First results from PARSE.Insight

The HEP survey on
data preservation, re-use and (open) access

- **Background information**
- **The HEP Survey**
- **Next steps**

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

First Workshop on Data Preservation and Long Term Analysis in HEP
[Open Access and Long Term Collaborative Governance]
DESY - 26-28 January 2009

Actors

Alliance for Permanent Access

Association of stakeholders with an interest in digital preservation of publications and research data.

- Research institutes (CERN, ESA, STFC, Helmholtz, MPG)
- Publishers (STM association)
- Libraries (DNB, KB, BL)

CERN involvement as natural extension of Open Access mission and exploration of possible opportunities

After connectivity (GEANT) and Grid (EGEE), both seen as parts of a e-infrastructure, and development of “repositories” (=databases) for publications,FP7 might look at data and data preservation as next frontier and need information to make strategic funding decision

PARSE.Insight

http://www.parse-insight.eu/

- Small (1.2M€) FP7 project (CSA) with partners from science (CERN, ESA, STFC, MPG) corporations (STM) and libraries (KB,DNB,...). 3/2008 to 2/2010. Lead by STFC (CASPAR, OAIS)
- Interdisciplinary study to offer an insight on:
  - "who is doing what in the field of digital preservation [of research data] and why they are doing it that way"
  - "gap" between what should be done and what is done
- All-around approach and case studies. HEP is the largest case study.
- Main deliverables (to inform FP7 policies/strategies):
  - Insight and Roadmap on issues/threats/opportunities in digital preservation
  - (Gap analysis: what's there what should be there)
  - Example from some communities (N.B. HEP innovator elsewhere!)

The PARSE.Insight HEP Case Study

What it is not:

- Technical solutions for data formats
- Technical solutions for data migrations
- This workshop!

What it is:

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

Our target:

- Make the scientific case with FP7 to use HEP as a case study for a e-infrastructure pilot in preservation ?
- Gather evidence on attitudes of the community to be used when and if we will have to make policy decision on access

The PARSE.Insight HEP Case Study

Three-pronged approach (Sept. 2007)

- Large-scale survey of the HEP community
- Follow-up and ad-hoc interviews to go into
  - technical details
  - approaches of past and current experiments
  - "superusers" vision
- Tripartite workshop with data producers, consumers, IT support to expose "gaps" between wishes, needs and reality

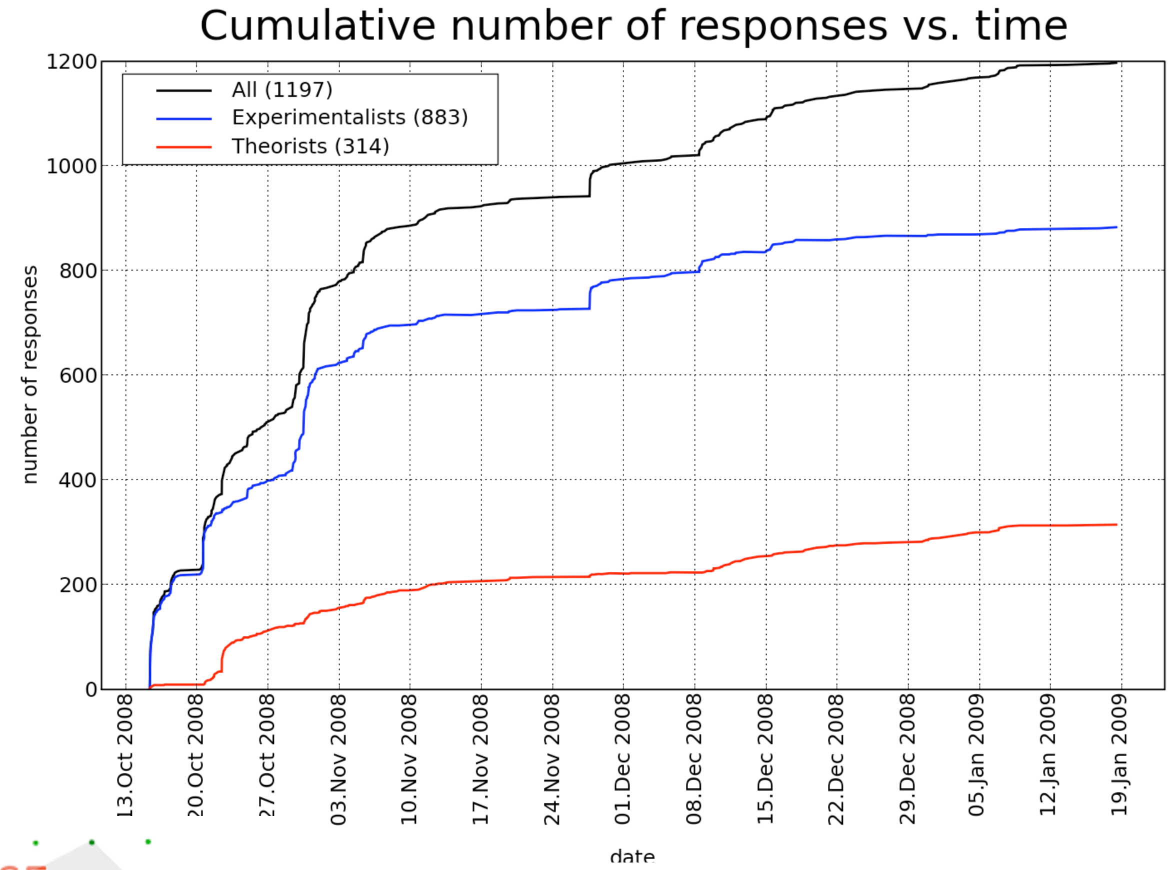
Re-scoping for synergy with this/these workshop(s)

- Large-scale survey untouched.
- On-demand and ad-hoc interviews to deepen and complete the results of the survey
- Workshop to be re-scoped following our discussions here

Today: first results from the survey

Survey strategy and response

- Livetime of 3 months 10/08-01/09
- Advertised on collaboration and theory mailing lists
- One or two reminders, according to response monitoring
- Advertised on SPIRES twice (reach more spread theorists)
- 1’200 answers (74% exp, 25% th). Target size ~20’000-30’000

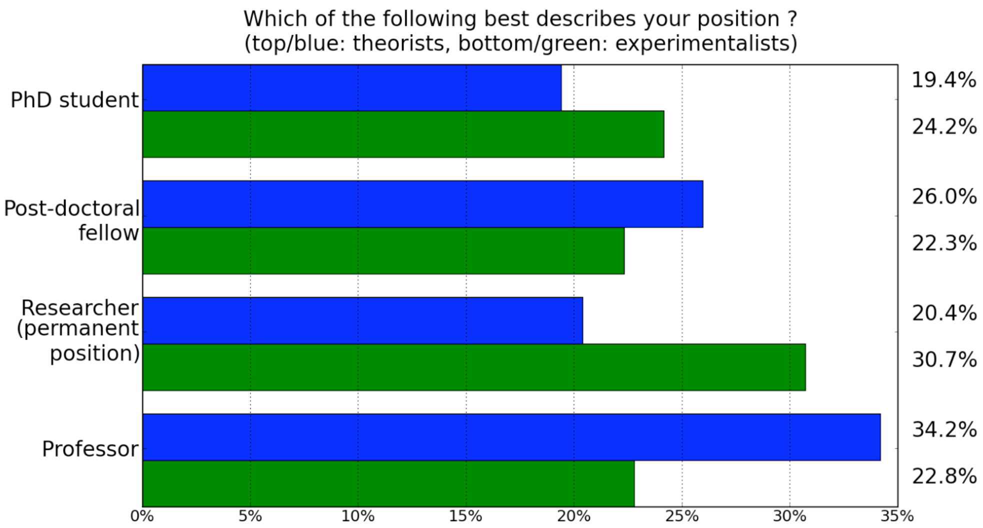


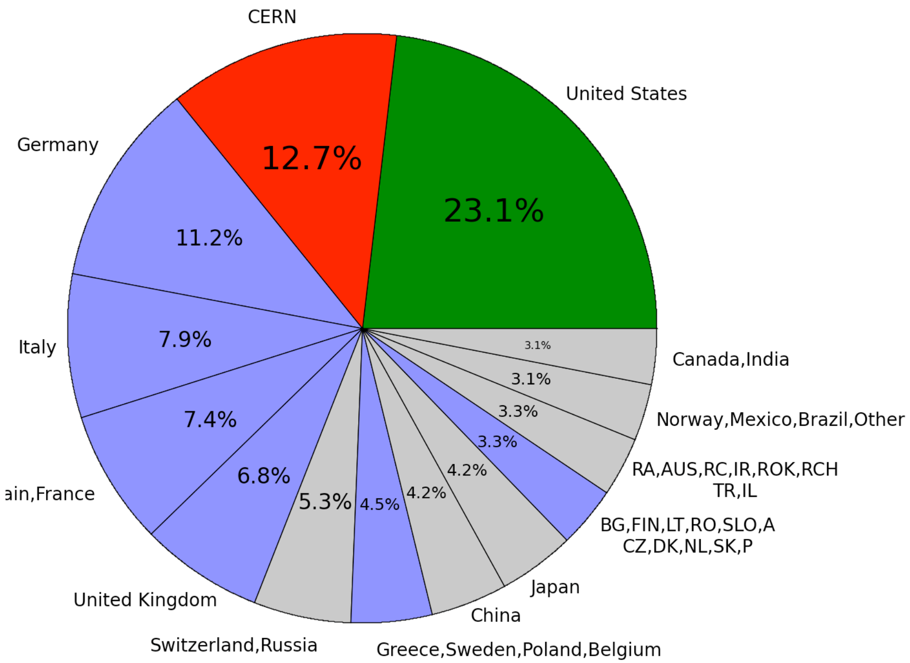
Survey structure

1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve it
5. Threats

Survey demographics

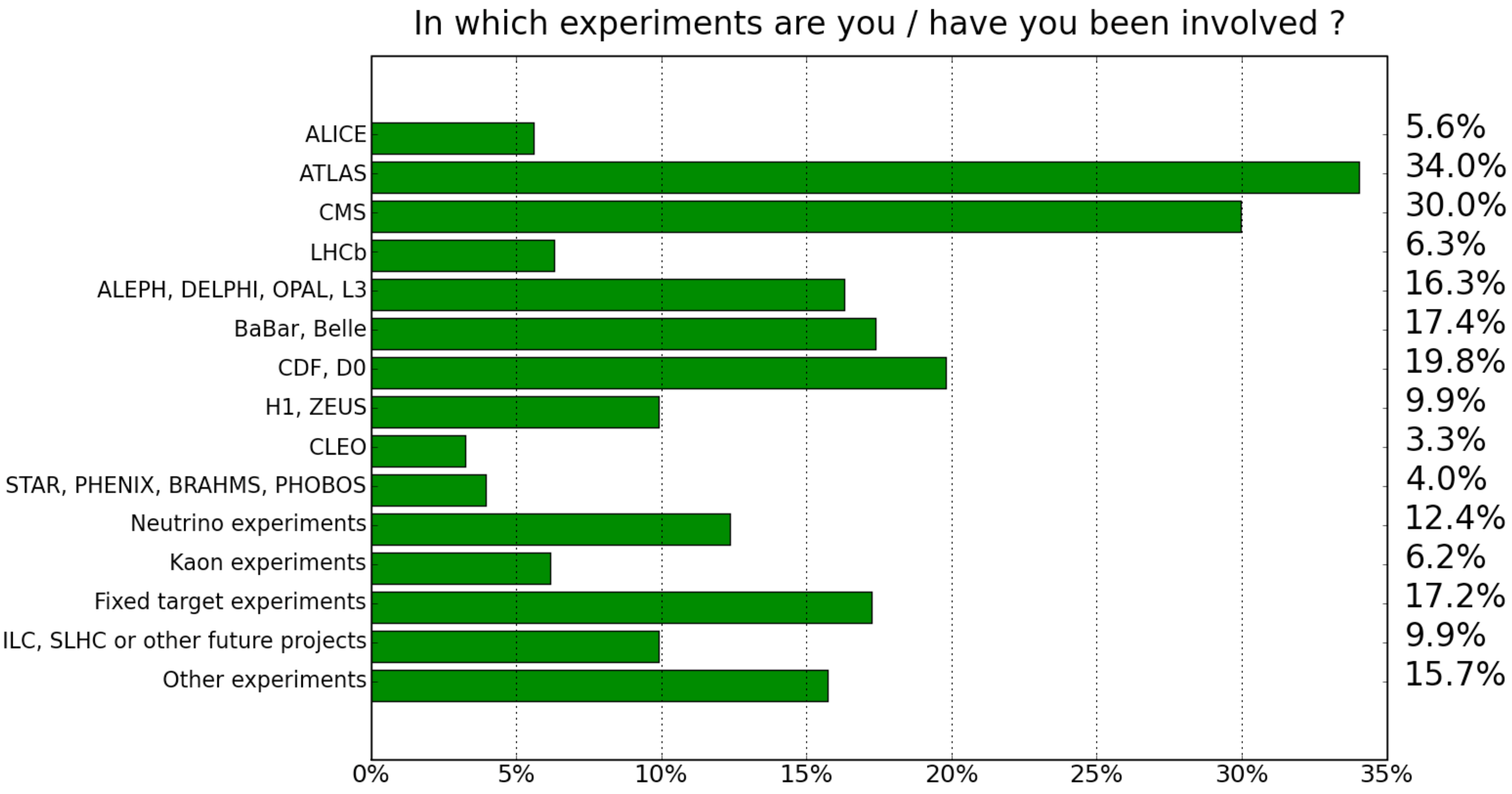
Reflects the demographics the community



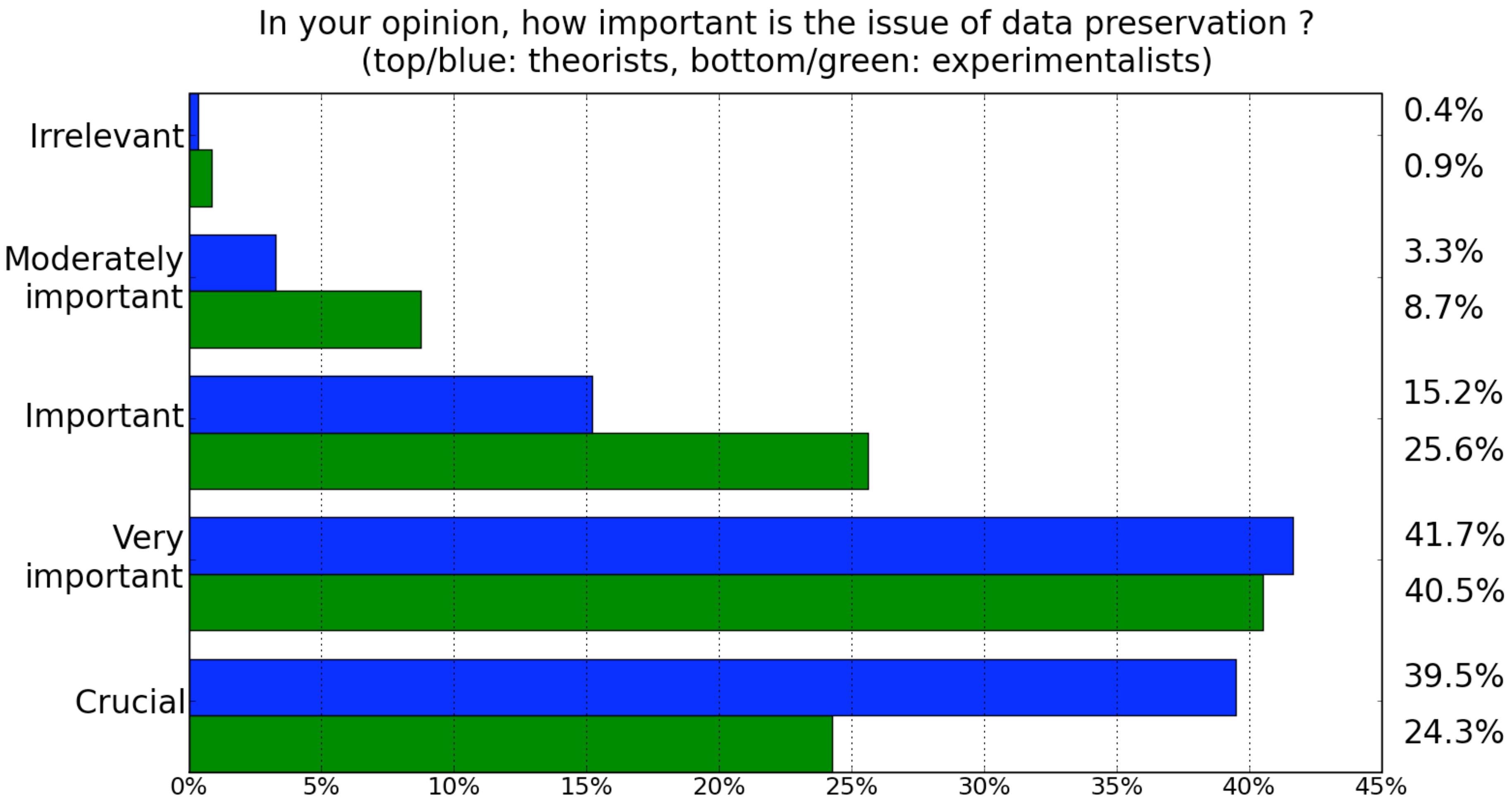


Survey demographics

Reflects the demographics of experiments



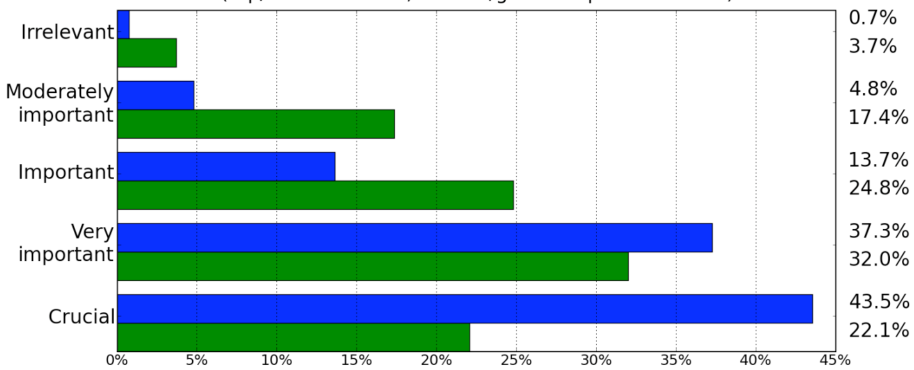
The importance of preservation



The importance of preservation

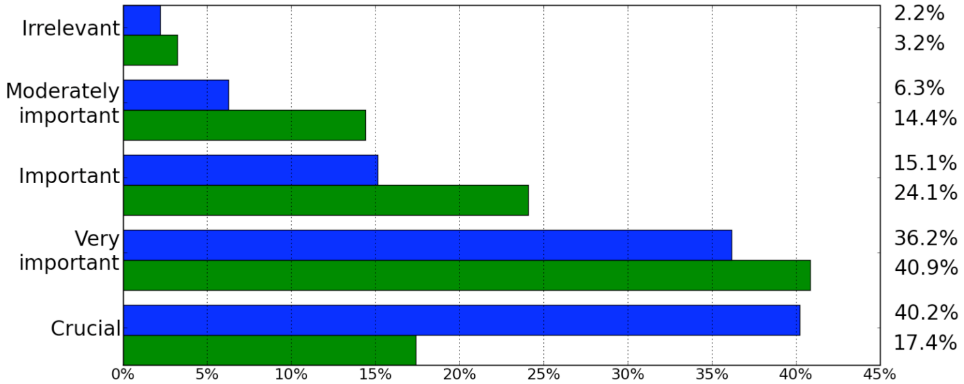
Future independent checks

b) Preserved data would allow future independent verification of results
(top/blue: theorists, bottom/green: experimentalists)



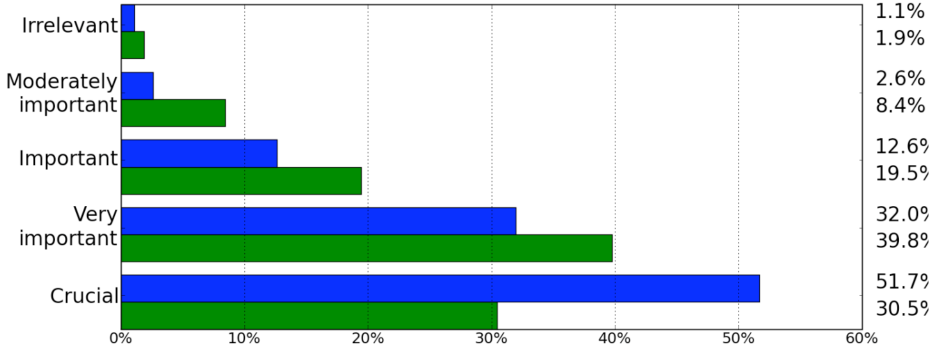
Combine with future data

c) Preserved data could be reused in combination with future data
(top/blue: theorists, bottom/green: experimentalists)



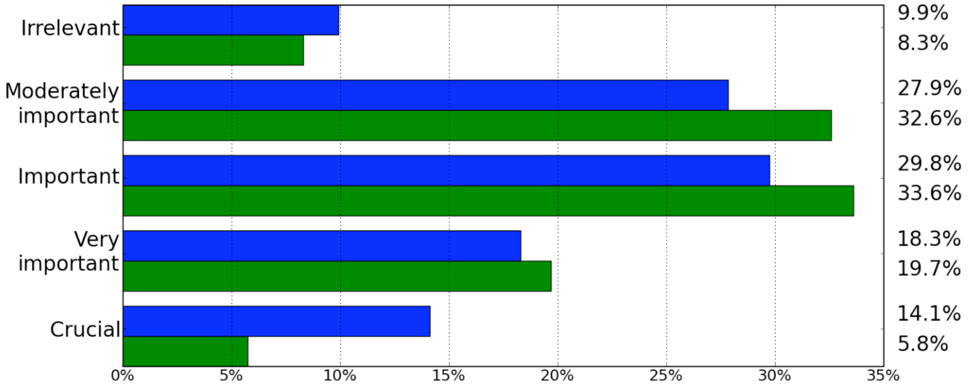
Re-analyse for future theories

d) Preserved data could be re-analysed in the light of new theories
or experimental results
(top/blue: theorists, bottom/green: experimentalists)



Teaching and outreach

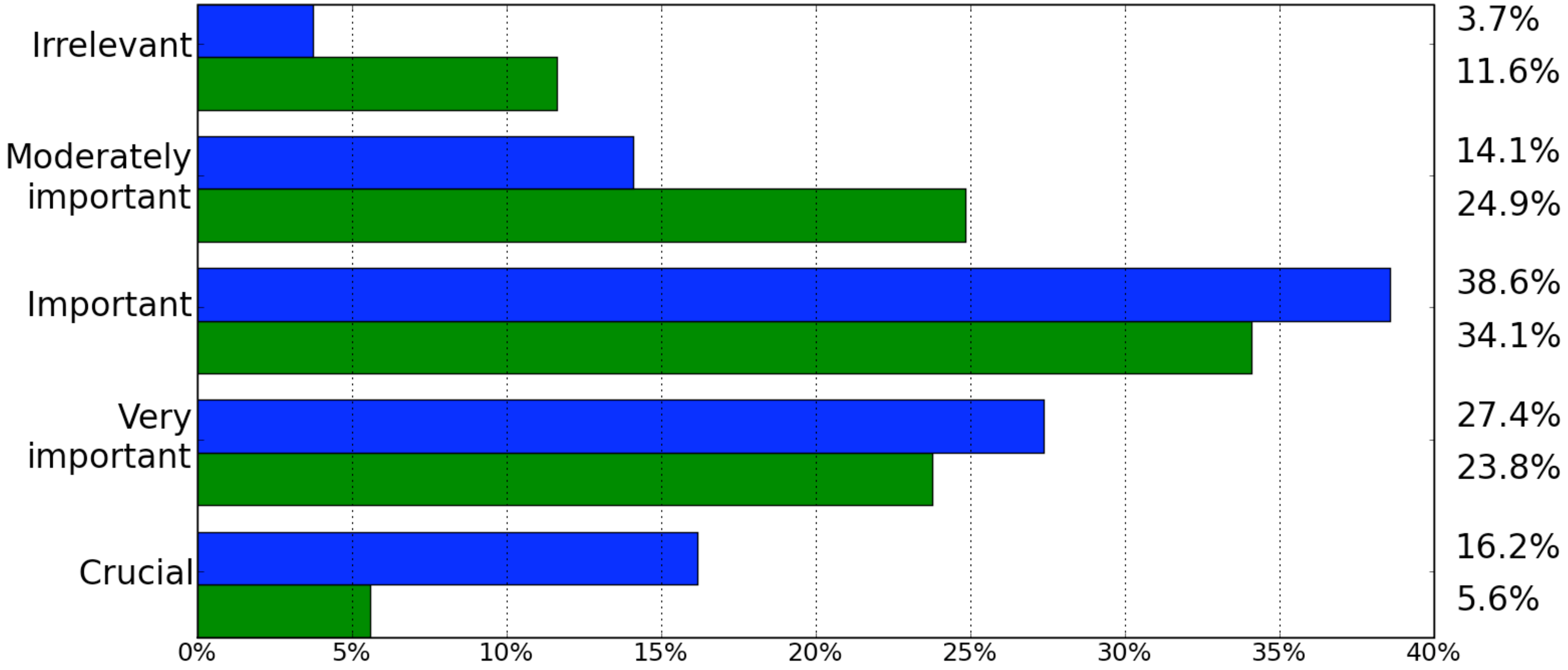
e) Preserved data could be used for teaching and outreach
(top/blue: theorists, bottom/green: experimentalists)



Why to preserve? - Compiling results

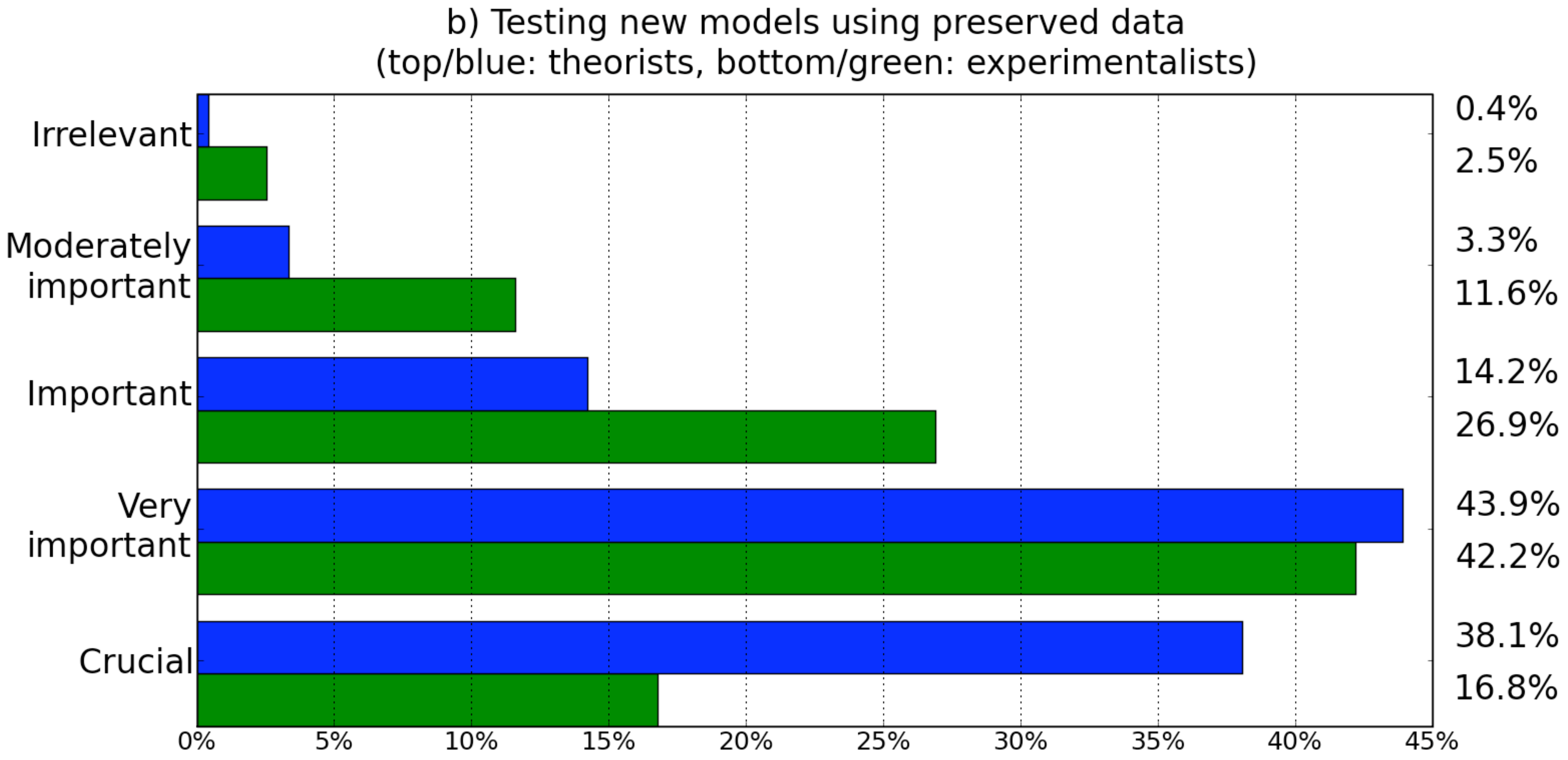
How much importance would you attach to the following uses of preserved data ?

a) Compiling published results on a given subject (e.g. for a review)
(top/blue: theorists, bottom/green: experimentalists)

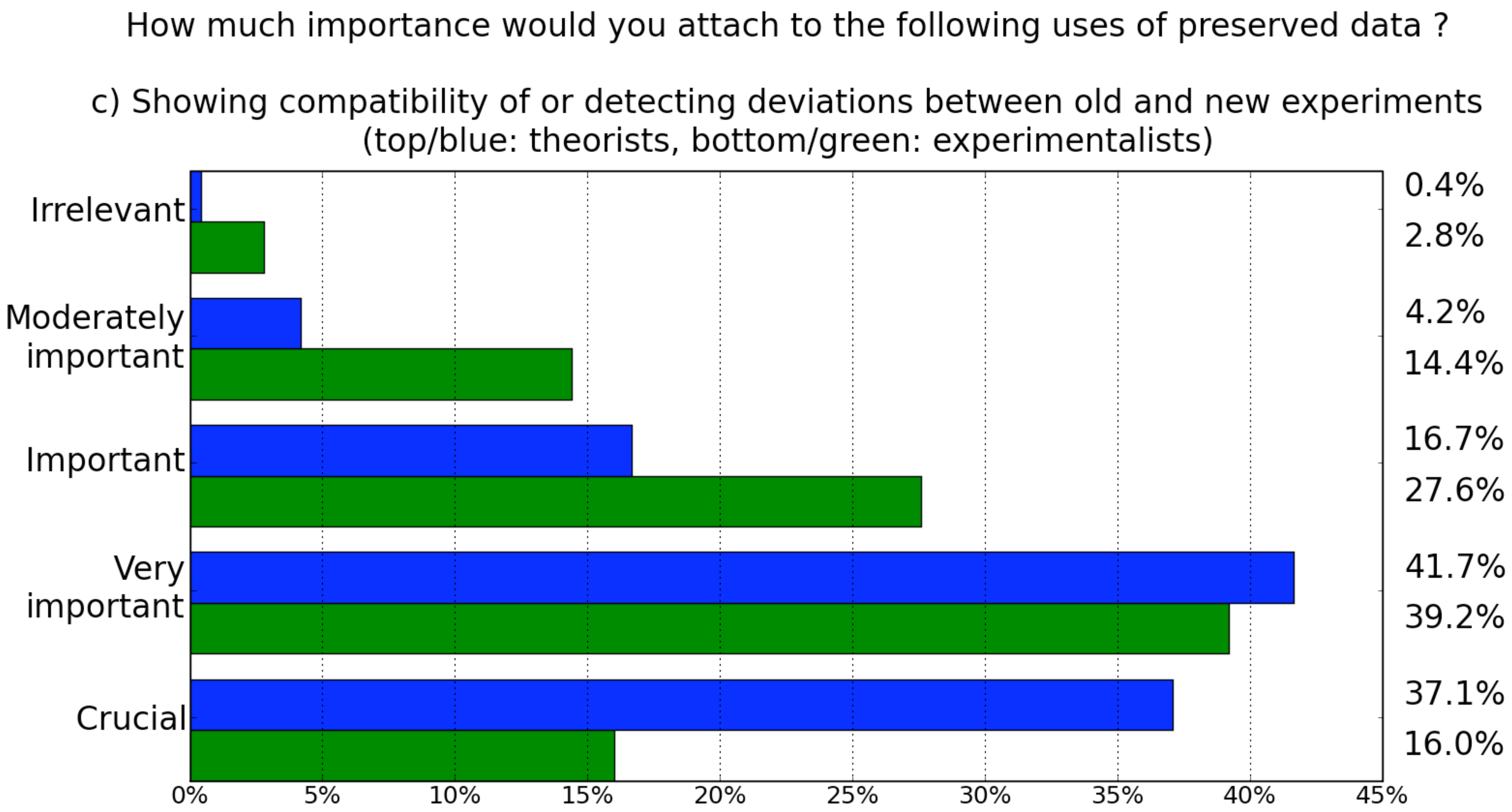


Why to preserve? - Testing new models

How much importance would you attach to the following uses of preserved data ?

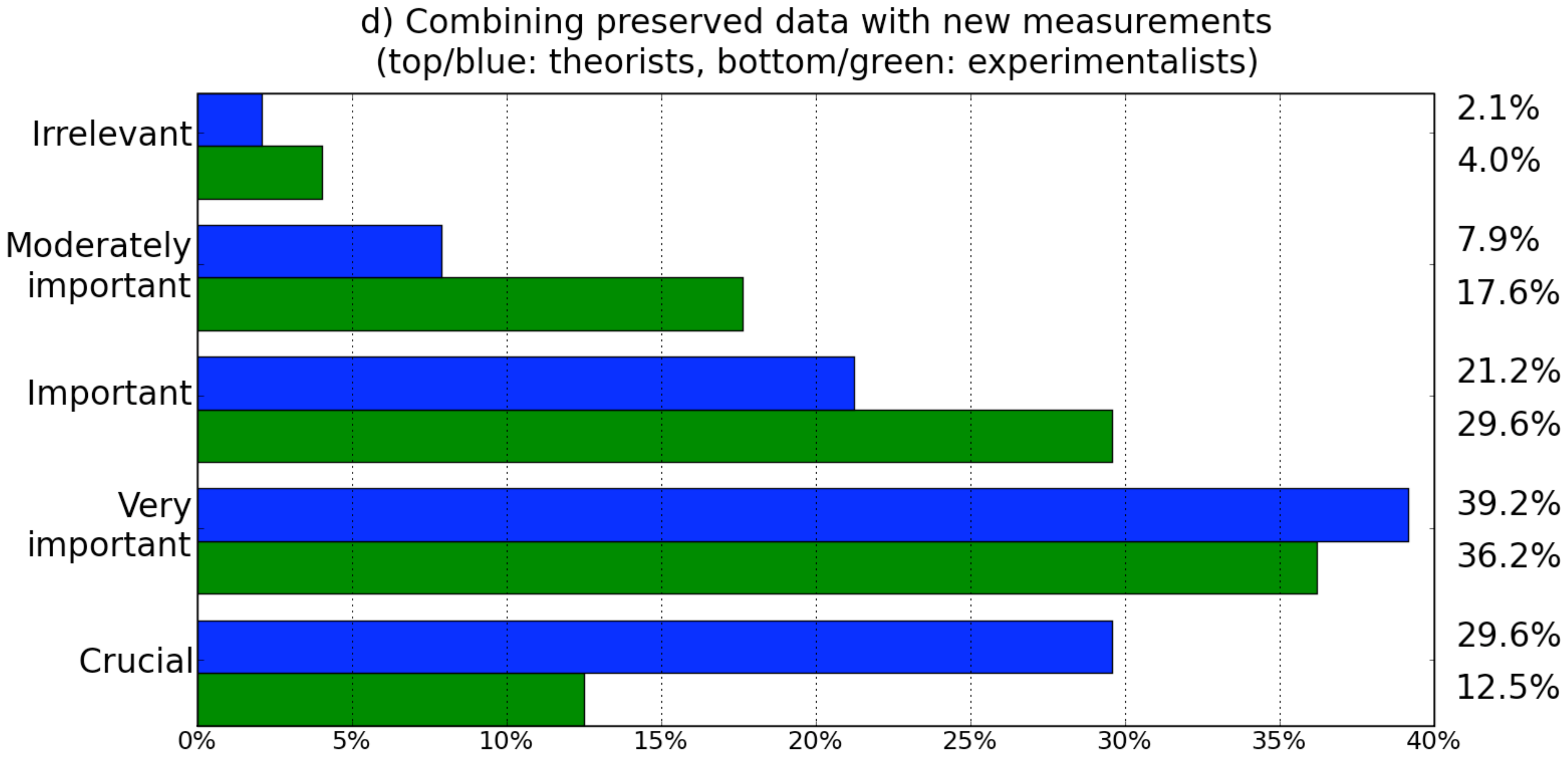


Why to preserve? - Comparing past and future

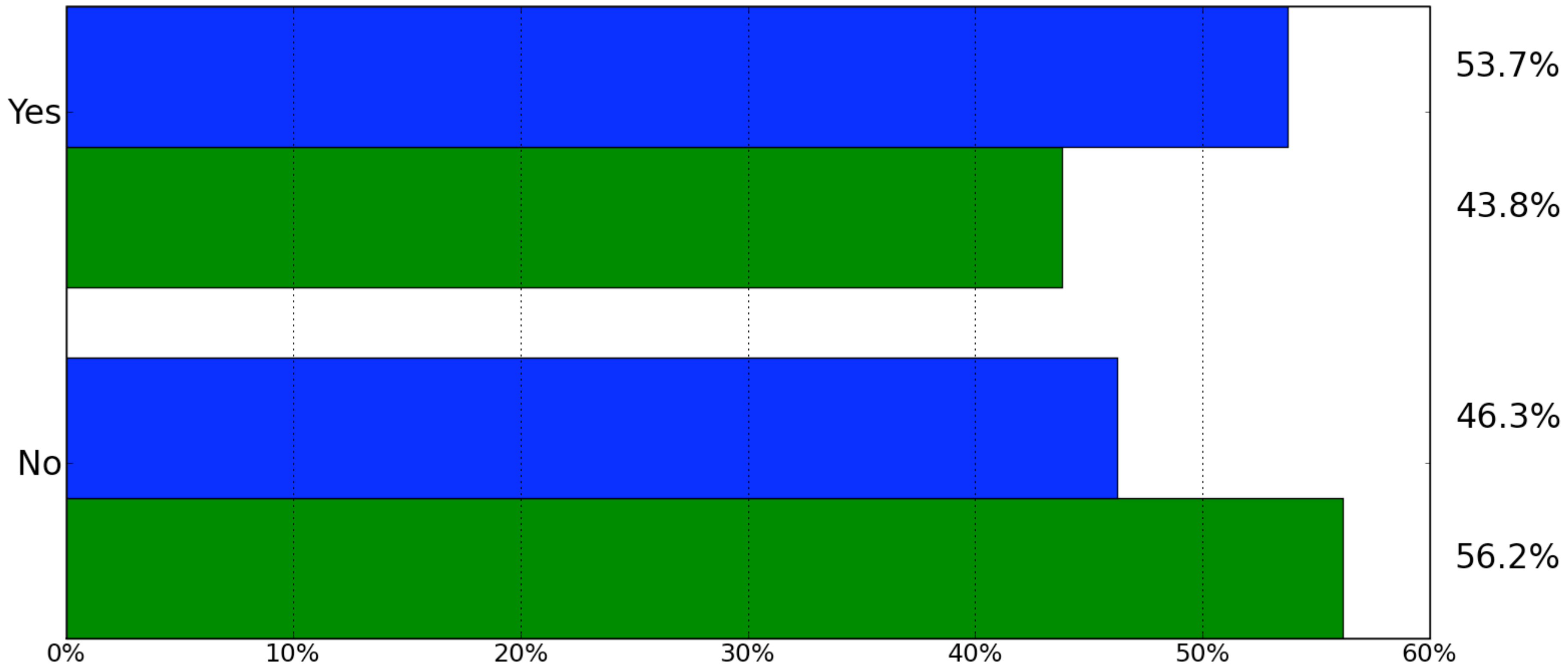


Why to preserve? - Combining past and future

How much importance would you attach to the following uses of preserved data ?



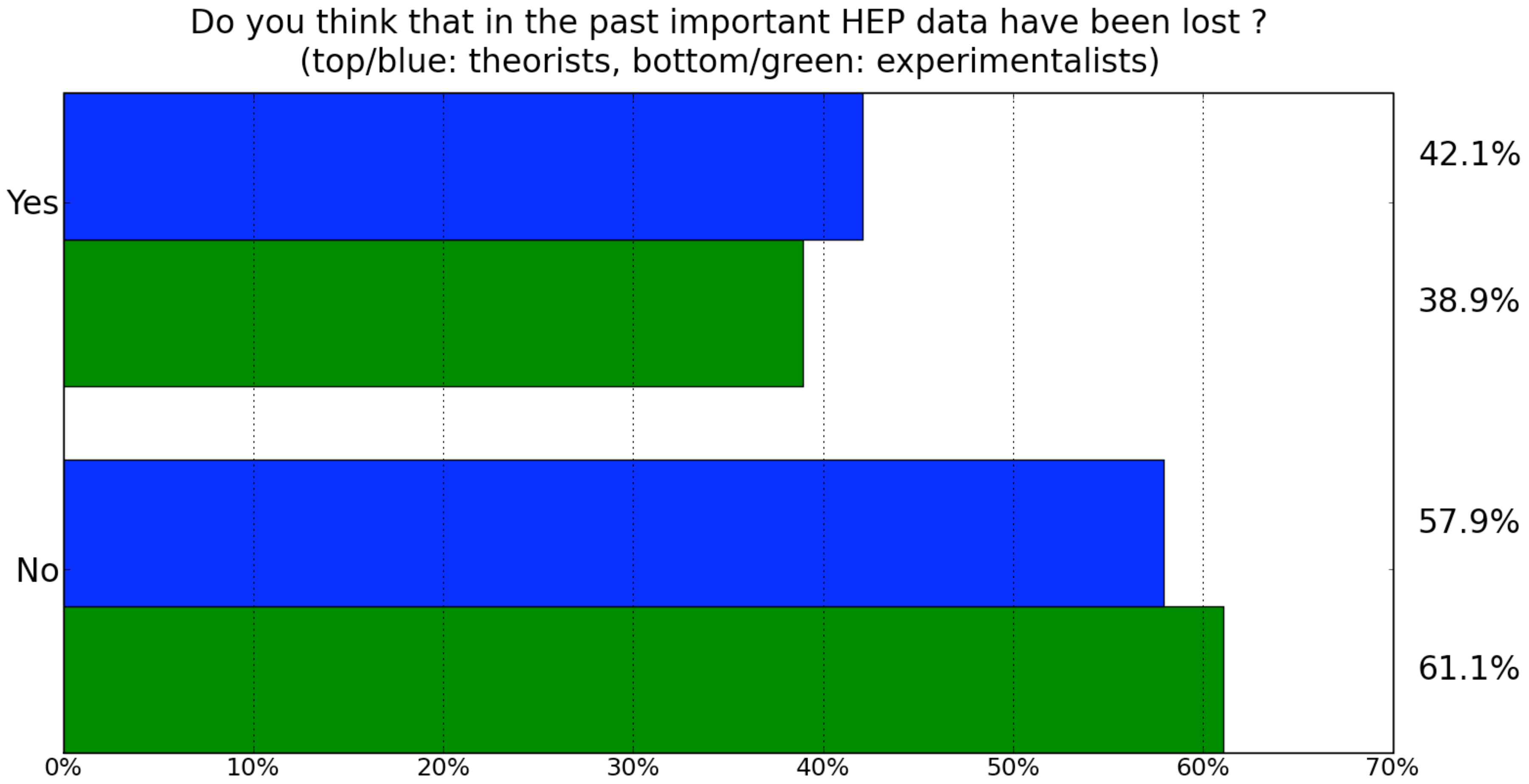
Should we have started long ago?



Do you think that access to data from past experiments could have improved your scientific results ?
(top/blue: theorists, bottom/green: experimentalists)

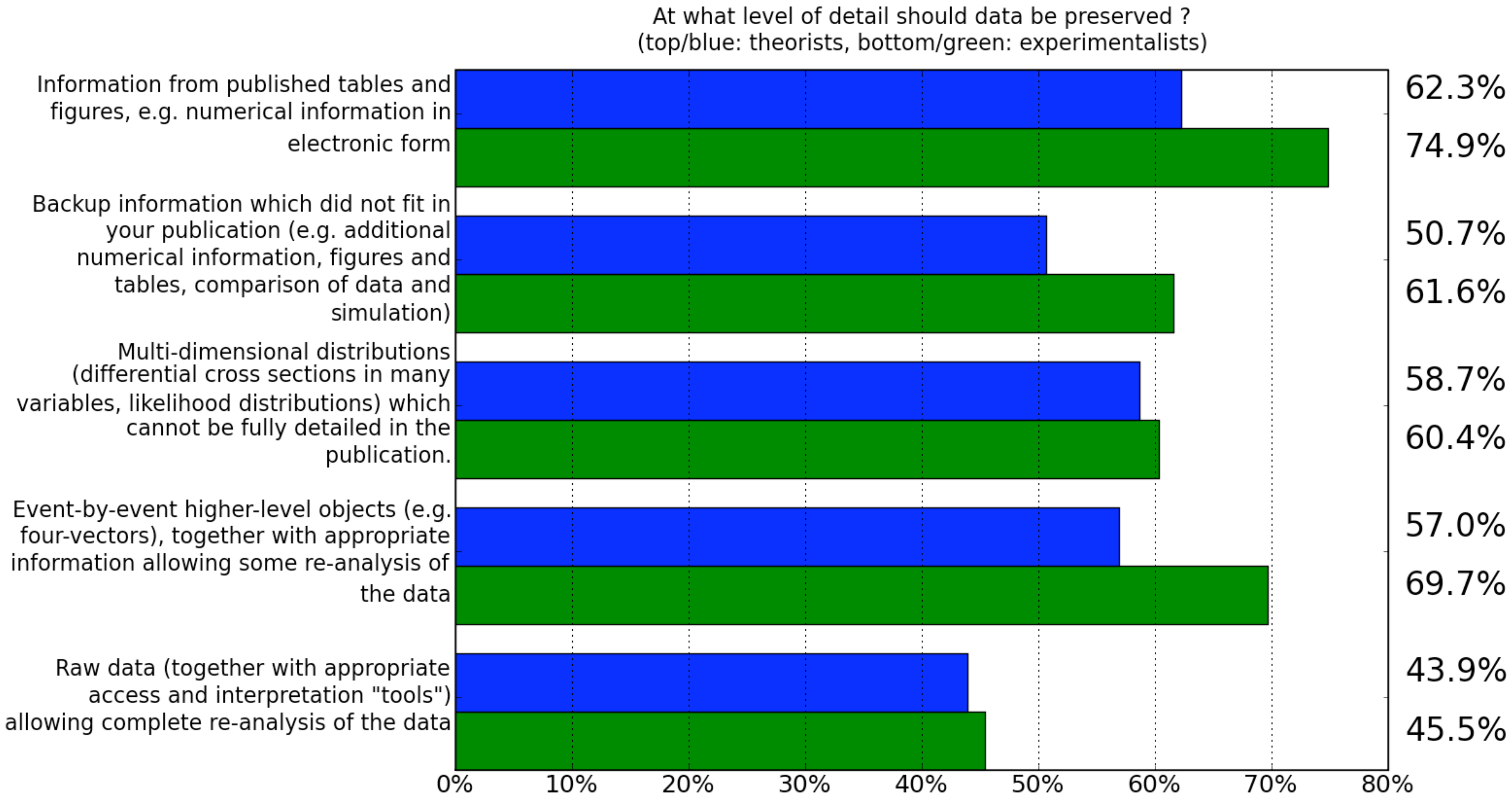
1. Strong argument to plea for support to preserve
2. Demonstrate that preservation, re-use and (open) access cannot be divided

Did anything go wrong so far?

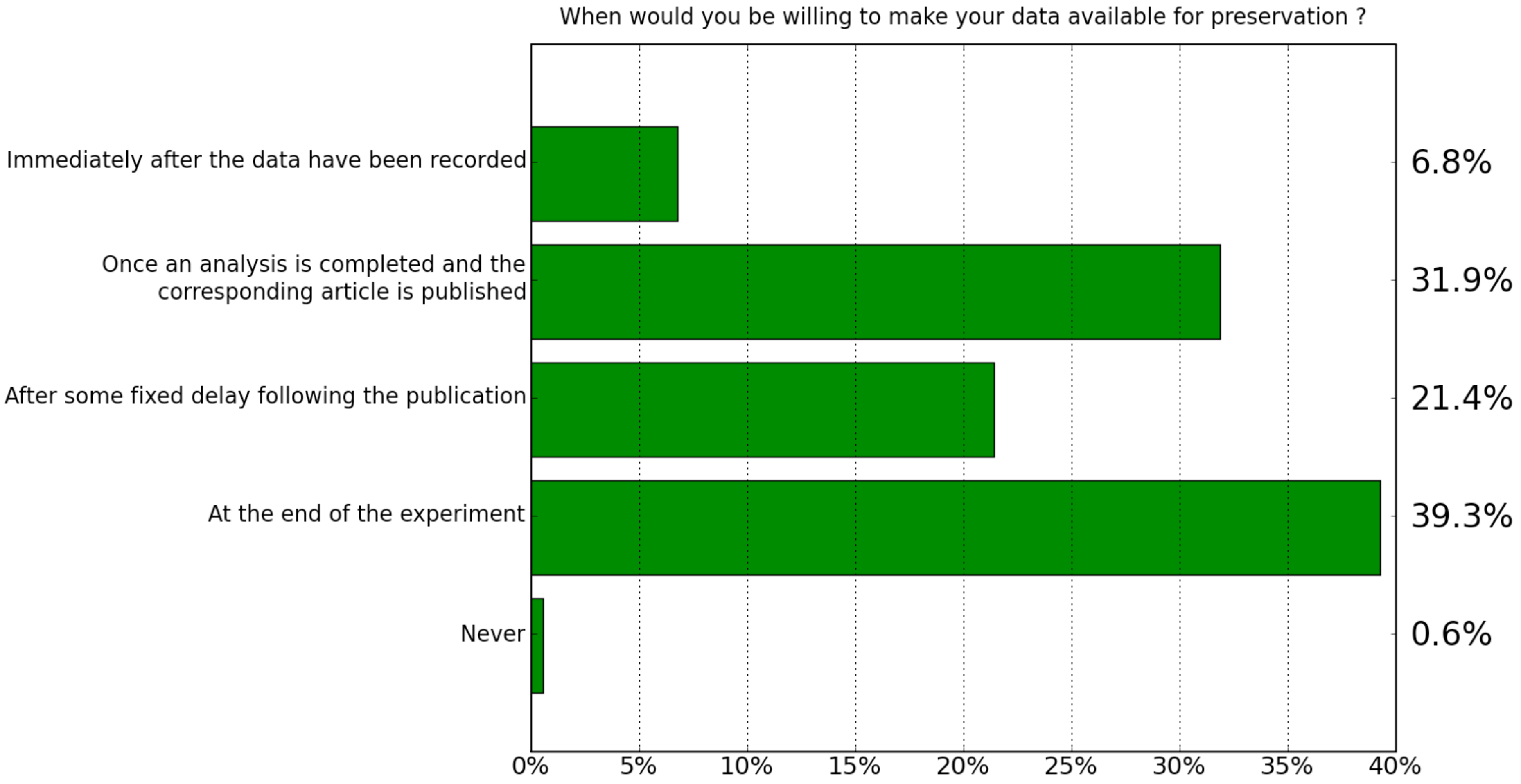


Over optimistic? Over pessimistic?

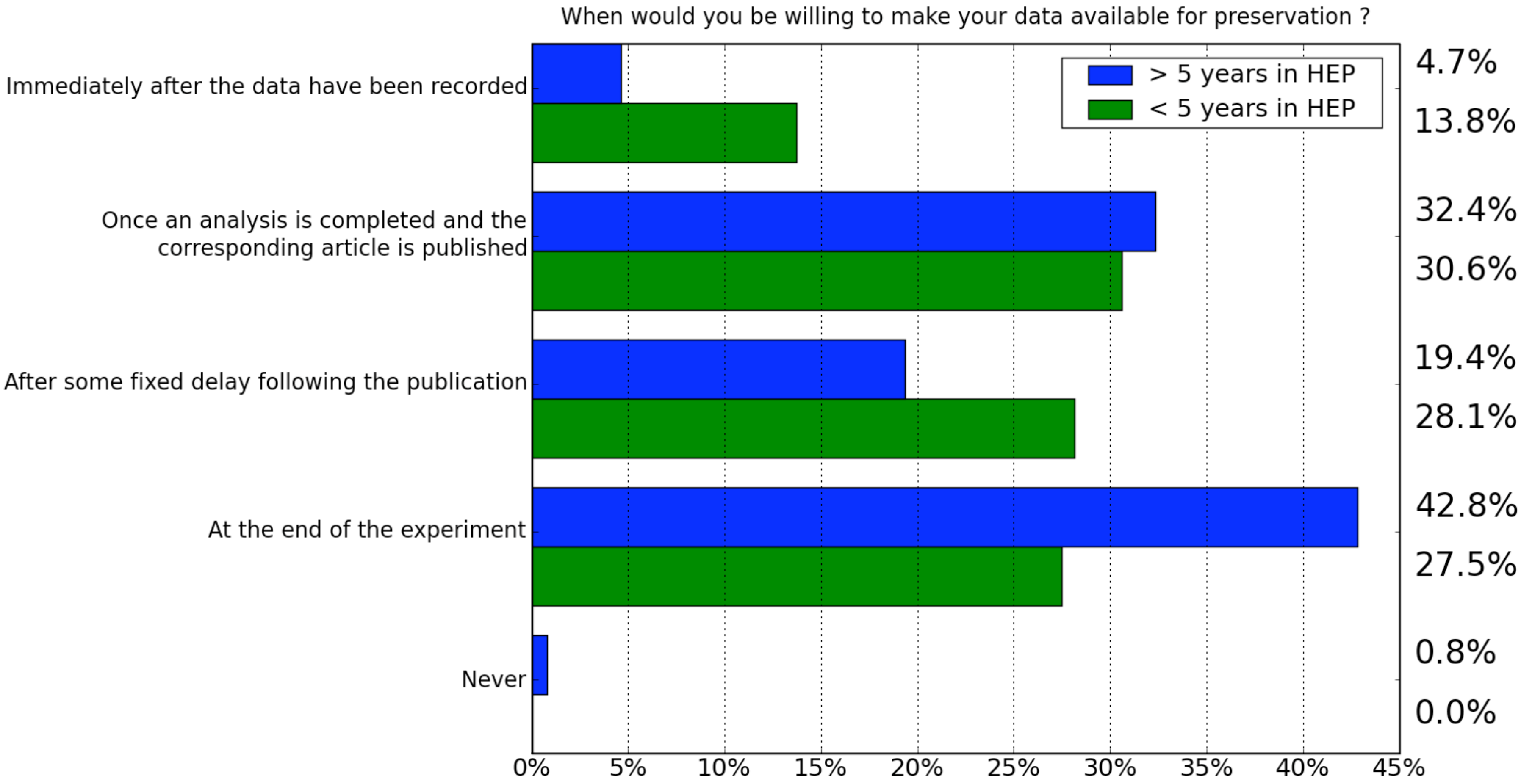
What to preserve?



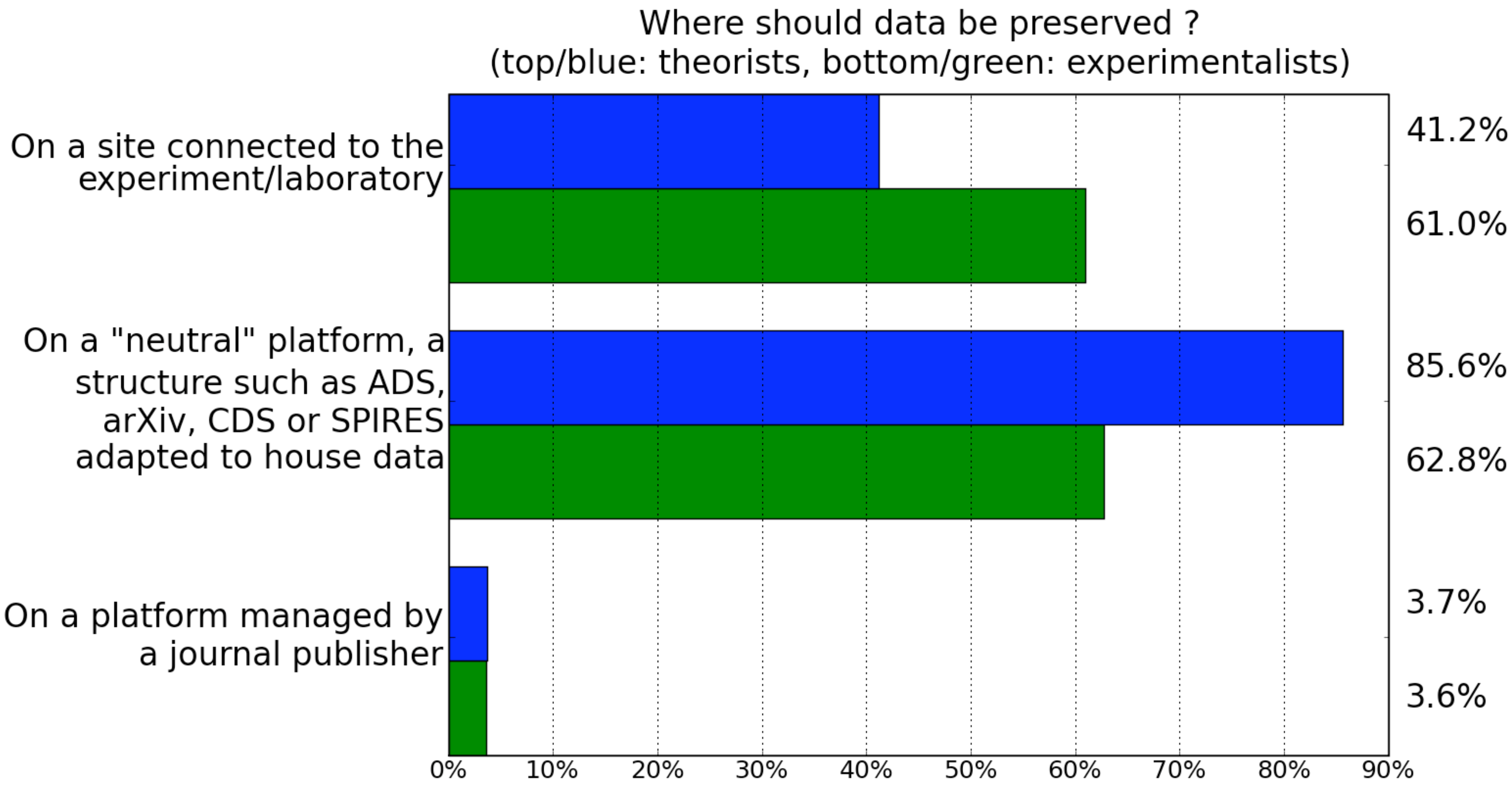
When to preserve it?



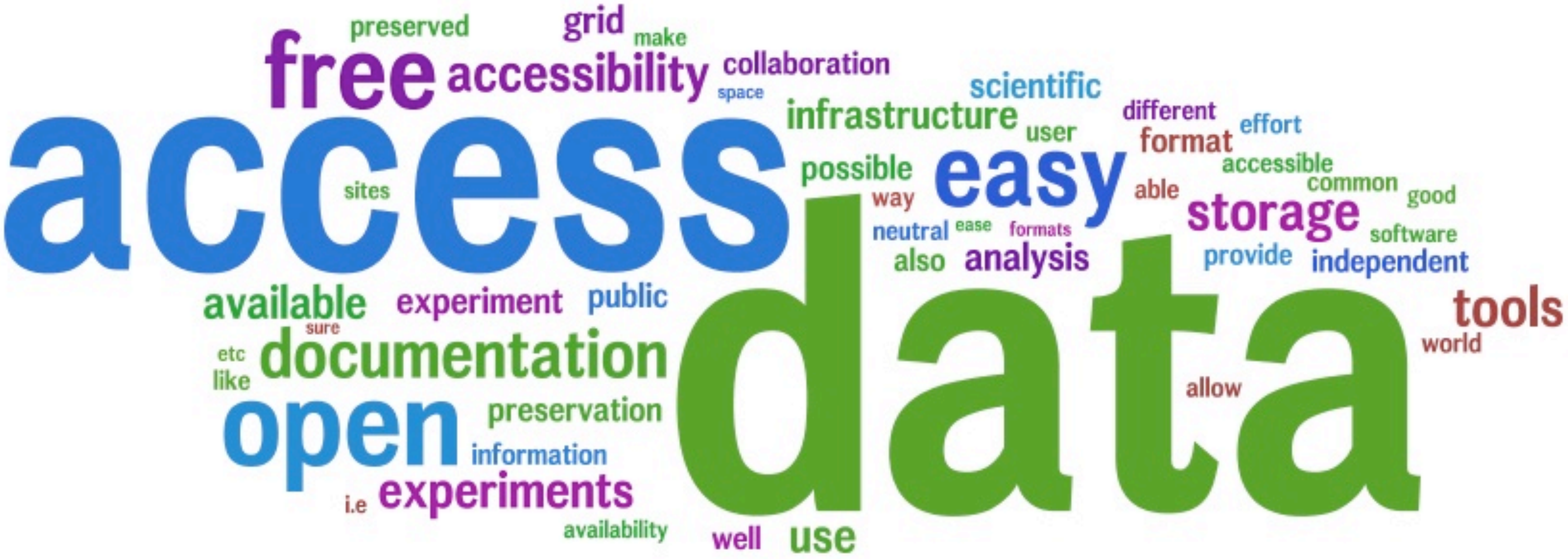
When to preserve it?



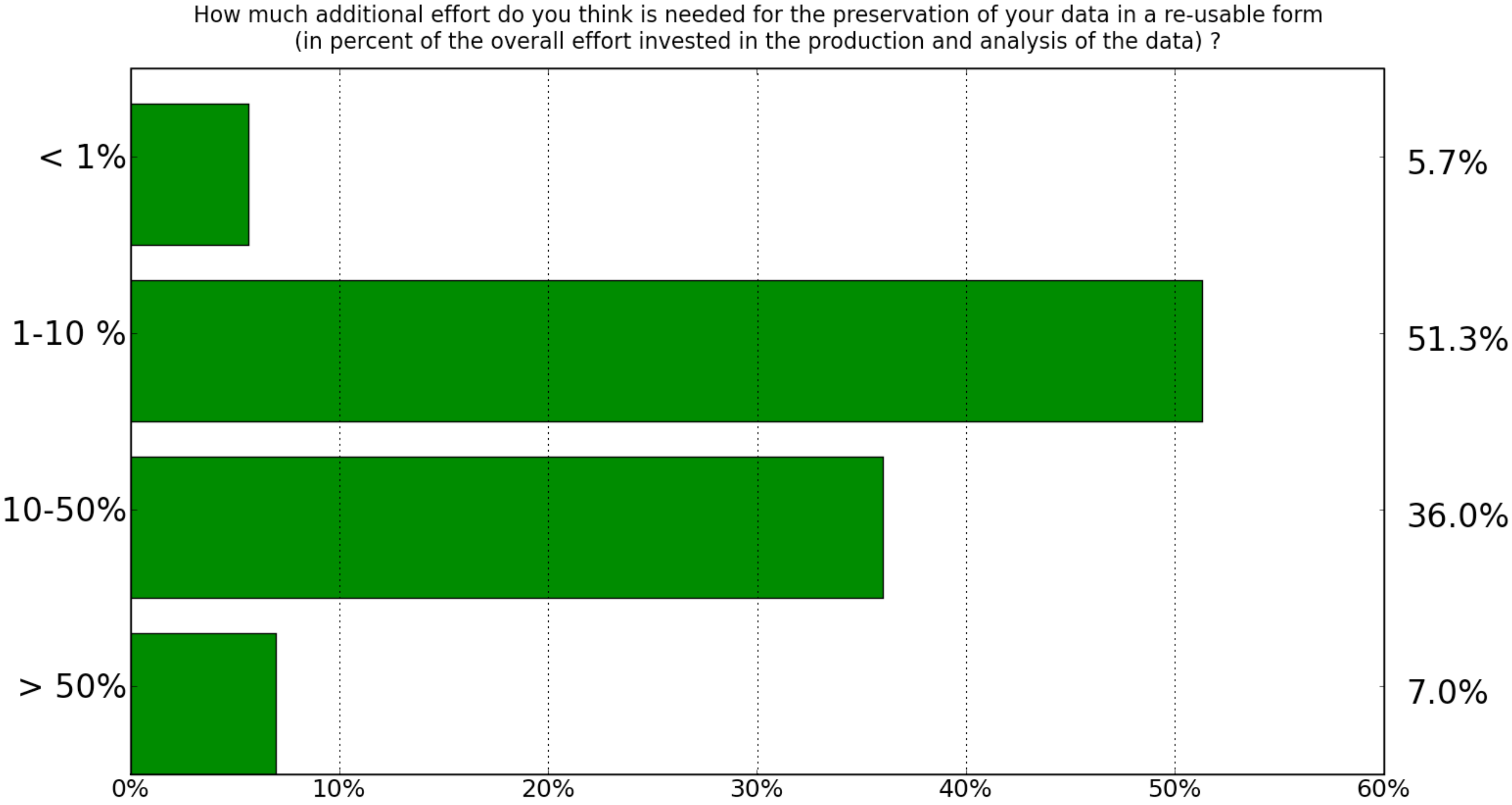
Where to preserve?



Where to preserve? Attributes of e-infrastructure

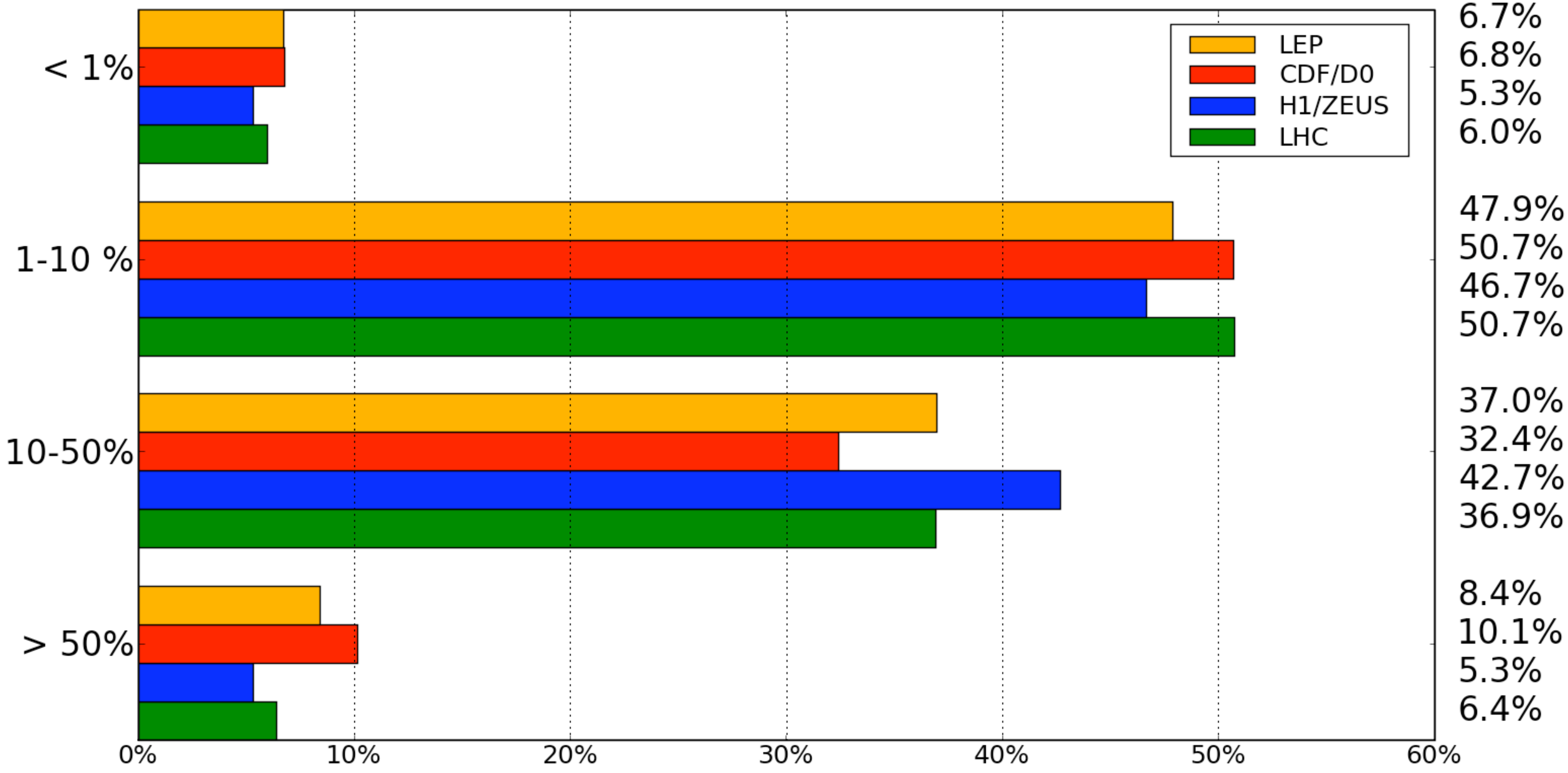


Reality check #1: how though is it to preserve?

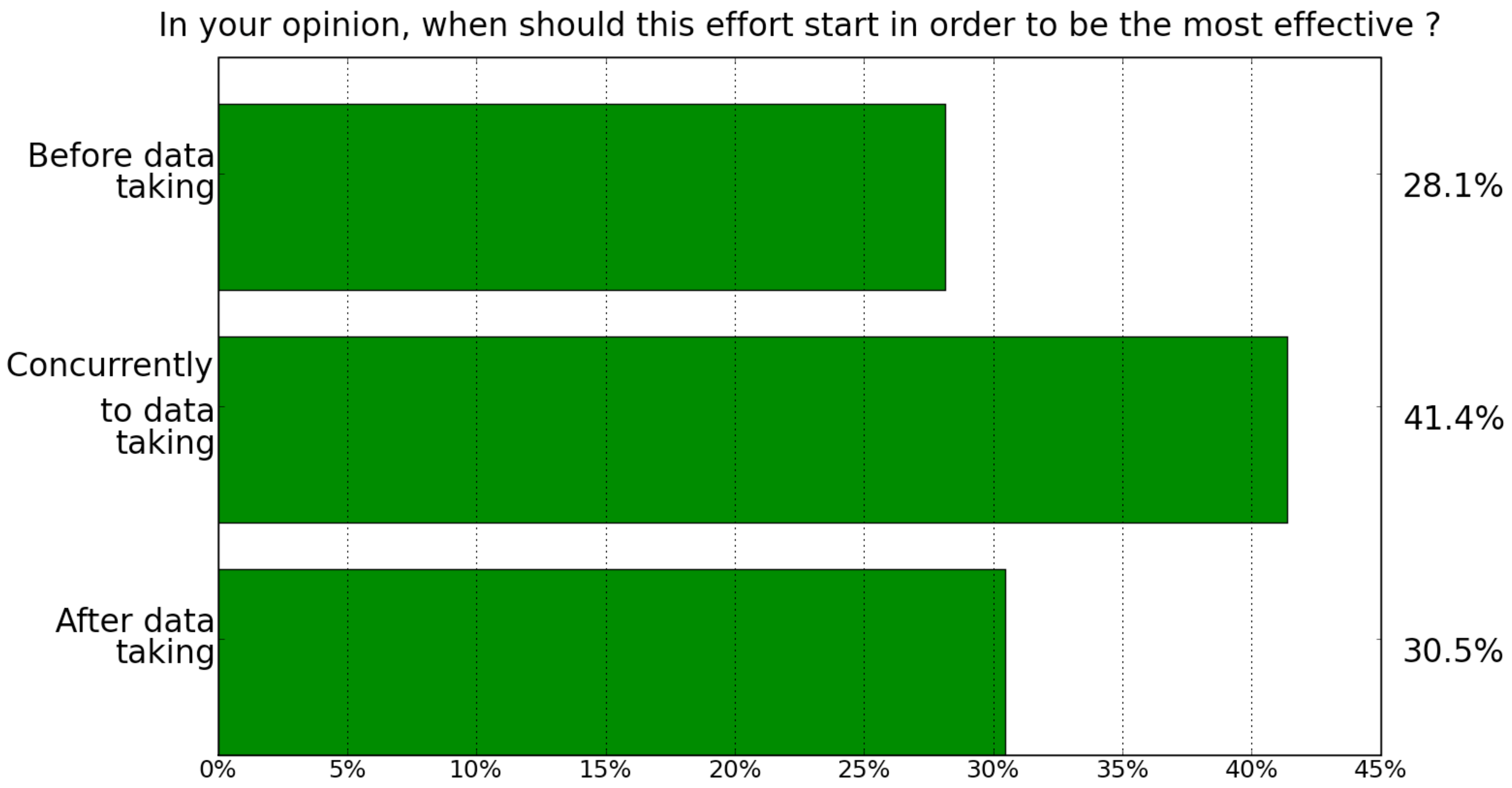


Reality check #1: how though is it to preserve?

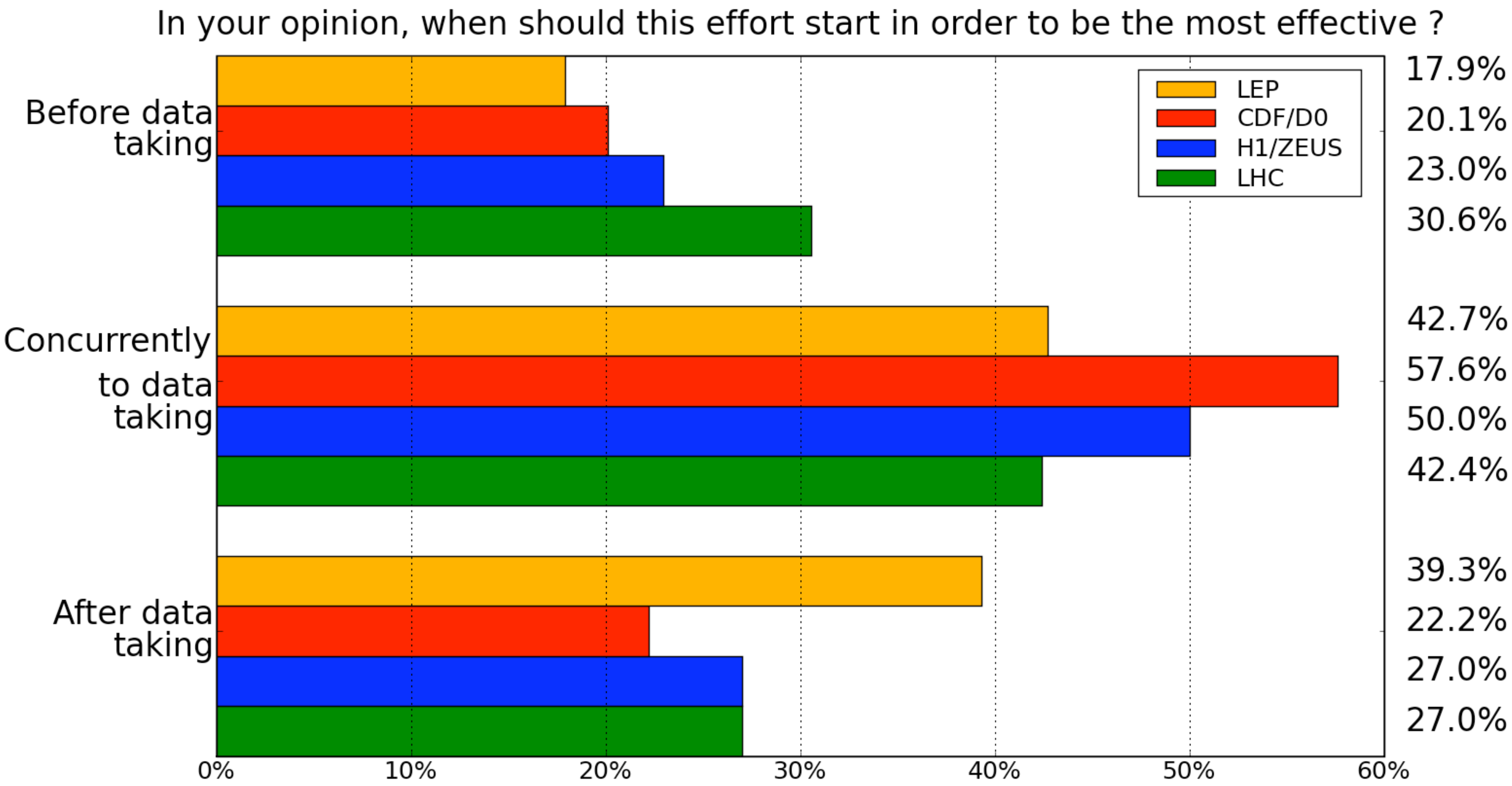
How much additional effort do you think is needed for the preservation of your data in a re-usable form (in percent of the overall effort invested in the production and analysis of the data) ?



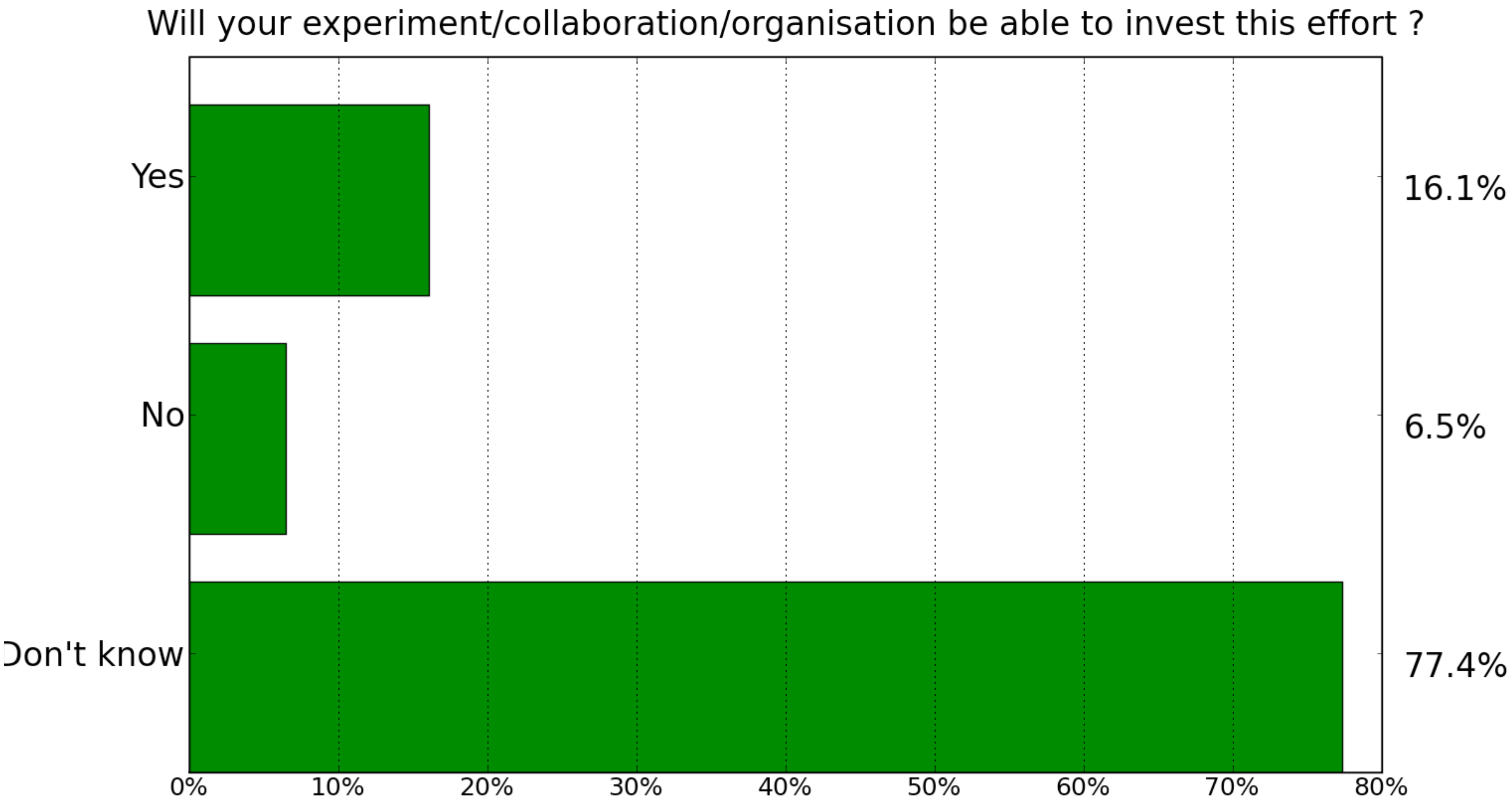
Reality check #2: when to preserve?



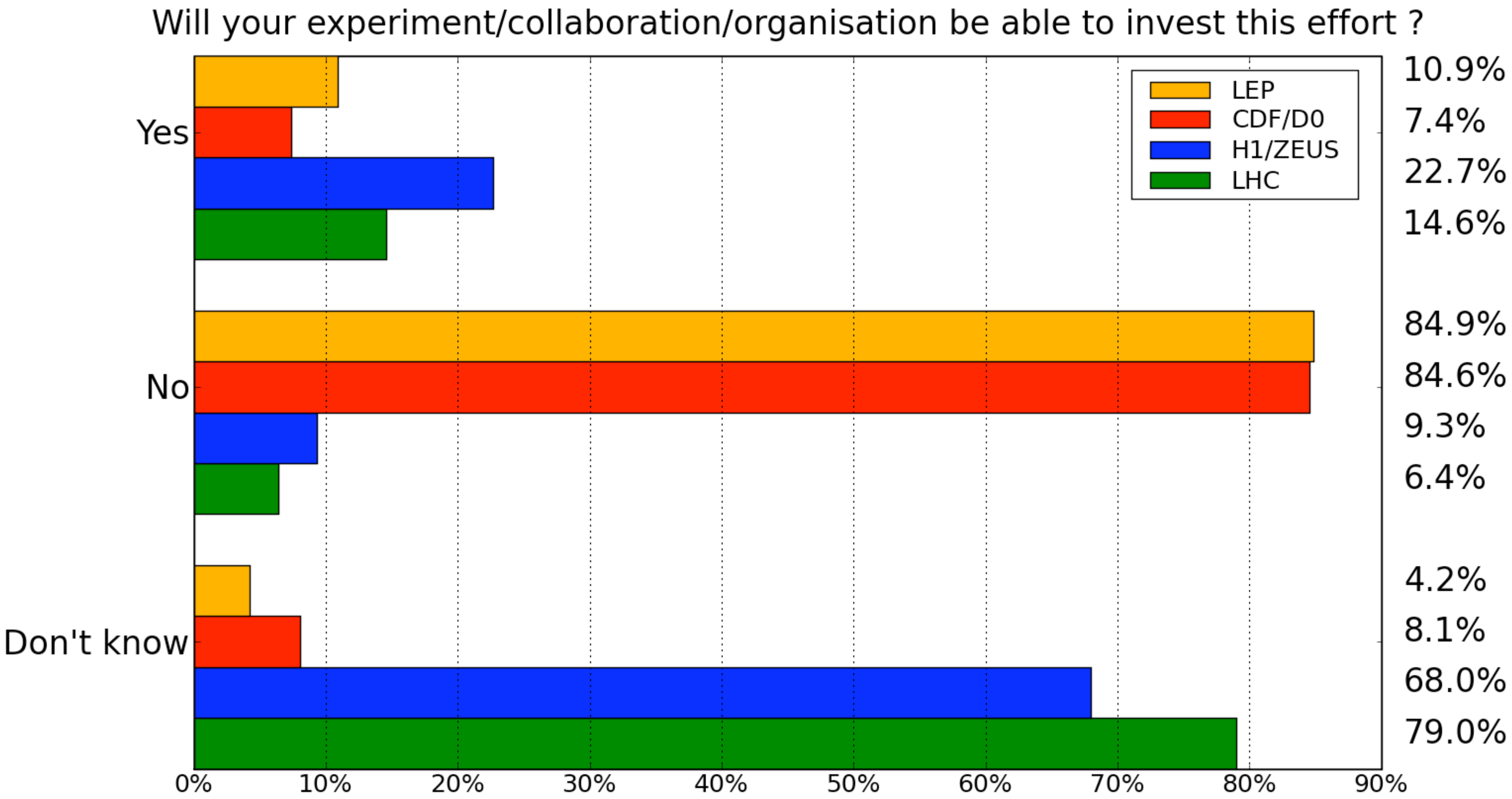
Reality check #2: when to preserve?



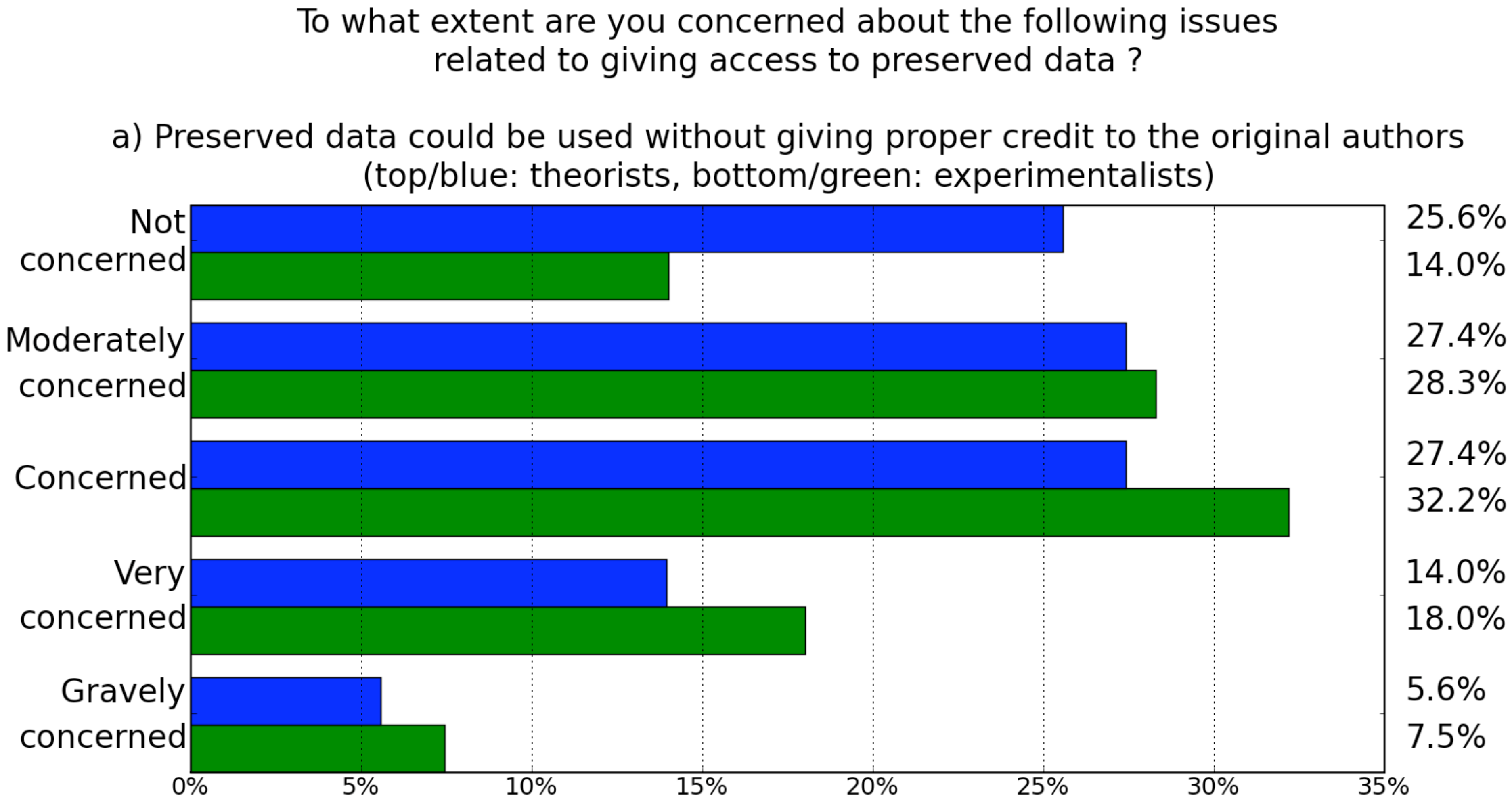
Reality check #3: is it doable?



Reality check #3: is it doable?



Ideal-case worries: getting credit



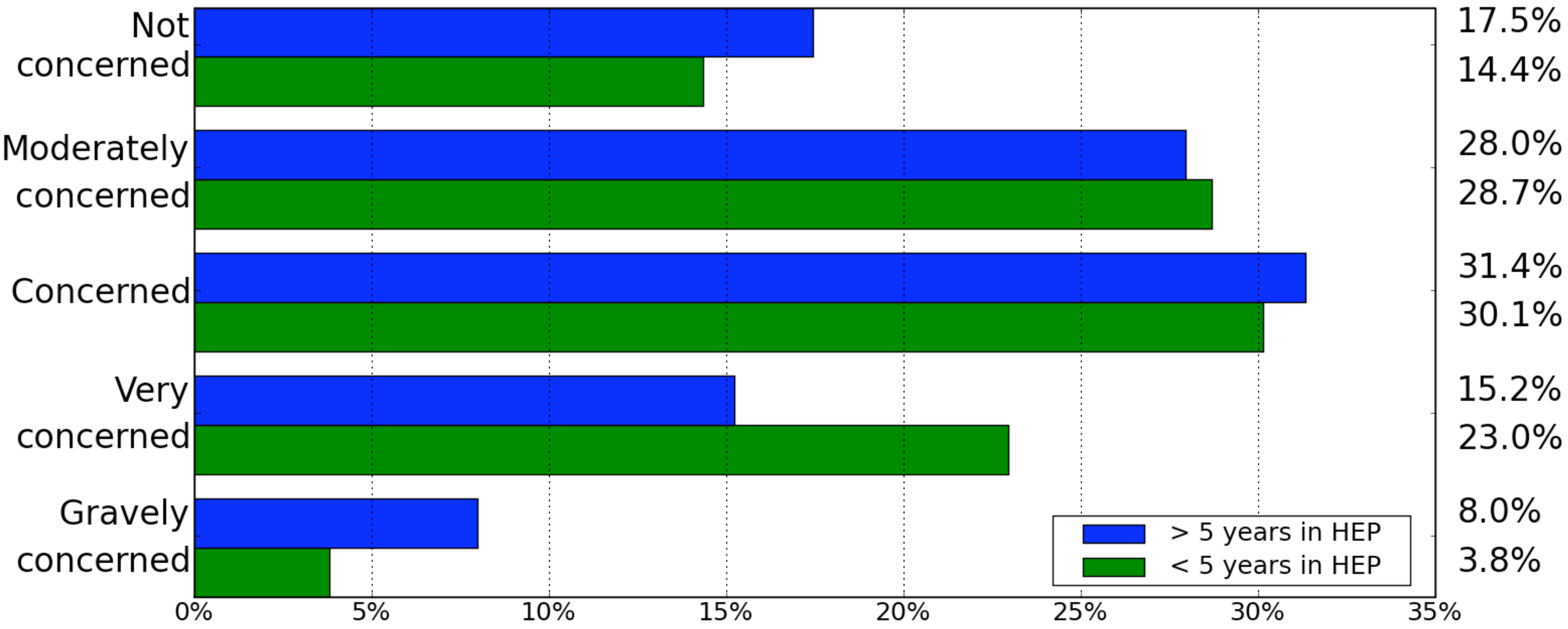
To what extent are you concerned about the following issues related to giving access to preserved data ?

a) Preserved data could be used without giving proper credit to the original authors (top/blue: theorists, bottom/green: experimentalists)

Ideal-case worries: getting credit

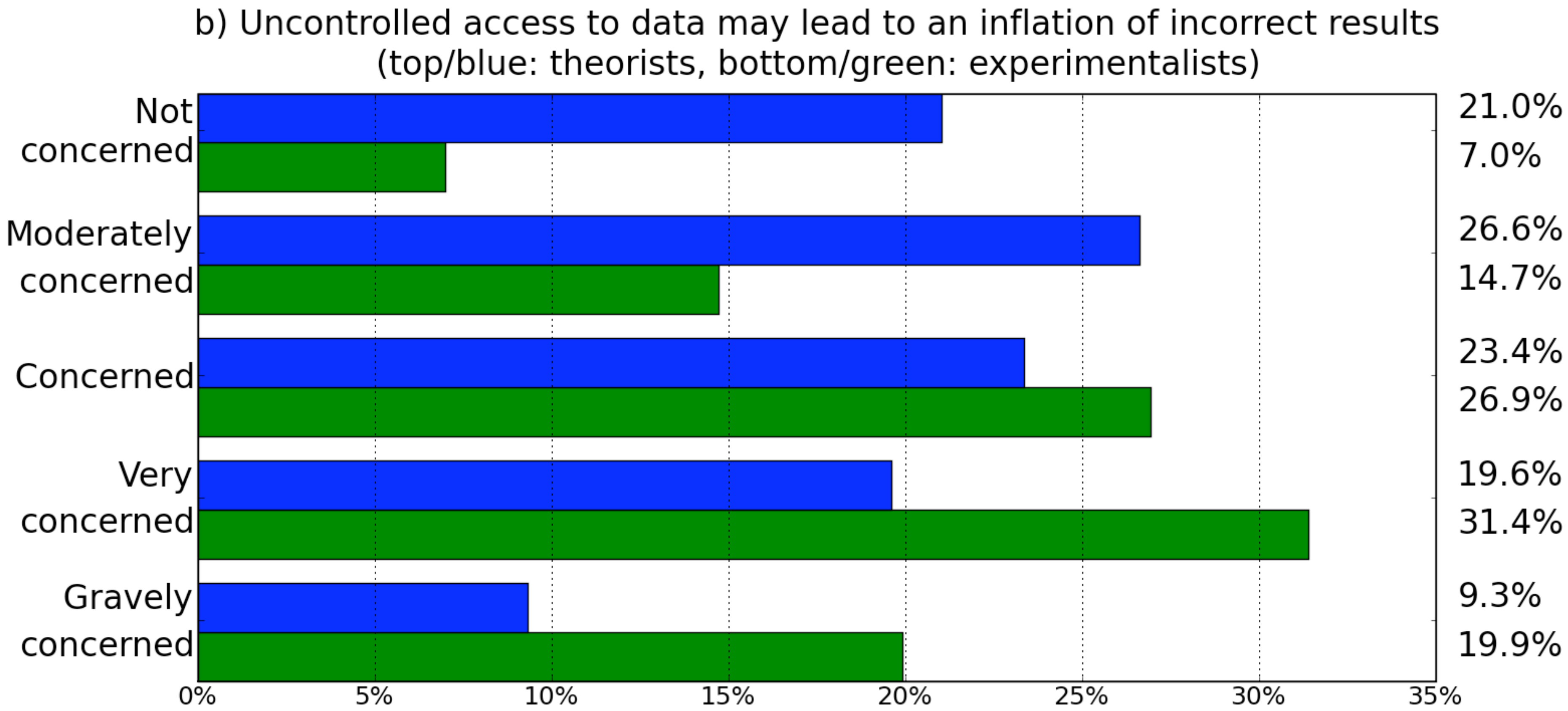
To what extent are you concerned about the following issues related to giving access to preserved data ?

a) Preserved data could be used without giving proper credit to the original authors

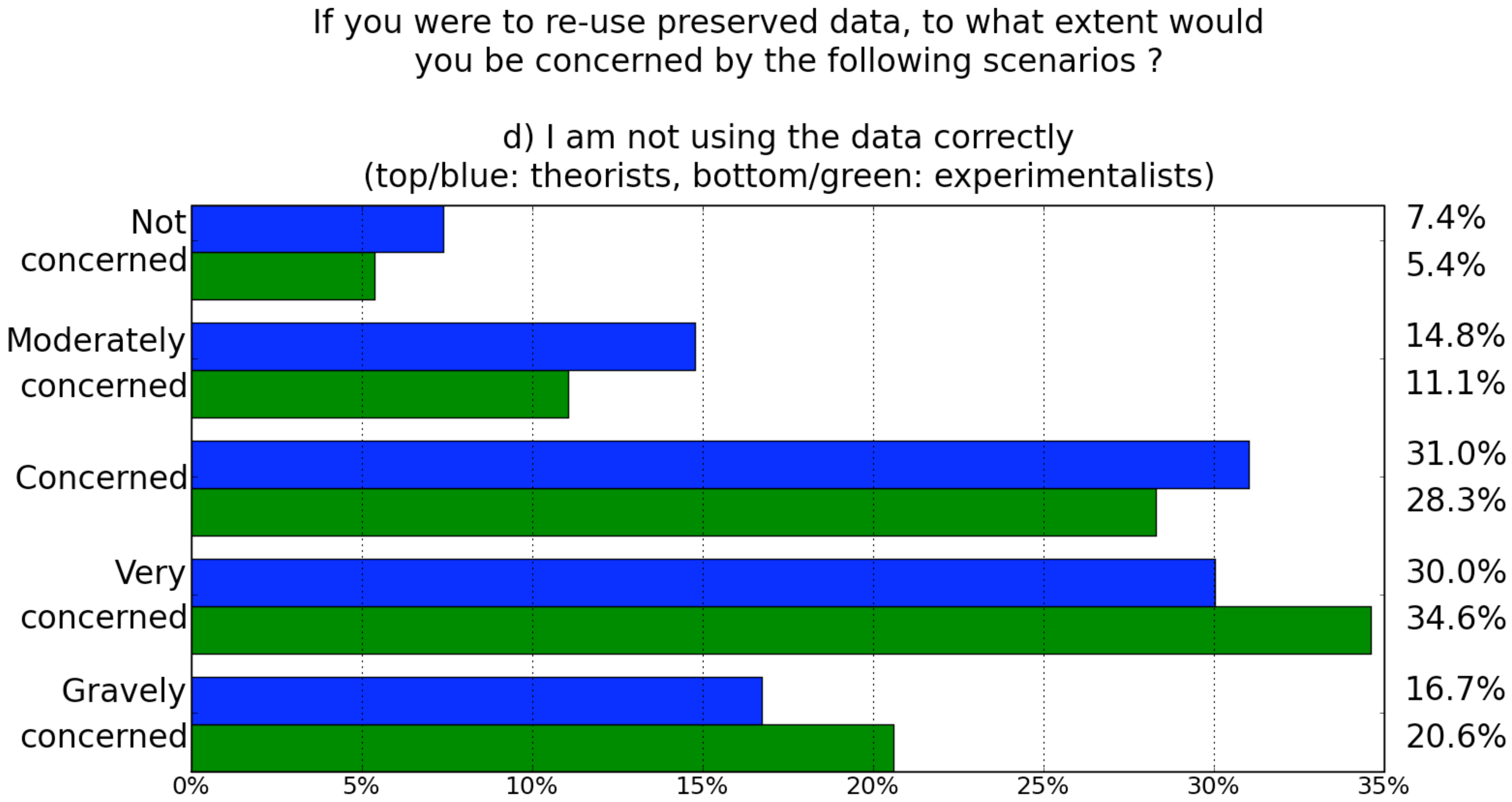


Ideal-case worries: inflation/noise

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



Ideal-case worries: documentation



PARSE.Insight HEP Case Study: next steps

- Many free-text open-ended questions still to analyse concerning: threats, opportunities, feasibility, access regulation, trustworthiness... Compile relevant excerpts and quantification of results

- 182 respondent (15%) made themselves available for an interview to express more opinion. Identify relevant subjects and run interviews

- Go to known “superusers” with same interview questions

Any inputs?