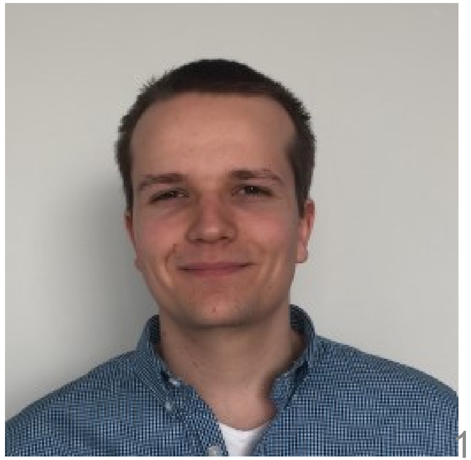
Methodological Principles for Reproducible Performance Evaluation in Cloud Computing

Alessandro Papadopoulos[1], **Laurens Versluis[2]**, André Bauer[3], Nikolas Herbst[3], Jóakim von Kistowski[3], Ahmed Ali-Eldin[4], Cristina Abad[5], José Amaral[6], Petr Tůma[7] and Alexandru Iosup[2]

[1]Mälardalen University, [2]**Vrije Universiteit Amsterdam**, [3]University of Würzburg, [4]Umeå university, [5]Escuela Superior Politecnica del Litoral, [6]University of Alberta, [7]Charles University

More detail in our Technical Report

L.Versluis@atlarge-research.com



What is Reprodicibility?

“Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results. [...] Reproducibility is a minimum necessary condition for a finding to be believable and informative.”

“What does research reproducibility mean?”, Goodman et al., 2016

Why (should you care about) Reproducibility?

- 70% of 1,500 surveyed articles could not reproduce work by others, over 50% couldn't reproduce their own work. [2]

- Also holds for **computer science**: more than 50% of work was unreproducible due to missing or un-compatible code. [3]

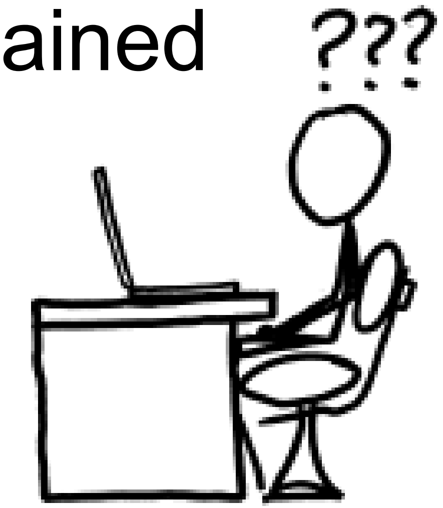
- Focus on reproducibility is increasing
  - Examples: SC19, ACM badges, etc.

[2] M. Baker, "Is there a reproducibility crisis?" Nature, 2016.

[3] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research," Commun. ACM, 2016.

This work: Three Research Questions

1. What methodological principles are needed for sound experimental evaluation of cloud performance?
2. Can the methodological principles be applied in common practice? (not in this talk)
3. How are cloud performance results currently obtained and reported?



- Context of this work: cloud computing, most principles also apply elsewhere

https://what-if.xkcd.com/14/

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:
  1. Repeated experiments
  2. Workload and configuration coverage
  3. Experimental setup description
  4. Open access artifact

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:
  1. Repeated experiments
  2. Workload and configuration coverage
  3. Experimental setup description
  4. Open access artifact

- **Number of repetitions**
- **Explanation 'why'**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:
  1. Repeated experiments
  2. Workload and configuration coverage
  3. Experimental setup description
  4. Open access artifact

- **(Software) parameters**
- **Hardware configuration**
- **Resources allocated**
- **Explore impact (possibly randomized)**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:
  1. Repeated experiments
  2. Workload and configuration coverage
  3. Experimental setup description
  4. Open access artifact

- **Software + version**
- **Data**
- **Hardware**
- **Any other important info, e.g. topology,**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:
  1. Repeated experiments
  2. Workload and configuration coverage
  3. Experimental setup description
  4. Open access artifact

- **Software**
- **(Representative subset of) data**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:

  5 Probabilistic result description of measured performance

  6 Statistical evaluation

  7 Measurement units

  8 Costs

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:

  5 Probabilistic result description of measured performance

  6 Statistical evaluation

  7 Measurement units

  8 Costs

- **Range of data (min, max, standard dev., etc.)**
- **Probablistic distribution**
- **Etc.**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:

  5 Probabilistic result description of measured performance

  6 Statistical evaluation

  7 Measurement units

  8 Costs

- **Statistical validation**
- **ANOVA test**
- **T-test**
- **Etc.**

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:

  5 Probabilistic result description of measured performance

  6 Statistical evaluation

  7 Measurement units

  8 Costs

- Graphs
- Tables

What methodological principles are needed for sound experimental evaluation of cloud performance?

- Eight basic principles:

  5 Probabilistic result description of measured performance

  6 Statistical evaluation

  7 Measurement units

  8 Costs

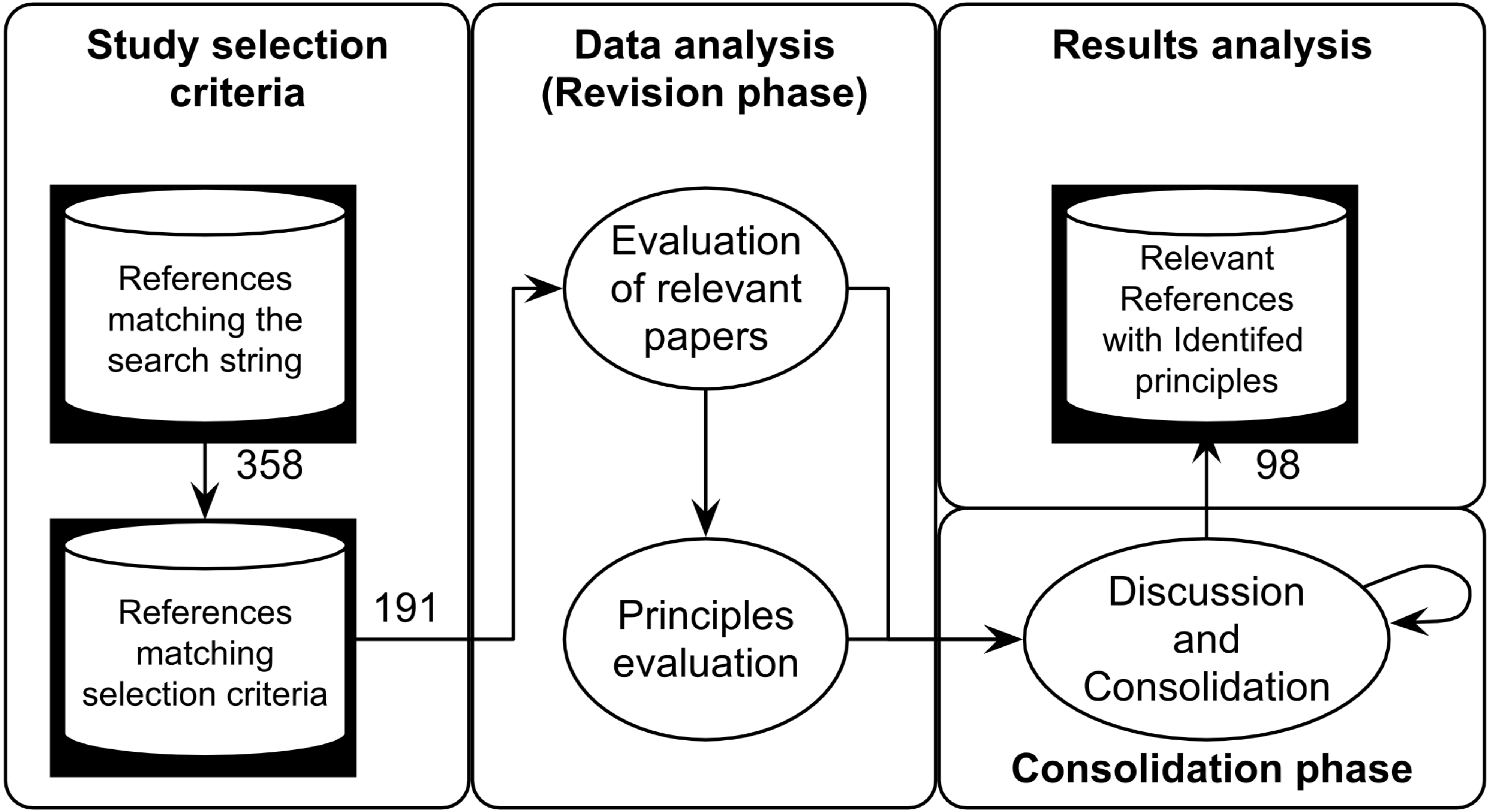
  - **Cost**
  - **Cost model**

How are cloud performance results currently obtained and reported? - Survey

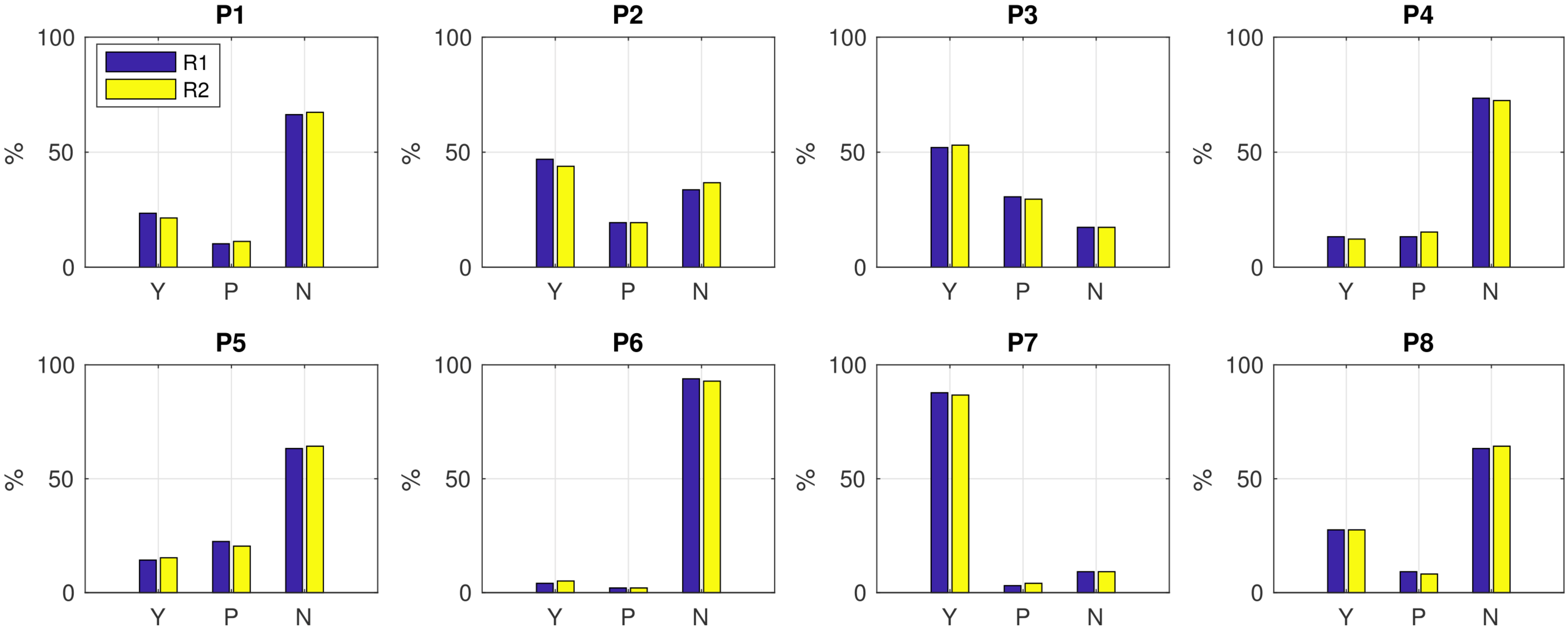
1. Gathered paper meta-data from DBLP and Semantic Scholar
2. Query papers of interest from **16 leading venues**

   "cloud" AND "management" AND NOT("security") AND (YEAR >= 2012)
2. Manually check if they use experimentation
3. Two authors assess which principles are adhered to
4. Consolidation when authors disagree

The Review Process

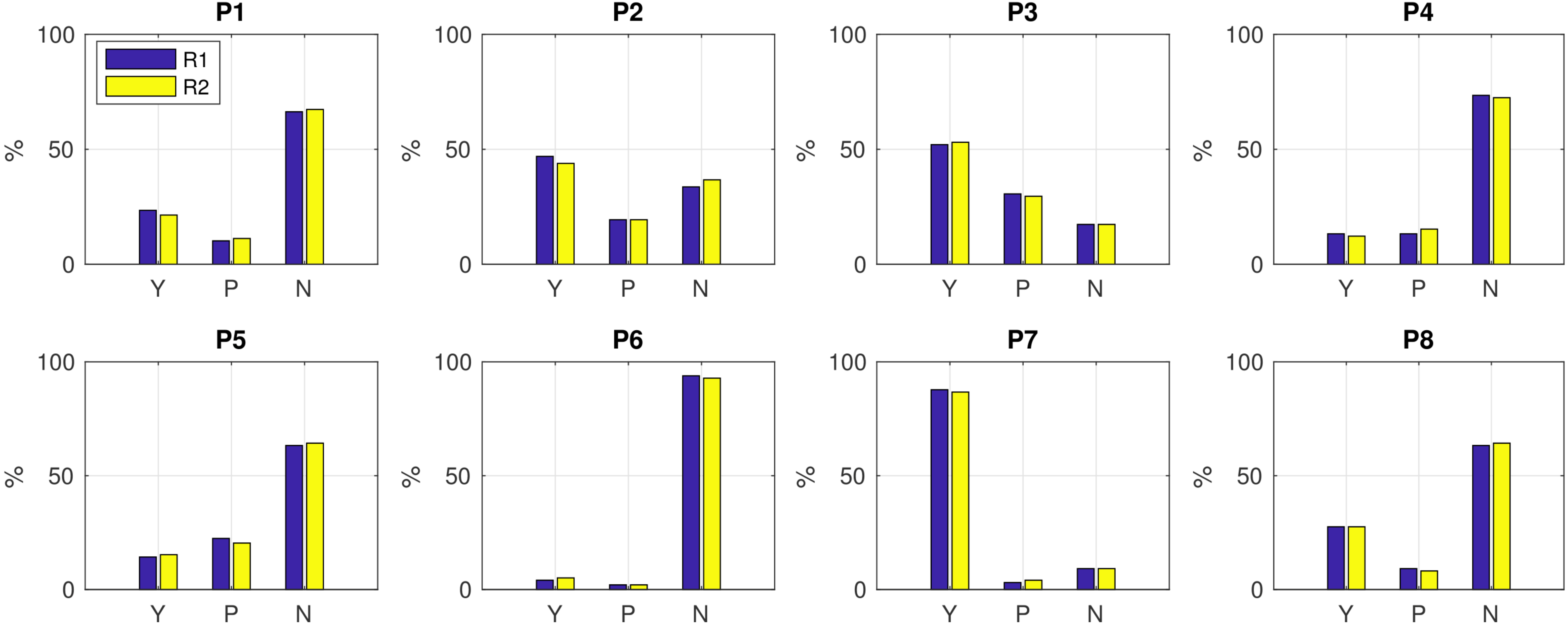

Study selection criteria
Data analysis (Revision phase)
Results analysis
References matching the search string
358
References matching selection criteria
191
Evaluation of relevant papers
Principles evaluation
Relevant References with Identifed principles
98
Discussion and Consolidation
Consolidation phase

Survey results



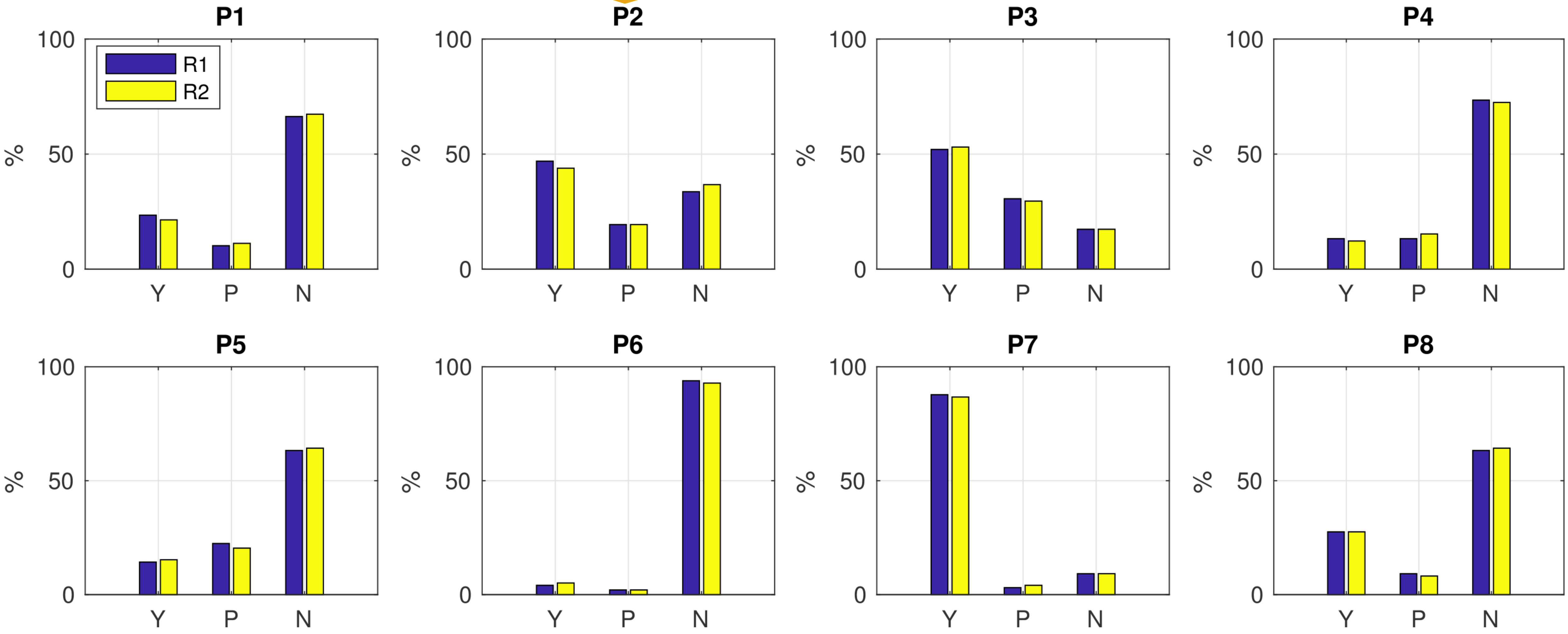
Survey results

P1: Repeated experiments.



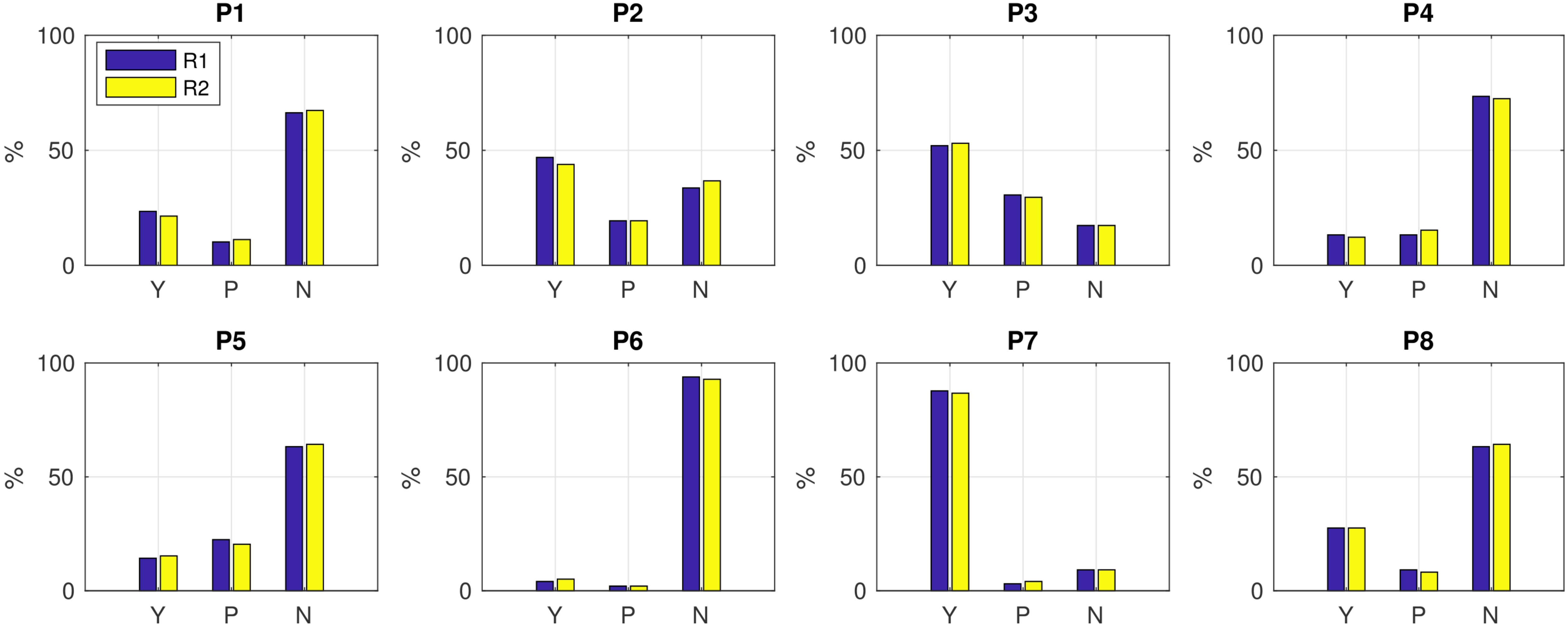
Survey results

P2: Workload and configuration coverage.



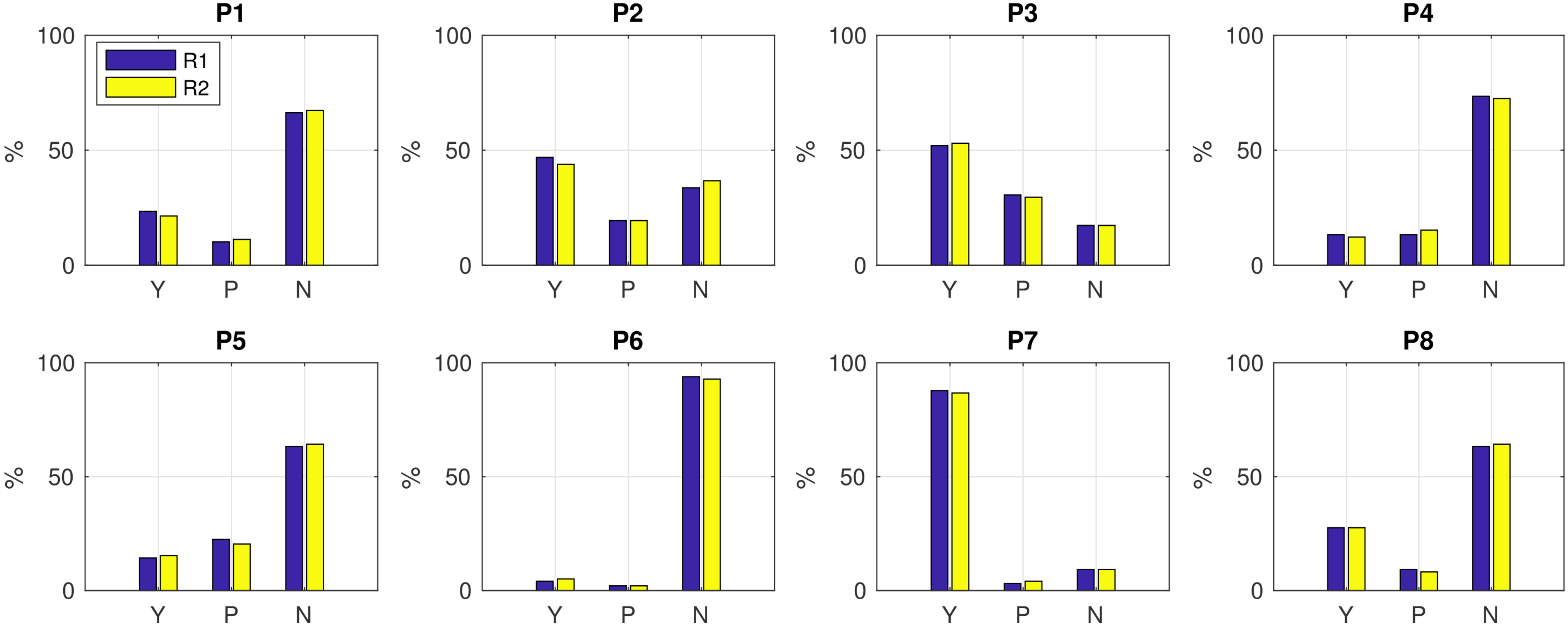
Survey results

P3: Experimental setup description.

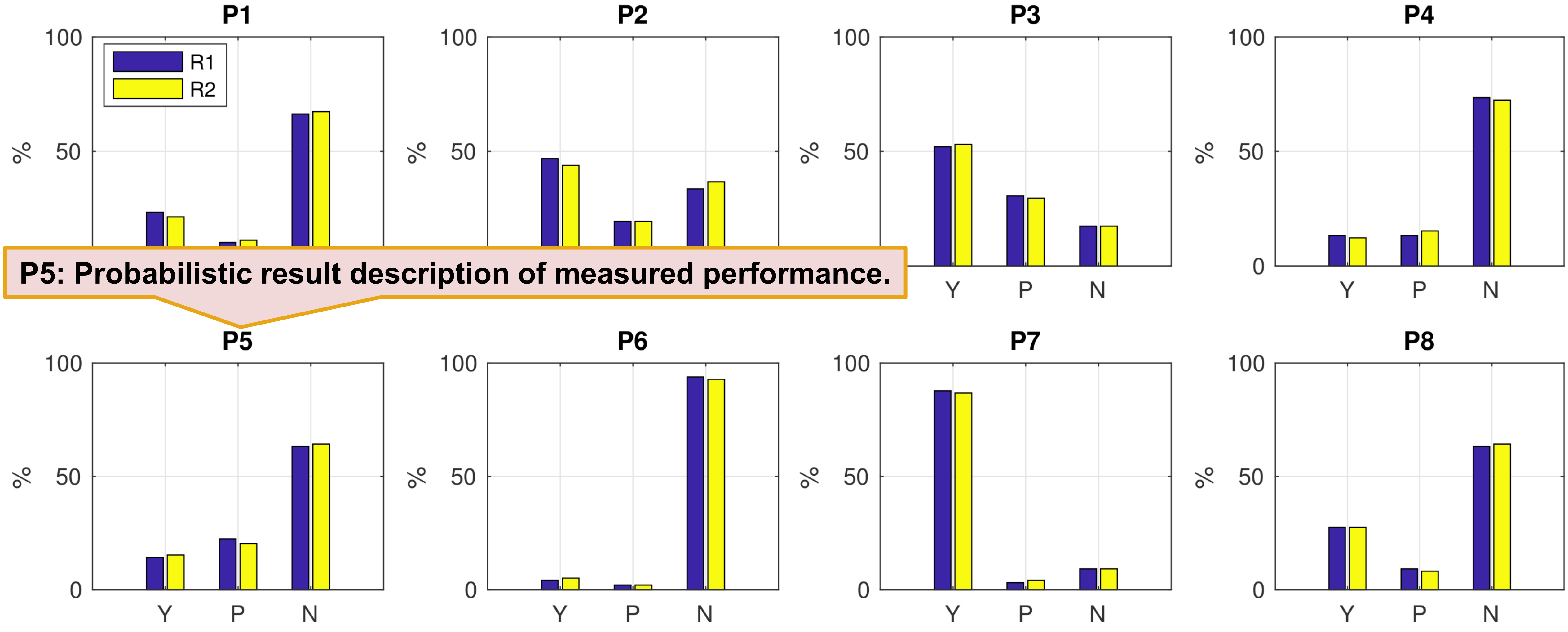


Survey results

P4: Open access artifact.

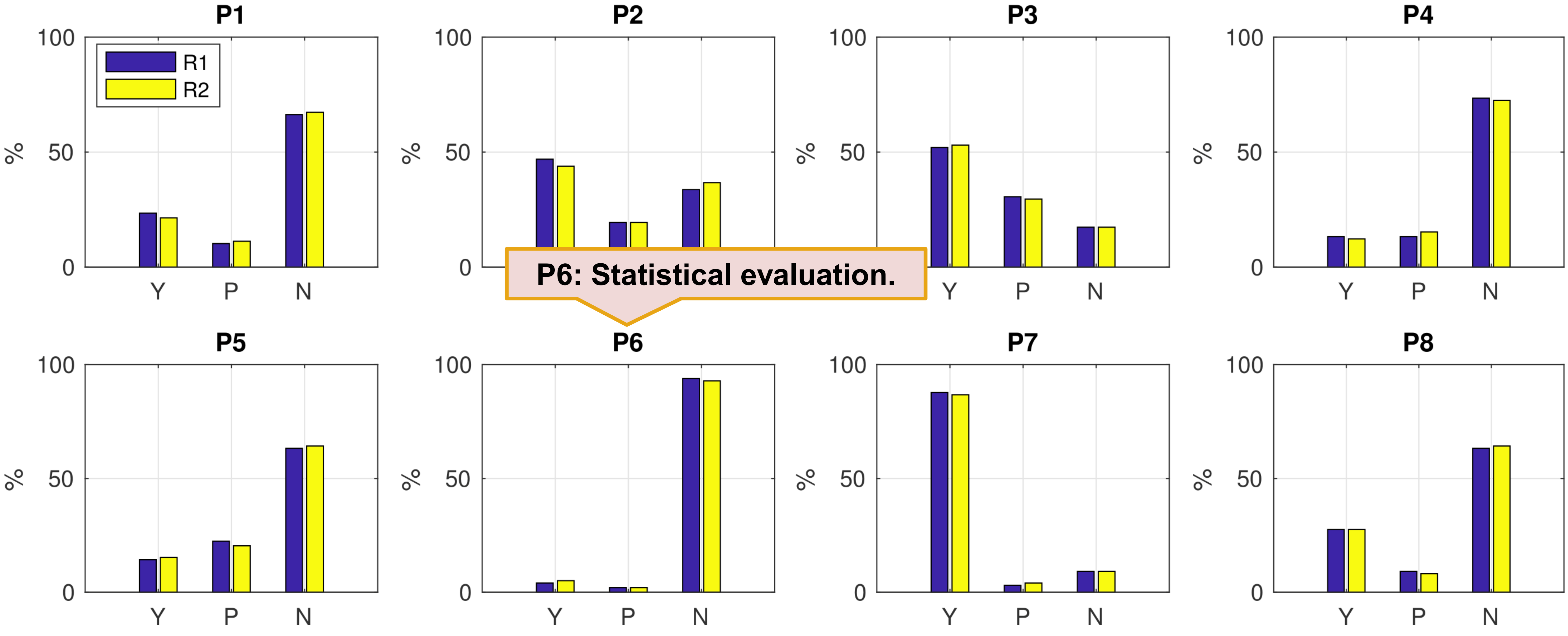


Survey results



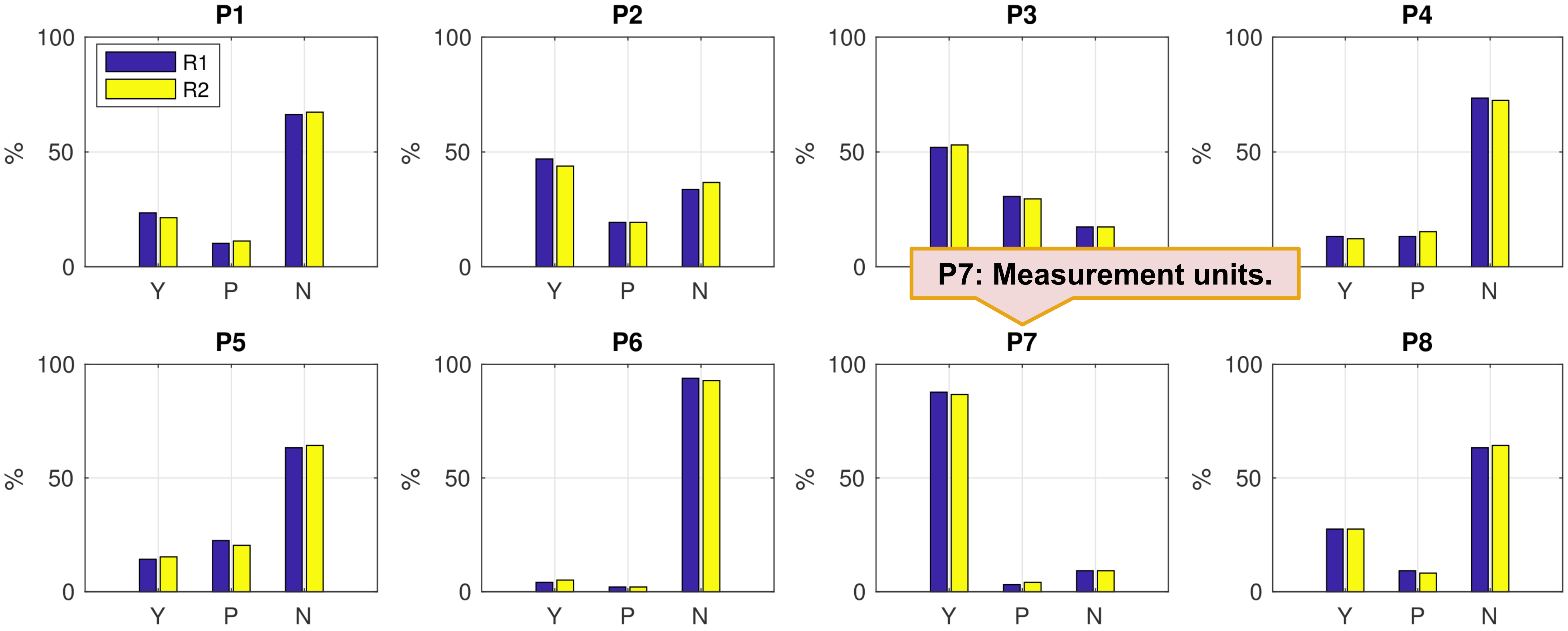
P5: Probabilistic result description of measured performance.

Survey results



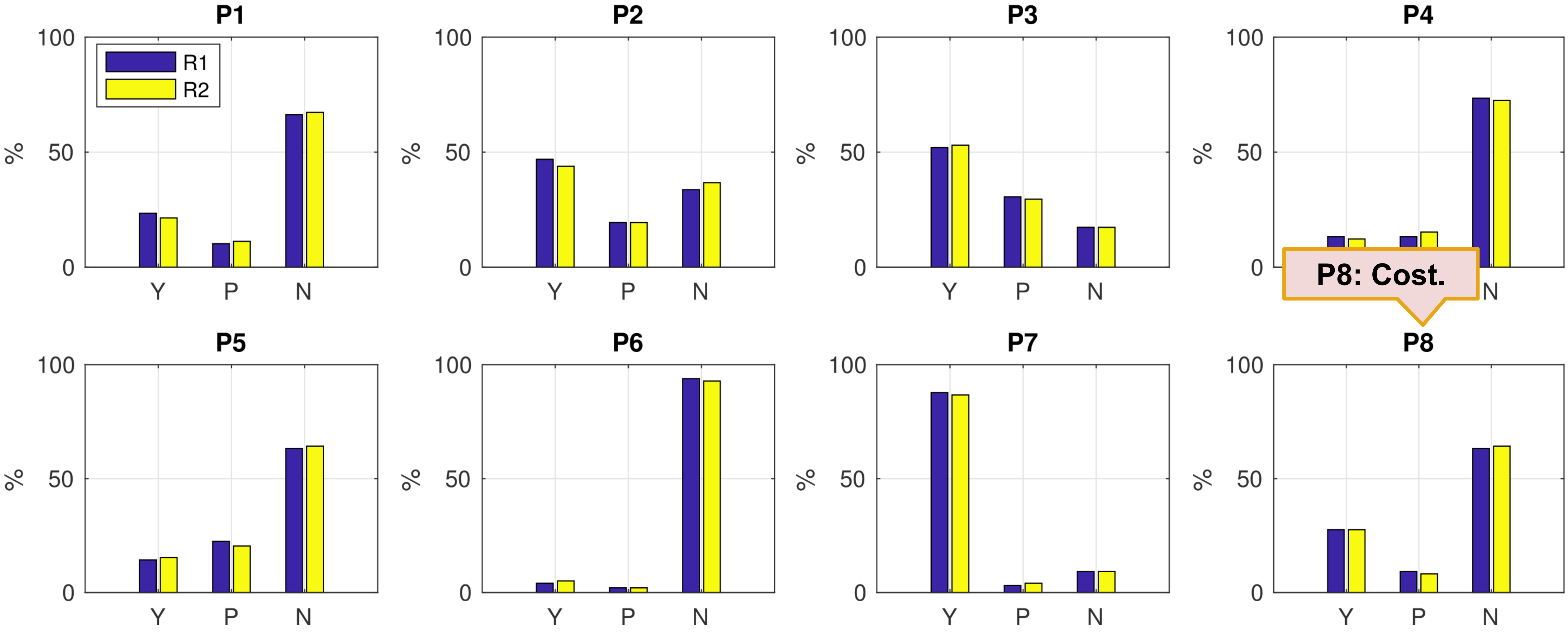
P6: Statistical evaluation.

Survey results



P7: Measurement units.

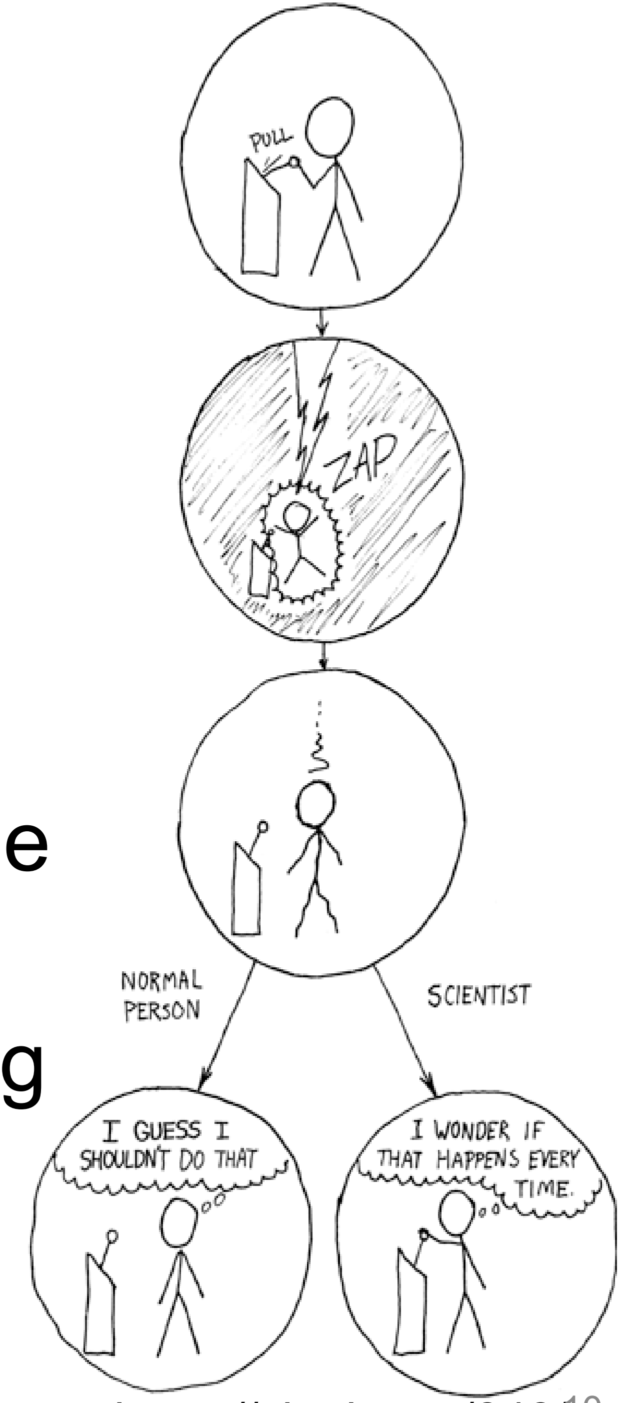
Survey results



P8: Cost.

Wrapping up

- We demonstrated there is a dire need for reproducibility
- To aid researchers, we introduced eight basic principles
- Demonstrated the community adheres to these eight princples in various degrees
- On a positive note: the community is becoming more aware and are tackling this problem


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

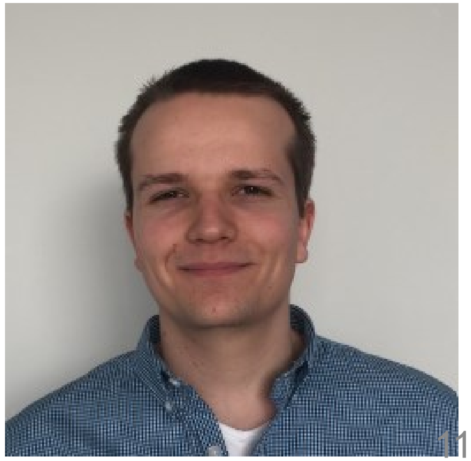
https://xkcd.com/242/

Methodological Principles for Reproducible Performance Evaluation in Cloud Computing

Alessandro Papadopoulos[1], **Laurens Versluis[2]**, André Bauer[3], Nikolas Herbst[3], Jóakim von Kistowski[3], Ahmed Ali-Eldin[4], Cristina Abad[5], José Amaral[6], Petr Tůma[7] and Alexandru Iosup[2]

[1]Mälardalen University, [2]**Vrije Universiteit Amsterdam**, [3]University of Würzburg, [4]Umeå university, [5]Escuela Superior Politecnica del Litoral, [6]University of Alberta, [7]Charles University

More detail in our Technical Report



L.Versluis@atlarge-research.com

P1: Repeated experiments. After identifying the sources of variability, decide how many repetitions with the same configuration of the experiment should be run, and then quantify the confidence in the final result.

P2: Workload and configuration coverage. Experiments should be conducted in different (possibly randomized) configurations of relevant parameters to cover a representative sample of the space of the controlled variables, such as, the workload, especially parameters that exhibit stochastic behaviour in real scenarios, and are thus not completely under control or those that may interact with the platform in unexpected ways, e.g., workloads can change from diurnal to bursty. Parameter values should be randomized according to realistic probabilistic distributions or using historical data. The confidence in the final result should be quantified.

P3: Experimental setup description. Description of the hardware and software setup used to carry out the experiments, and of other relevant environmental parameters, must be provided. This description should include the operating system and software versions, and all the information related to the configuration of each experiment. In addition, the description should clearly state the objective of each experiment.

P4: Open access artifact. At least a representative subset of the developed software and data (e.g., workload traces, configuration files, experimental protocol, evaluation scripts) used for the experiment should be made available to the scientific community. The metadata of the released artifact should uniquely identify the artifact, including timestamping and version in a versioning system.

P5: Probabilistic result description of measured performance. Report a characterization of the empirical distribution of the measured performance, including aggregated values and variations around the aggregation, with the confidence that the results lend to these values.

P6: Statistical evaluation. When making conclusions from experimental data, provide a statistical evaluation of the significance of the obtained results.

P7: Measurement units. For all the reported quantities, report the corresponding unit of measurement.

P8: Cost. Every cloud experiment should include (i) the cost model used or assumed for the experiment; (ii) accounted resource usage (per second), independently of the model; and (iii) charged cost according to the model.