DESY/HERA data preservation

Achim Geiser, DESY Hamburg, Germany

for the H1, ZEUS and HERMES collaborations + DESY IT

DPHEP coll. workshop, CERN, Geneva, 14. 3. 2017

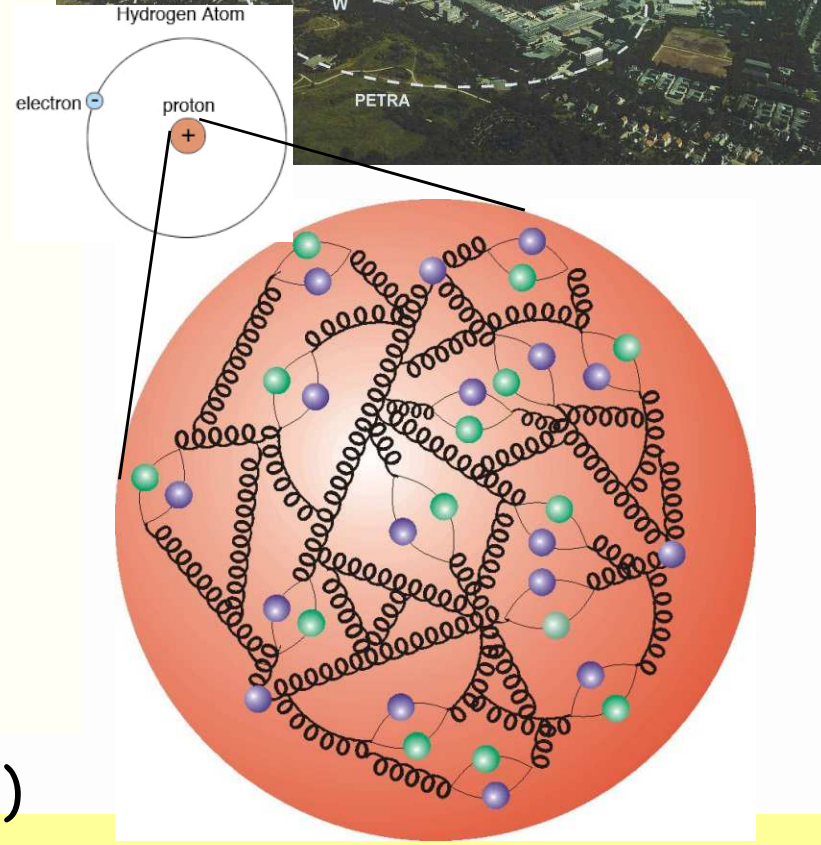
- Why? (motivation)
- How? (challenges)
- What? (achievements)

What is HERA?

- **International Particle Physics Experiment** which recorded high energy electron-proton collisions at the world's (so far) unique lepton-proton collider **HERA** at DESY in Hamburg, Germany

- **Physics data taking: 1992-2007**

- one of main physics goals:
  measure structure of the proton to
  ~10^{-18} m, i.e. 1/1000 of proton size
  ("X ray" of proton with electrons)
  used e.g. in measurements of Higgs properties at LHC



- also well suited to study **general QCD** and electroweak physics + **proton spin** (hermes)

HERA data preservation

- Data preservation project internally started within HERA experiments between 2006 (ZEUS) and 2009 (H1, HERMES)

- HERA experiments + DESY/IT are core co-authors of 2012 DPHEP study group document (end of HERMES funding)

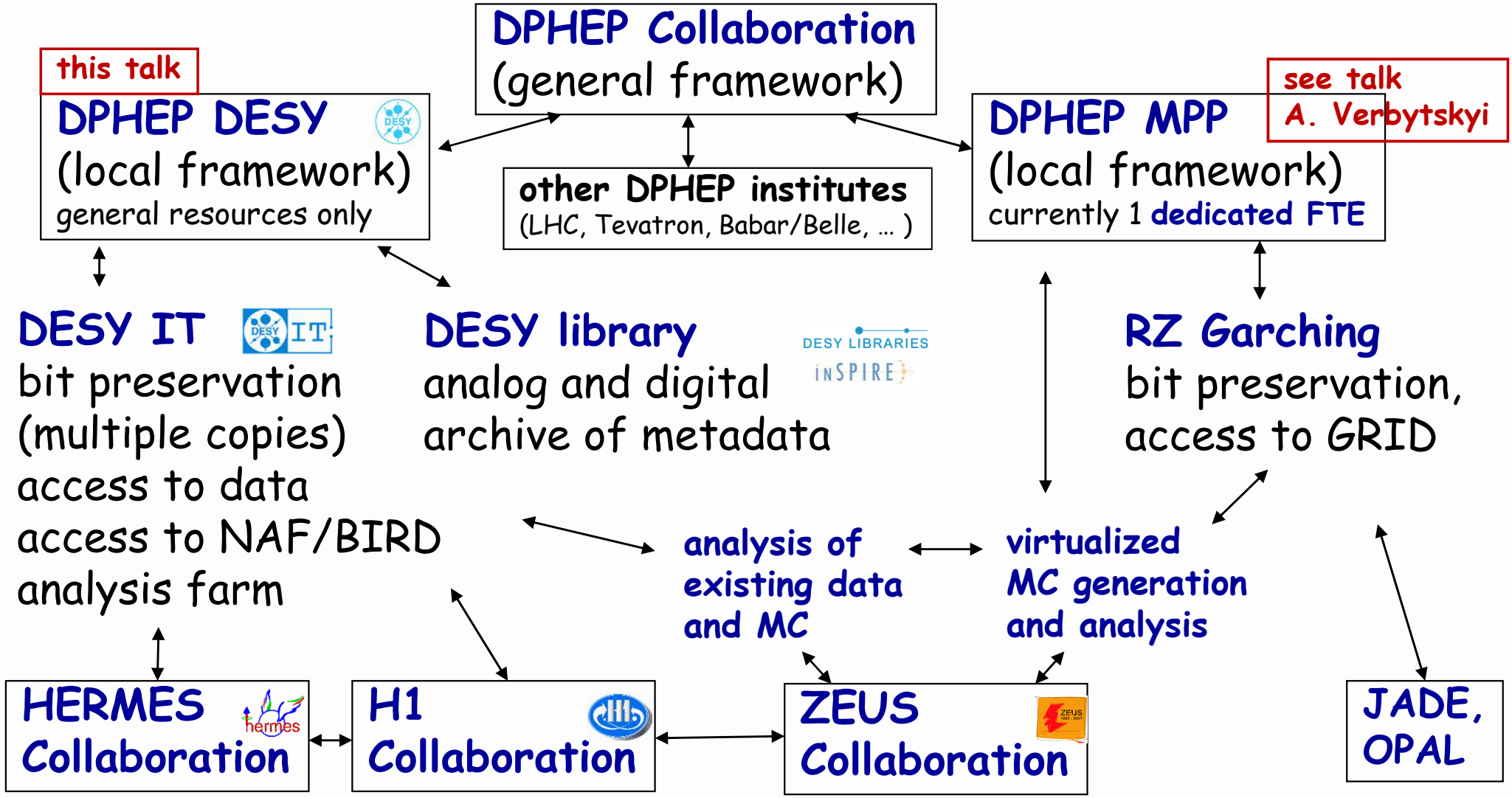
world DPHEP community

- DESY and MPP are co-founding members of Collaboration Agreement for the DPHEP project supported by ICFA (May 2014)

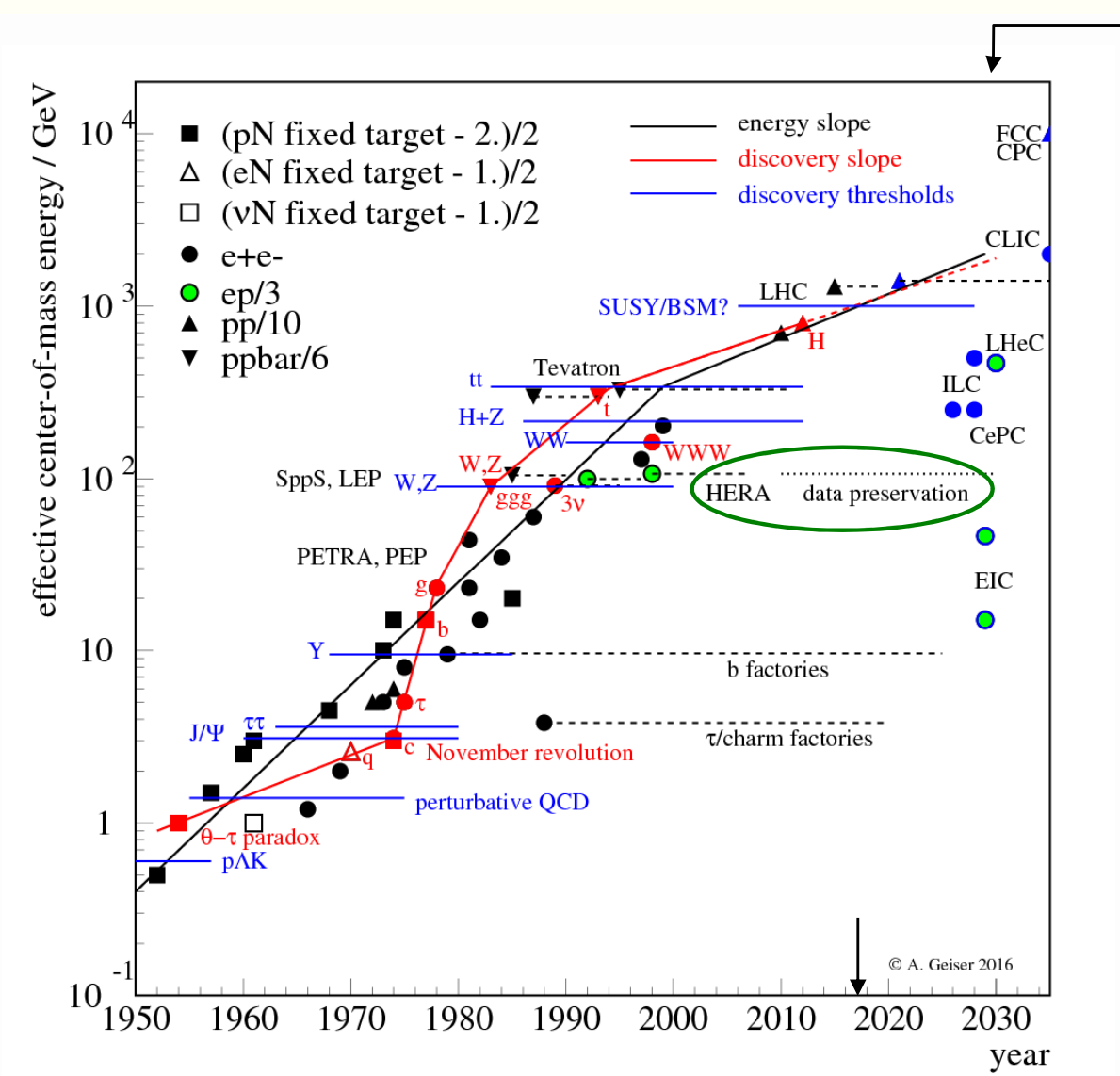
- workshop on Future Physics with HERA Data at DESY (Nov. 2014, end of H1/ZEUS funding)

world HEP community

HERA Data Preservation Challenge: How to organize the Management?



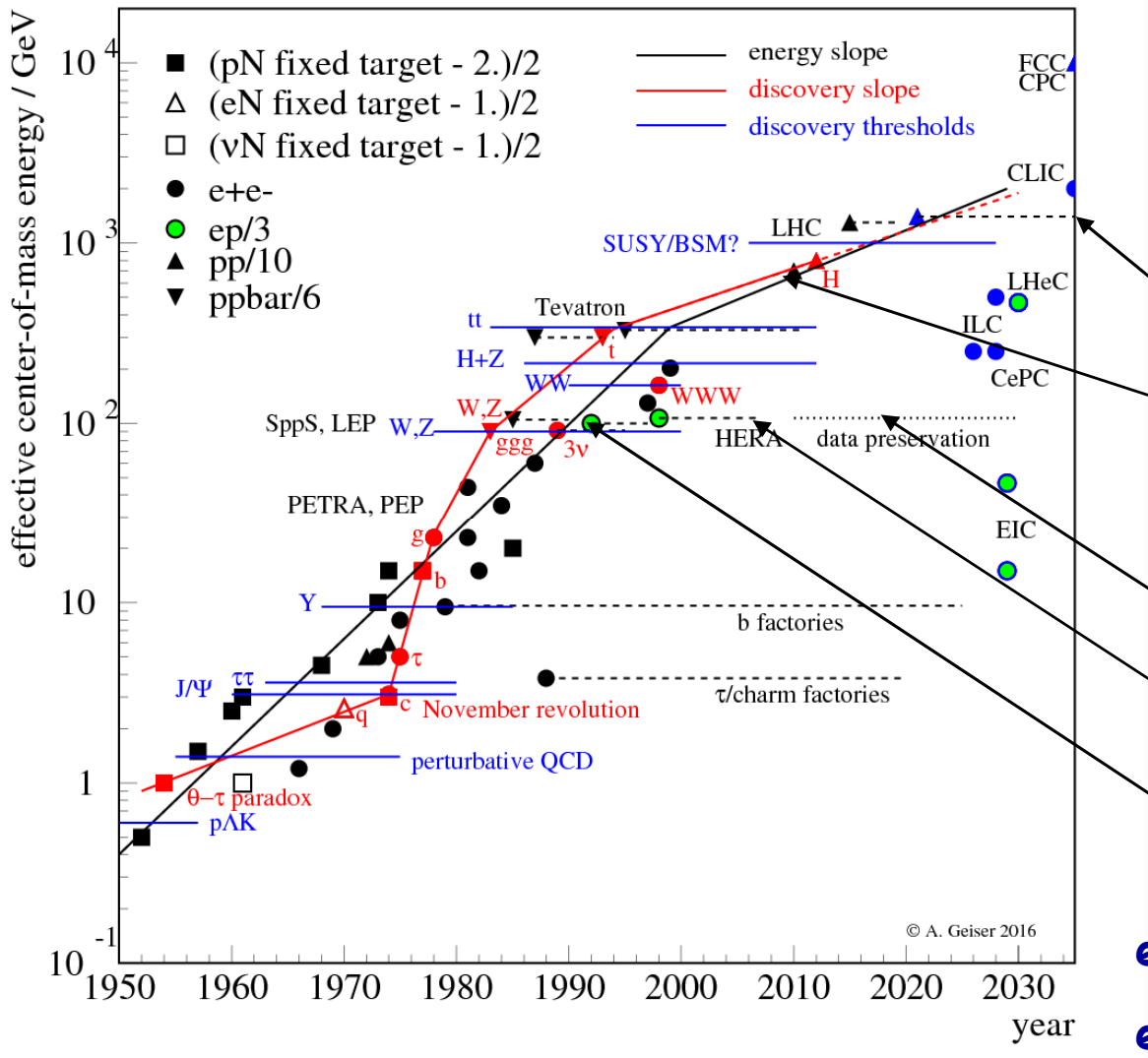
Why to preserve HERA data?



so far only ep collider:

HERA data are unique!

Why to preserve HERA data?



planned new projects

equivalent LHC future

possible end of LHC

start of LHC data taking

now (HERA data preservation)

end of HERA data taking

start of HERA data taking

end of LEP a few years earlier

end of Tevatron a few years later

Workshop:

- What do the HERA data still have to say and how are they relevant to other facilities?

- two days with lively discussions and almost 30 presentations
https://indico.desy.de/event/futurehera

- ~ 70 participants, both experimentalists and theorists from across the globe

- -> list of dozens of subjects that are still to be investigated or exploited fully, using the preserved data sets (proceedings in arXiv:1601.01499, arXiv:1512.03624)

DPHEP data preservation levels

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses -> education
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

Table 3: Various preservation models, listed in order of increasing complexity.

- ZEUS: level 3 (data and existing Monte Carlo (MC) data),
  level 4 (additional Monte Carlo data)
- H1 and HERMES: level 4

Challenge: What is the "Data"?

- "Data" = recorded events, simulated events,
  + related software, knowledge, and documentation

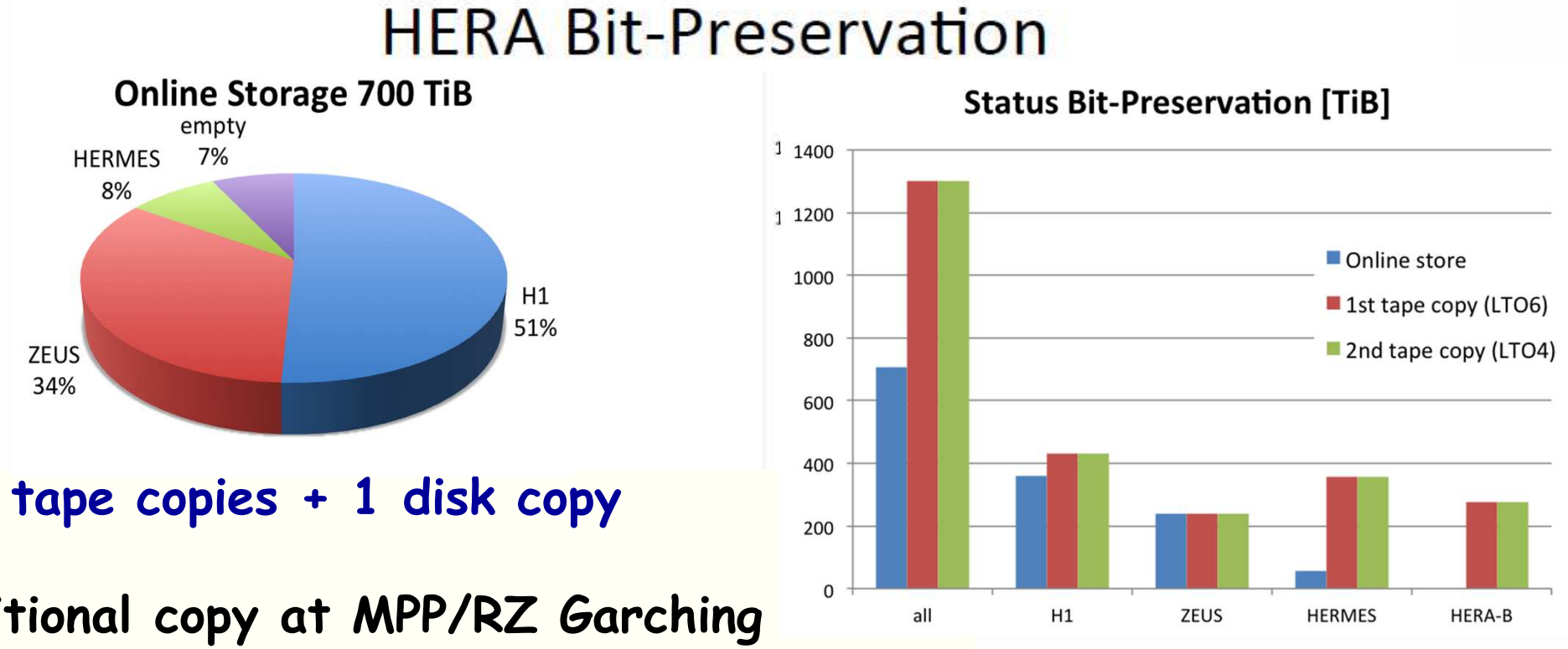
- Bit preservation and data access (computing):
  existing data and MC samples

- Software preservation:
  simulation, reconstruction, analysis, event display

- Documentation:
  analog and digital archives, web pages

Challenge: Bit preservation

- at DESY: common approach for all three HERA experiments

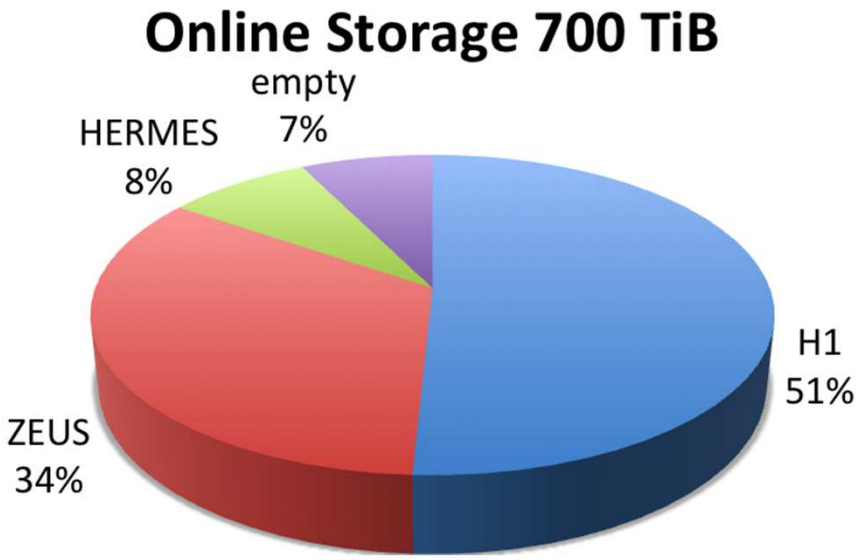


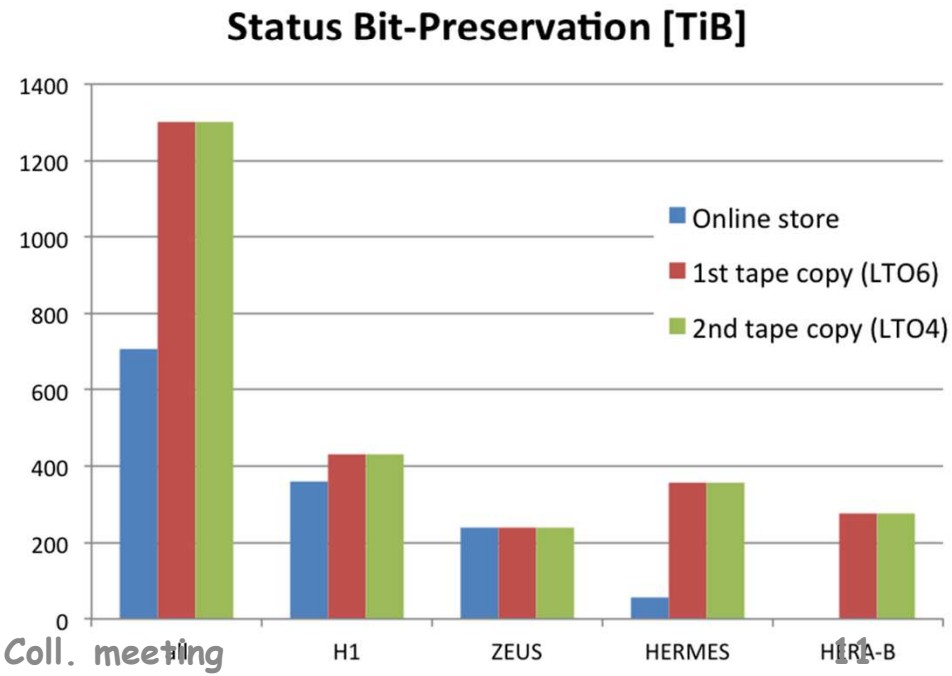
2 tape copies + 1 disk copy

+ additional copy at MPP/RZ Garching (for ZEUS part) -> talk A. Verbytskyi

HERA Bit-Preservation

- The HERA data archive is finalized
- The online (disk) store is filled and 2 tape copies are written
- Small additions to the heritage data are possible - details about the procedure will be defined in agreement with the experiments
  - First cases now
- The content of the archive and the procedures how to add and restore data had been documented
- Restoring data from the tape archive to the online store had been successfully exercised





For the Statistics Enthusiasts: final storage content

H1	HERMES	ZEUS	HERA-B	Type
983398	6557725	1183157	846059	single files
11111	9179	7318	4110	archive (tar) files
810316	774032	1182941	0	files online
359	57	239	0	TiB online
464	581	368	392	# LTO4 (800G) tapes
134	174	104	110	# LTO6 (2.4T) tapes
430	358	239	276	TiB on LTO4/LTO6 tapes

- In summary: 1.3 PB and 10 million files
- In addition there are 10 TB data of polarimeter data/simulations included

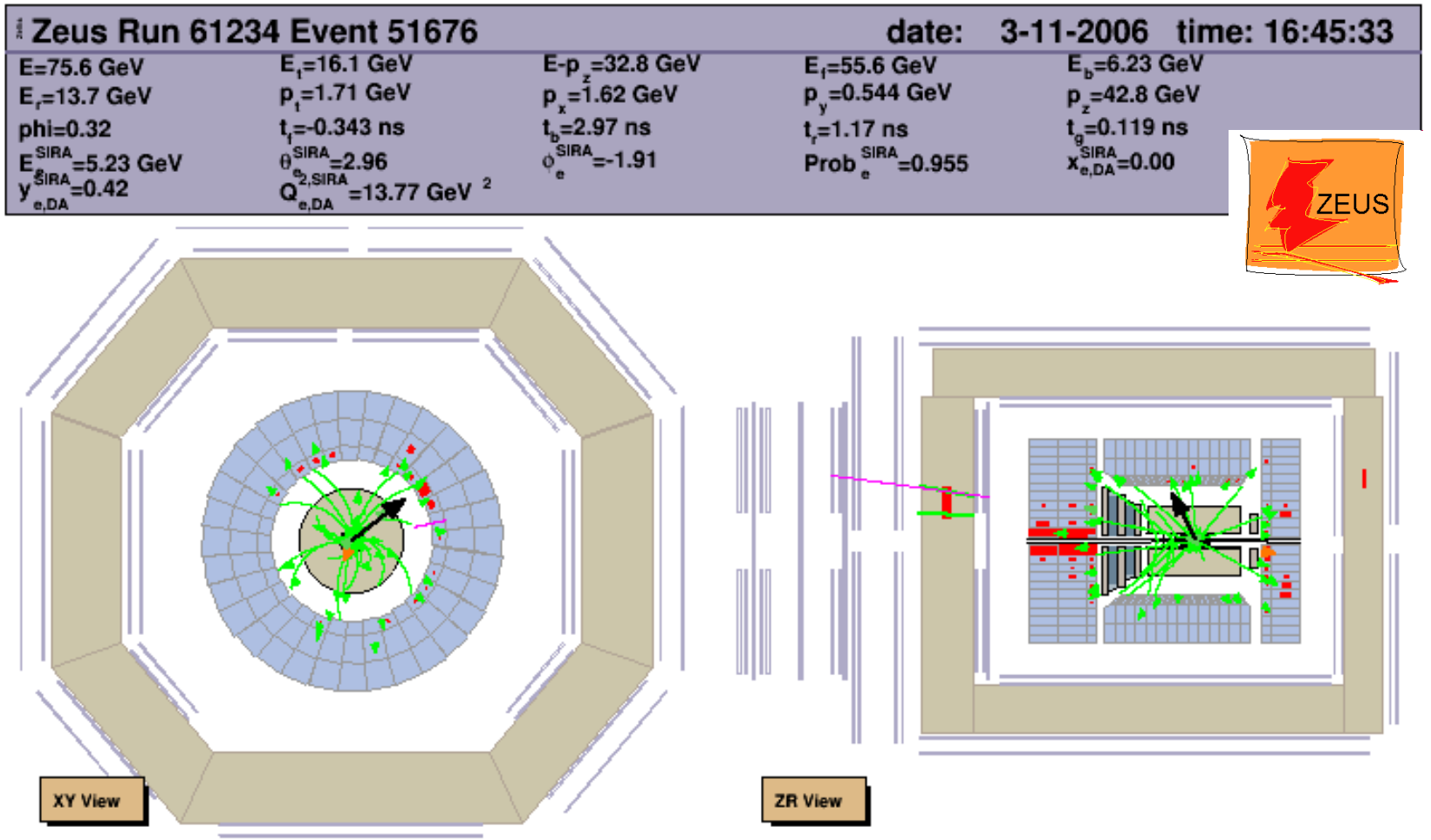
Challenge: Computing

- all remaining dedicated hardware for all three HERA experiments decommissioned since 2014/15.
- **long term data access guaranteed by DESY IT.**

- currently access to preserved data at DESY on generic "BIRD" batch farm (National Analysis Facility, NAF), e.g. ~30 ZEUS users (integrated).
- shared opportunistically with LHC and other experiments but fully sufficient for relatively modest HERA needs.
- job submission via dedicated servers (SL6) maintained by DESY IT. Can also be used for interactive debugging and event display.

- access to ZEUS data also at MPP Munich (see talk A. Verbytskyi)

What do HERA data look like?



event display

from ZEUS

"Common Ntuple"

complicated data format and content: for useful analysis, need

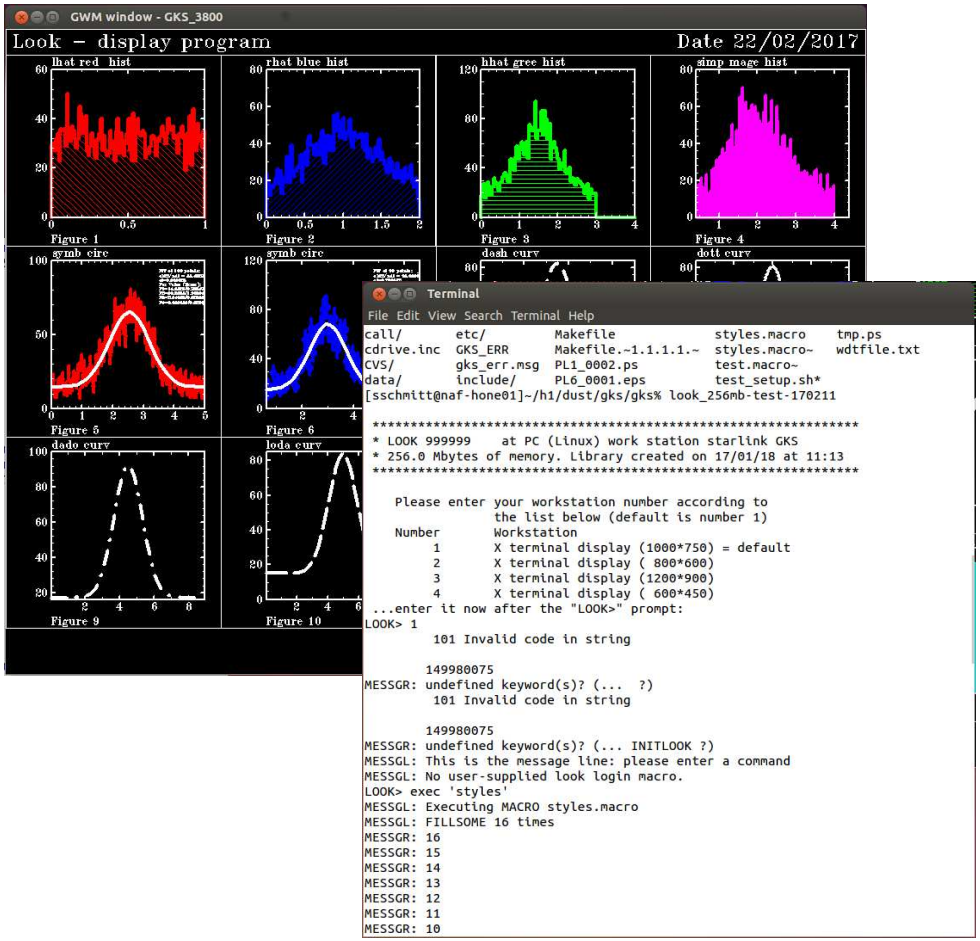
significant expert knowledge + documentation + guidance how to use it

Software preservation

- **H1: already switched to OO/C++ around 2001. Continuously port and maintain all related software. Porting to SL6 completed.**
- **ZEUS:** unmaintainable (manpower) **software from 1990's completely replaced by simplified ROOT common ntuple approach** for analysis; **SL5 -> SL6 "transparent".**
- **virtualization approach based on frozen SL5 executables** (MPP) for new MC (-> talk A. Verbytskyi).
- **HERMES: freeze software and executables with SL6 and keep running as long as possible. Currently working fine.**
- **Long term manpower remains a critical concern**

H1 software

- Customized GKS libraries installed on SL6 → it works
- H1 software depending on GKS:
  - "Look" (histogramming and analysis framework) by V.Blobel → working
  - "old" event display: supports some features not available on the new (root-based) tools
  - Special event display for drift-chamber analysis (hit-level)



Not critical for "standard" H1 analyses but nice to have

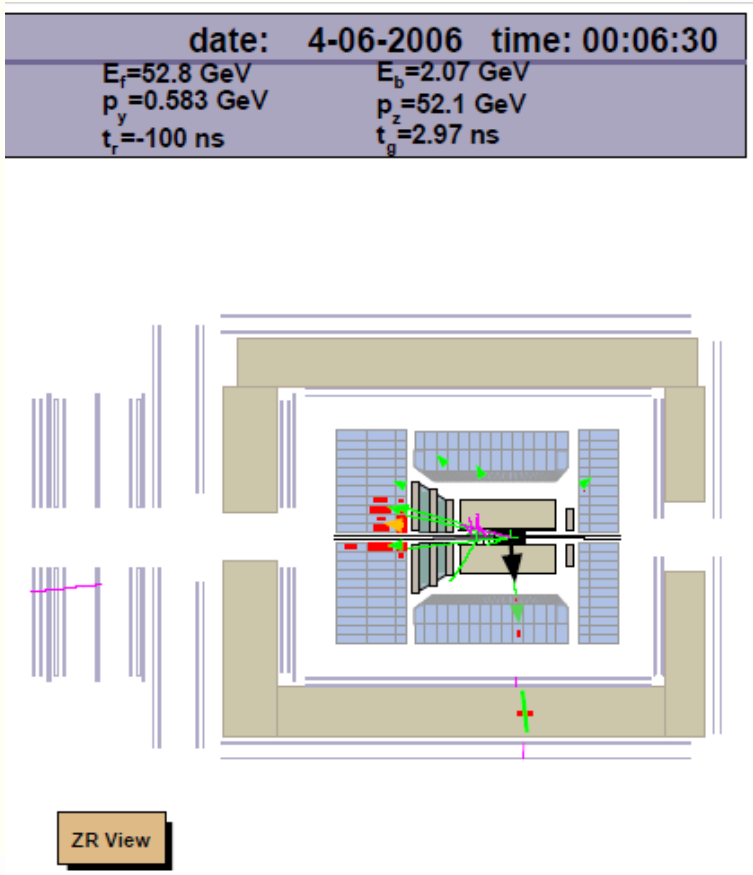
Common Ntuple analysis model

- **ZEUS Common Ntuple:** **Motto: keep it simple!**
  flat (simple) ROOT-based ntuple (same format as PAW ntuple converted with h2root)
  containing high level objects (electrons, muons, jets, energy flow objects, …)
  as well as low level objects (tracks, CAL cells, …)

- **Well tested !**
  almost all recent ZEUS papers (10 out of 13)
  based on Common Ntuples, already 19 total

- **"Easy" to use**
  several recent ZEUS papers based on results produced by master students from remote institutes, using resources at DESY:
  analysis on DESY NAF/BIRD computing farm

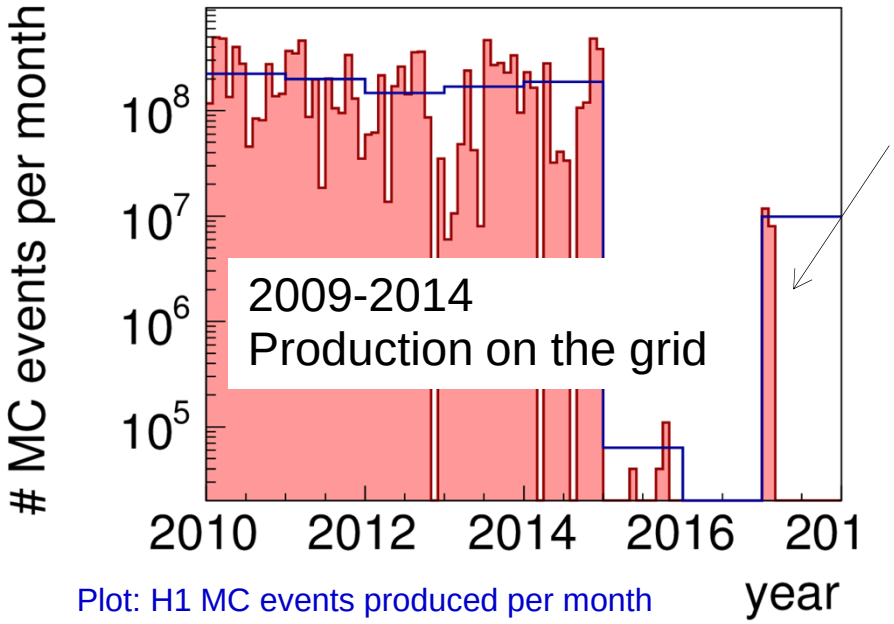
  PhD students can produce a ZEUS paper within only a fraction of their PhD time
  (e.g. ~6 months - 1 year)



Challenge: new MC production

- **H1**: Recent progress in MC production: production on the NAF system is working

2017 production using DESY batch system for new analyses



Plot: H1 MC events produced per month
(Thin line: average over one year)

- **HERMES**: production of new MC on NAF/Bird works

- **ZEUS**: new MC successfully produced via MPP system

  (-> talk A. Verbytskyi)

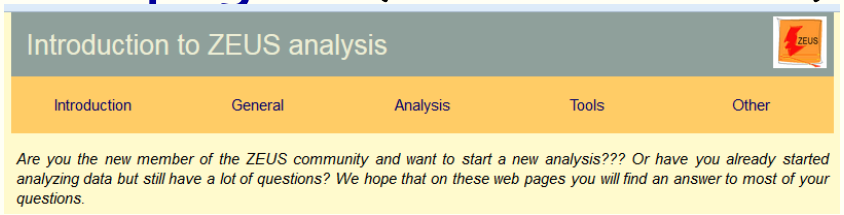
- being used in two current physics results

Analog and digital archive

- full analog archive in DESY library, partially digitized (HERMES)

- all ZEUS technical notes digitized on INSPIRE (via DESY library)
- plain html documentation web pages (DESY web office)
  - ZEUS since 2014

    meeting management -> Indico



  - H1 public web server now also in plain html mode

    Many H1 collaborative tools based on cgi-scripts for accessing oracle.
    Work-around: for critical tools -> local web-server using port 8080 which is not reachable outside firewall.
    Longer term: have to seek for another solution.

  - HERMES web server: wikis, cgi scripts, hosted on virtual machine
- knowledge preservation also in "human neural networks" (collaboration members)

Publicly available information on DPHEP and HERA data preservation

5. **The DPHEP Study Group: Data Preservation in High Energy Physics**
David M. South (DESY). Feb 14, 2013. 6 pp.
ICHEP-2012
e-Print: arXiv:1302.3379 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service
Details des Eintrags

6. **The H1 Data Preservation Project**
H1 Collaboration (David M. South *et al.*). Jun 2012. 6 pp.
Published in J.Phys.Conf.Ser. 396 (2012) 062019
DOI: 10.1088/1742-6596/396/6/062019
Conference: C12-05-21.3 Proceedings
e-Print: arXiv:1206.5200 [physics.data-an] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service
Details des Eintrags

+ DPHEP@DESY documents

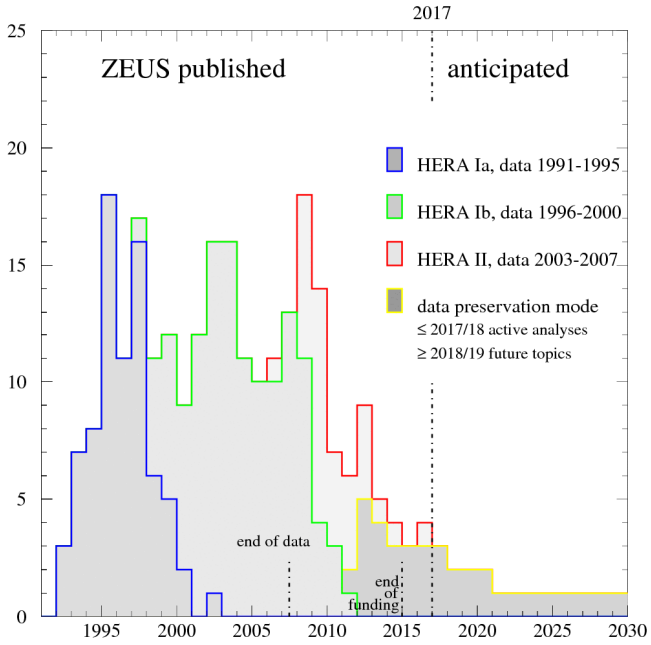
7. **Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable**
DPHEP Study Group (Zaven Akopov (DESY) *et al.*). May 2012. 93 pp.
DPHEP-2012-001, FERMILAB-PUB-12-878-PPD
e-Print: arXiv:1205.4667 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server; ADS Abstract Service; OSTI Information Bridge Server; Fermilab Library Serv
Details des Eintrags - Zitiert von 21 Datensätzen

8. **The ZEUS data preservation project**
ZEUS Collaboration and DESY DPHEP Group (J. Malka
DOI: 10.1109/NSSMIC.2012.6551468
Conference: C12-10-29, p.2022-2023 Proceedings
References | BibTeX | LaTeX(US) | LaTeX(EU) | Ha
Details des Eintrags

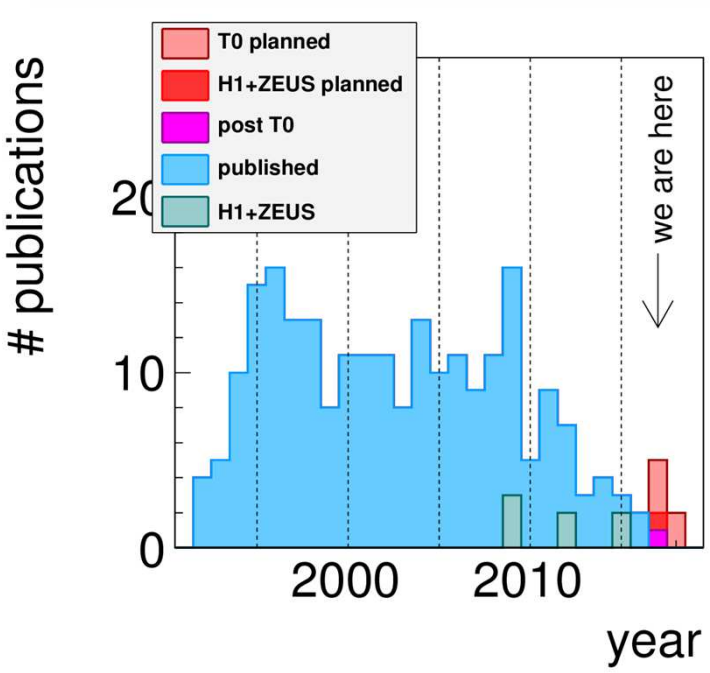
INSPIRE itself is a "level 1 data preservation project"

HERA papers, past and future

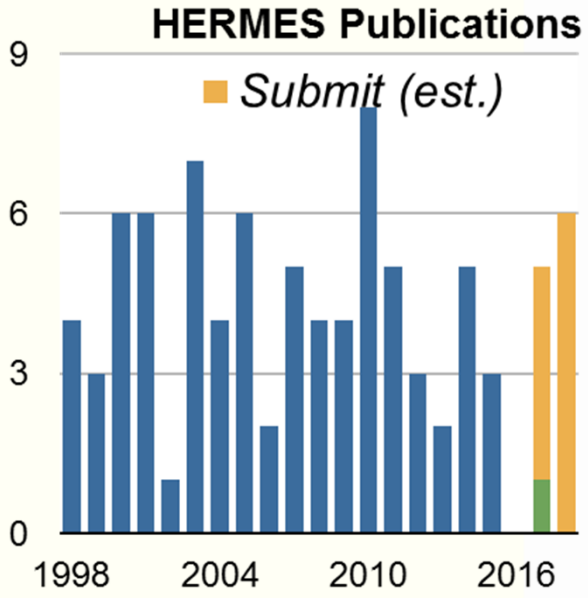
ZEUS



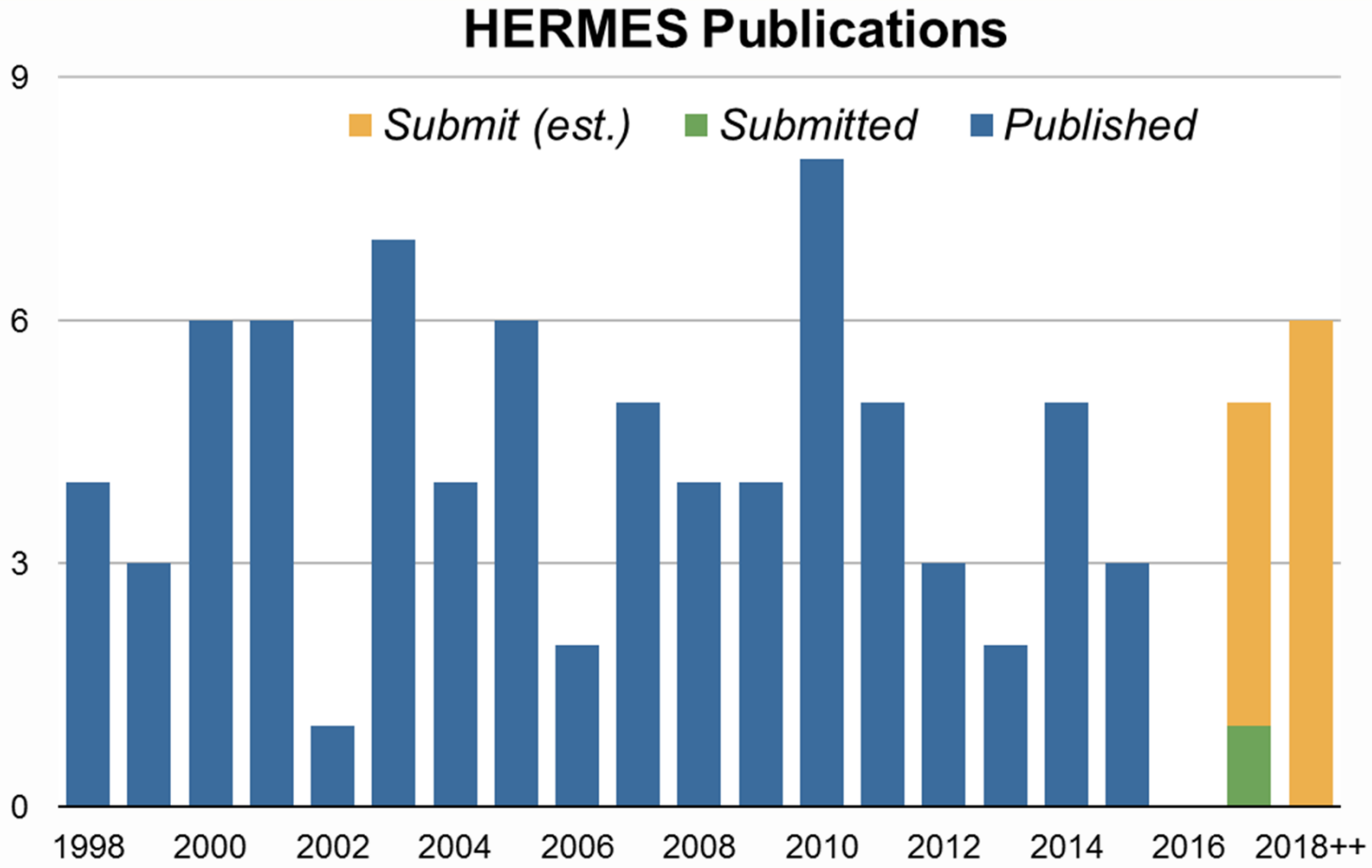
H1



HERMES



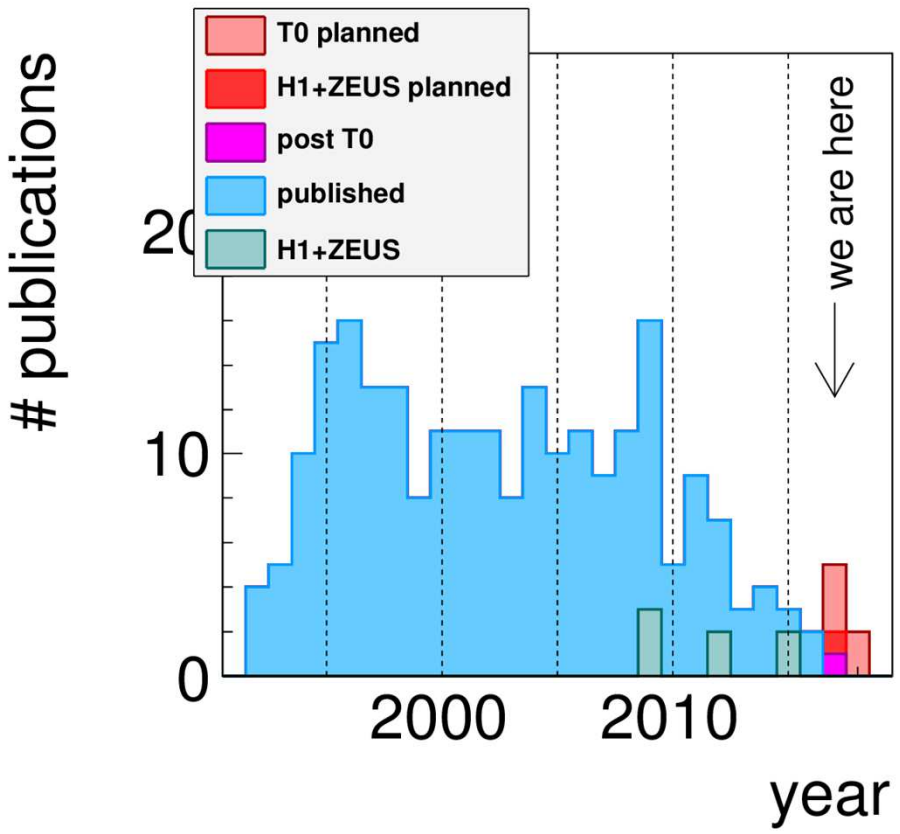
HERMES physics papers



- # of publications almost constant after end of funding 2012 (except 2016)
- substantial number of papers still in pipeline

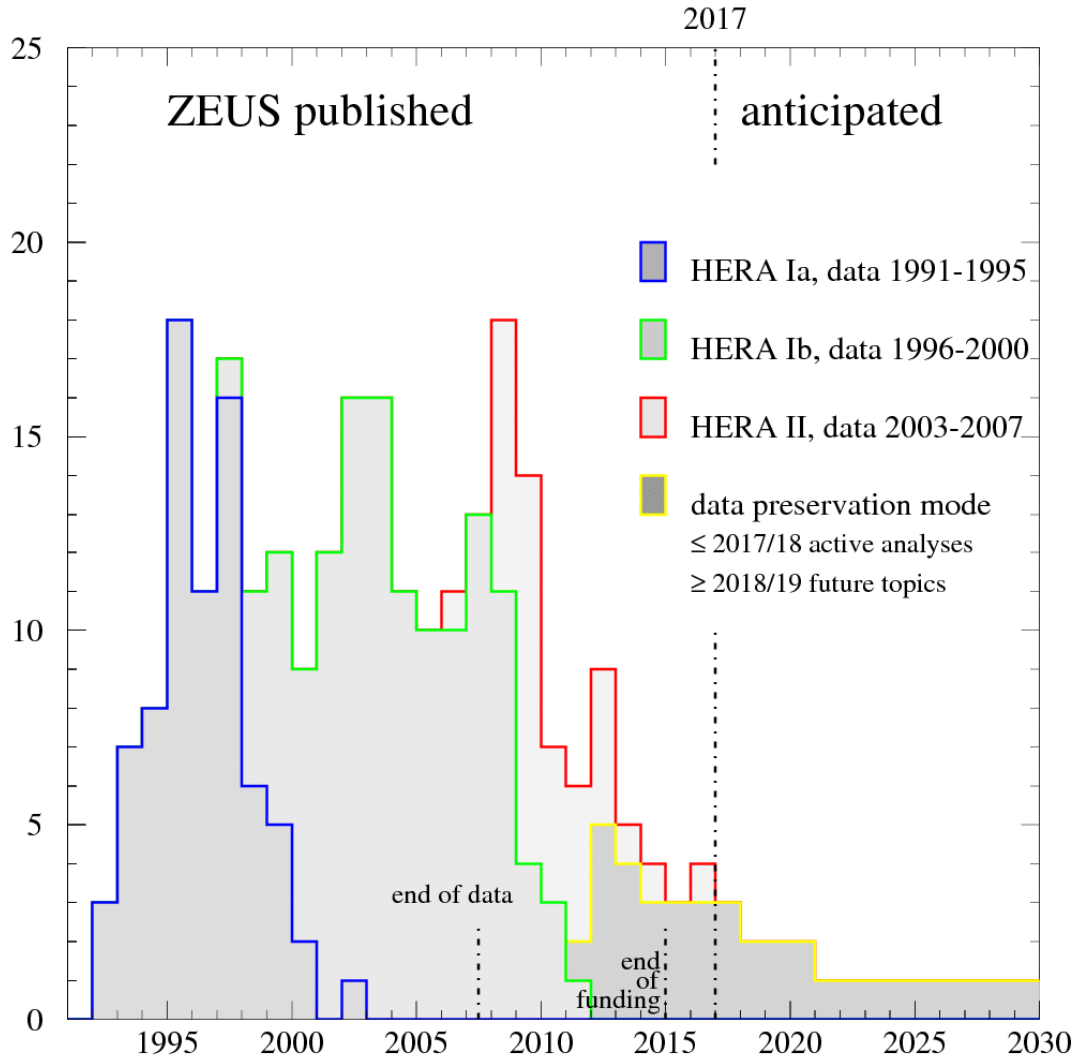
H1 Paper production summary

- Paper production 2016: two papers
  - QCD instanton search
  - Jet production at low Q^2
- 2017: hope to have five publications
  - Diffractive D* in DIS
  - α_S at NNLO from jets
  - H1/ZEUS F2b,F2c
  - ρ meson diffractive photoproduction
  - H1 electroweak fit



Analyses marked in red: **using preserved H1 data**

ZEUS physics papers



majority of papers produced

in "data preservation mode"

already since 2012 (19 papers)

since end of funding 2014:

2015: 3 papers
2016: 4 papers
2017: expect 2-4 papers

long term: ~1 paper/year -> ~2030

expect ~10% of total ZEUS output

~80% of these would never exist without dedicated data preservation

Synergy with current experiments: LHC

- LHC collides protons on protons
- detailed knowledge of proton structure is crucial for many LHC physics topics, e.g. for measurement of Higgs boson properties
- in general, many common physics topics

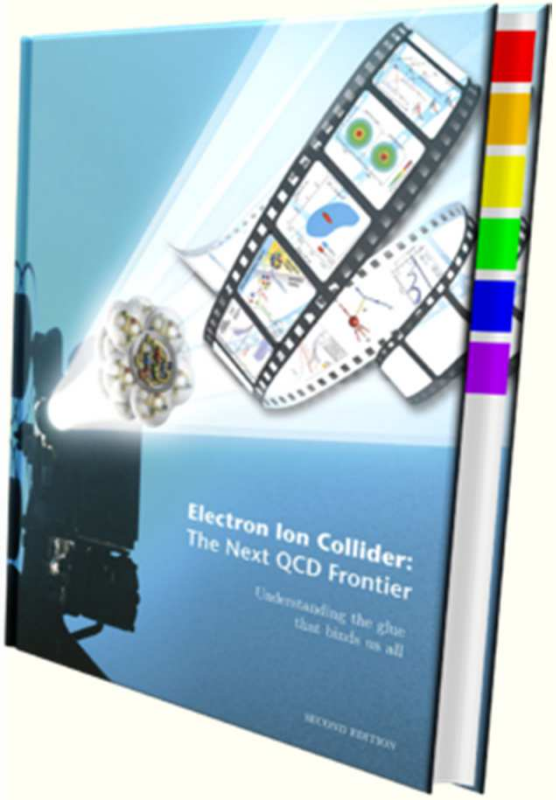
see also

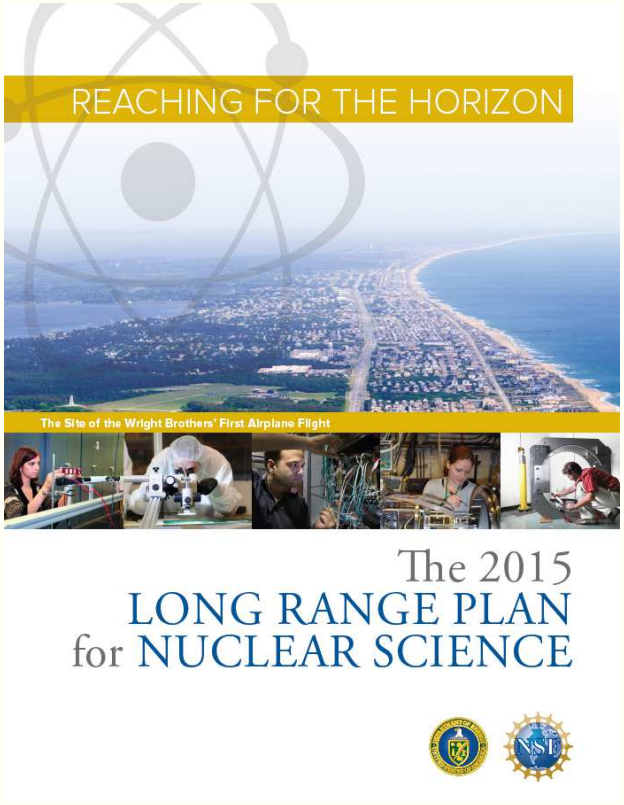
- HERA-LHC workshops, DESY and CERN
- workshop on Future Physics with HERA Data,
- DESY, November 2014, https://indico.desy.de/conferenceDisplay.py?confId=10523

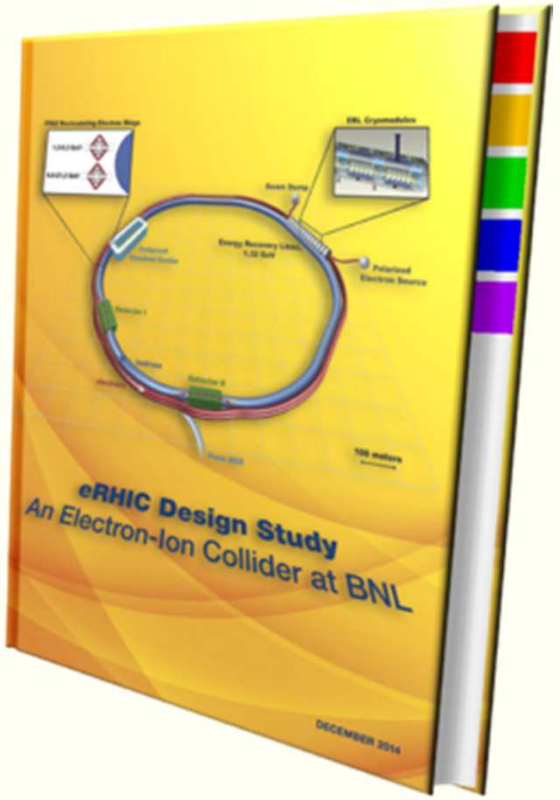
- some LHC Heavy Ion people have recently joined ZEUS to work on common analysis topics with ZEUS data in the context of ISOQUANT SFB1225

Synergy with future experiment: EIC

- many EIC topics common with HERA







- some EIC members have recently joined ZEUS to work on common analysis topics with real ZEUS data

Challenge:
"When will the project be finally done?"

- my answer:

  (usually hard to digest for host labs, funding agencies, committees …)

  if taken serious, a data preservation project will never be "done", unless and until one gives up on useability of the data

  (or the data get completely superseded by similar newer, better data sets)

Challenge:
"When is the best time to start?"

answer from HERA data preservation experience:

- the earlier, the better!
- the earlier one starts (with appropriate manpower, e.g. O(1%)? of running project) the more (data, documentation, expertise) precious information gets saved usefully, and the larger the resulting extra benefit will be.
  (we have achieved a lot, but we could have achieved even more)
- extra benefit >> extra cost

Conclusions and Outlook

- HERA data are scientifically unique and worth preserving !
- large parts of the original data preservation plans have been successfully implemented and are in active use.
- almost 10 years after end of data taking in 2007, thanks to data preservation, HERA scientific output continues at a significant rate, for very little cost.
- about 30% of total number of HERA papers produced after end of data taking. Made possible through substantial support by collaborations, host lab (DESY, IT), and external institutes!
- expect ~10% of total scientific output to originate from data preservation efforts (i.e. after end of funding), if long term sustainability is achieved.
  (1/3 of that already done!)
- Bottleneck: Long term data preservation needs long term manpower: don't need "much" (~O(‰) of original project investment, spread over 20 years), but 0 will not do …

Backup

Possible HERA collider physics topics

as discussed at Future Physics with HERA Data workshop

- BSM:
  - Provide standard candles against which new physics searches can be calibrated
- Proton structure:
  - FL combination, integration of high x results into PDF fit, finalize heavy flavour combinations and fit, improved transverse momentum dependent PDFs, investigation of low x phenomenology, …

  -> understand the proton, understand QCD, provide detailed descriptions for other colliders
  - Are we starting to hit the nonperturbative limit?
  - Can we make further decisive measurements from existing data?
  - Can we achieve improved theoretical interpretations from existing results?
  - Can statements about new physics at high scales be made from the low energy data?
- Diffraction and DVCS
  - Finalize inclusive diffractive measurements, make them more differential
  - Finalize measurements of elastic vector meson production and compare to improved theory models and to other experiments
  - Measure elastic scalar model production, test odderon hypothesis
  - Finalize measurements of DVCS

Possible HERA collider physics topics

as discussed at Future Physics with HERA Data workshop

- Jets:
  - Finalize (ZEUS) measurements, combine,
  - make more differential measurements, event shape measurements,
  - apply NNLO theory, remeasure alphas
- Hadronic final states:
  - Study multiparton interactions and other nonperturbative effects
  - (re)measure photon structure
  - (re)measure QCD instanton production
  - Search for exotic resonances
  - Complete total gamma-p cross section
- Heavy Flavours:
  - Intrinsic charm
  - NNLO measurements of c- and b-masses
  - Multi-differential heavy-flavour cross sections
  - More cross-section combinations
  - Improved measurements of charm fragmentation functions

ZEUS software approach

- original ZEUS data format and core software from 1990's
- maintenance of software, simulation and analysis framework needed **~4 FTE/year** (experiment) **+ IT**
- e.g. porting from SL4 to SL5 took about 2 years

-> not sustainable long term

-> go for simplified ZEUS data format:

"Common Ntuples" = flat ROOT ntuples

- almost no dedicated software maintenance needed

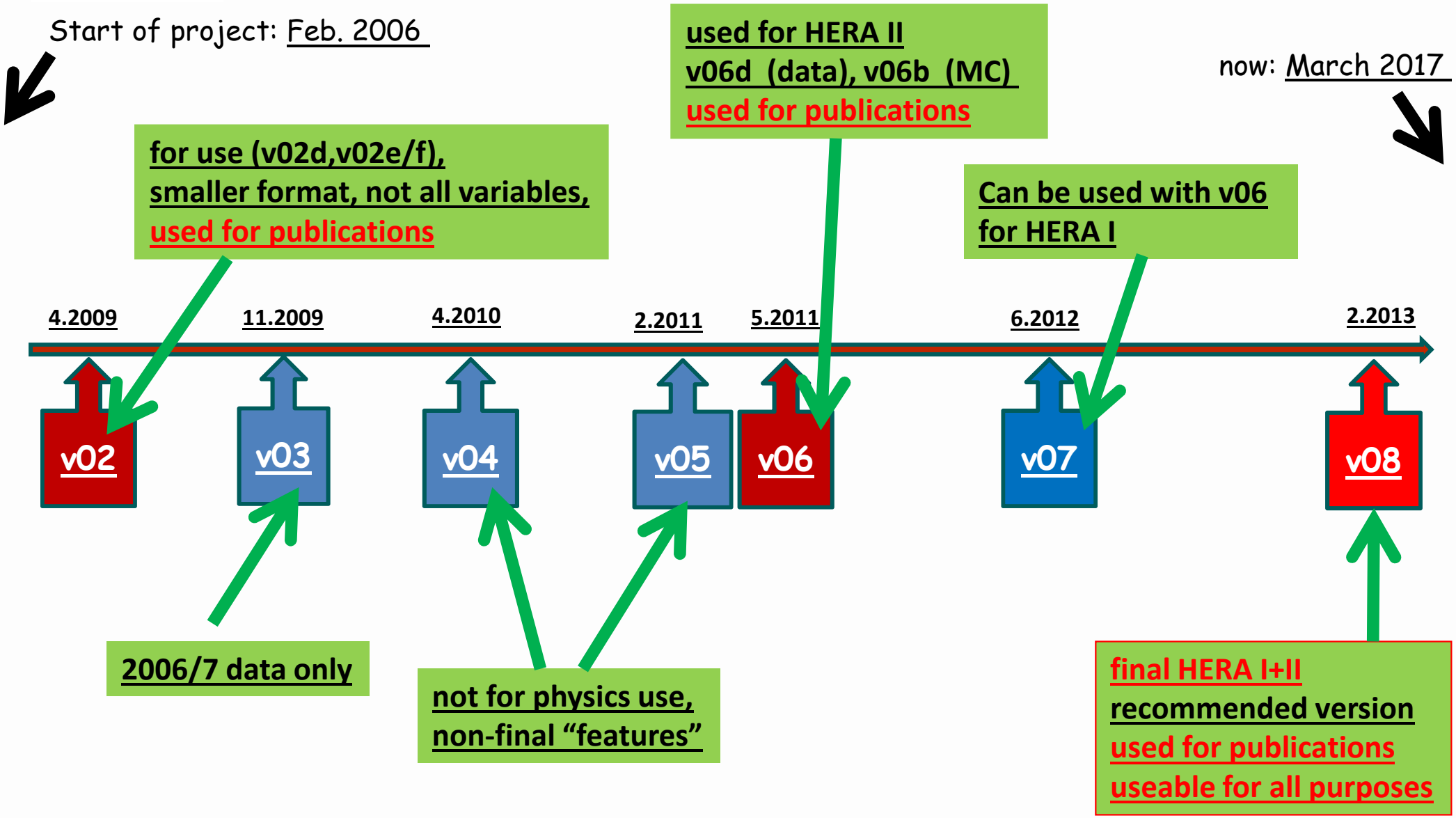
-> for new simulation: freeze software and run compiled executables in virtualized environment

- see also https://wwwzeus.mpp.mpg.de

managed at MPP

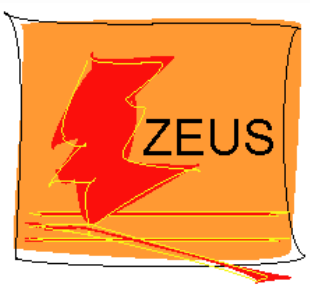
Available Common Ntuples

compiled by D. Szuba



Size of data sets

compiled by D. Zotkin/A.G.



Root files (officially preserved) units: Tb (status 4.9.13)

HERA II	v02	v06	v08	HERA I	v08 +v07	total	
Data	1.9	5.2	7.0	1.7+1.		17.	
MC	10.5	64.0	70.	4.8+4.		153.	+30 for future MC

~ 100 million inclusive DIS events (Q^2>5 GeV^2, triggered almost bias-free)

~ 100 million semi-inclusive photoproduction events (mainly via p_T>4 GeV dijet trigger)

smaller sets of more specialised triggers/samples (e.g. heavy flavours, vector mesons, …)

~ equal sample sizes for e+, e-, righthanded/lefthanded polarisation

~ 4 billion MC events, for almost any analysis

generation of additional MC samples might be possible (see talk A. Verbytskyi)

can technically read/analyze full ZEUS data set on NAF/BIRD at DESY within ~1 day

(for even faster access, many analyzers produce their own mini-ntuples for analysis)

Some ingredients for success of actual project

- Make sure you start the 'user mode' well (>~ 2 years) before the temporary manpower ends (-> need to be able to fix "hickups" !) ☺

  ZEUS: user data preservation mode gradually started 2011-2013

- Ensure strong support of host lab or other funding body during the 'long term benefit' phase 😐

  ZEUS: scientific support OK, long term manpower/minimal funding support more difficult than expected/hoped for

- Make sure to get the necessary **dedicated long term manpower** (and funding!) going along with this support ☹

  ZEUS need: ~2/3 short term ~1/3 long term (~20 year integral)

  people understand the need to maintain storage, networks and tape vaults, and to provide some minimal CPU power, but rarely understand the (size of the) manpower need for **knowledge preservation, software preservation, and user support …**

  -> this is the main point upon which some (parts of) current projects risk to fail

personal view