# Brainstorming about WG4 Themes

Nicola Ferro and Paulo Rupino da Cunha


Keystone Winter WG Meeting 2017
20-21 February 2017, Belgrade, Serbia

# WG4 Goals

- Develop Scenarios
- Define Requirements and Functionalities
- Identify solutions and prototypes
- Benchmark

# Benchmark Challenges: Datasets and Corpora

- Corpora (datasets + topics + relevance judgements)
  - They should target clearly identified scenarios and functionalities

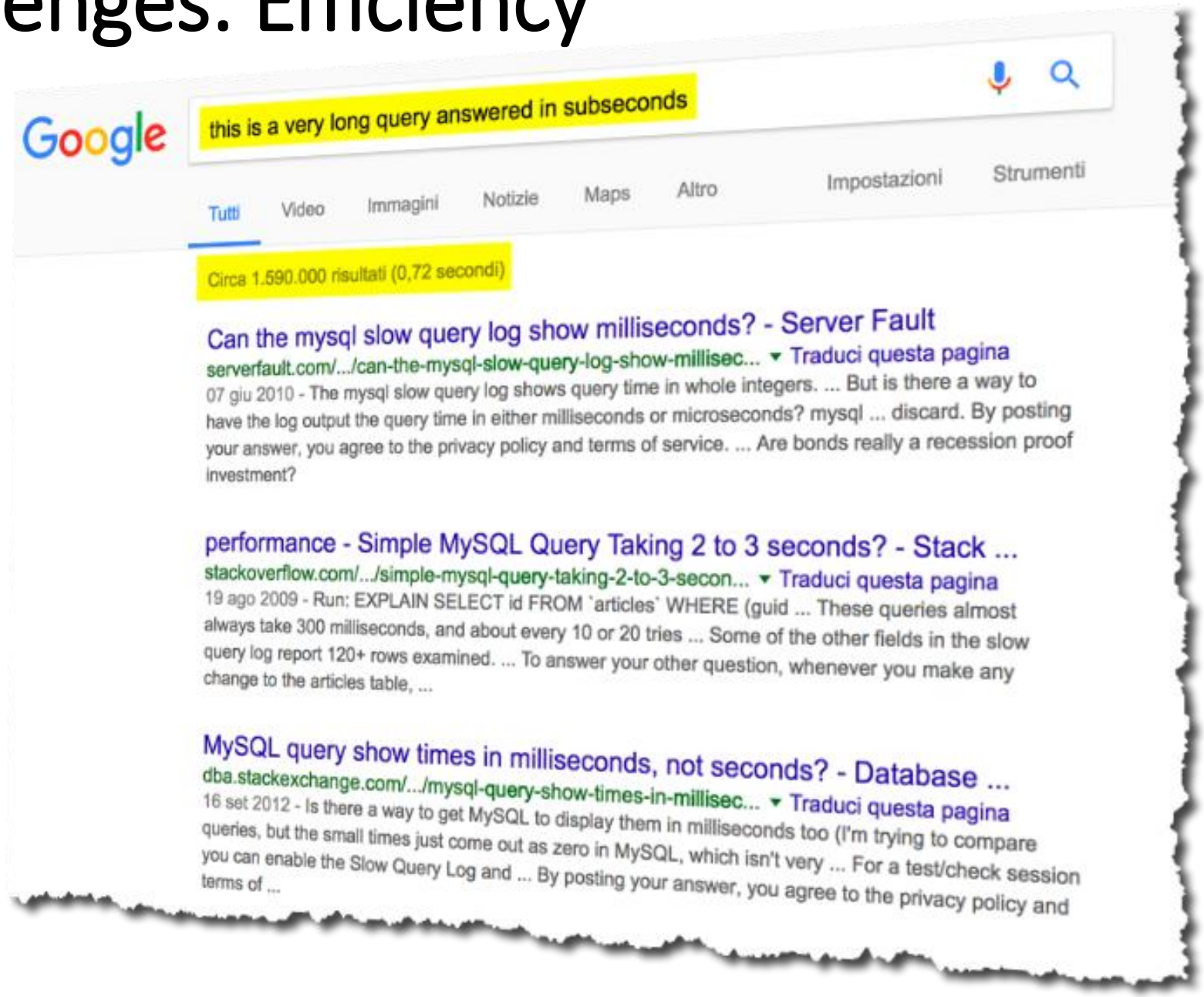
- Datasets – How to define/create/choose them?
  - What extent of “structuredness”?
  - What complexity of?
  - What scale/size?
  - What about vagueness?

- Topics – How to create/choose them?
  - Queries vs user information needs

- Relevance Judgements – How to create them?
  - What is the unit of retrieval?
  - How do we match the output of algorithms to units of retrieval?
  - How many relevant units of retrieval per topic?

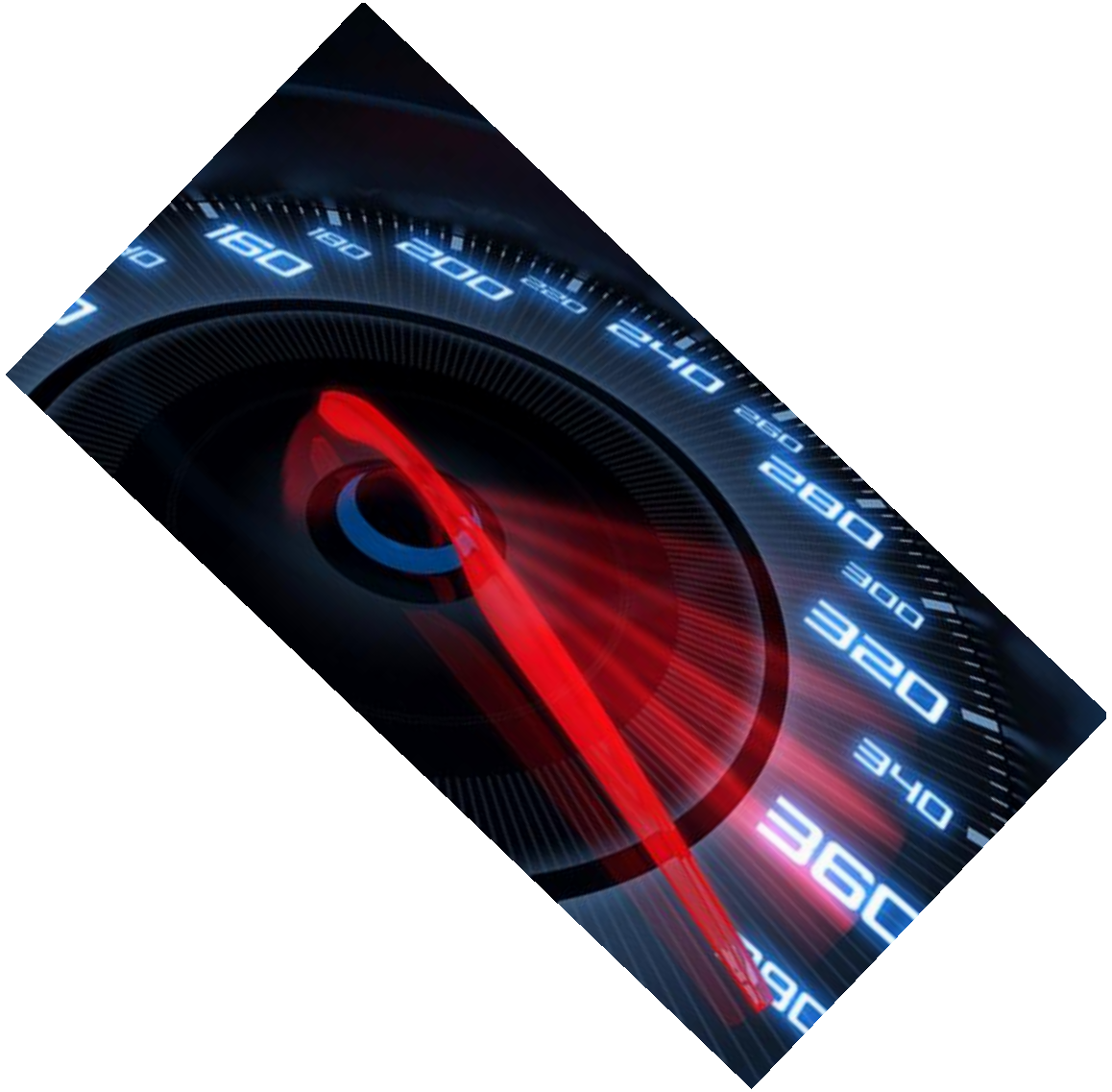
# Benchmark Challenges: Efficiency

- 2-3 query terms
  - From seconds to minutes
- 4-6 query terms
  - From minutes to hours
- Long queries
  - Unmanageable
- Memory occupation



# Benchmark Challenges: Effectiveness

- What is our perception of effectiveness?
  - Nearly 100% of precision is not very realistic
- How many units of retrieval returned?
  - How does this relate to the number of relevant units of retrieval?
  - Not all the measures are equally good



# Benchmark Challenges: Reproducibility

- No open source full-fledged systems
  - Like Terrier or Lucene
- Techniques difficult to implement
  - Many omitted details in the papers
- Not all the techniques can work on all the datasets
  - Sometimes very strong assumptions not matched by a schema (e.g. BLINKS)

# Reproducibility: advertising

- ACM Journal of Data and Information Quality (JDIQ)
- Special Issue on Reproducibility in Information Retrieval
- Submission deadline
  8 September 2017

CALL FOR PAPERS

## ACM Journal of Data and Information Quality

Special Issue on Reproducibility in Information Retrieval

**Guest editors**
Nicola Ferro, University of Padua, Italy, ferro@dei.unipd.it
Norbert Fuhr, University of Duisburg-Essen, Germany, norbert.fuhr@uni-due.de
Andreas Rauber, Technical University of Vienna, Austria, rauber@ifs.tuwien.ac.at

**Aim**

Information Retrieval is a discipline that has been strongly rooted in experimentation since its inception. Experimental evaluation has always been a strong driver for IR research and innovation, and these activities have been shaped by large scale evaluation campaigns such as TREC, CLEF, NTCIR and FIRE.

IR systems are getting more and more complex. They need to cross language and media barriers; they span from unstructured, to semi-structured, to highly structured data; and they are faced with diverse and complex user information needs, search tasks, and societal challenges. As a consequence, evaluation and experimentation, which has remained a fundamental element, has in turn become increasingly sophisticated and challenging.

In this context, *repeatability*, *reproducibility*, and *generalizability* of experiments and results cannot be taken for granted. Indeed we need to emphasize these aspects as key requirements if we wish to continue to reliably and durably advance research and technology in the field. In turn, we need to actively pursue them as a core part of our experimental methodology and practice.

In this special issue of JDIQ, we aspire to provide an overview of innovative research at the *intersection of information retrieval and data quality*, from theory to practice, with a focus on challenges, solutions, and experiences in *reproducibility of IR experimental results*.

**Topics -** Specific topics within the scope of the call include, but are not limited to, the following:

- Analysis of reproducibility challenges in system-oriented evaluation
- Analysis of reproducibility challenges in user-oriented evaluation
- General reproducibility frameworks for IR
- Lessons learned in reproducing third-party experiments
- Reproducibility of query results
- Reproducibility challenges on private or proprietary data
- Reproducibility challenges on ephemeral data, like streaming data, tweets, etc.
- Reproducibility challenges on online experiments, e.g., A/B testing
- Reproducibility in evaluation campaigns
- Evaluation infrastructures and Evaluation as a Service (EaaS)
- Experiment data management, data curation, and data quality
- Data models, semantic or not, for IR experimental data
- Reproducible experimental workflows: tools and experiences
- Quality of IR experimental data
- Data Citation: citing experimental data, dynamic data sets, samples, and statistical analyses

**Expected contributions -** We welcome the following two types of contributions:

- Research manuscripts reporting mature results [25+ pages].
- Experience papers that report on lessons learned from addressing specific issues towards improved quality and reproducibility of experimental results [12+ pages plus an optional appendix].

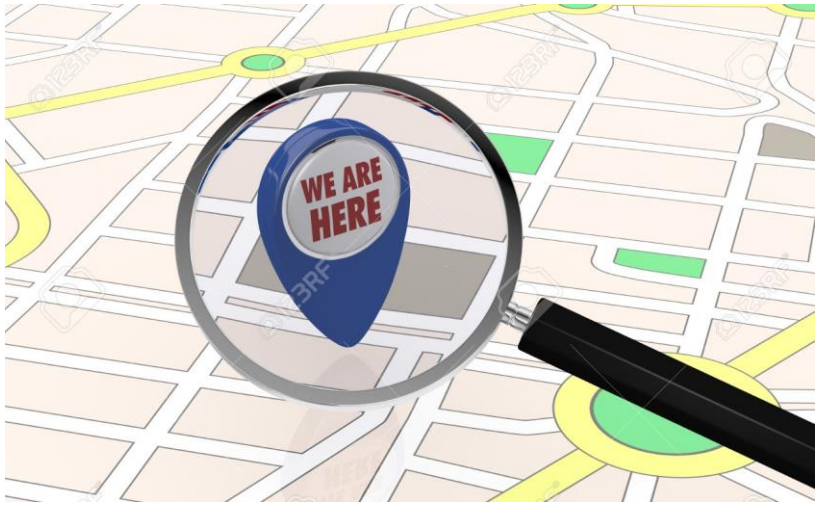
If this is an extension of prior published work, then submitted manuscripts must contain at least 30% new material, and the significant new contributions must be clearly identified in the introduction.

**Submission guidelines** with Latex (preferred) or Word templates are available here:
http://jdiq.acm.org/authors.cfm#subm

**Important dates:**

| | | | |
|---|---|---|---|
| Initial submission: | Friday, September 8, 2017 | Second review: | Friday, May 11, 2018 |
| First review: | Thursday, December 7, 2017 | Camera-ready manuscripts: | Friday, July 13, 2018 |
| Revised manuscripts: | Friday, March 9, 2018 | Publication: | Late October 2018 |

# Where KEYSTONE WG4 is?




- Who is doing what?
- What are the main achievements?
- What are the available resources (datasets, topics, ground-truth, prototypes)?
- What are still open challenges?

**Outcomes survey:**
**https://tinyurl.com/keystoneWG4**

**WG4 chapter**
**in the final KEYSTONE Book**