable text- and data-mining applications

- Detect relations between documents carrying similar information
- Create datasets to exercise new metrics to measure the impact of articles, authors and groups
- Extract numerical information from figures and tables within published articles.

### 3. Demonstrate and deploy Web2.0 applications

- Engage readers/authors in subject tagging, altering automatically assigned classifications
- Enable the possibility to review and comment on articles, adding links to additional documents or other digital objects
- Community-based aggregation of related objects (articles, preprints, conferences, lectures)

Many (all?) of those already exist... with little buy-in

**Aim for a production system containing the entire corpus of a discipline, used by all practitioners.**





### 3. Demonstrate and deploy Web2.0 applications

- Engage readers/authors in subject tagging, altering automatically assigned classifications
- Enable the possibility to review and comment on articles, adding links to additional documents or other digital objects
- Community-based aggregation of related objects (articles, preprints, conferences, lectures)

*The mid-term future*

Many (all?) of those already exist... with little buy-in

**Aim for a production system containing the entire corpus of a discipline, used by all practitioners.**



## 4. Preservation and re-use of research data

- Natural evolution of repositories
- Aim to access data, simulations, computer programs behind each repository object
- Not a technological/archival problem: our computing centres routinely copy old tapes onto new facilities
- Partly a (not insurmountable) software problem: however, experiment life-cycle longer than computing environment life-cycle, migrations can and do occur
- HEP data from facilities recently stopped or about to be discontinued is vaguely readable but not re-usable



## 4. Preservation and re-use of research data

- Natural evolution of repositories
- Aim to access data, simulations, computer programs behind each repository object
- Not a technological/archival problem: our computing centres routinely copy old tapes onto new facilities
- Partly a (not innumerable) software problem: however, experimental life-cycle longer than computing environment life-cycle, migrations can and do occur
- HER data from facilities recently stopped or about to be discontinued is vaguely readable but not re-usable

Long-term target



# The next frontier: Research data

## Goals:

- long-term preservation
- re-usability
- accessibility

## Obstacles:

- sheer size
- complexity
- funding

# Preservation, re-use and (open) access continua (who and when)

- The same researchers who took the data, after the closure of the facility (~1 year, ~10 years)
- Researchers working at similar experiments at the same time (~1 day, week, month, year)
- Researchers of future experiments (~20 years)
- Theoretical physicists who may want to re-interpret the data (~1 month, ~1 year, ~10 years)
- Theoretical physicists who may want to test future ideas (~1 year, ~10 years, ~20 years)



# Much ado about nothing?

Strong force: gets weaker the closer the quarks get.

**Most counter-intuitive idea of contemporary physics**

Idea 1972, Nobel prize 2004

To verify it, start pulling quarks far apart:

- 1) Produce quark at accelerators
- 2) Put more and more energy in
- 3) Do quark pull each other more?



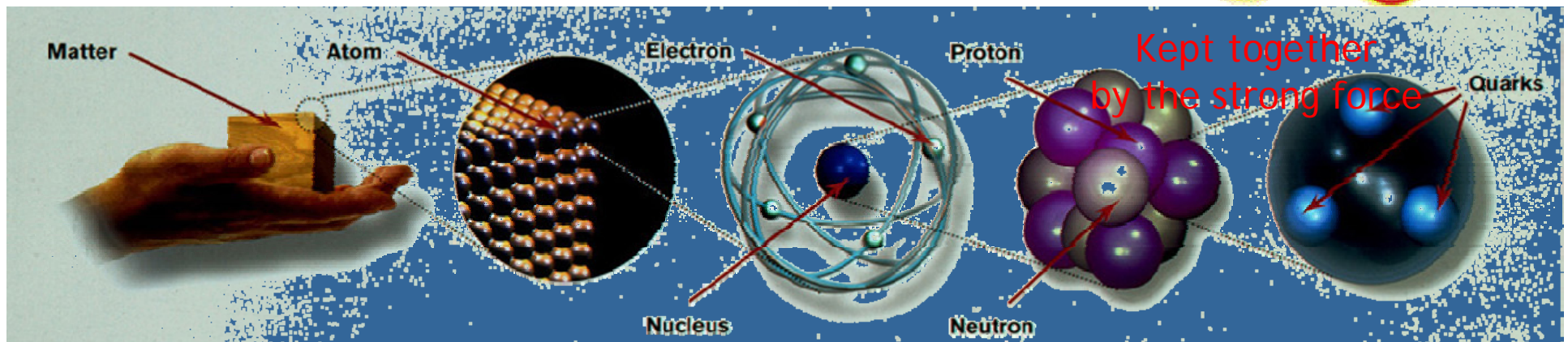
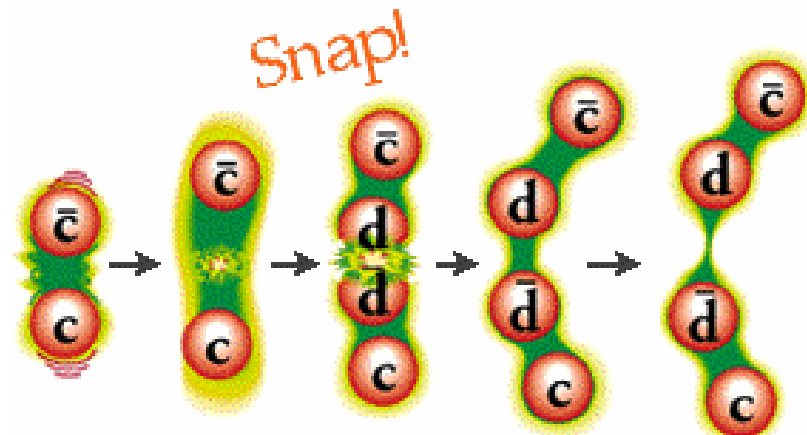
David J. Gross



H. David Politzer



Frank Wilczek



# Measuring the strong force

Need theory to analyse data, theory improves with *in-silico* experiments, which improve with computing power, which grows with time.

Need to re-analyse data with time!

EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH

CERN-EP/99-175  
13th December 1999

arXiv:hep-ex/0001055v1 24 Jan 2000

## QCD Analyses and Determinations of $\alpha_s$ in $e^+e^-$ Annihilation at Energies between 35 and 189 GeV

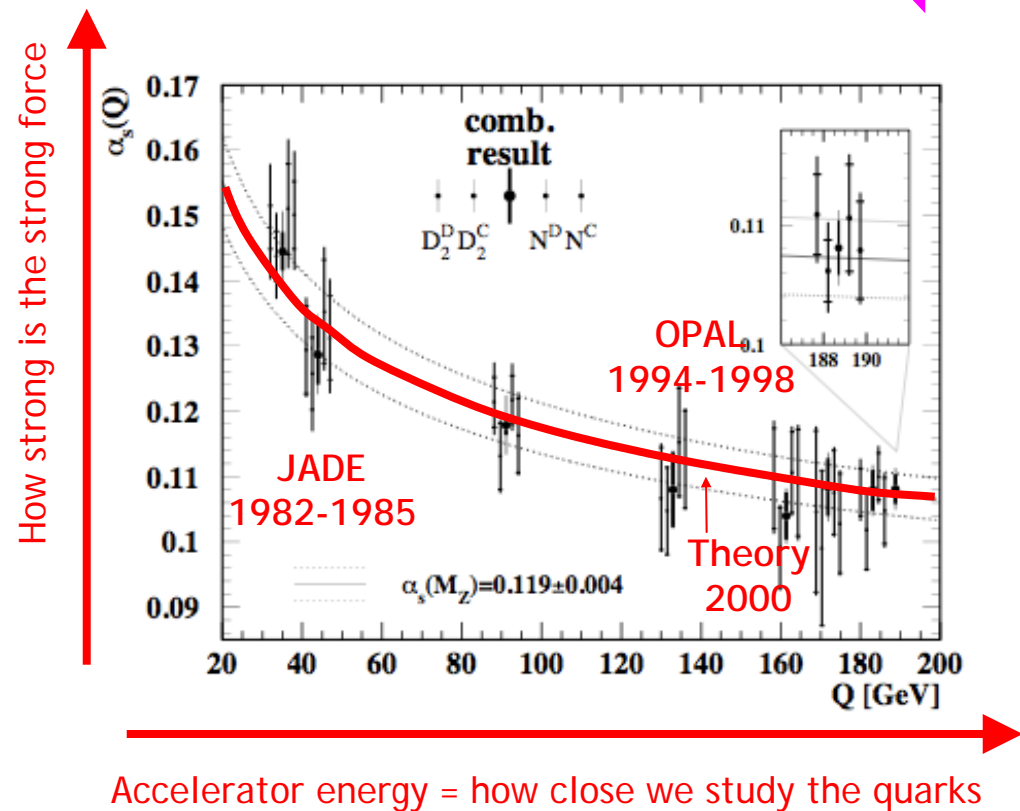
The **JADE** (\*) and the **OPAL** (\*\*) Collaboration

**Abstract:**  
We employ data taken by the JADE and OPAL experiments for an integrated QCD study in hadronic  $e^+e^-$  annihilations at c.m.s. energies ranging from 35 GeV through 189 GeV. The study is based on jet-multiplicity related observables. The observables are obtained to high jet resolution scales with the JADE, Durham, Cambridge and cone jet finders, and compared with the predictions of various QCD and Monte Carlo models. The strong coupling strength,  $\alpha_s$ , is determined at each energy by fits of  $\mathcal{O}(\alpha_s^2)$  calculations, as well as matched  $\mathcal{O}(\alpha_s^2)$  and NLLA predictions, to the data. Matching schemes are compared, and the dependence of the results on the choice of the renormalization scale is investigated. The combination of the results using matched predictions gives

$$\alpha_s(M_{Z^0}) = 0.1187^{+0.0034}_{-0.0019}$$

The strong coupling is also obtained, at lower precision, from  $\mathcal{O}(\alpha_s^2)$  fits of the c.m.s. energy evolution of some of the observables. A qualitative comparison is made between the data and a recent MLLA prediction for mean jet multiplicities.

To be submitted to European Physical Journal C

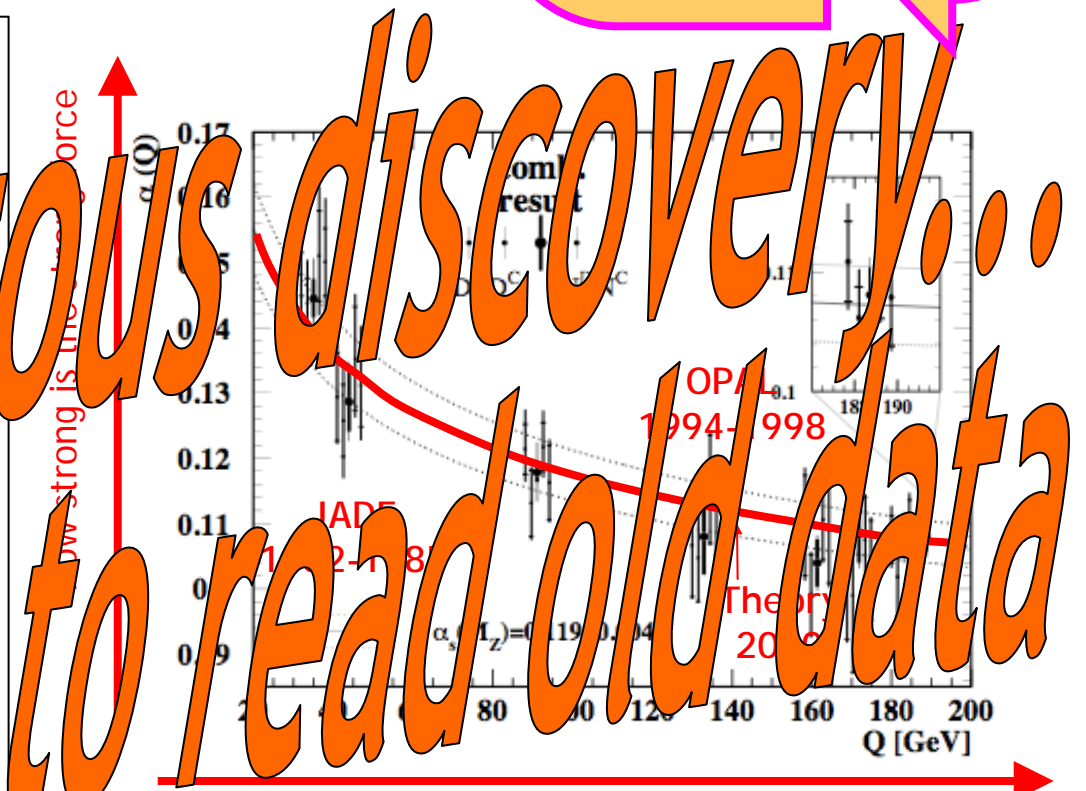
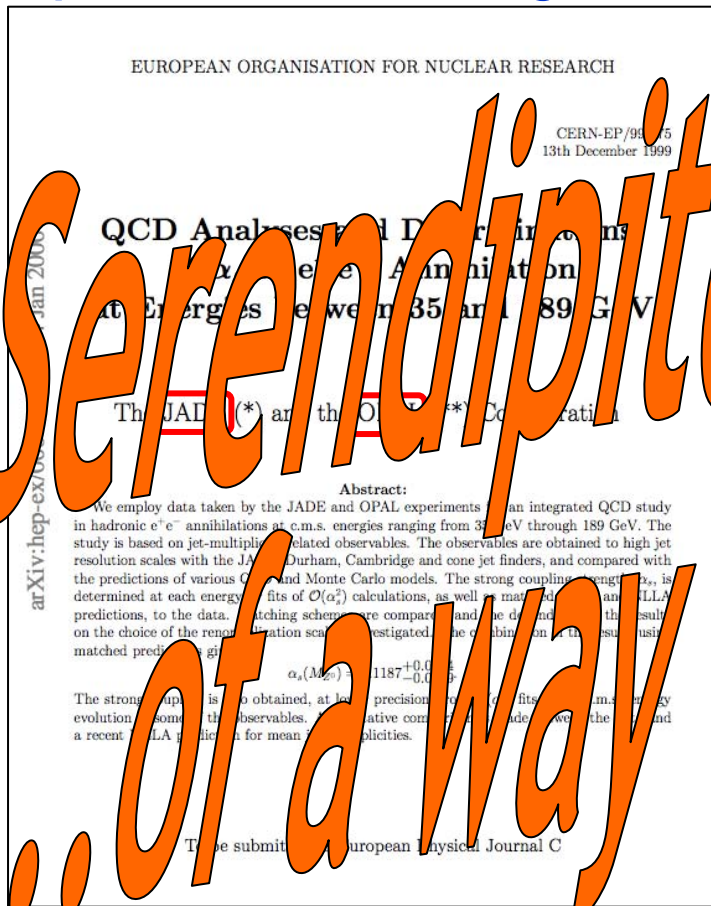


# Measuring the strong force

Need theory to analyse data, theory improves with *in-silico* experiments, which improve with computing power, which grows with time.

Need to re-analyse data with time!

Serendipitous discovery...  
...of a way to read old data



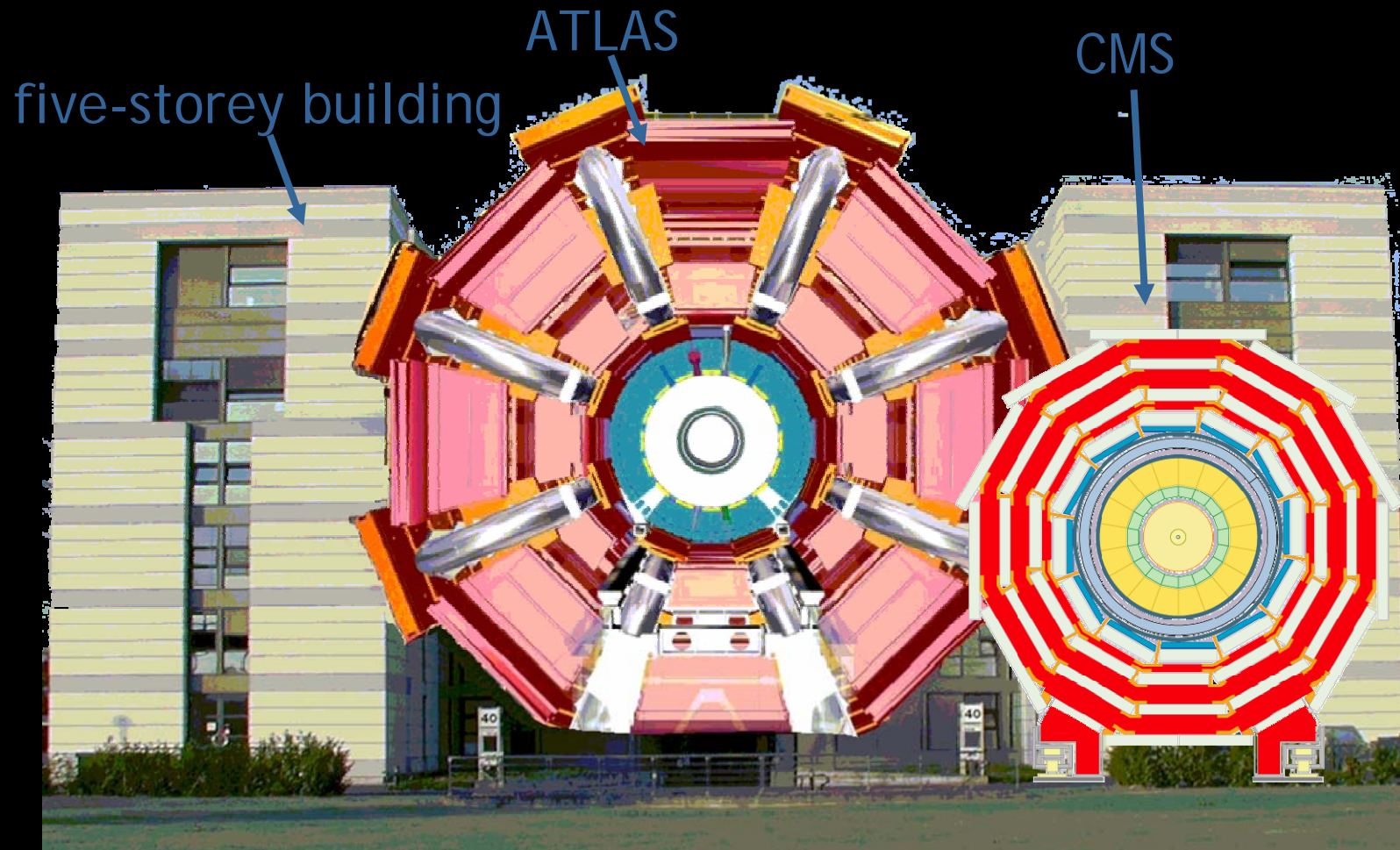
Accelerator energy = how close we study the quarks



# The Large Hadron Collider

- 
- Largest scientific instrument ever built, 27km of circumference
  - The “coolest” place in the Universe  
-271 °C
  - 10000 people involved in its design and construction
  - Worldwide budget of 6bn€
  - Collides protons to reproduce conditions at the birth of the Universe...  
...40 million times a second

The LHC experiments:  
about 100 million "sensors" each  
[think your 6MP digital camera...  
...taking 40 million pictures a second]



# The LHC data

- 40 million events (pictures) per second
- Select (on the fly) the ~200 interesting events per second to write on tape
- “Reconstruct” data and convert for analysis:  
“physics data” [inventing the grid...]

(x4 experiments x15 years)

Per event

Per year

Raw data

1.6 MB

3200 TB

Reconstructed data

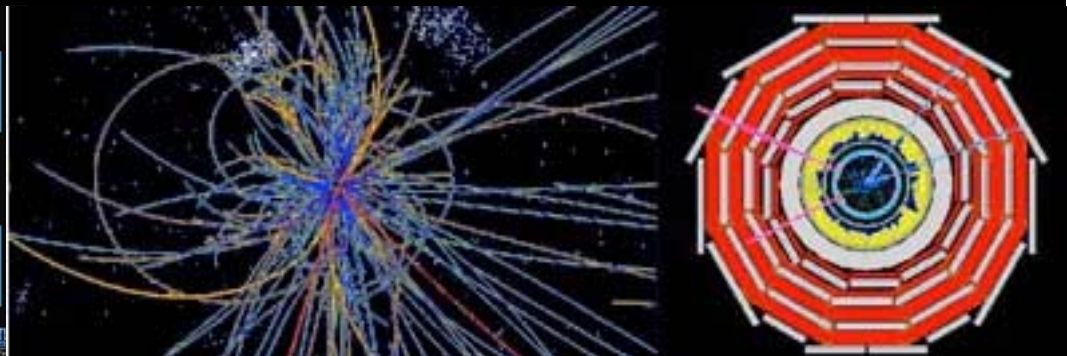
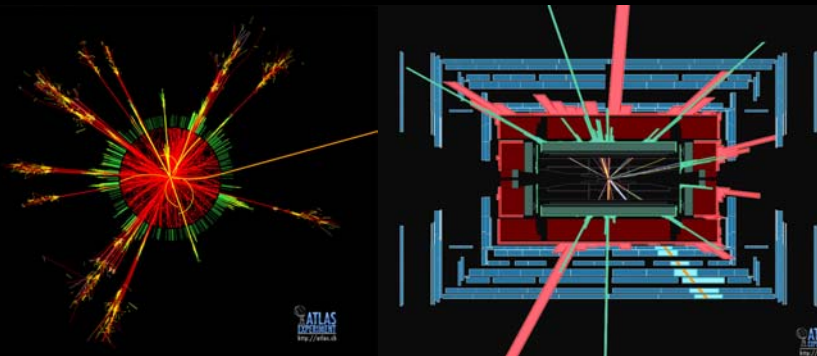
1.0 MB

2000 TB

Physics data

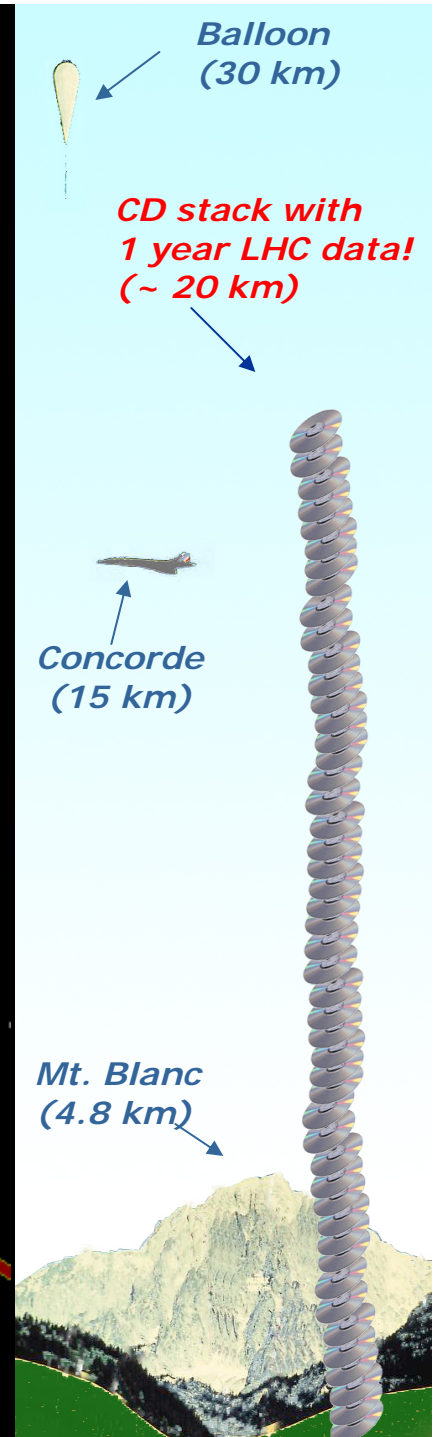
0.1 MB

200 TB



# Preserving HEP data?

- The HEP data model is highly complex. Data are traditionally not re-used as in Astronomy or Climate science.
- Raw data → calibrated data  
→ skimmed data → high-level objects  
→ physics analyses → results.
- All of the above needs duplication for *in-silico* experiments, necessary to interpret the highly-complex data.
- Final results depend on the grey literature on calibration constants, human knowledge and algorithms needed for each pass...oral tradition!
- Years of training for a successful analysis



# Data archival and re-use

Billions of funds are invested in colliders and experiments all over the world.

If data can not be re-used after the experiment stopped this investment is not exploited to its full capability.

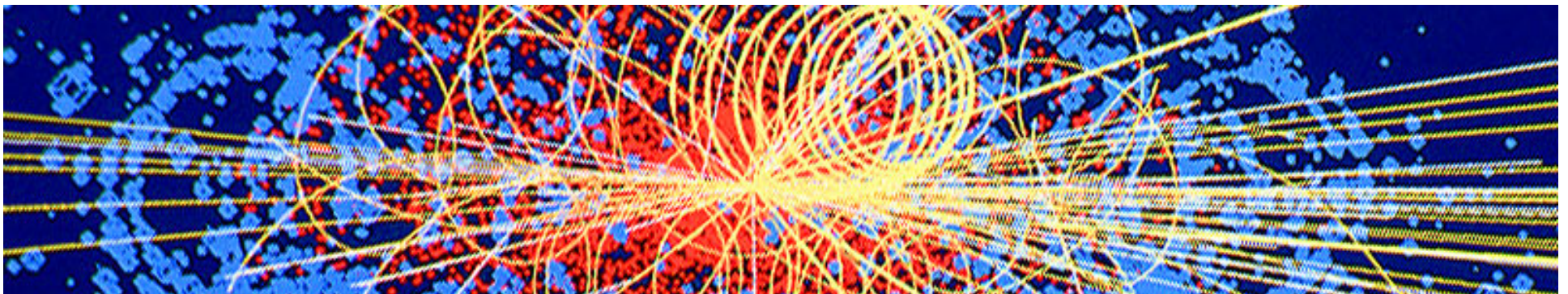
- Everything one hasn't thought of or known (new models, better parametrization)
- Combination with future experiments

LEP@CERN  
HERA@DESY  
TEVATRON@FNAL  
KLOE@LNF  
BABAR@SLAC  
BELLE@KEK

An additional relatively small fraction of the funds preserves a large fraction of the knowledge.

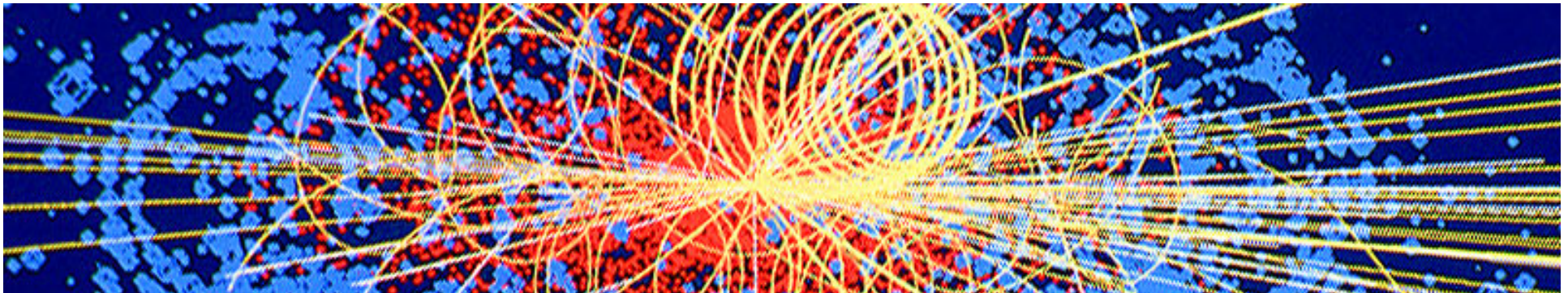
# HEP data: The “parallel way” to publish/preserve/re-use/OpenAccess

- In addition to experiment data models, elaborate a parallel format for (re-)usable high-level objects
  - In times of need (to combine data of “competing” experiments) this approach has worked
  - Embed the “oral” and “additional” knowledge
- A format understandable and thus re-usable by practitioners in other experiments and theorists
- Start from tables and work back towards primary data
- How much additional work? 1%, 5%, 10%?



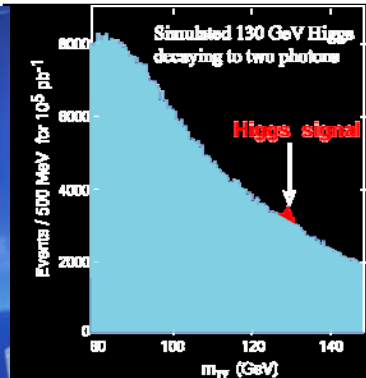
# HEP data: The “parallel way” to publish/preserve/re-use/OpenAccess

- In addition to experiment data models, elaborate a parallel format for (re-)usable high-level results
  - In times of need (to combat “publishing” experiments that “never worked”)
  - From “additional” knowledge
- A f understandable and thus re-usable by practitioners in other experiments and theorists
- Start from tables and work back towards primary data
- How much additional work? 1%, 5%, 10%?



# Issues with the “parallel” way

- A small fraction of a big number gives a large number
- Activity in competition with research time
- 1000s person-years for parallel data models need enormous (impossible?) academic incentives for realization ...or additional (external) funds
- Need insider knowledge to produce parallel data
- Address issues of (Open) Access, credit, accountability, “careless measurements”, “careless discoveries”, reproducibility of results, depth of peer-reviewing
- A monolithic way of doing business needs rethinking





# Conclusions

- With 50 years of preprints and 16 years of repositories and the web, HEP has spearheaded (Open) Access to Scientific Information
- Next step: SCOAP3 model for Open Access Publishing
- Time is ripe for an e-Infrastructure for HEP Scientific Communication
  - Build a complete HEP information platform
  - Enable text- and data-mining applications
  - Demonstrate and deploy Web2.0 applications
- The next challenge is the preservation of HEP data

**Exciting times are ahead!**

# Thank you !

Rolf-Dieter.Heuer@desy.de

[scoap3.org](http://scoap3.org)

[scoap3.org/files/Scoap3WPReport.pdf](http://scoap3.org/files/Scoap3WPReport.pdf)

[scoap3.org/files/Scoap3ExecutiveSummary.pdf](http://scoap3.org/files/Scoap3ExecutiveSummary.pdf)

