EDA'2013

Gathering Real OLAP Analysis Sessions: A Feedback

Julien Aligon
Université François Rabelais Tours
Laboratoire Informatique
France

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

LI Laboratoire d'Informatique EA 6300

EDA'2013

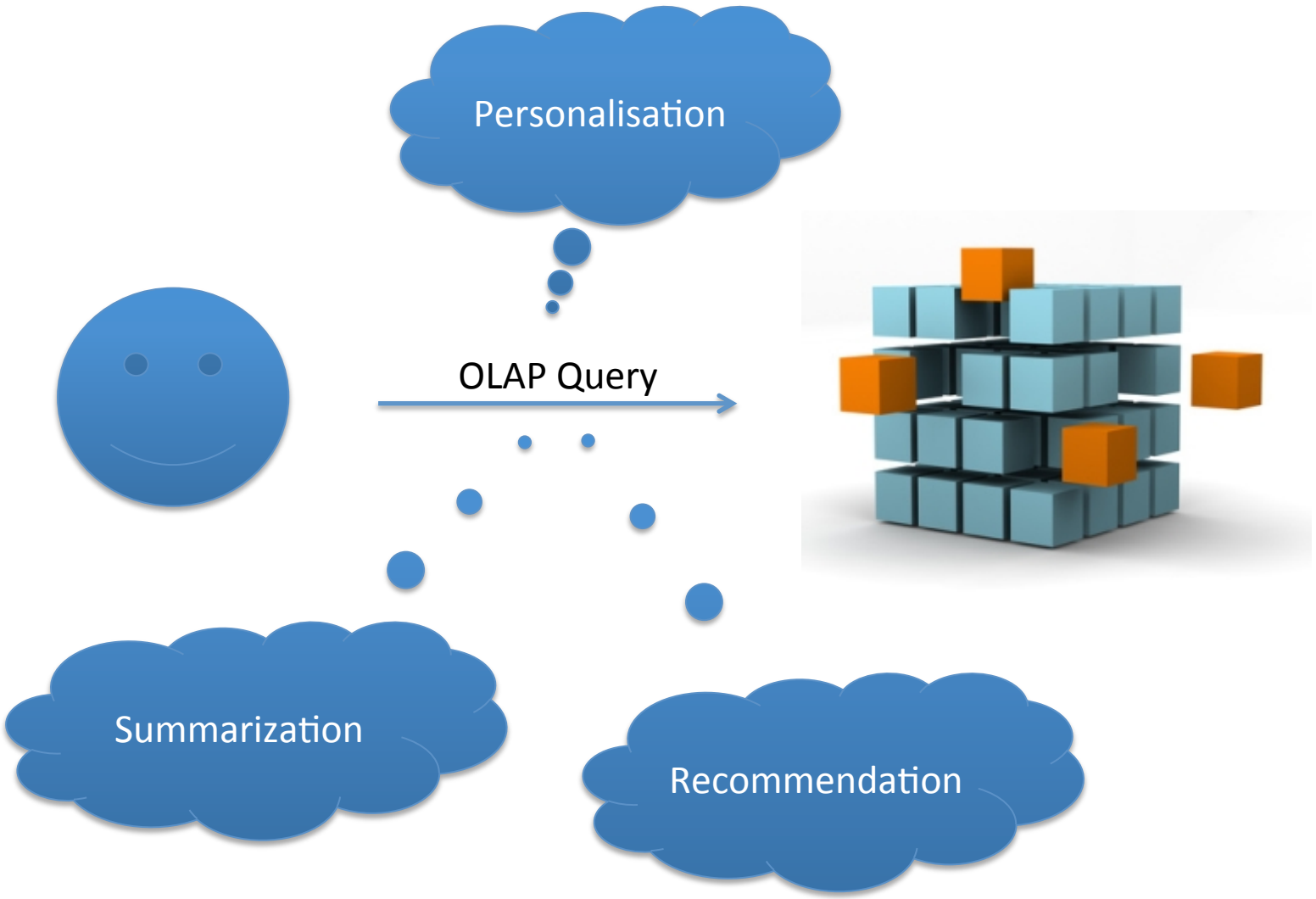
Content

- Purpose of the work
- Obtaining logs
- The Feedback
- Statistical results
- Conclusion & Discussion

Purpose of the work

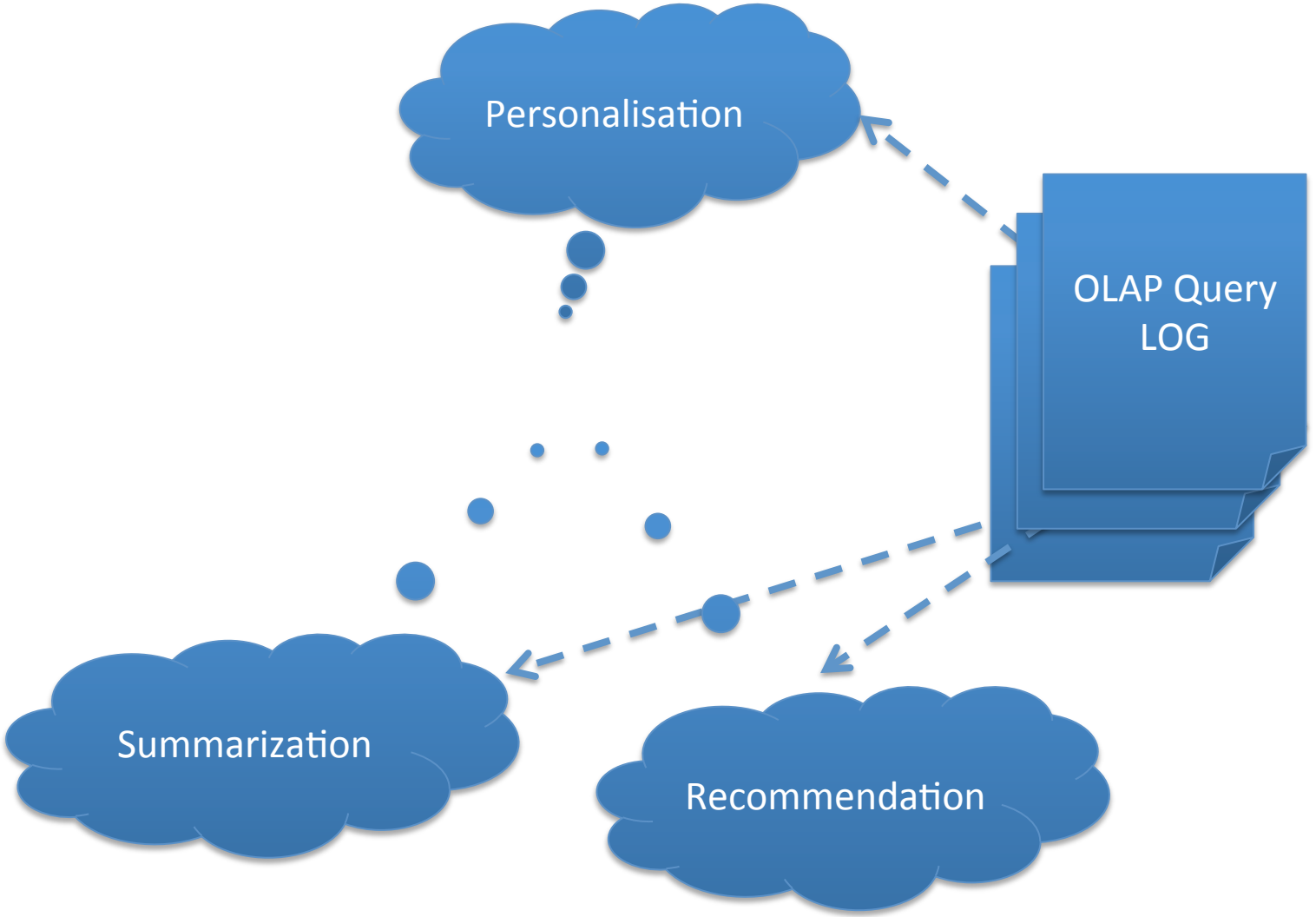
Gathering Real OLAP Analysis Sessions: A Feedback

Purpose of the work



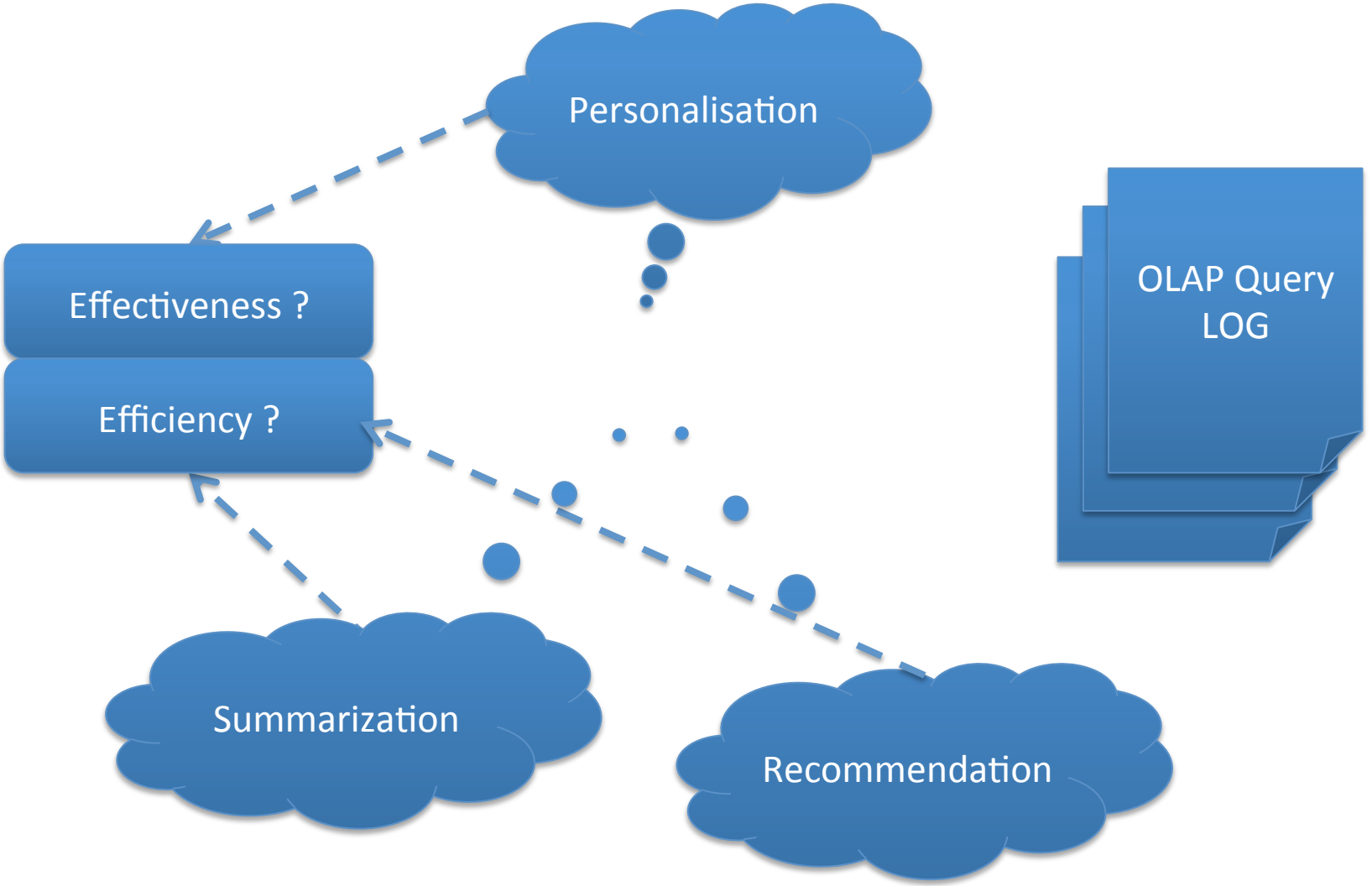
Gathering Real OLAP Analysis Sessions: A Feedback

Purpose of the work



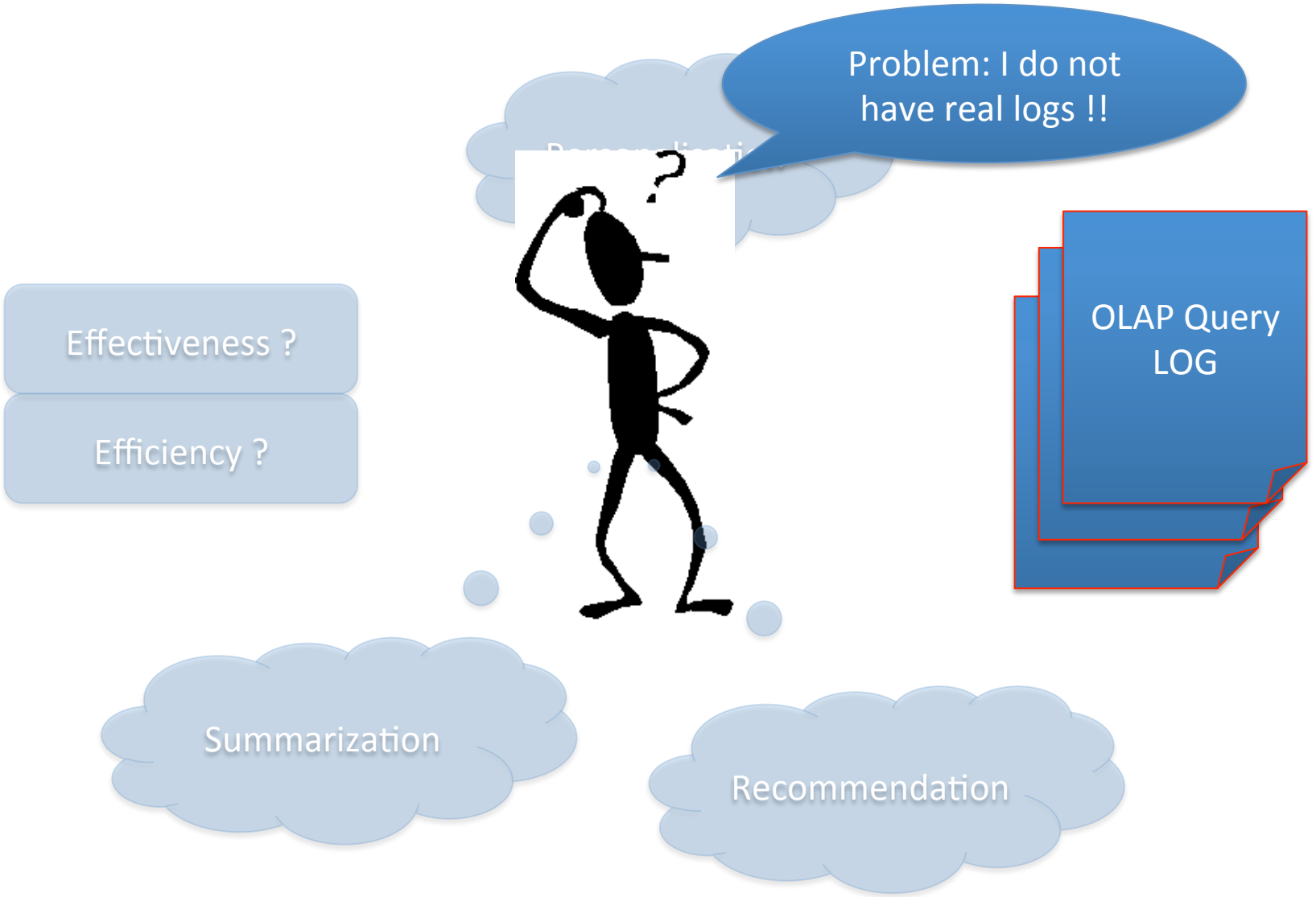
Gathering Real OLAP Analysis Sessions: A Feedback

Purpose of the work



Gathering Real OLAP Analysis Sessions: A Feedback

Purpose of the work



Trivial workloads

« Academic research on Big Data is excessively based on boring data and nearly trivial workloads. On the other hand, Big Data research aims to obtain insights from interesting data and cope with demanding workloads. This is a striking mismatch. »

Gerhard Weikum, Sigmod Blog, 2013

Obtaining Logs

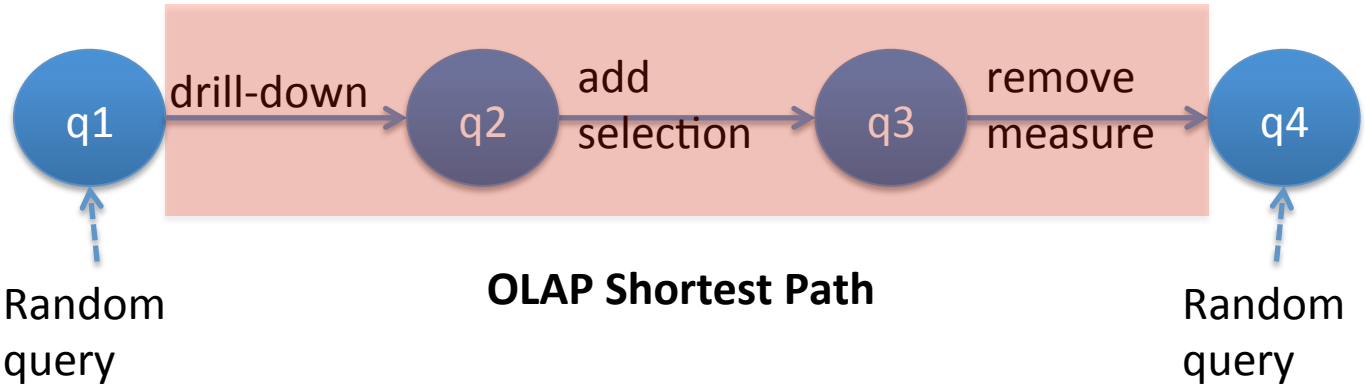
How to obtain logs ?

- from companies ? → too difficult (sensitive data, etc…)
- from synthetic generation (imitating analysis behavior)

Gathering Real OLAP Analysis Sessions: A Feedback

Solutions imitating OLAP sessions

Example:



Example:

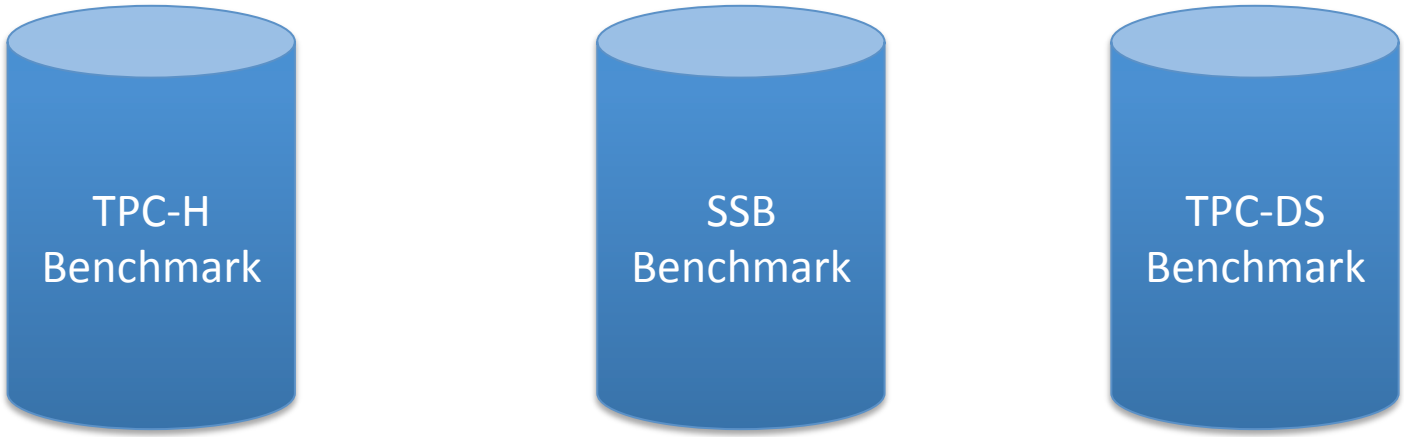
Operators for aing data-cube: ***[Sarawagi in VLDB 1999, 2000 and 2001]***

How to obtain logs ?

- from companies ? → too difficult (sensitive data, etc…)
- from synthetic generation (imitating analysis behavior)
- from public data

Gathering Real OLAP Analysis Sessions: A Feedback

Solutions from public data

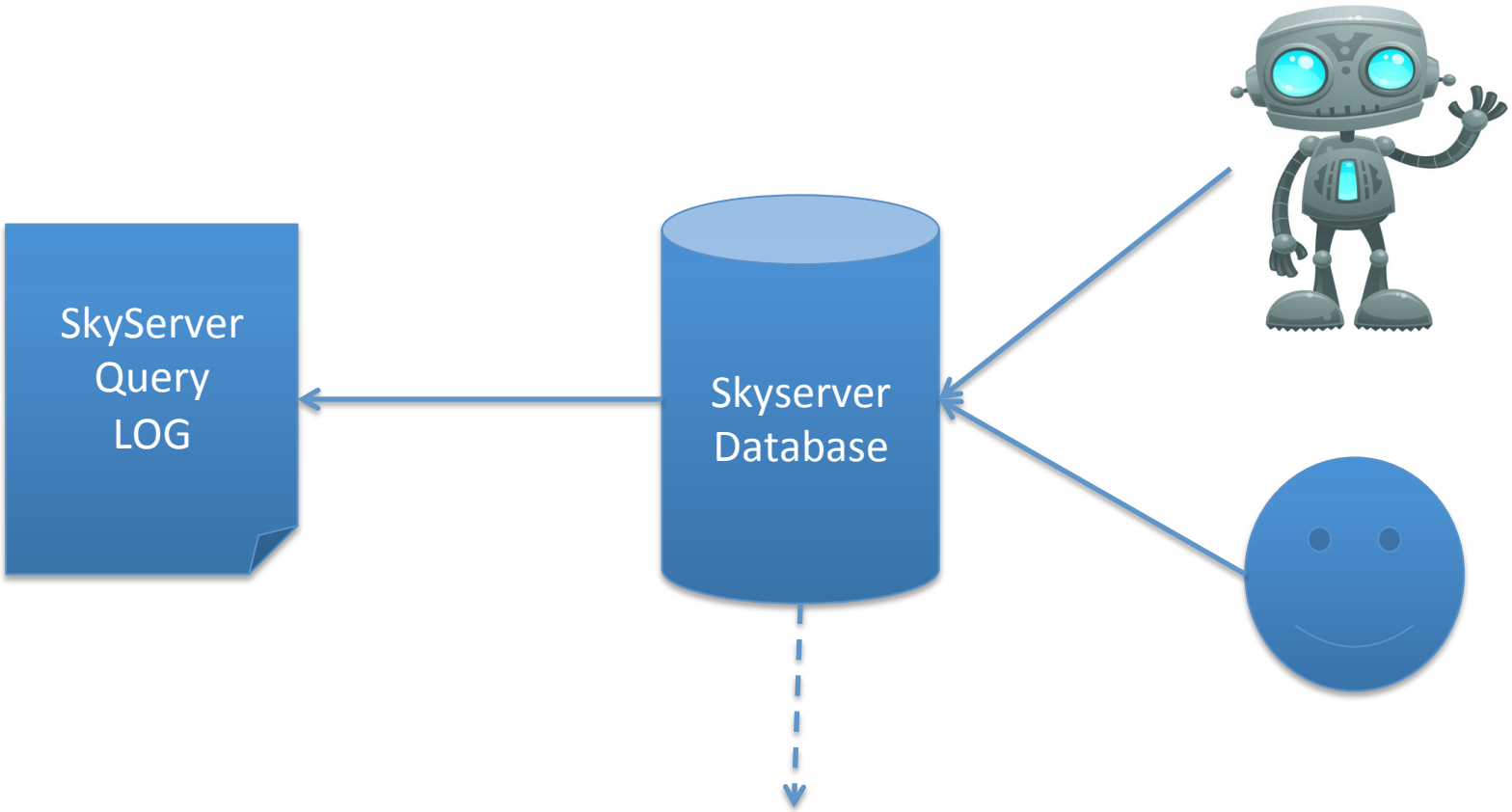


Gathering Real OLAP Analysis Sessions: A Feedback

Solutions from public data

Gathering Real OLAP Analysis Sessions: A Feedback

Solutions from public data



No multidimensional !

How to obtain logs ?

- from companies ? → too difficult (sensitive data, etc…)
- from synthetic generation (imitating analysis behavior)
- from public data
- from analysis reports (retro-engineering)
  → **to be investigated !**

Gathering Real OLAP Analysis Sessions: A Feedback

Obtaining Logs

How to obtain logs ?

- from graduate students in Business Intelligence ?
  →why not !

  Advantage:
  - potentially many students (→ many sessions)
  - knowledge in OLAP
  - already done in previous works (***[Aligon&al. to appear in KAIS], [Khoussainova&al. CIKM'2011]***)

  Disadvantage:
  - potentially beginner?

Note: This work is a feedback, not a proposal for a benchmark !

The feedback

Framework of the test

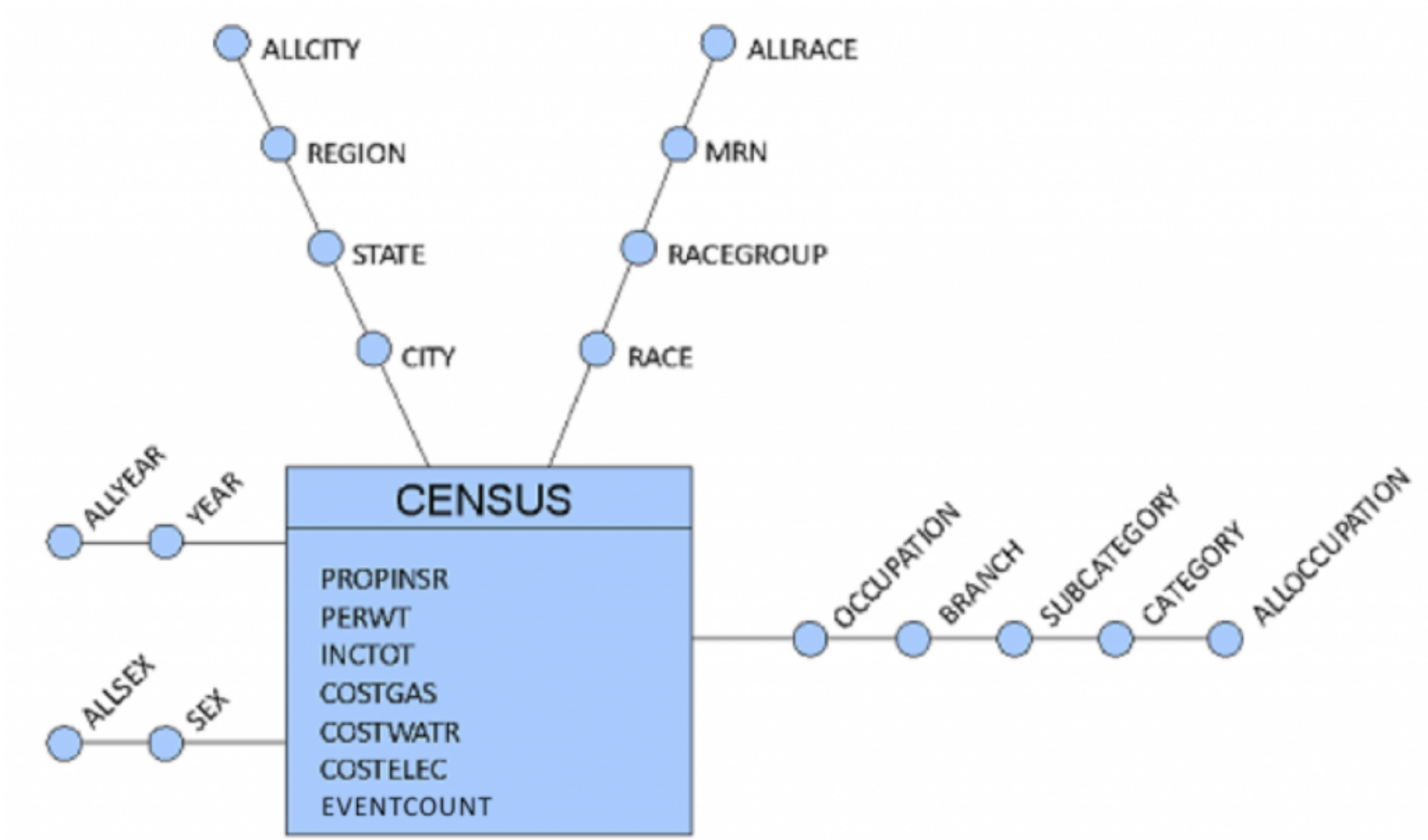
- Using a simple OLAP schema
- Defining several questionnaires including different degrees of difficuties with various questions
- Hiding the query language with a GUI

Gathering Real OLAP Analysis Sessions: A Feedback

Tests conducted with students from:

- the University François Rabelais Tours (France, 18 students)
- the University of Bologna (Italy, 22 students)
- **IT4BI students (Erasmus Mundus Program) → not included yet (24 sessions)**

OLAP schema


ALLCITY
REGION
STATE
CITY
ALLRACE
MRN
RACEGROUP
RACE
ALLYEAR
YEAR
ALLSEX
SEX
CENSUS
PROPINSR
PERWT
INCTOT
COSTGAS
COSTWATR
COSTELEC
EVENTCOUNT
OCCUPATION
BRANCH
SUBCATEGORY
CATEGORY
ALLOCCUPATION

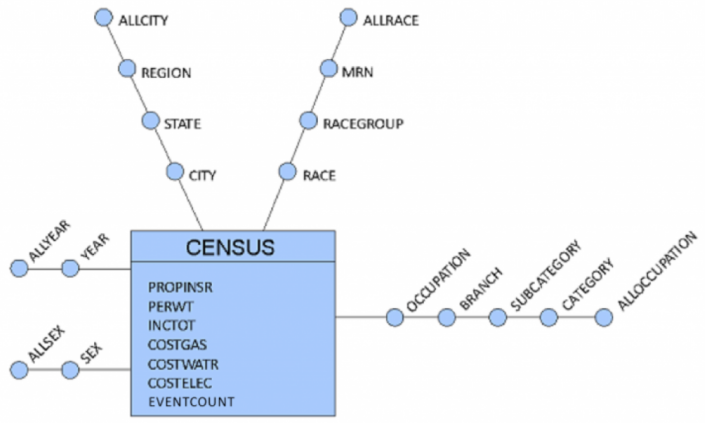
EDA'2013

Query Definition

- a fragment-based form with:
  - Group-by set (G)
  - Set of predicates (P)
  - Measure set (M)

Example:

<{Sex, Year, AllCity, AllRace, AllOccupation},
{Sex='Female', Year={'2000', '2001'}},
{INCTOT}>



Questionnaires

3 types of needs:

- Individual profile (Sex + Race dimensions)
- Occupation
- Mixed (Individual profile + Occupation)

2 sub-types for each needs:

- INCTOT measure
- Energy measures (COSTGAS, COSTWATR and COSTELEC)

6 questionnaires developed

Questionnaires

3 levels of difficulties:

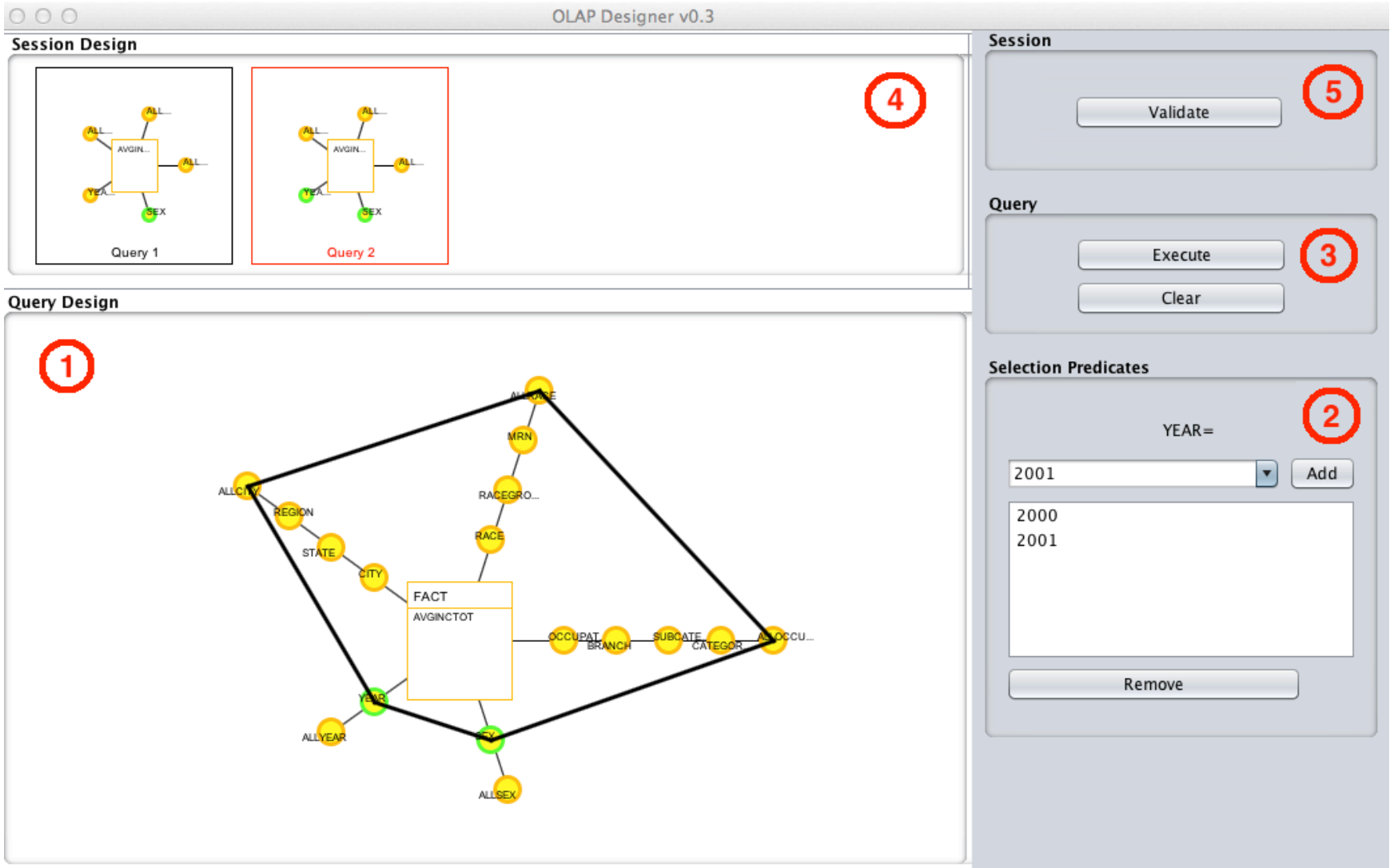
- Basic need: *Is there a trend in the evolution of the average cost of gas for some profiles?*
- Intermediate need: *Compare the evolution of the minimum of energy costs, for the highest income, with the evolution of the maximum energy costs for the lowest incomes.*
- Advanced need: *Where is it better to live in terms of incomes, for an occupation?*

5 questions per questionnaire

Gathering Real OLAP Analysis Sessions: A Feedback

Feedback

GUI



Statistical Results

Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

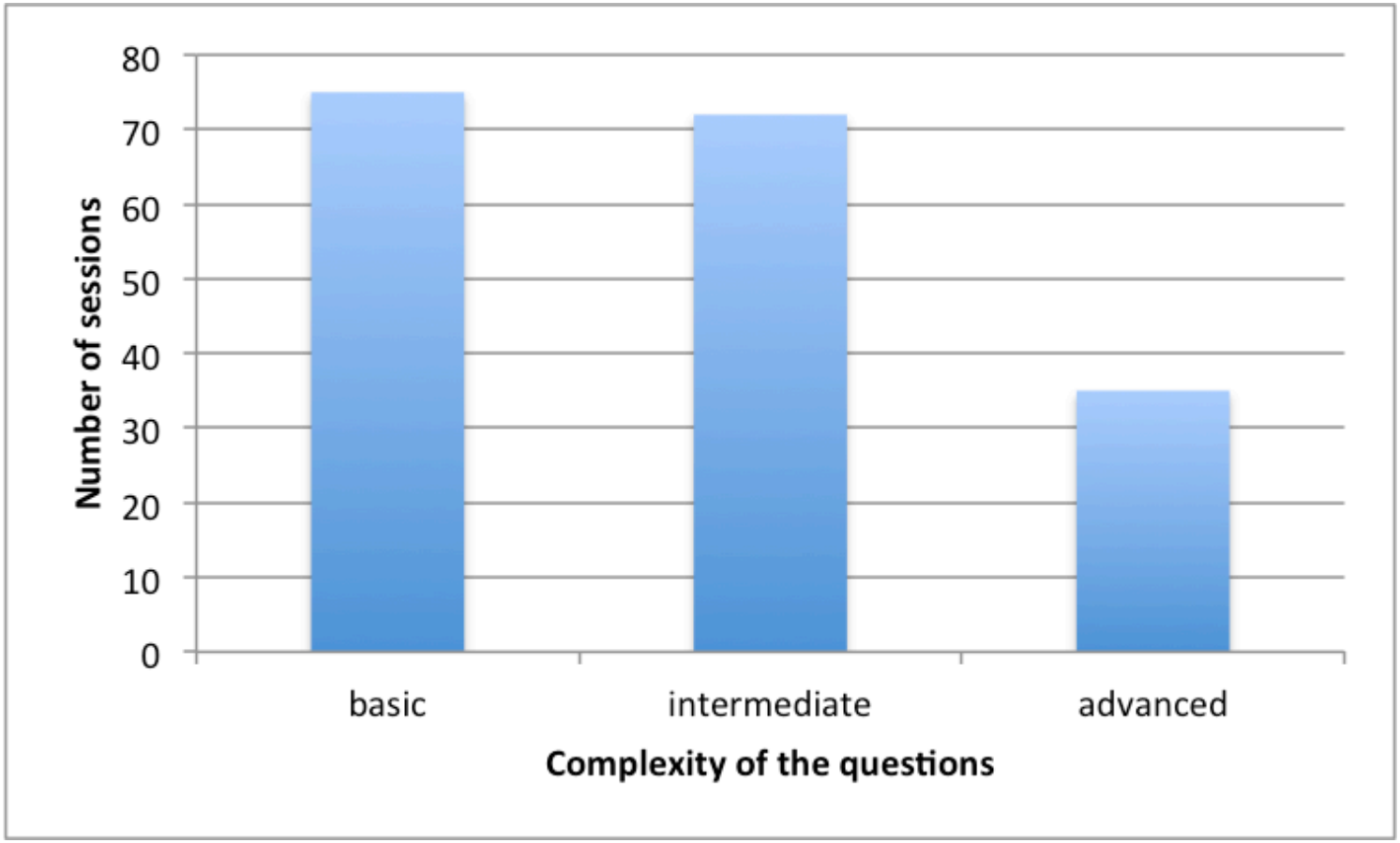
Over logs:

- 182 sessions (85 from France, 97 from Italy)
- 810 queries
- Each questionnaire has been done 4-5 times

Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

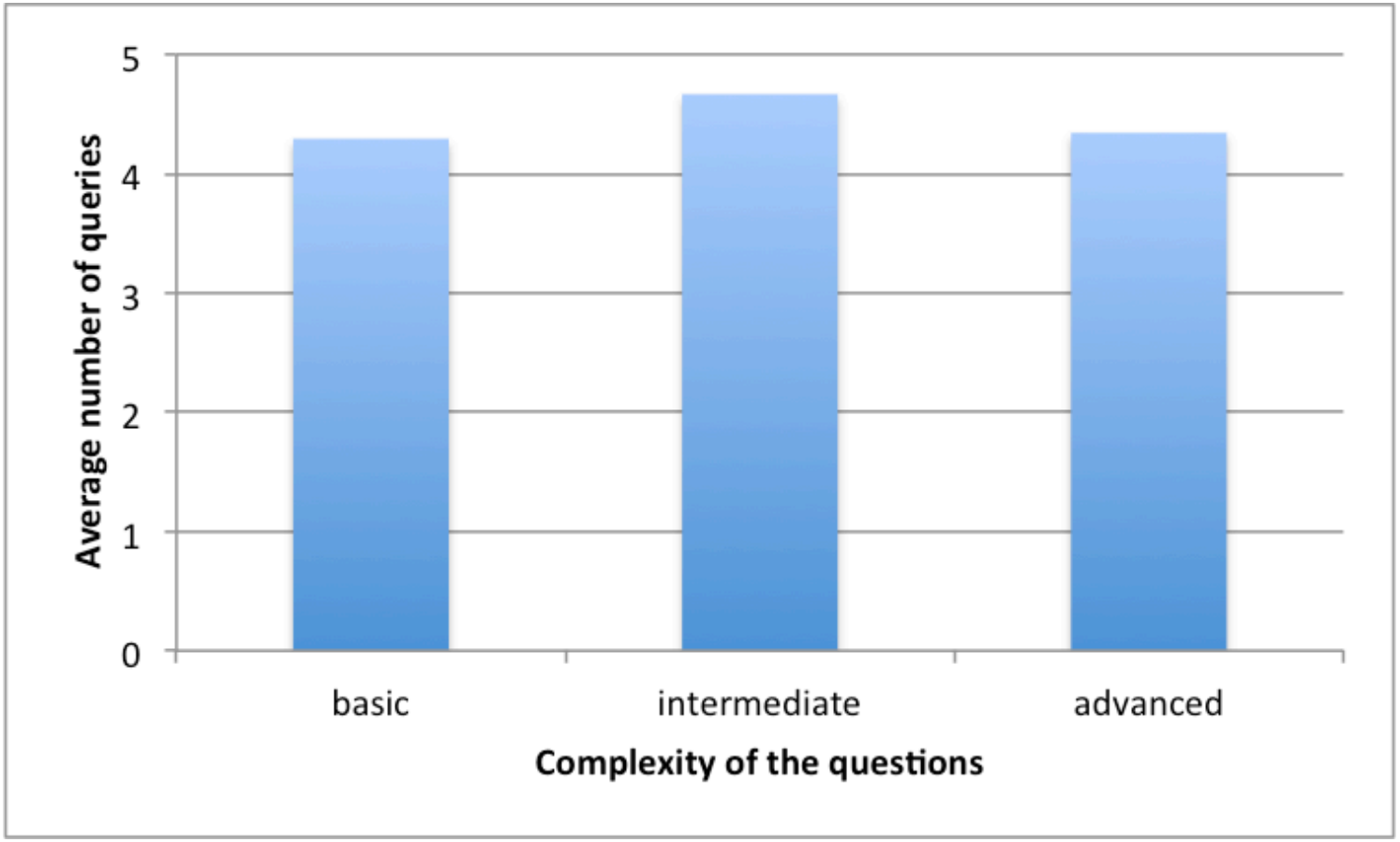
Over sessions:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

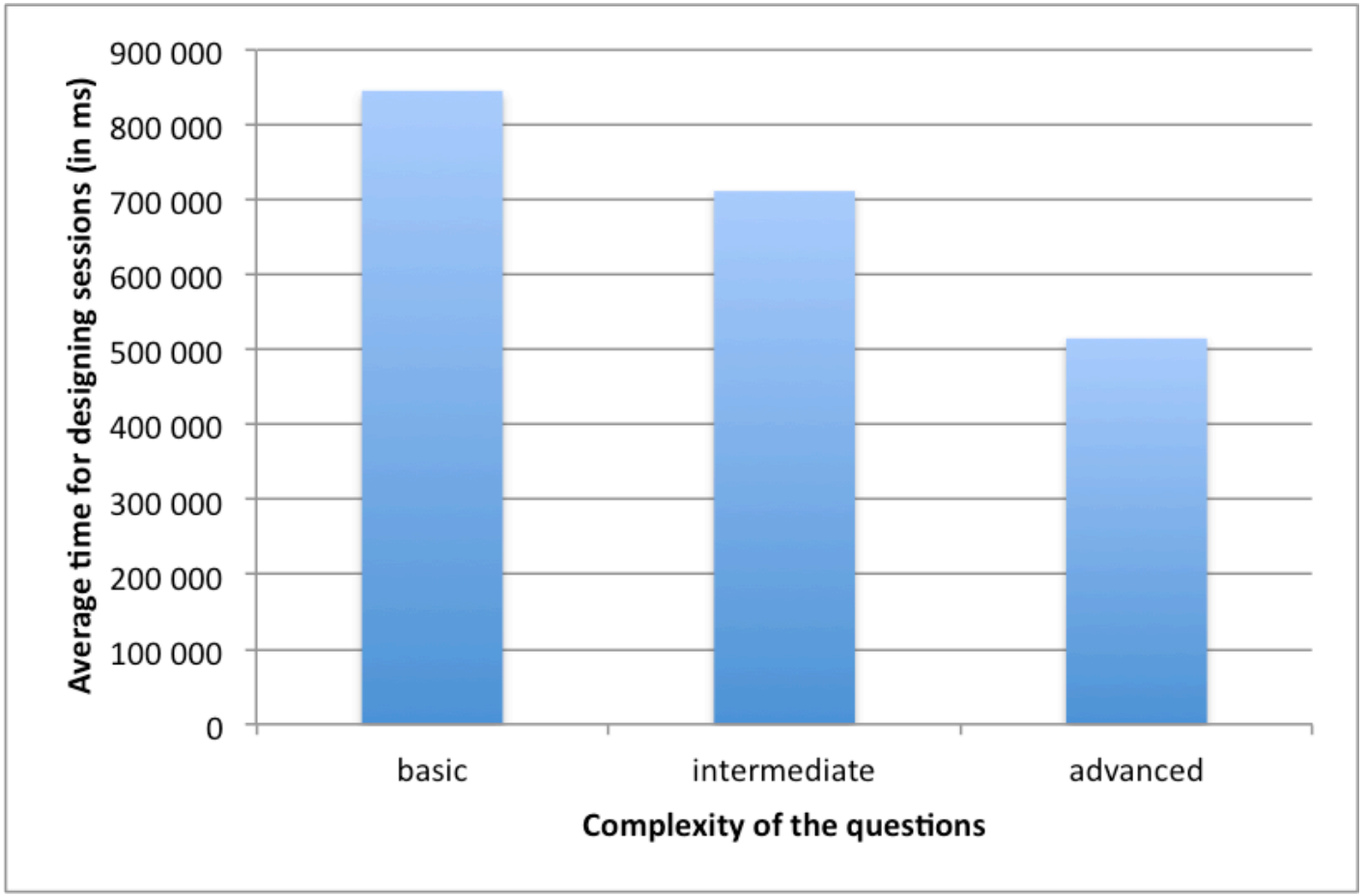
Over sessions:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

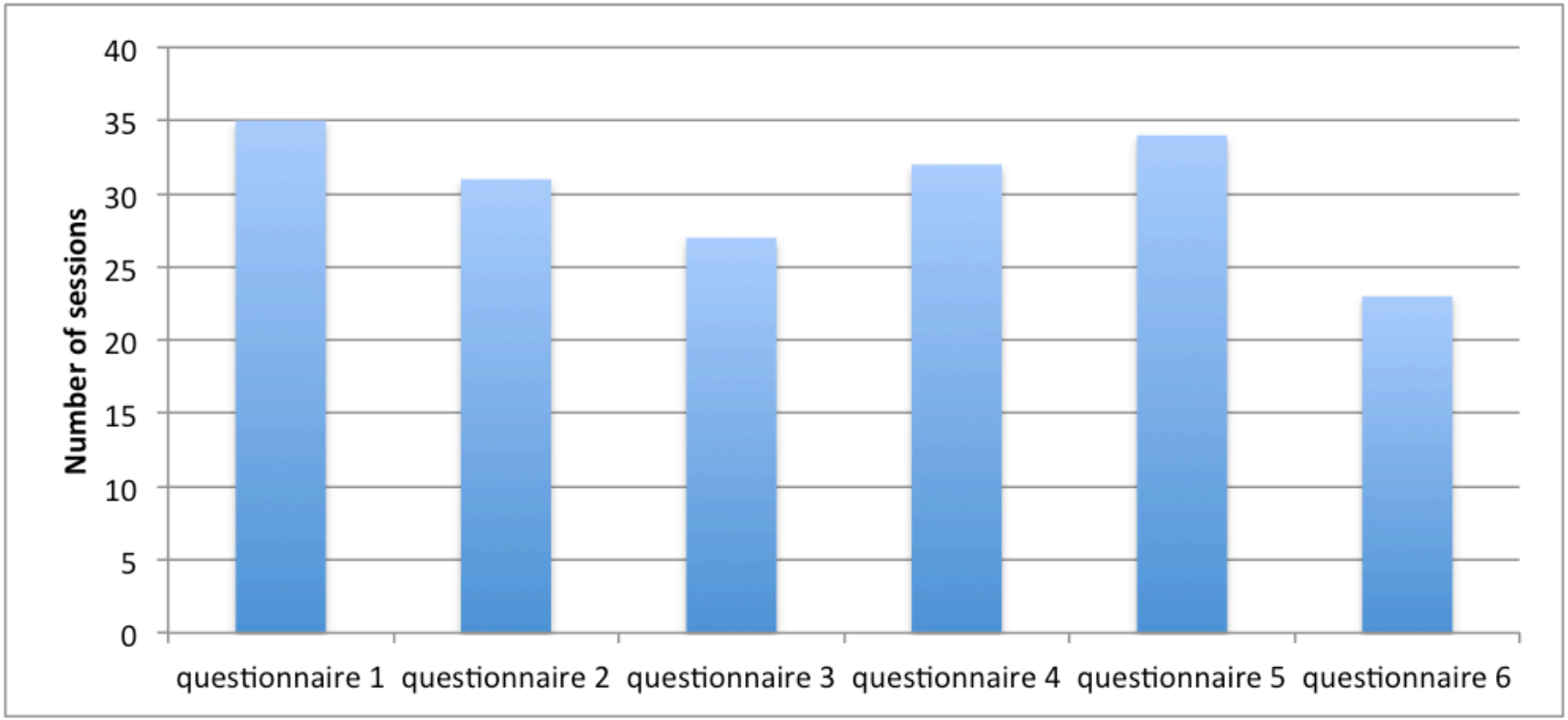
Over sessions:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

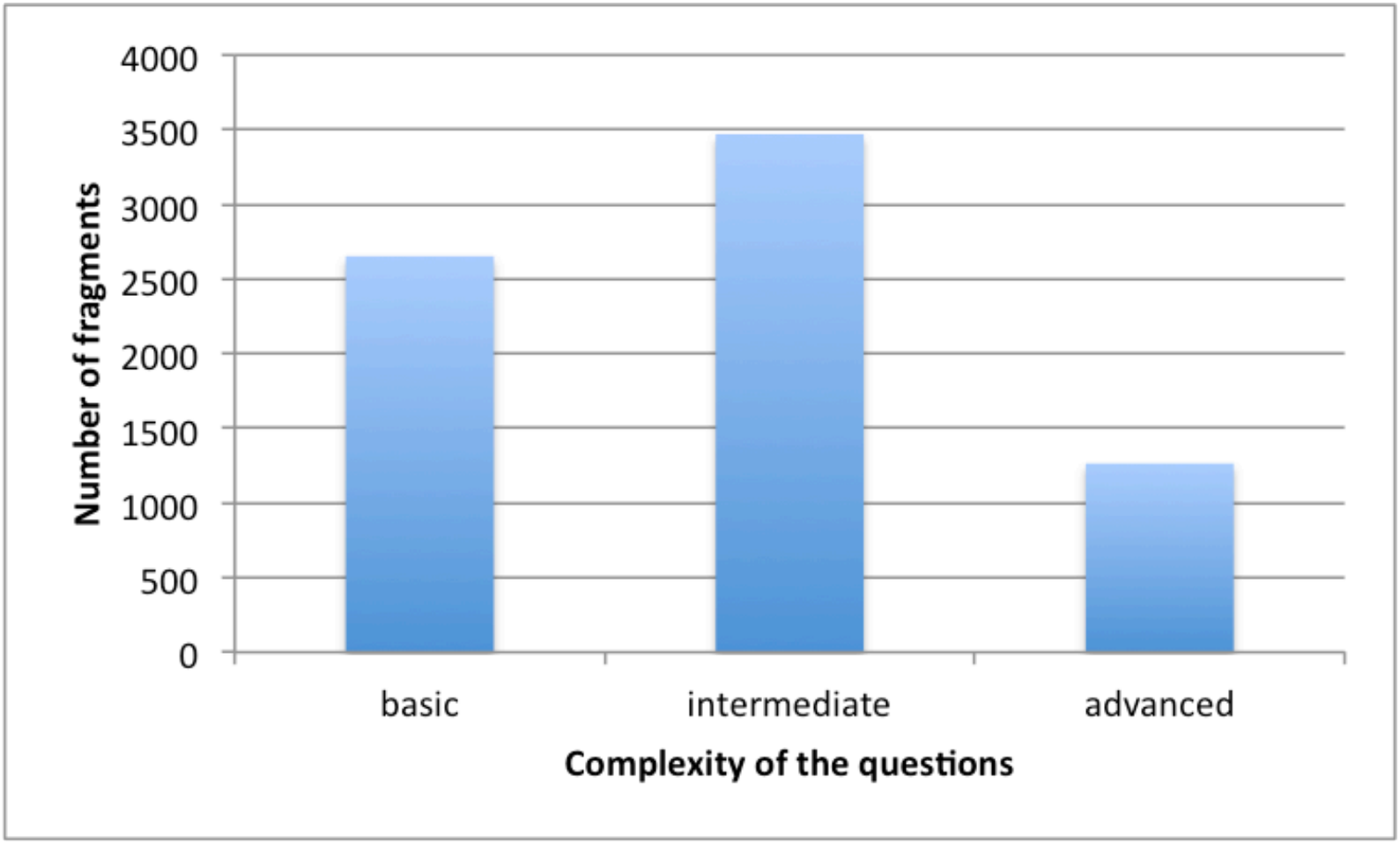
Over sessions:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

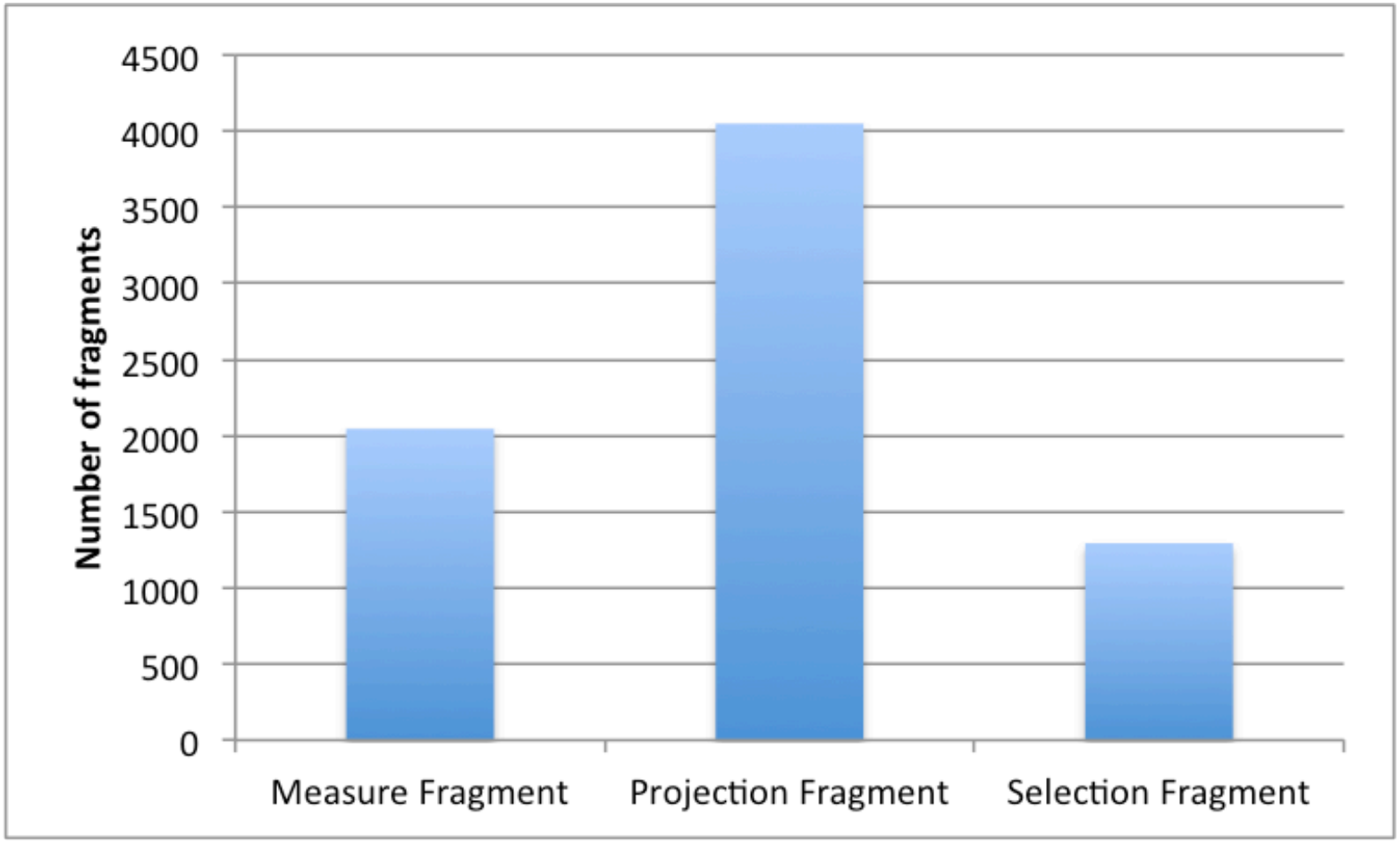
Over fragments:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

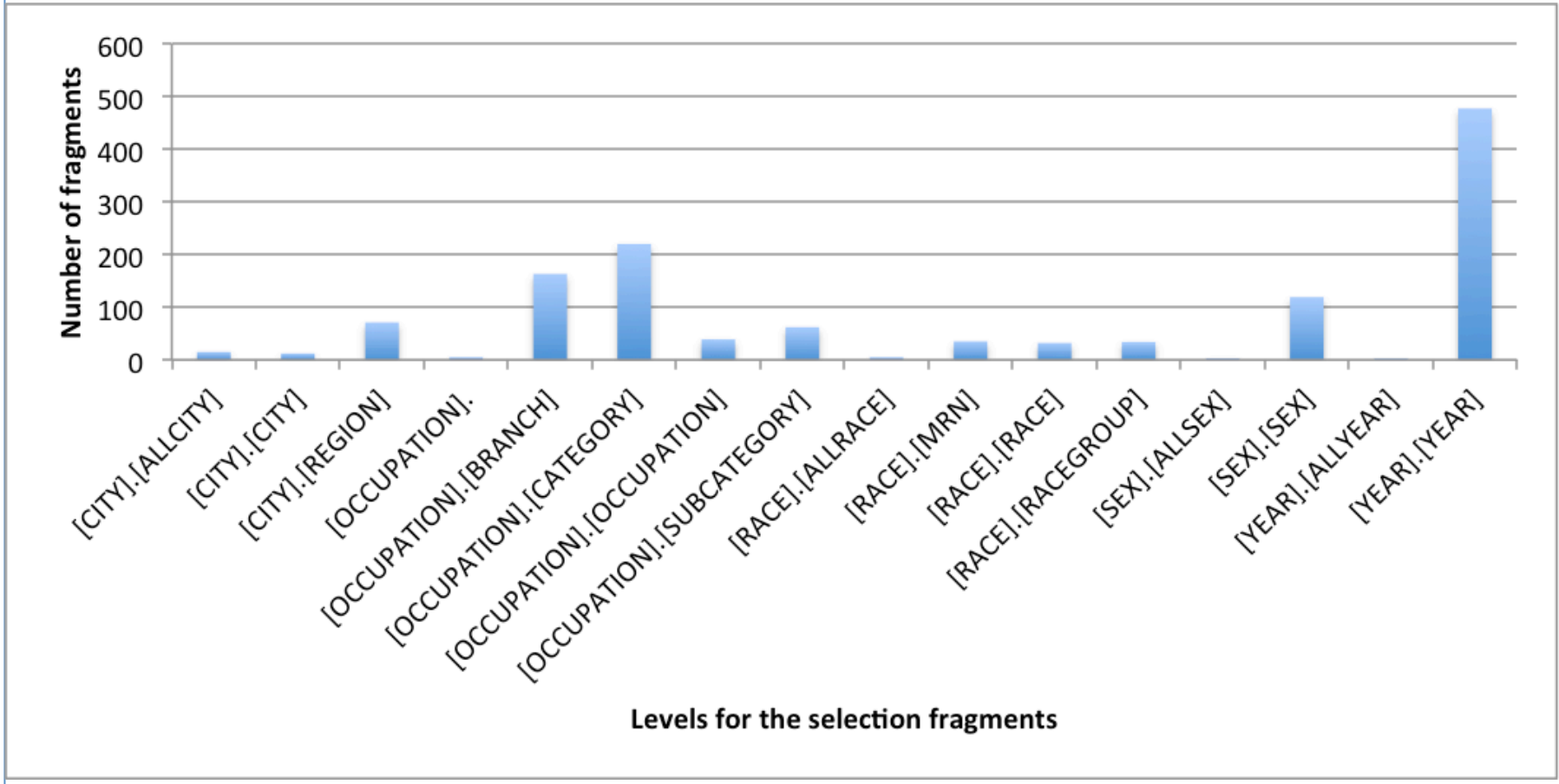
Over fragments:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

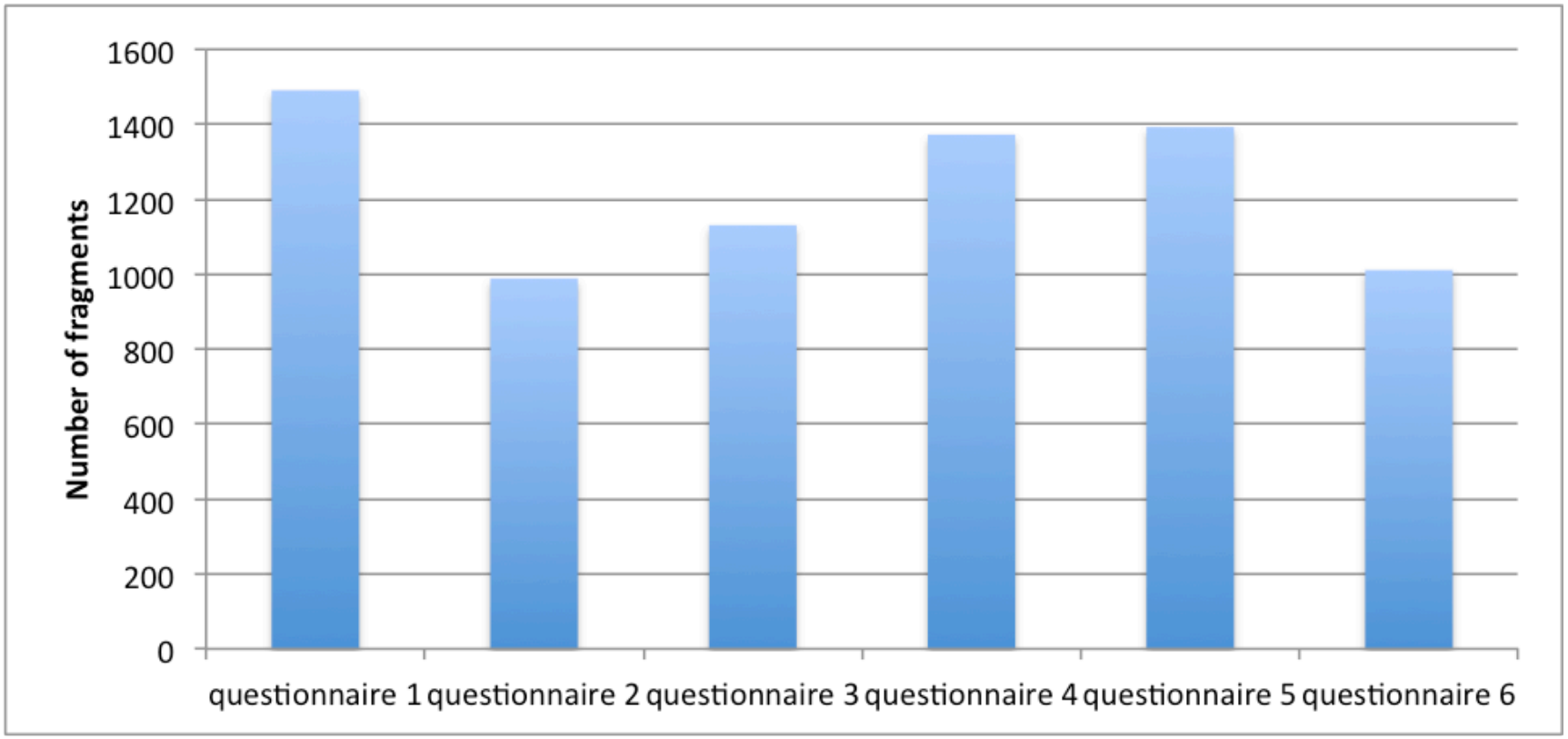
Over fragments:



Gathering Real OLAP Analysis Sessions: A Feedback

Statistical Results

Over fragments:



Gathering Real OLAP Analysis Sessions: A Feedback

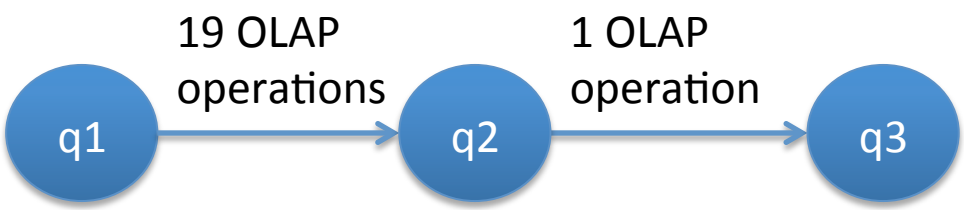
Statistical Results

Strange behaviors:

25 sessions (14%) less than 3 queries

<{Sex, Year, AllCity, AllRace, AllOccupation},
{Sex={'Female', 'Male'}},
{INCTOT}>

128 queries (15%)



Conclusion & Discussion

Gathering Real OLAP Analysis Sessions: A Feedback

Conclusion

- sessions are workable (paying attention with the advanced questions)
- filtering the logs → unavoidable

Gathering Real OLAP Analysis Sessions: A Feedback

- **easy** to identify criteria for identifying **no relevant** sessions
- **difficult** to identify criteria for identifying **relevant** sessions
  - pattern (profile) of sessions for each question ?
  - identifying navigational behavior? (***[Sapia in Dawak 2000]***)
- Long term perspective: a benchmark of OLAP sessions
  - Definition of OLAP session ?
  - Metrics for measuring the quality of sessions (like in exploratory search)

EDA'2013

Thanks for your attention !