


Data preservation, re-use and (open) access.
A case study in High-Energy Physics

- **Background information**
- **Highlights from a survey**
- **Food for thoughts**

Andre Holzner (CERN), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)

PARSE.Insight Workshop
Darmstadt, June 21st, 2009

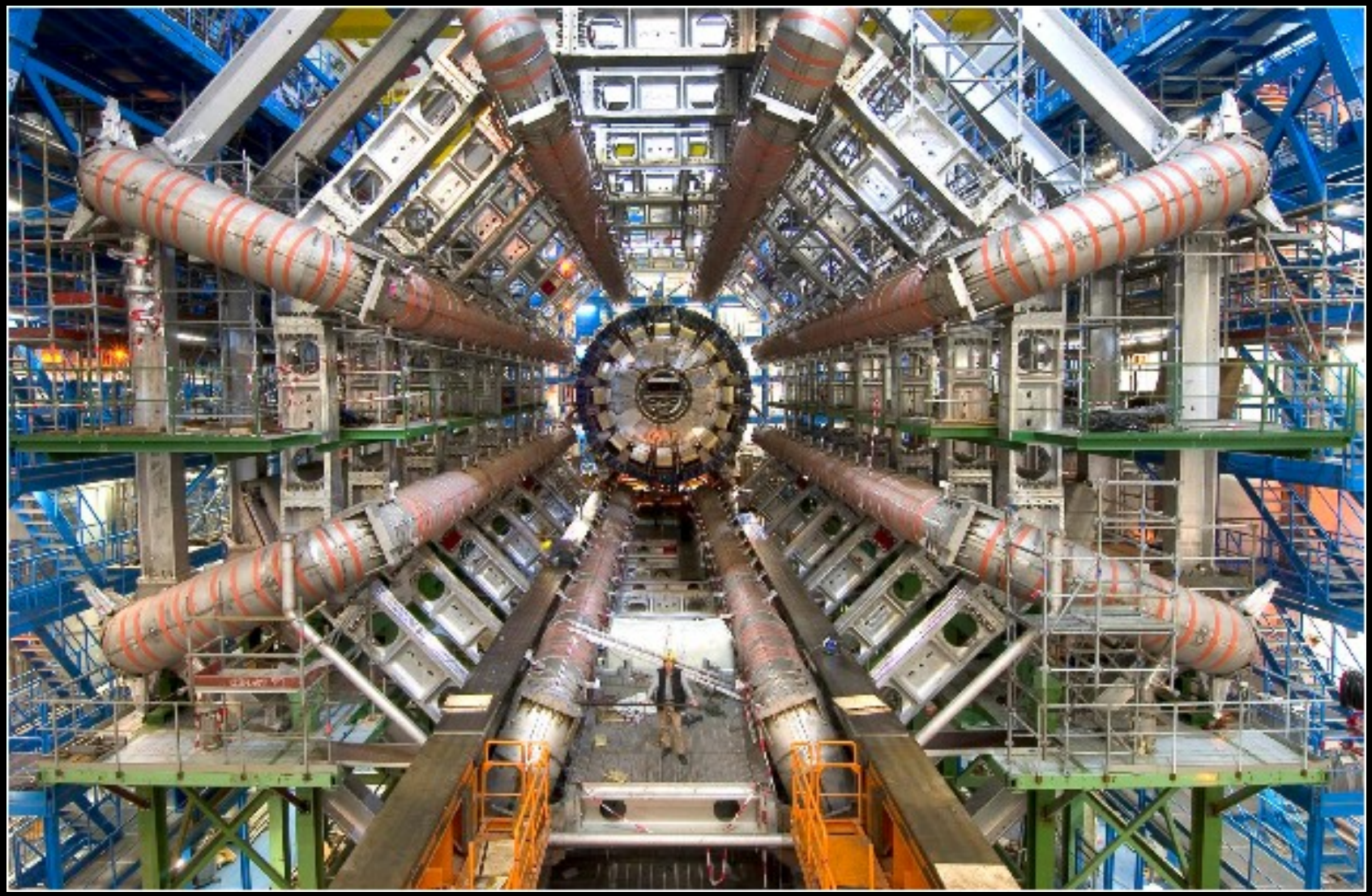
Data preservation, re-use and (open) access. A case study in High-Energy Physics

- **Background information**
- Highlights from a survey
- Food for thoughts

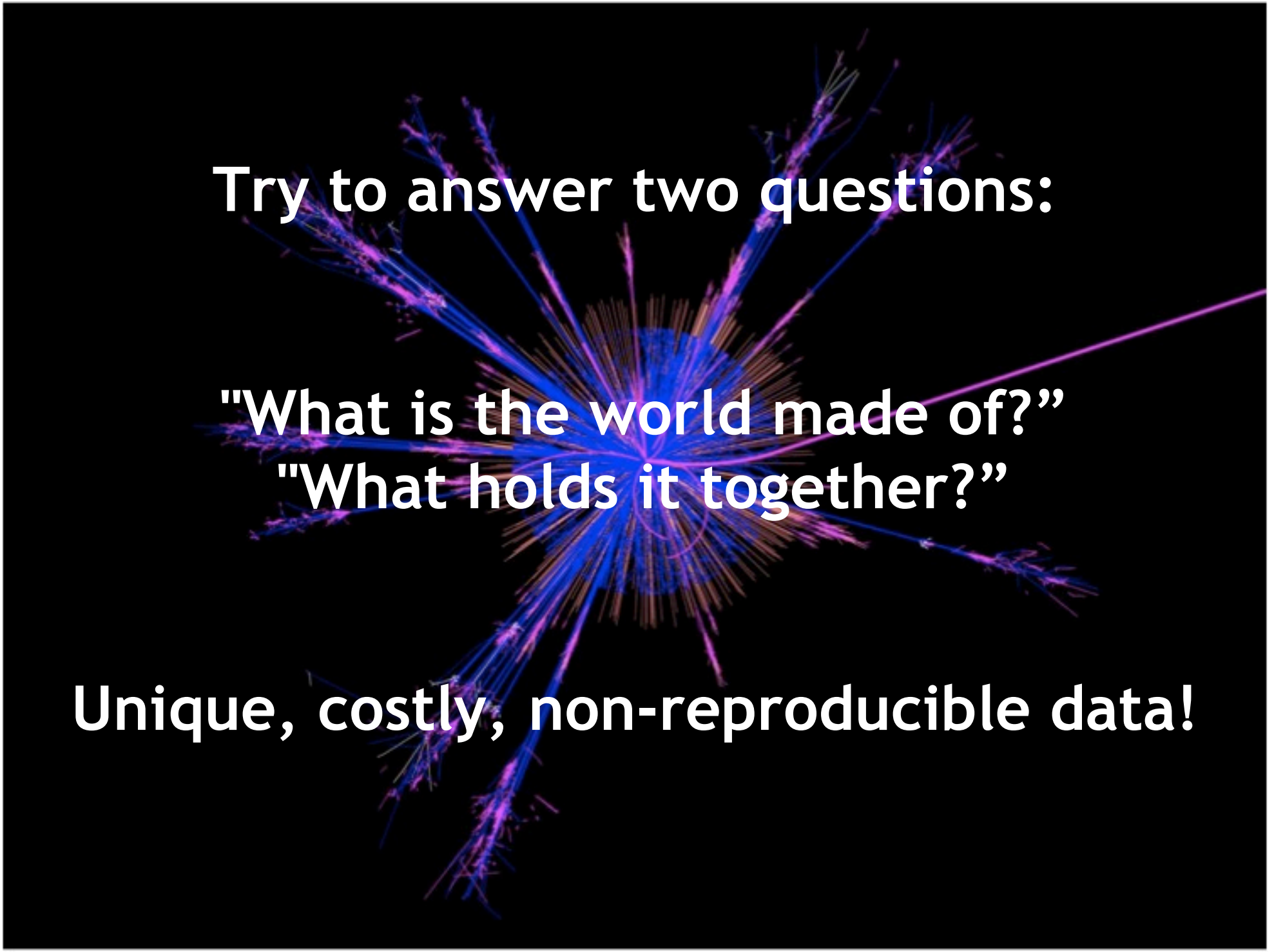
Andre Holzner (CERN), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)

PARSE.Insight Workshop
Darmstadt, June 21st, 2009

~15'000 High Energy Physics (HEP) scientists smash stuff at the speed of light to produce new stuff



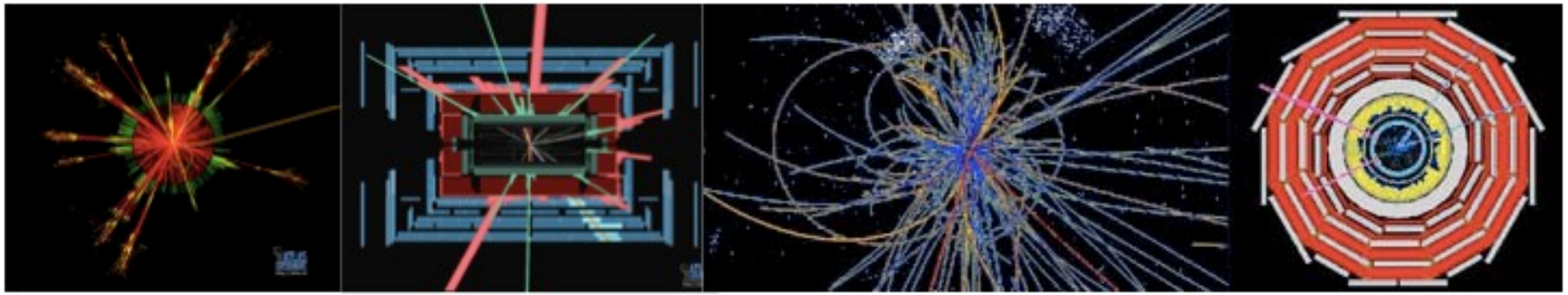
~15'000 HEP theorists scratch their heads to make sense of all that stuff and then some more


Try to answer two questions:
"What is the world made of?”
"What holds it together?”
Unique, costly, non-reproducible data!

The LHC data (2009-202?)

- 40 million events/second ("100MP pictures")
- ~200 interesting events/second on tape
- "Reconstruct" raw data
- Study "physics data" [grid anyone?]

(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



Preservation, re-use, (open) access continua

- Same researchers who took data, after the closure of the facility (~1 year, ~10 years)
- Researchers at similar facilities at same time (~1 day, ~1 week, ~1 month, ~1 year)
- Researchers of future (~20 years)
- Re-interpretation by theoretical physicists (~1 month, ~1 year, ~10 years)
- Theoretical physicists testing future ideas (~1 year, ~10 years, ~20 years)

The PARSE.Insight HEP Case Study

Who ?

- Researchers at facilities recently closed (2007-2010)

Why ?

- Pioneer community in e-Infrastructure/e-Science
- Perceived lack of action on data preservation

What ?

- Motivations *vs.* Concerns
- Opportunities *vs.* Threats
- Wishes *vs.* Obstacles

How ?

- Large scale survey (1200 respondents worldwide)
- Interviews. Synergy with other technical workshops

Data preservation, re-use and (open) access. A case study in High-Energy Physics

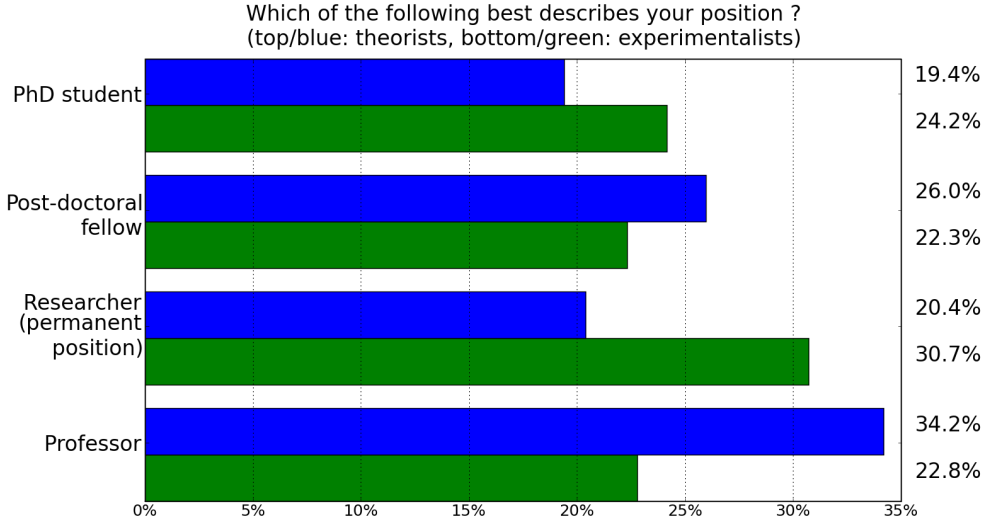
- Background information
- **Highlights from a survey**
- Food for thoughts

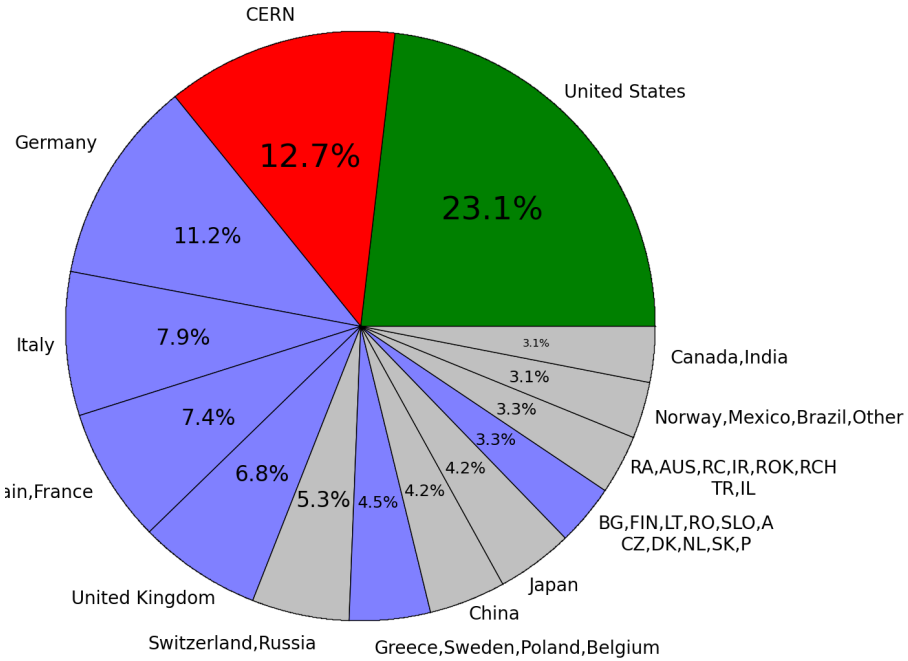
Andre Holzner (CERN), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)

PARSE.Insight Workshop
Darmstadt, June 21st, 2009

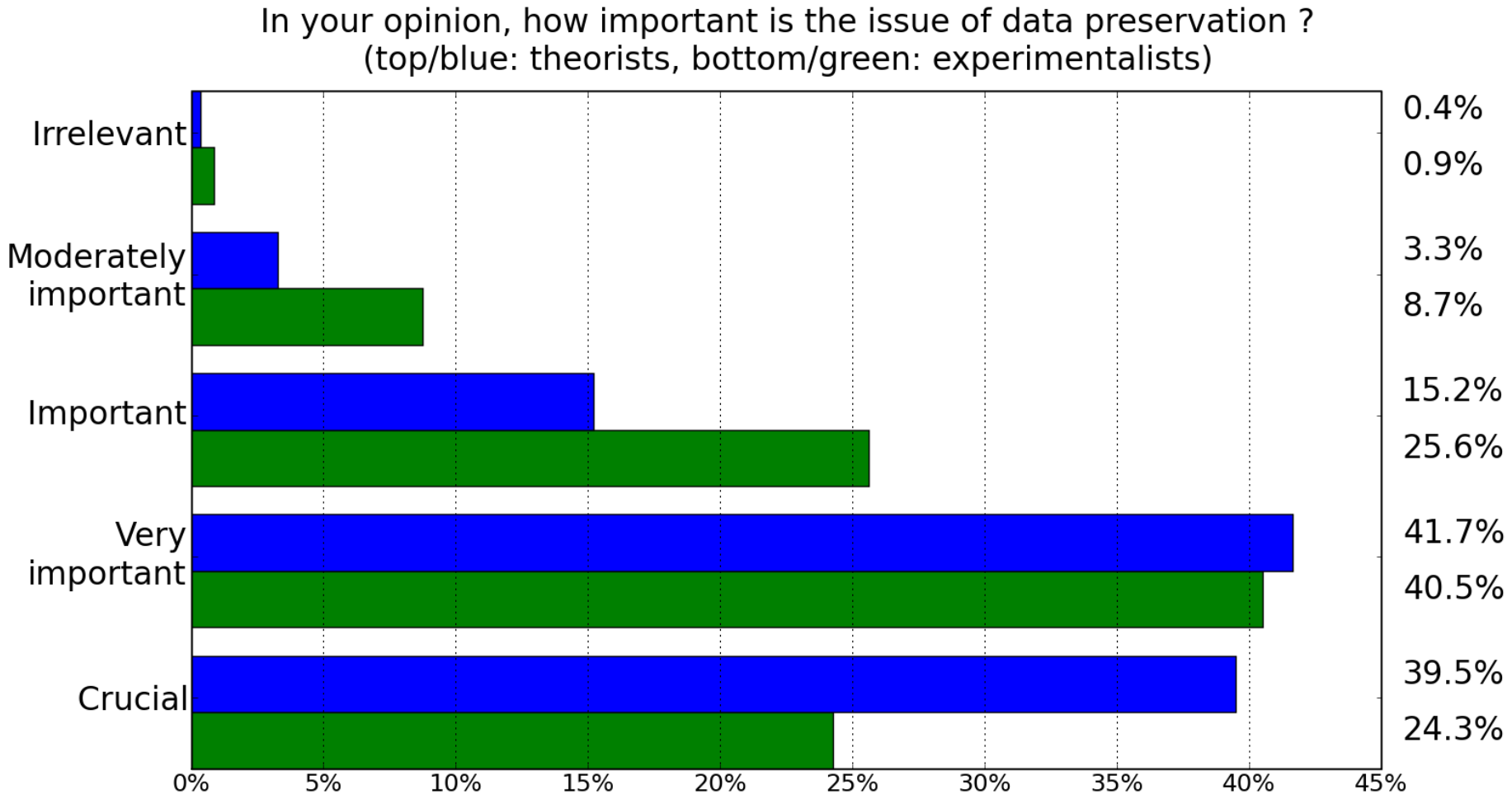
Survey demographics

- Livetime of 3 months 10/08-01/09
- Advertised on
  - mailing lists of large experimental teams
  - front page of main community digital library
- 1’200 answers (74% exp, 25% th). Target 20K-30K





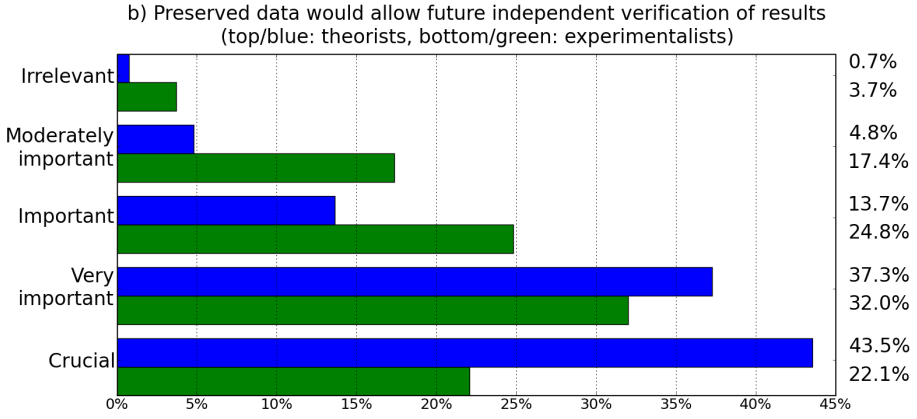
The importance of preservation



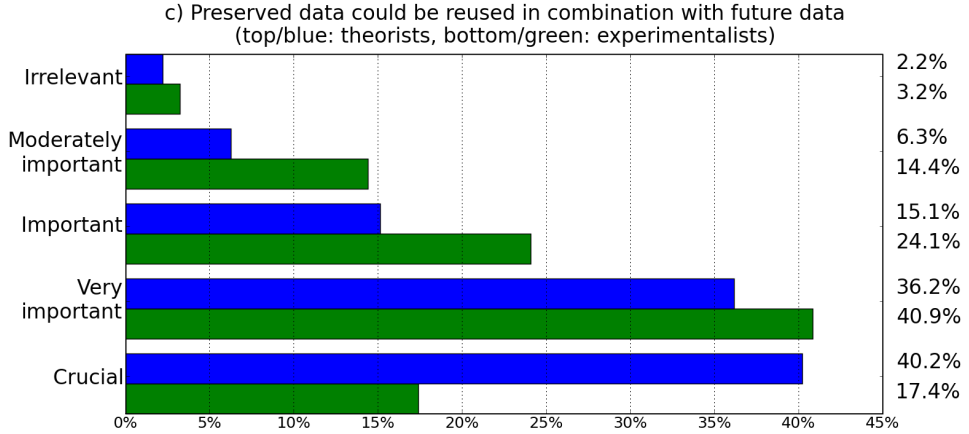
“Very important” + “Crucial” = 69%

The importance of preservation

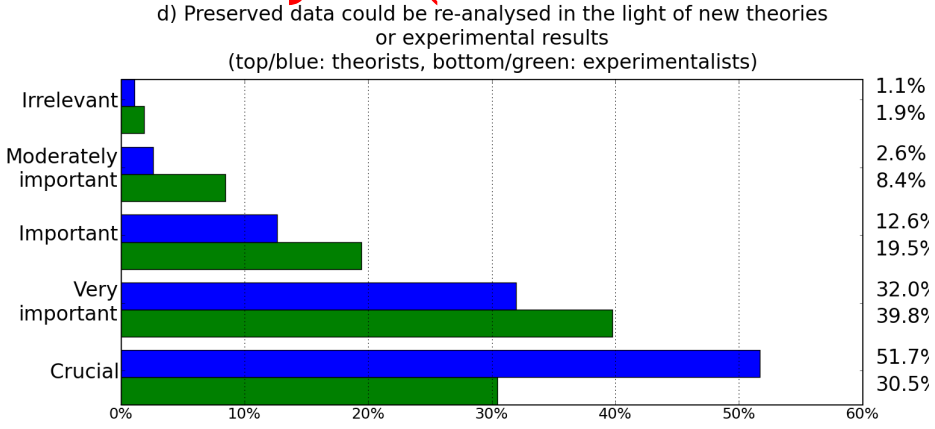
Future independent checks



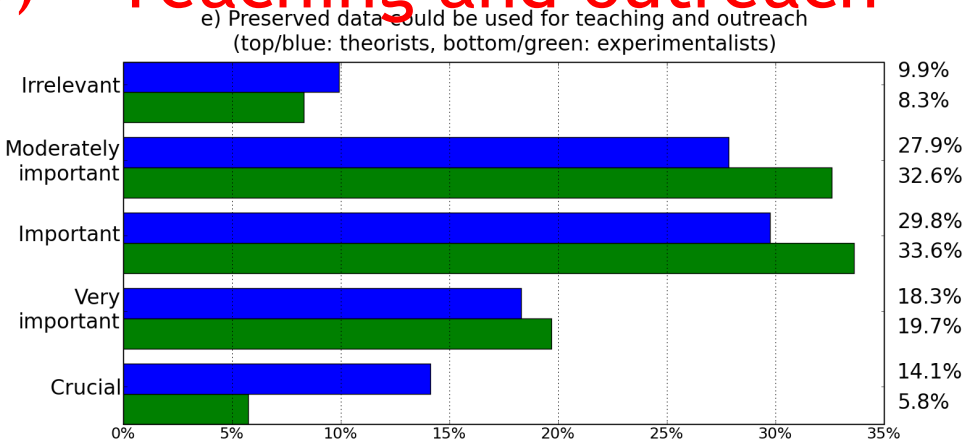
Re-use with future data



Re-analyse (future theories)



Teaching and outreach



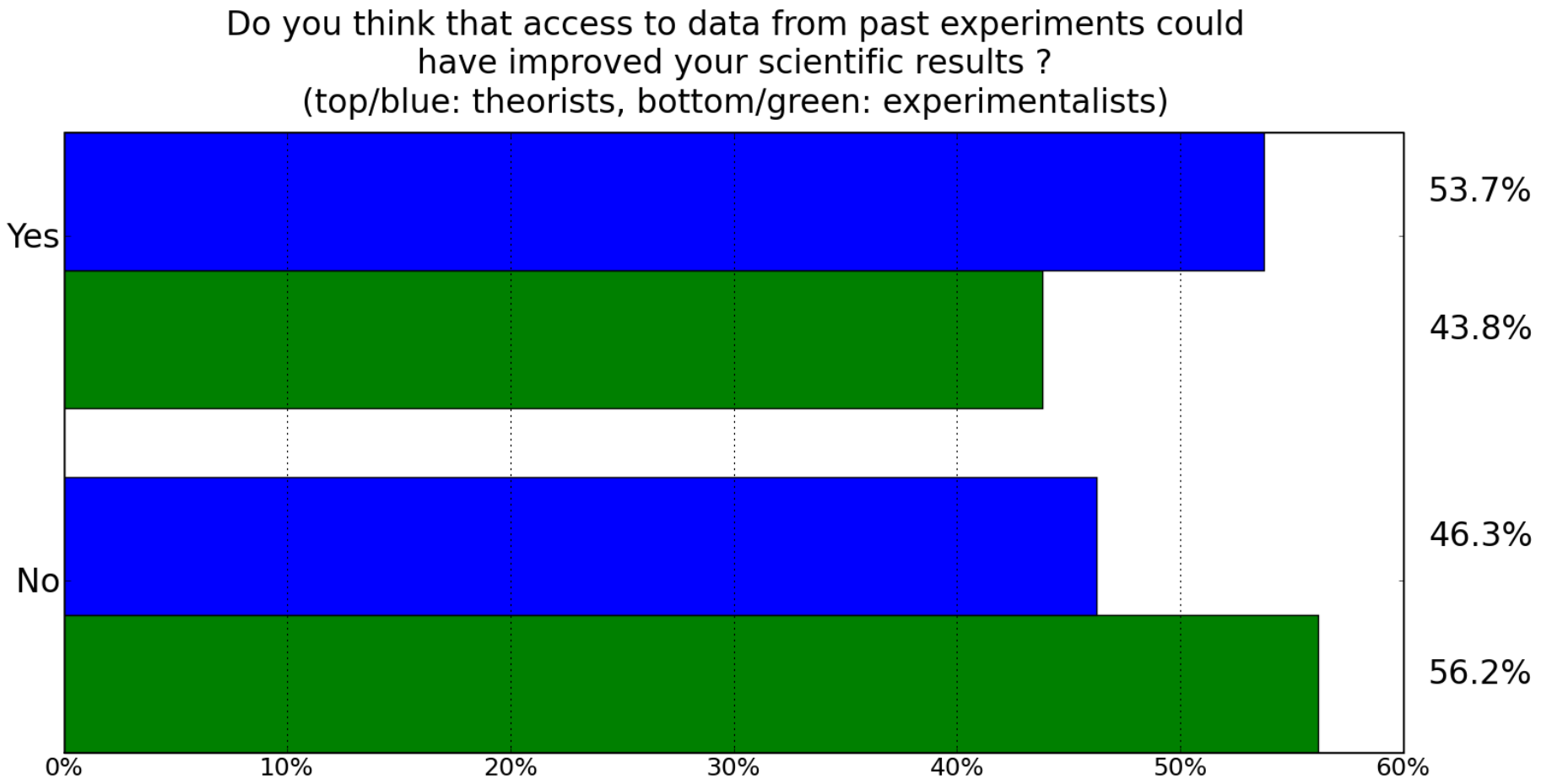
CERN

PARSE insight
Permanent Access to the Records of Science in Europe

e-infrastructure

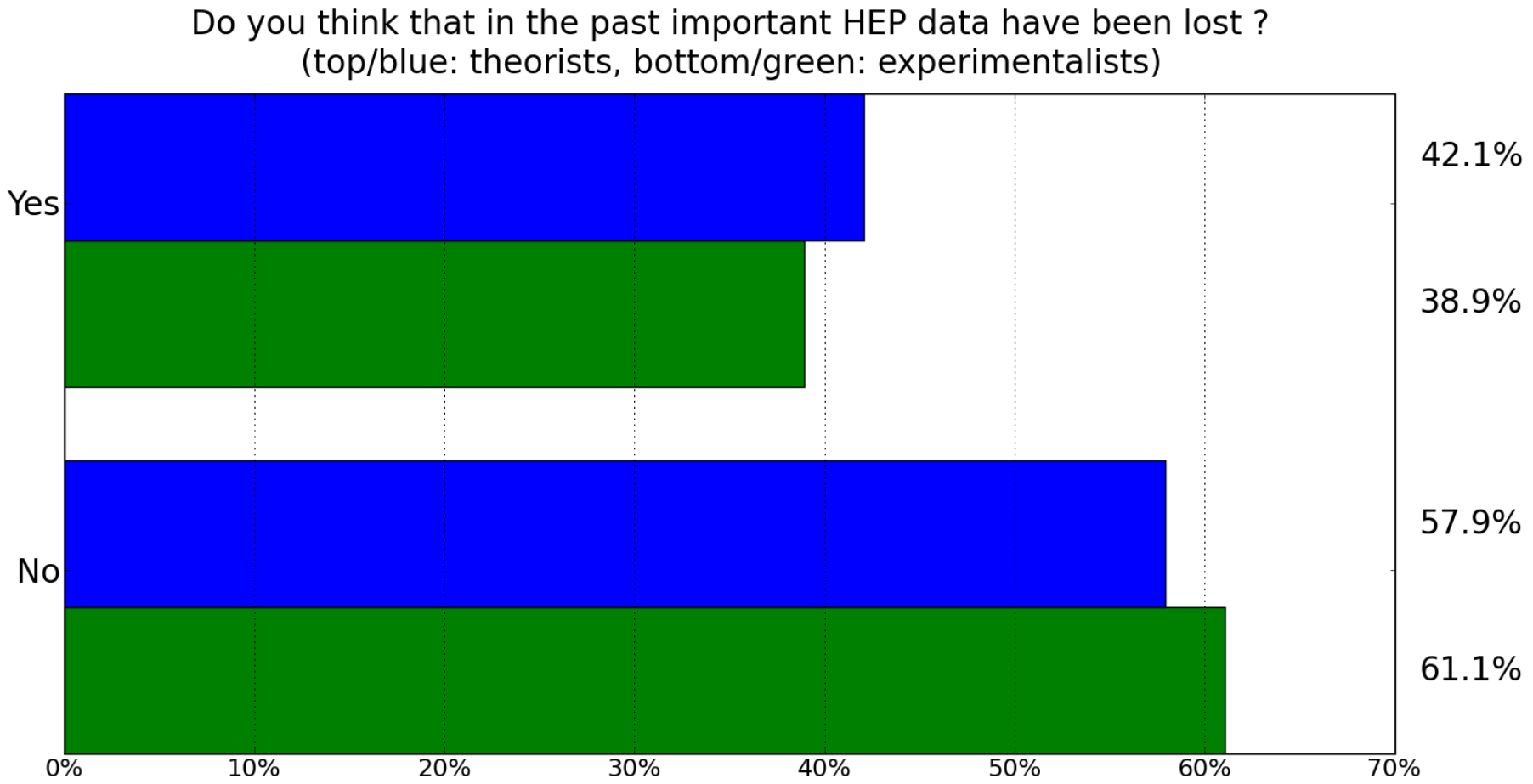
SEVENTH FRAMEWORK PROGRAMME

Should we have started long ago?



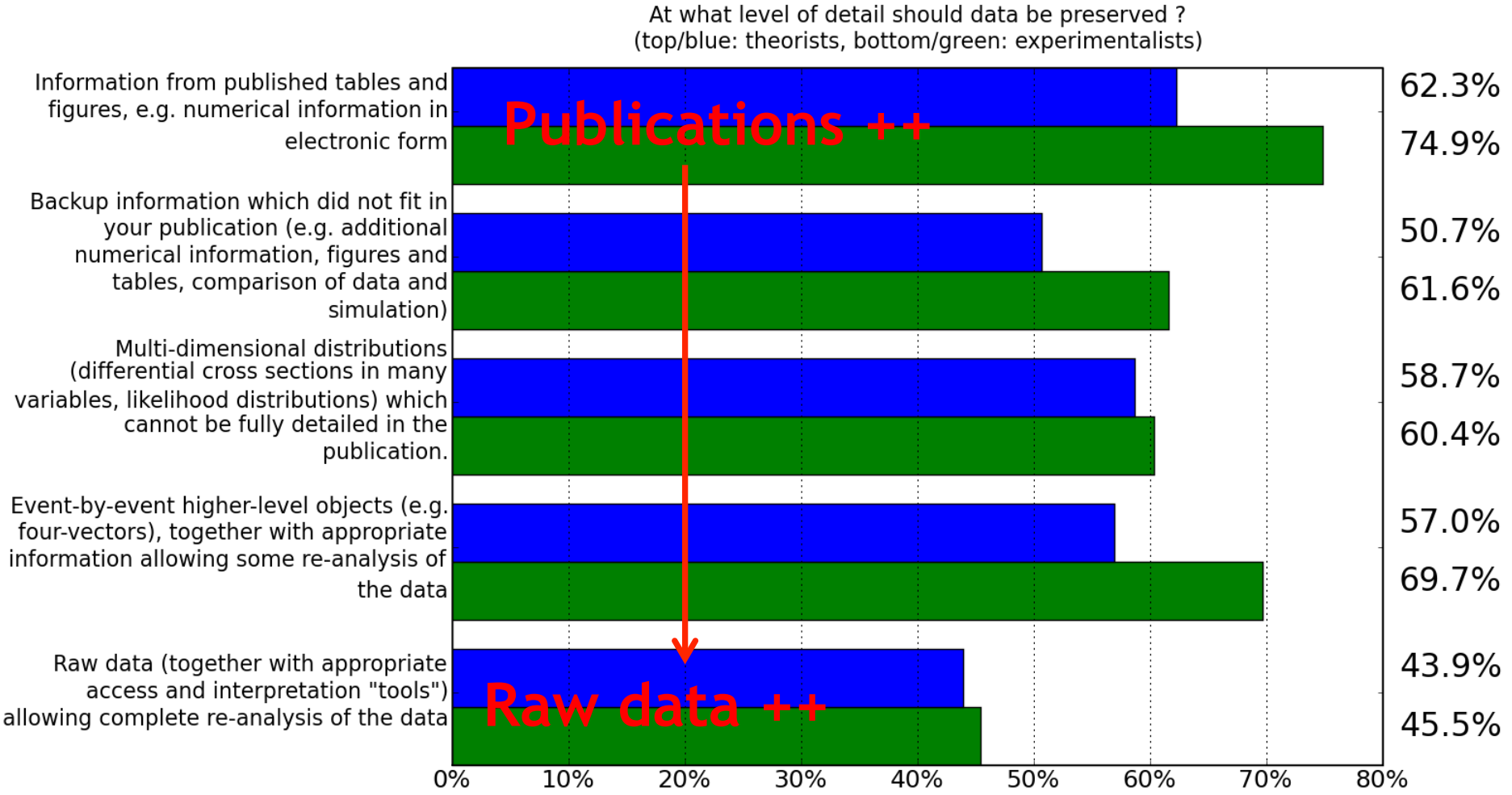
Preservation (and re-use) enables science !

Did anything go wrong so far?

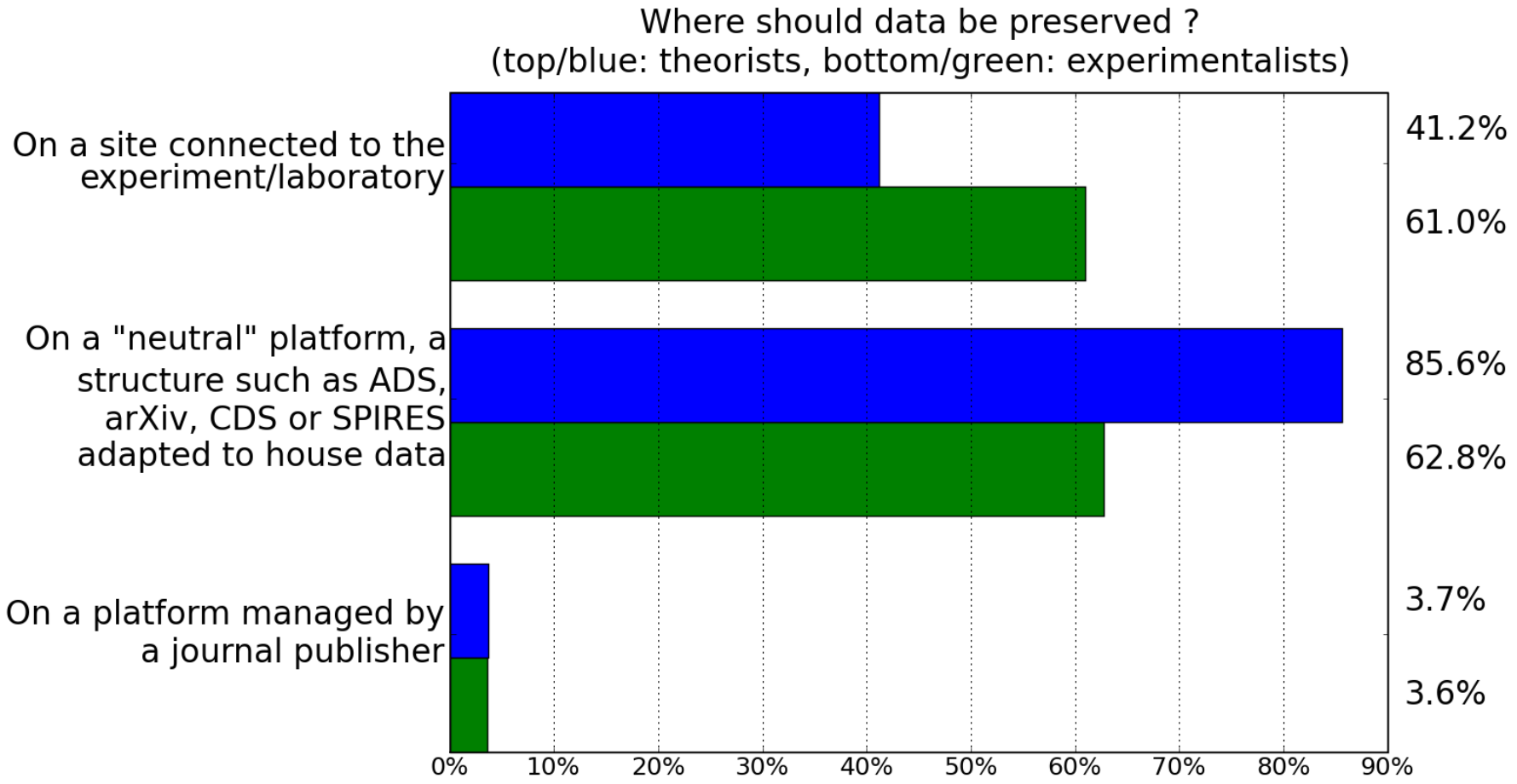


Over optimistic? Over pessimistic?

What to preserve?



Where to preserve?



Where to preserve? Attributes of e-infrastructure

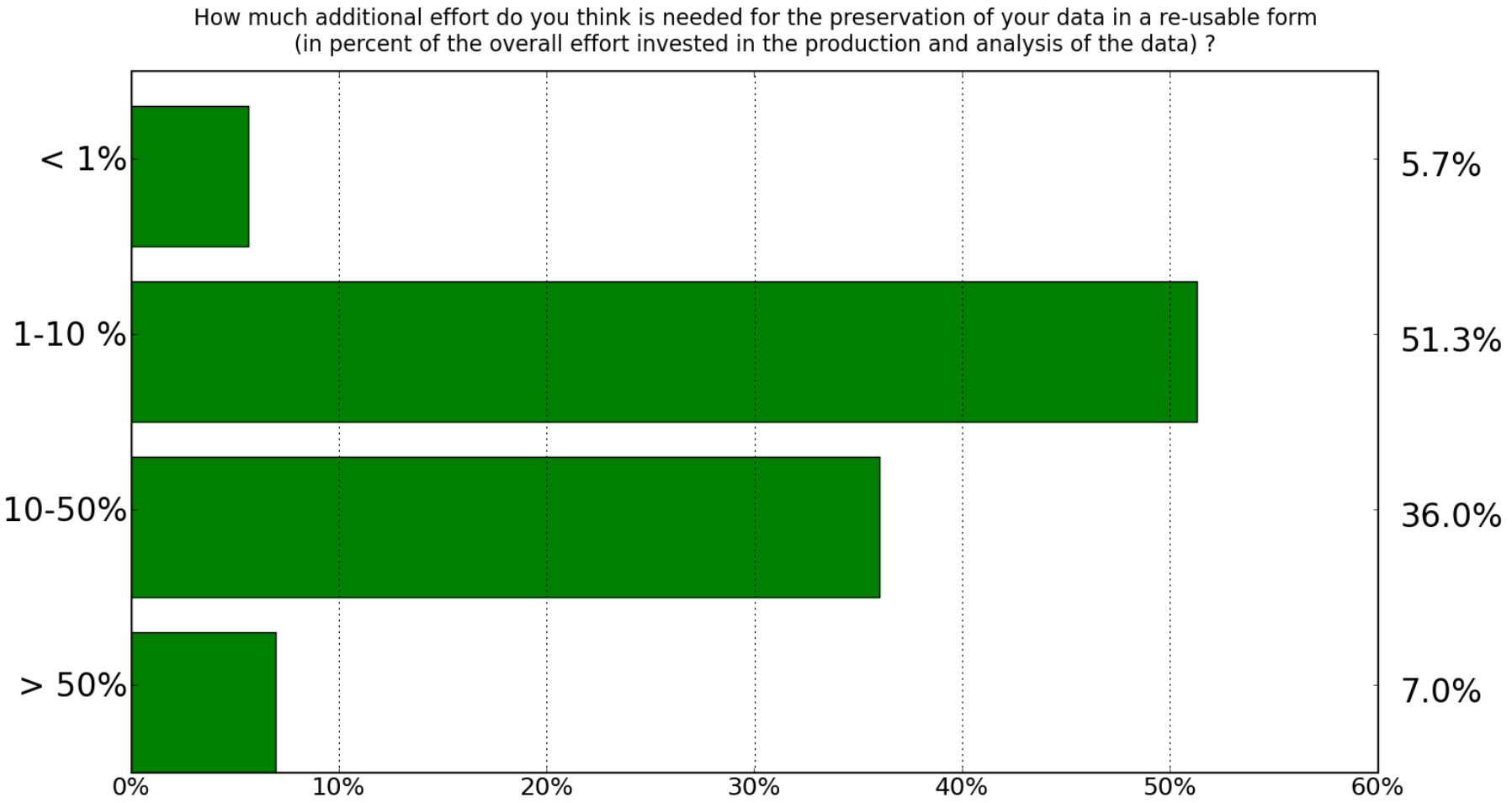
CERN

PARSE insight
Permanent Access to the Records of Science in Europe

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

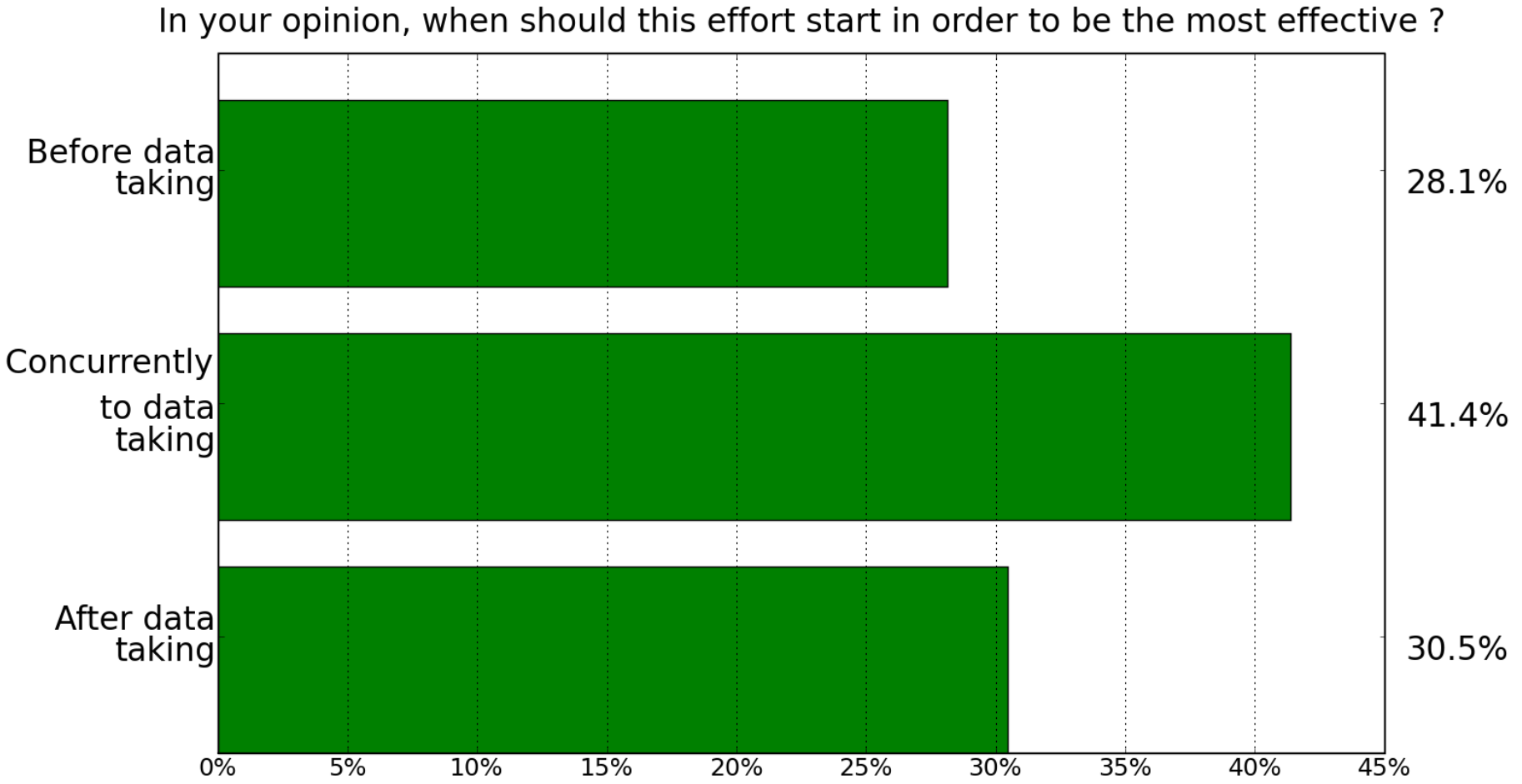
Reality check #1: how though is it to preserve?



Additional effort on top of taking data…
…what sums up to thousands of person-years!

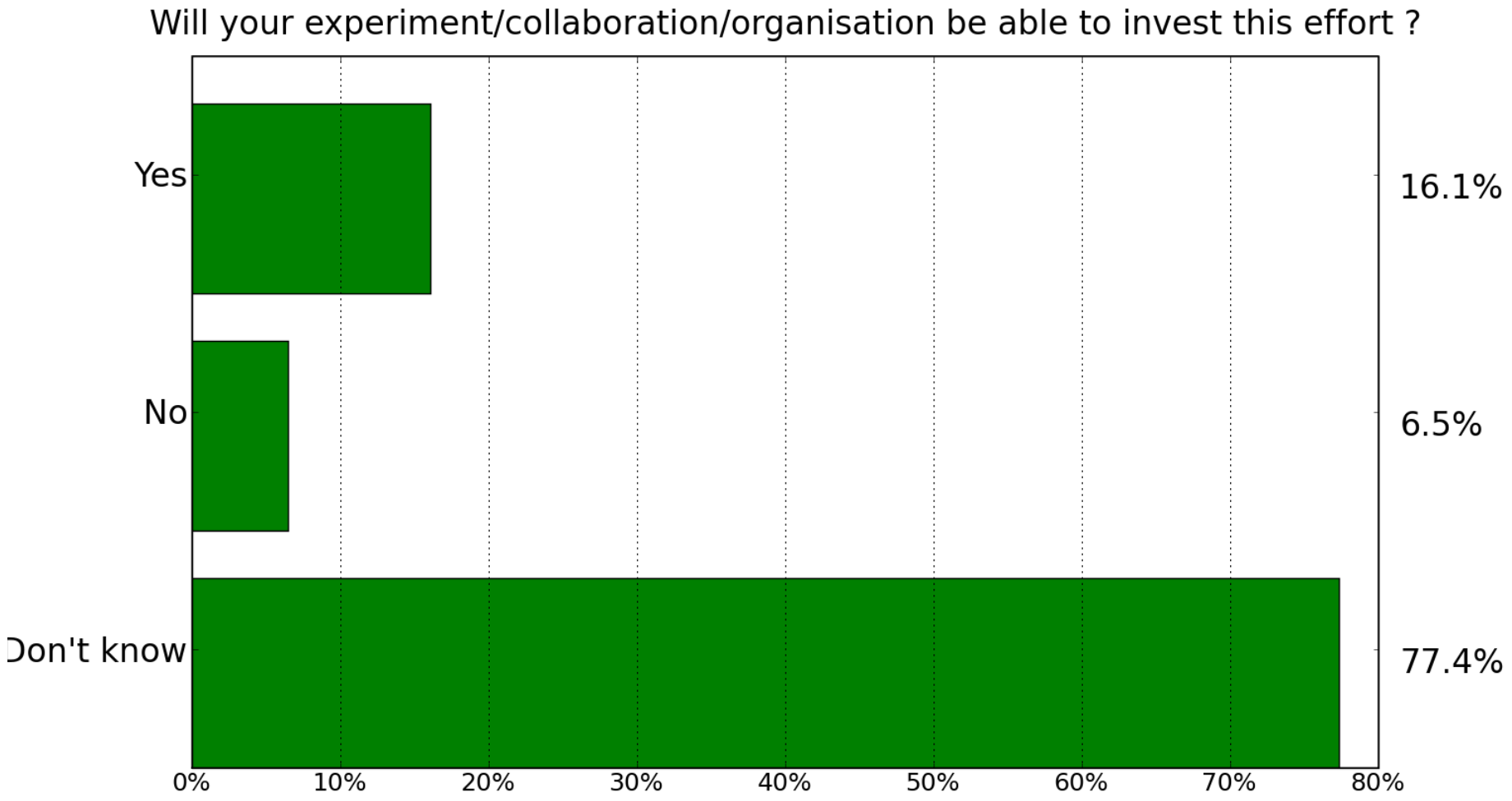
CERN PARSE insight Permanent Access to the Records of Science in Europe e-Infrastructure SEVENTH FRAMEWORK PROGRAMME

Reality check #2: when to preserve?

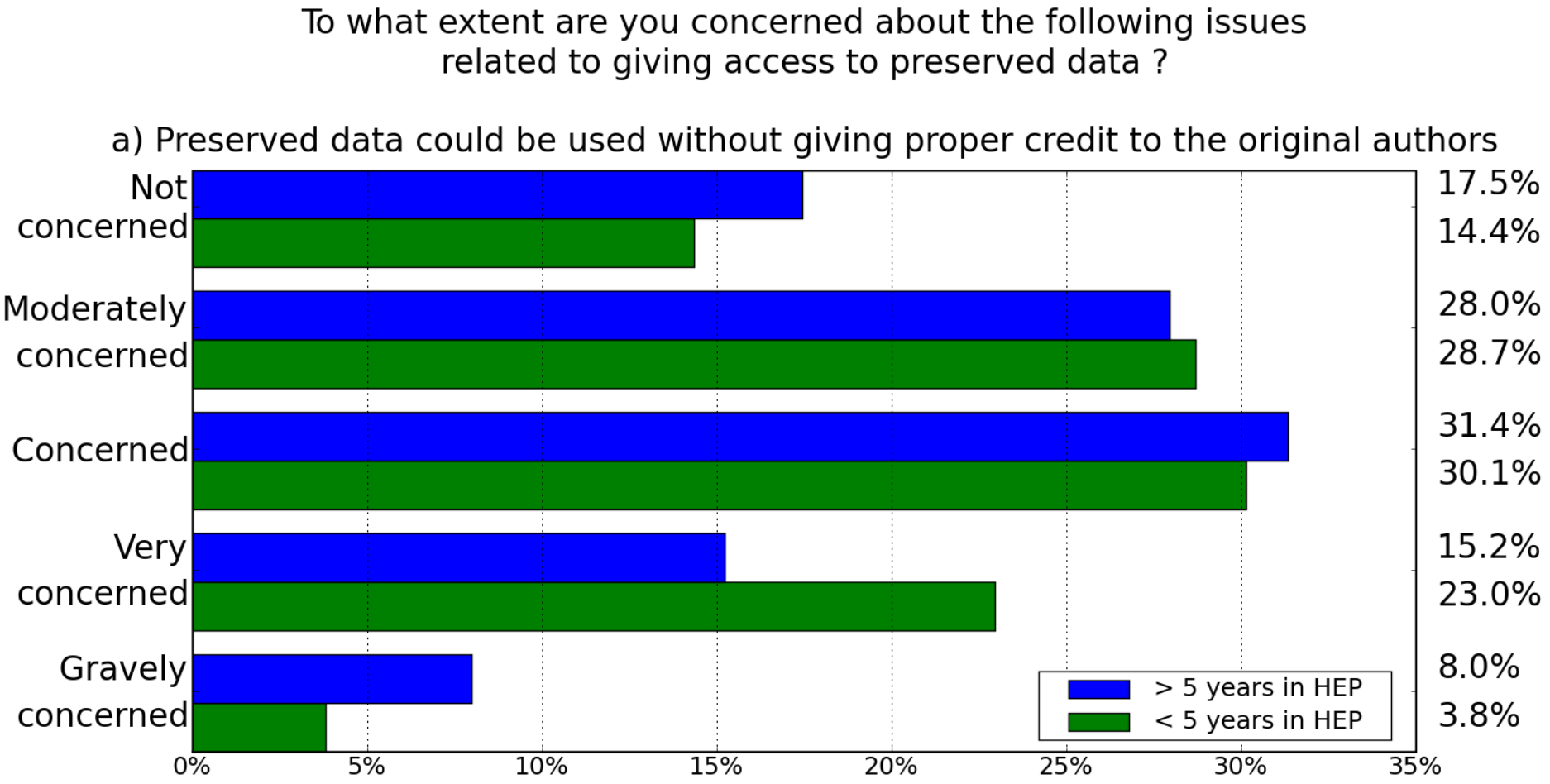


Note: ALL non-reproducible data already taken!

Reality check #3: is it doable?



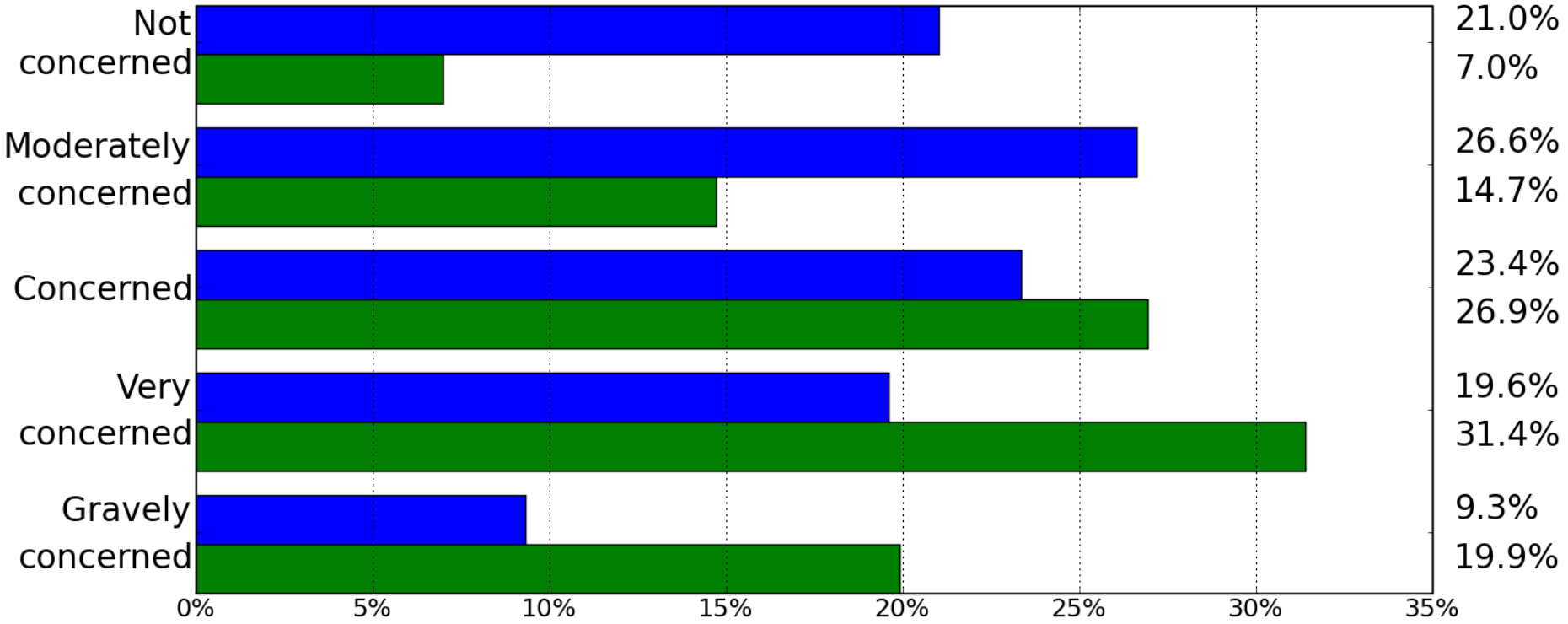
Ideal-case worries: getting credit



Ideal-case worries: inflation/noise

To what extent are you concerned about the following issues related to giving access to preserved data ?

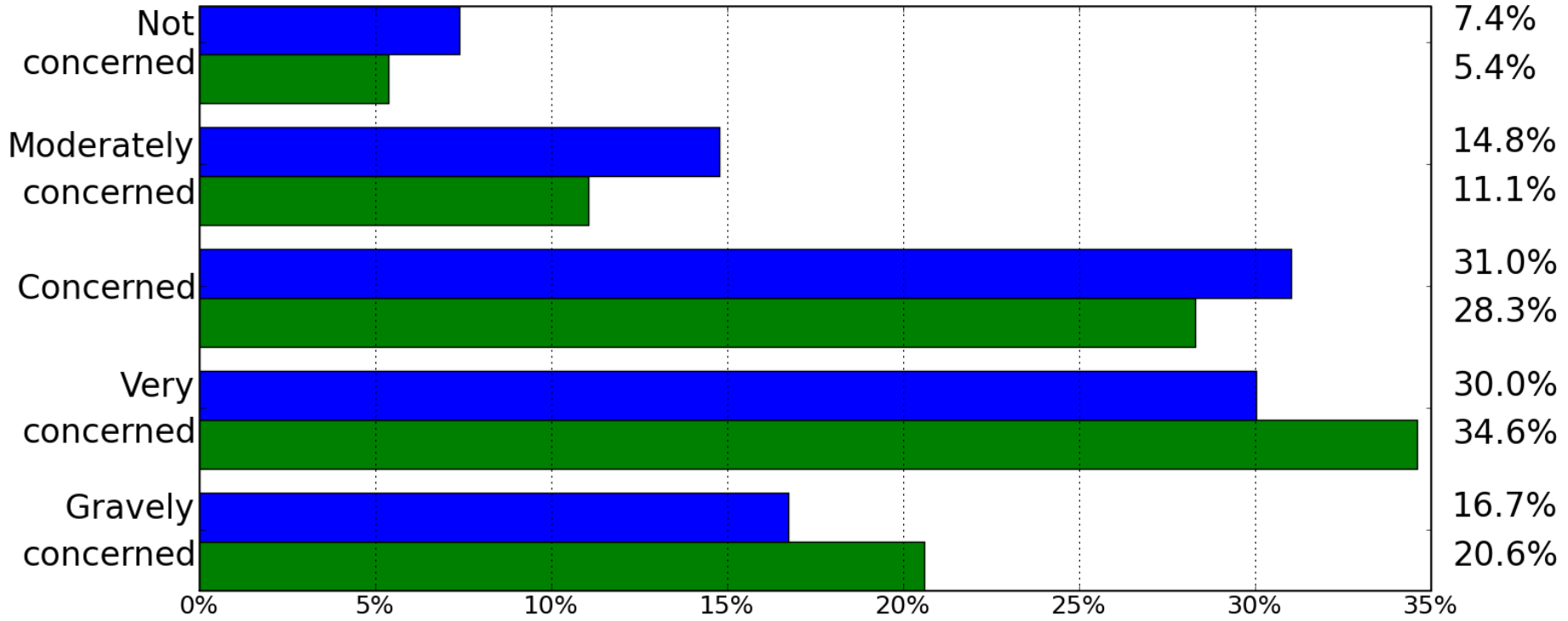
b) Uncontrolled access to data may lead to an inflation of incorrect results
(top/blue: theorists, bottom/green: experimentalists)



Ideal-case worries: documentation

If you were to re-use preserved data, to what extent would you be concerned by the following scenarios ?

d) I am not using the data correctly
(top/blue: theorists, bottom/green: experimentalists)



Data preservation, re-use and (open) access. A case study in High-Energy Physics

- Background information
- Highlights from a survey
- **Food for thoughts**

Andre Holzner (CERN), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)

PARSE.Insight Workshop
Darmstadt, June 21st, 2009

First results from in-depth interviews

Q: What would you do if you were in charge?
A: Have new (younger and hungry for data) people port and document software

Q: How to protect from (bona-fide) crackpots?
A: Non-issue. Increase in 'background noise' overcompensated by increases of 'signal': good ideas that would not be checked otherwise

Q: How to organize full access?
A: Political agreement first (European level/Open Access) and then a real agreement to standardize

In their own words (full-text answers to survey)

The most important reasons for preservation are the ones we do not see now

Often we do not know what the crucial data will turn out to be. Only with hindsight this becomes apparent

Each set of data is unique: once lost, lost forever

The fact that HEP data is closed is a historic accident

There are very few examples of preserved HEP data; almost all HEP data is lost in the sense of your survey

Why throw away something we might use later?

Not preserving data is simply a <expletive deleted>

In their own words (full-text answers to survey)

Documentation: that's the real bottle-neck. Need to explain the real limits of how data can be re-used

Data cannot be used properly at a later time by people not participating in the experiment, but also by those who did, once the infrastructure is dismantled

Fraud is related to ethics; the more open the data the harder to manipulate since they are easier to check.

Future publications should link to preserved data

Do NOT preserve in a <expletive deleted> journal!

Thank you!

Andre.Holzner@cern.ch
Peter.Igo-Kemenes@cern.ch
Salvatore.Mele@cern.ch

More on Digital Preservation in HEP:
http://arxiv.org/abs/0906.0485
http://arxiv.org/abs/0805.2739
http://dphep.org