Performance Evaluation and Experimental Assessment

Conscience or Curse of Database Research?

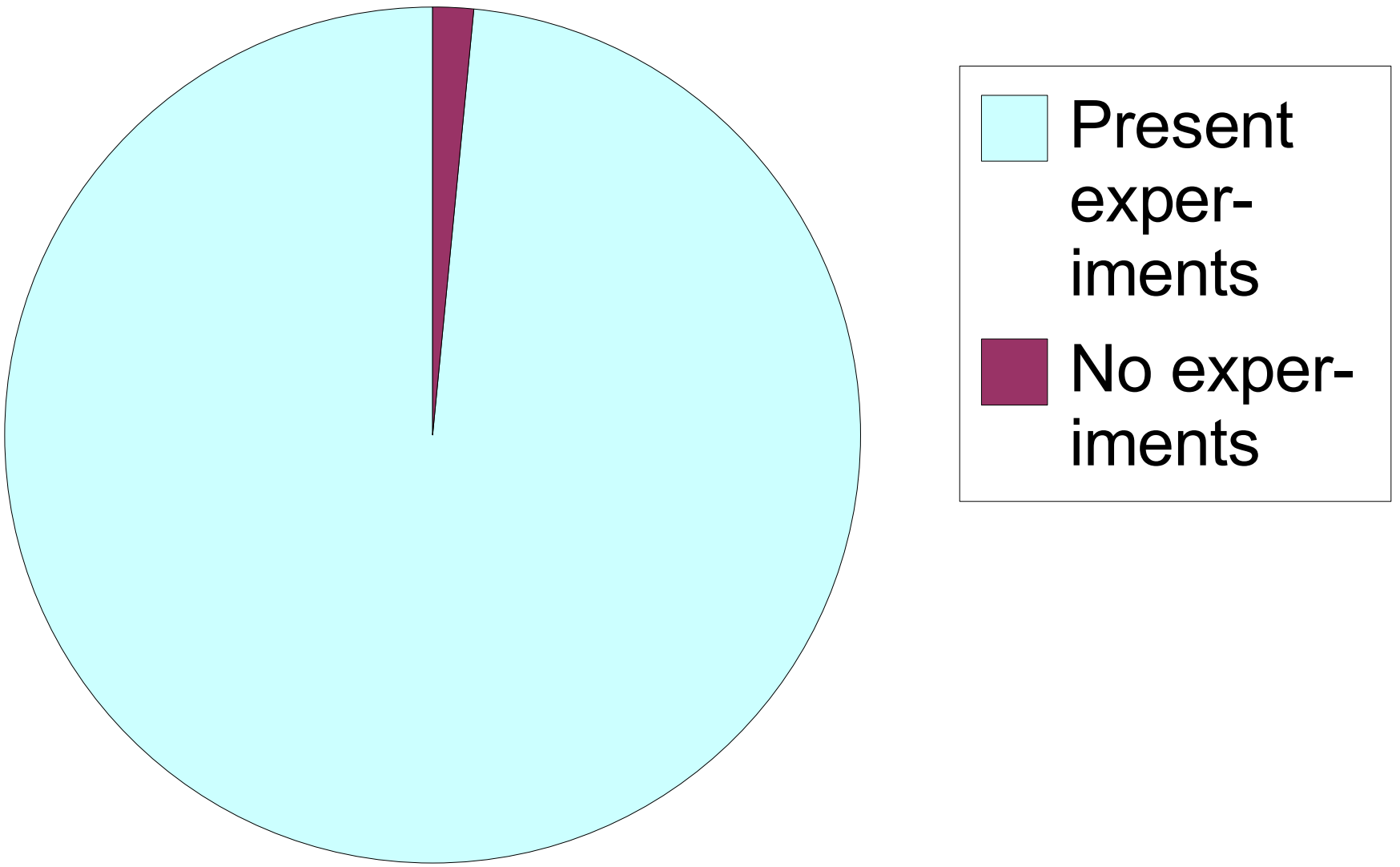
Panelists: **Torsten Grust** (Technische Universität München)
Martin Kersten (CWI, Amsterdam)
Paul Larson (Microsoft Research)
Guido Moerkotte (Universität Mannheim)
Yannis Papakonstantinou (UCSD)

Moderators: **Ioana Manolescu** (INRIA Futurs)
Stefan Manegold (CWI, Amsterdam)

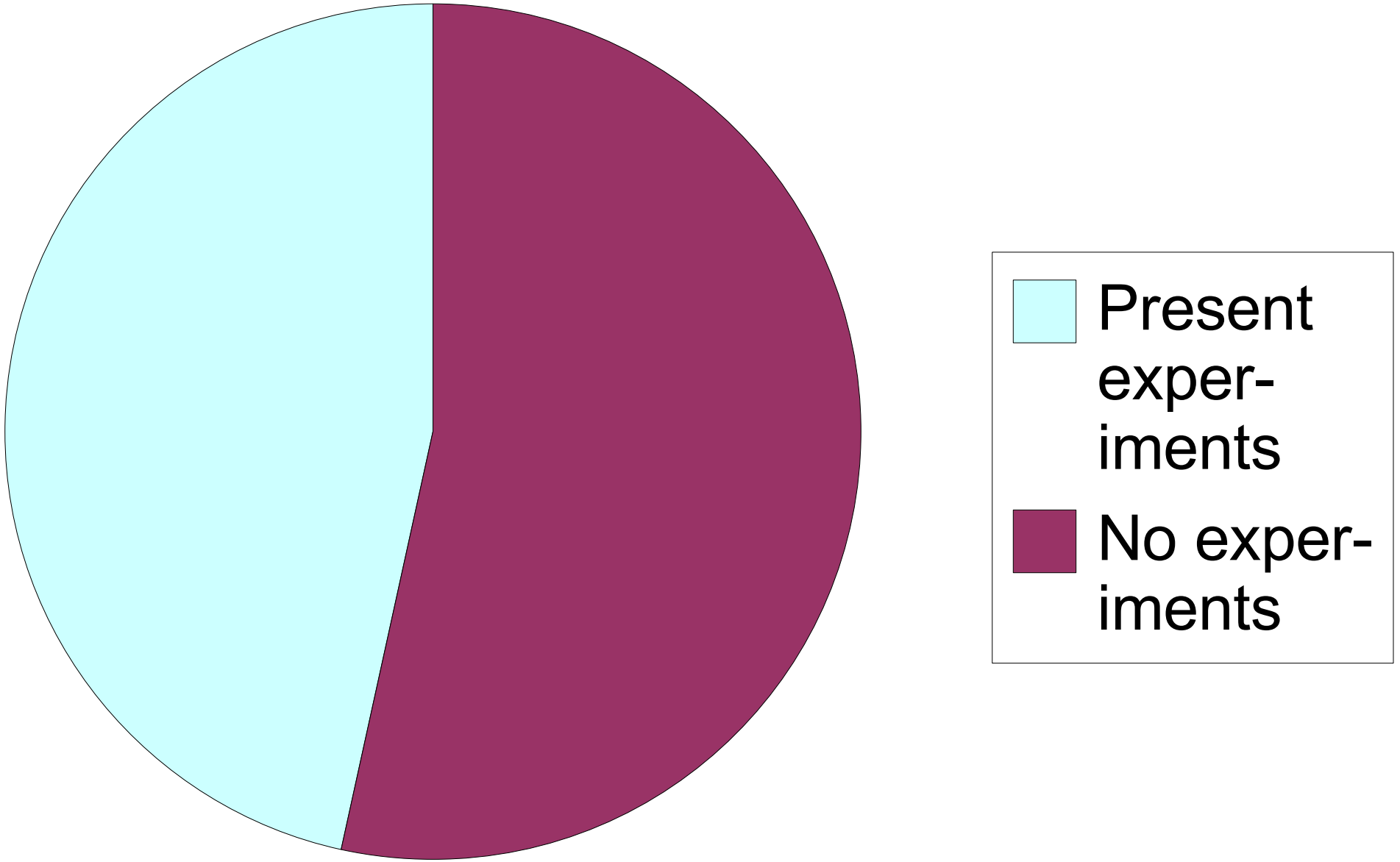
Experimental evaluation in SIGMOD 2007 accepted papers: some statistics

Ioana Manolescu (INRIA)
with help from Denilson Barbosa (U. Calgary)

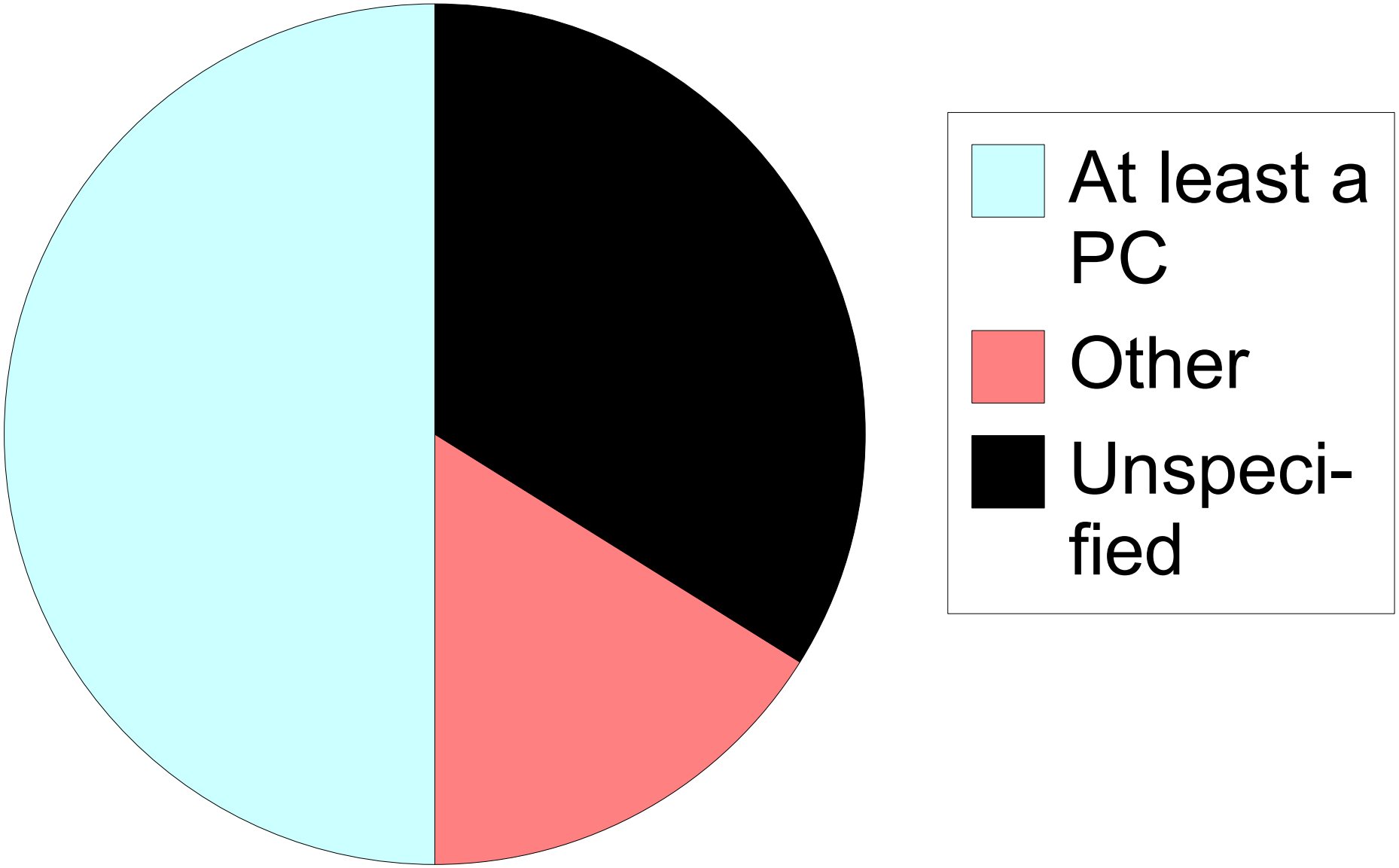
SIGMOD 2007 research papers (total: 68)



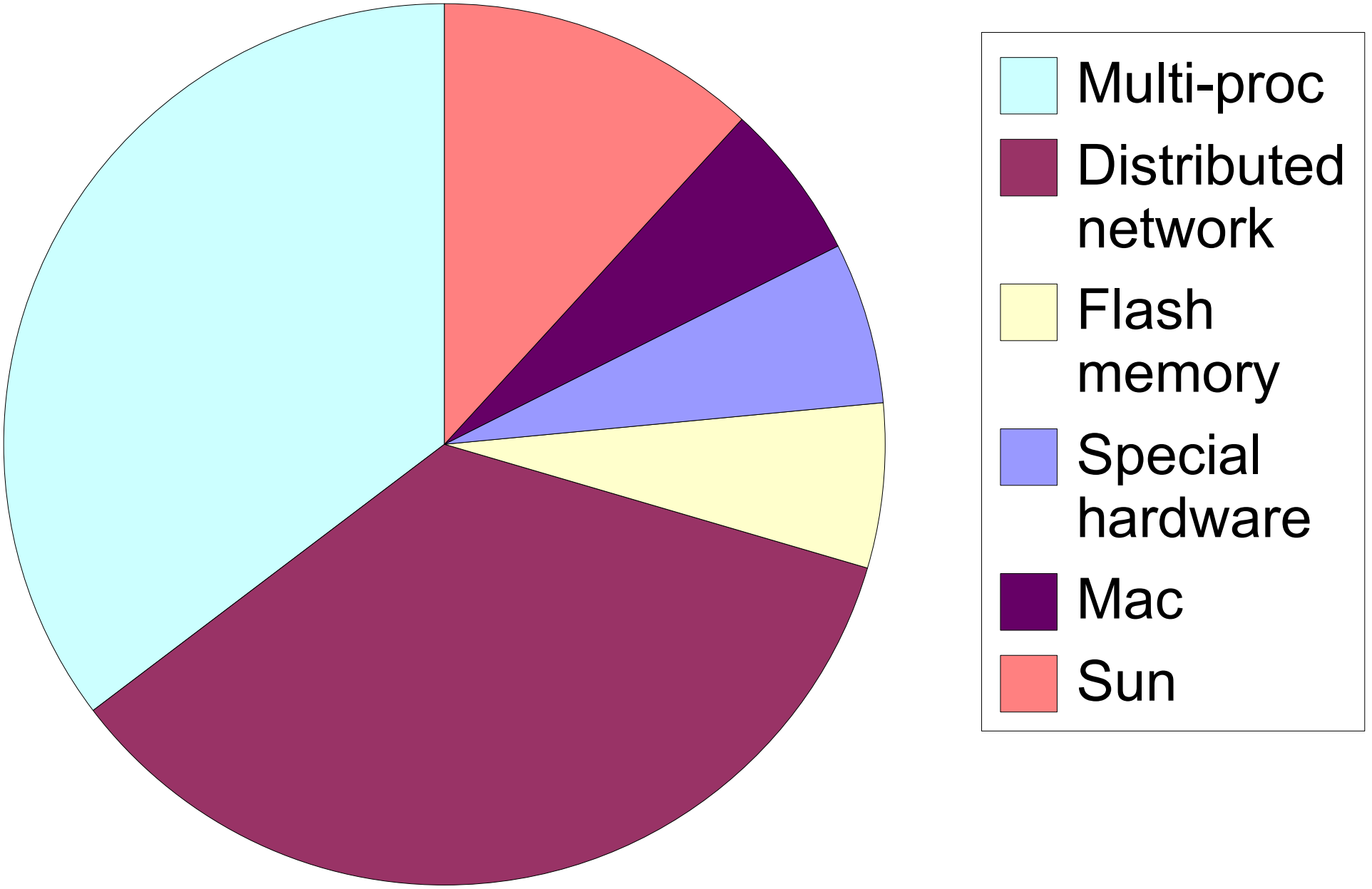
SIGMOD 2007 industrial papers (total: 15)



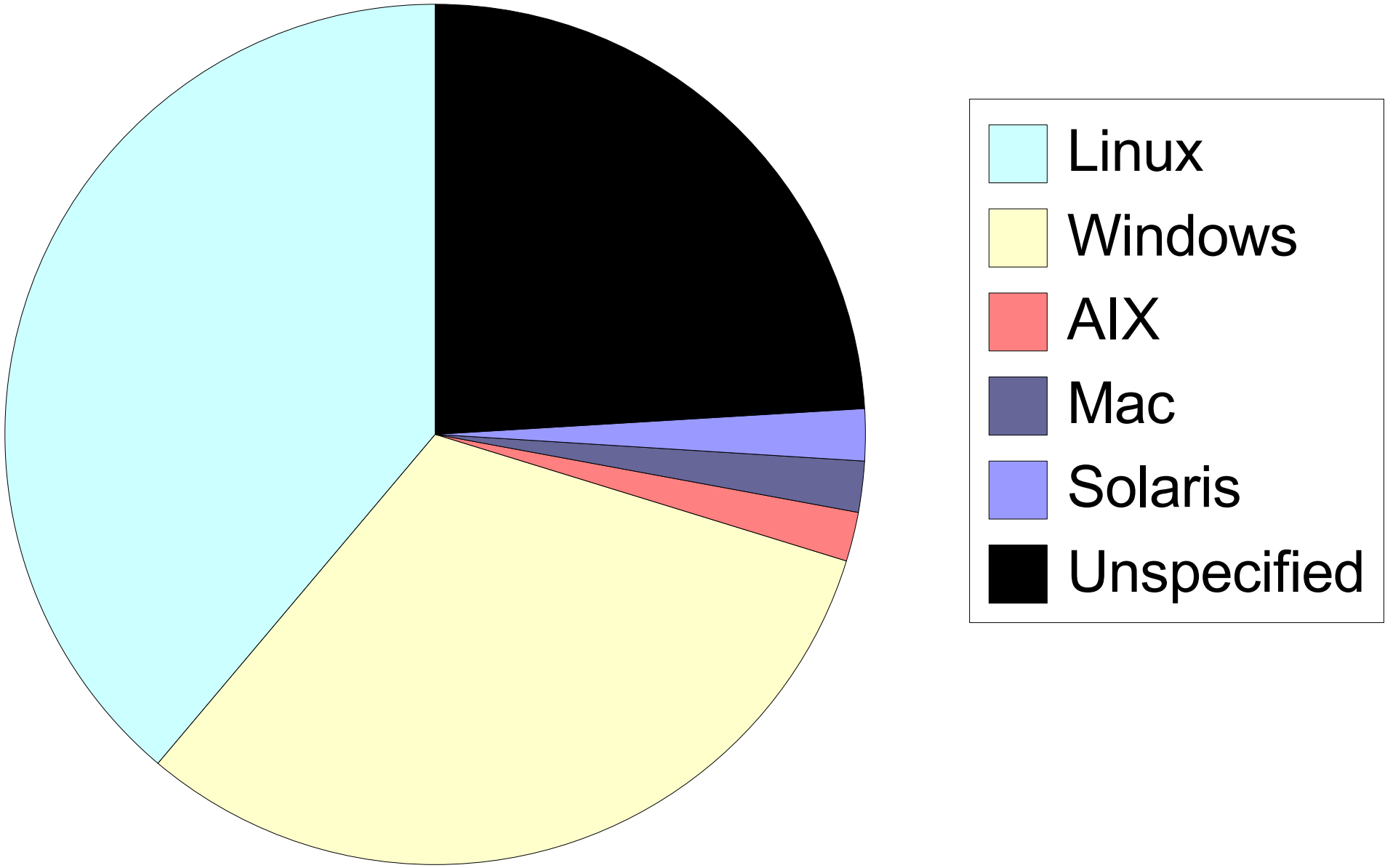
Hardware used in experiments



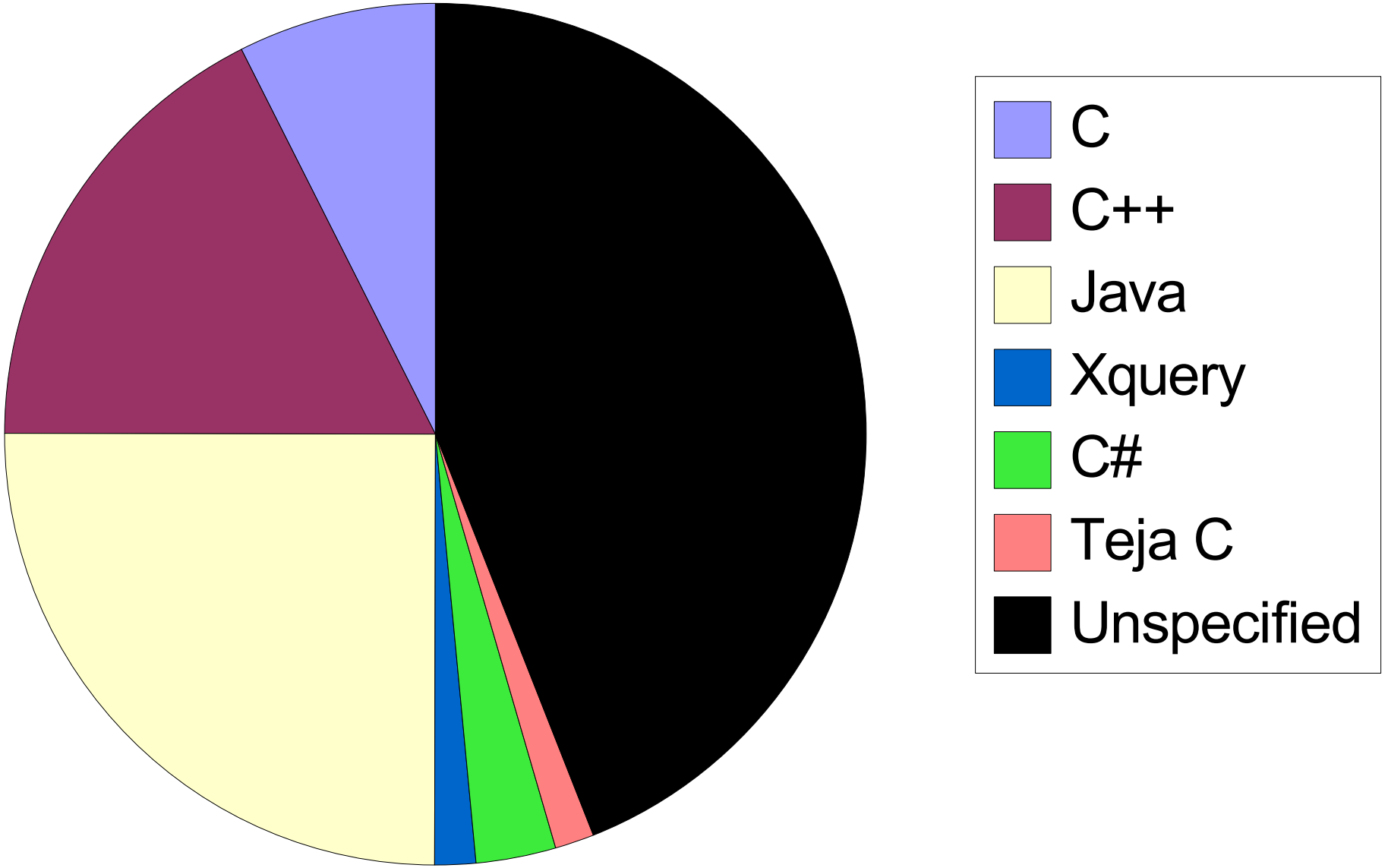
Hardware beyond 1 PC (total: 17)



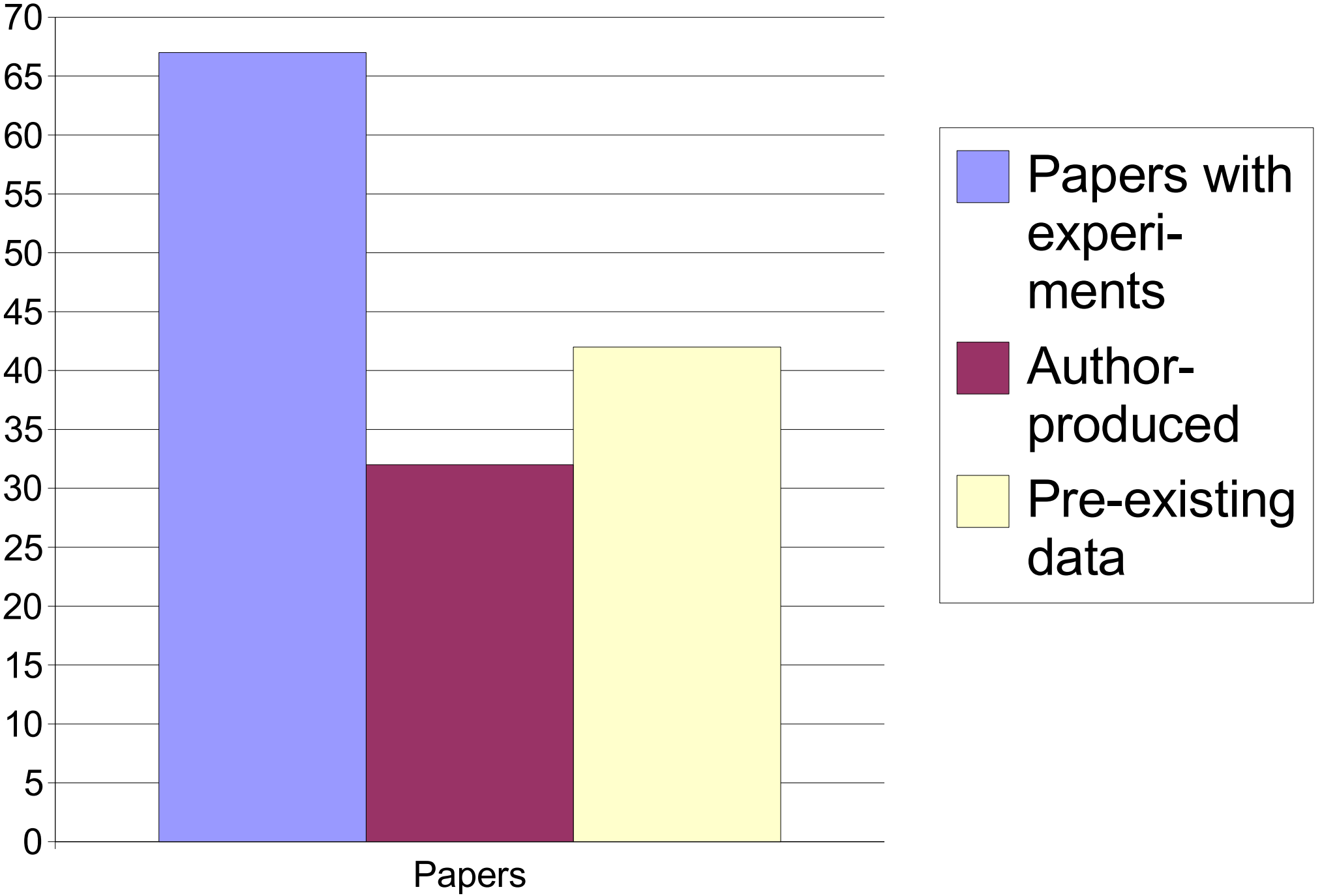
OS used in experiments



Programming language



Data sources used in experiments



Anecdote

"Note that we were not able to compare timing results directly with X since a working executable/code is not directly available.[1]

[1]Personal communication with the authors."

Experimental Assessments in Research Papers Today

A Little Shop of Horrors

Torsten Grust
Technische Universität München

http://www-db.in.tum.de/~grust/

“Let them figure out the correct syntax...”

combine aggregation-based, structure-based and value-based predicates by the logical operators, `not`, `or` and `and`, which allows more expressive queries. For example, in the following query, “`/descendant::a[[child::b = "B"] and [descendant::c]]`

“Let them figure out the correct syntax...”

combine aggregation-based, structure-based and value-based predicates by the logical operators, `not`, `or` and `and`, which allows more expressive queries. For example, in the following query, “`/descendant::a[[child::b = "B"] and [descendant::c]]`

- Apparent lack of language knowledge / care doesn’t help your case.

“Let them figure out the correct syntax...”

combine aggregation-based, structure-based and value-based predicates by the logical operators, `not`, `or` and `and`, which allows more expressive queries. For example, in the following query, “`/descendant::a[[child::b = "B"] and [descendant::c]]`

- Apparent lack of language knowledge / care doesn’t help your case.
- Show love respect for your object of study.

You *can* typeset {} in LaTeX

```
<RECORDLIST>
for $play in
        document("dxv.xml")/PLAY/ROW
Order by $play/POSITION/text()
return
 <PLAY>
  <BAND/>$play/BAND_PCDATA/text(),
  for $song in
       document("dxv.xml")/SONG/ROW
      [PID/text() = $play/IID/text()]
  order by $song/POSITION/text()
  return
   <SONG>
    $song/SONG_PCDATA/text()
   </SONG>
 </PLAY>
</RECORDLIST>
```

You *can* typeset {} in LaTeX

```
<RECORDLIST>
for $play in
      document("dxv.xml")/PLAY/ROW
Order by $play/POSITION/text()
return
 <PLAY>
  <BAND/>$play/BAND_PCDATA/text(),
  for $song in
     document("dxv.xml")/SONG/ROW
    [PID/text() = $play/IID/text()]
  order by $song/POSITION/text()
  return
   <SONG>
    $song/SONG_PCDATA/text()
   </SONG>
 </PLAY>
</RECORDLIST>
```

- Missing `{}` in node constructors,
- `document(…)`?,
- miXeD cASe KEyWorDS,
- empty `<BAND/>` tag suspicious....

You *can* typeset {} in LaTeX

```
<RECORDLIST>
for $play in
      document("dxv.xml")/PLAY/ROW
Order by $play/POSITION/text()
return
 <PLAY>
  <BAND/>$play/BAND_PCDATA/text(),
  for $song in
     document("dxv.xml")/SONG/ROW
    [PID/text() = $play/IID/text()]
  order by $song/POSITION/text()
  return
   <SONG>
    $song/SONG_PCDATA/text()
   </SONG>
 </PLAY>
</RECORDLIST>
```

- Missing `{}` in node constructors,
- `document(…)`?,
- miXeD cASe KEyWorDS,
- empty `<BAND/>` tag suspicious....
- This is **the running example** in this paper.

Beyond Syntax …

$XQuery_2$:

```
FOR $b in //B,
    $d in $b//D
LET $c := $b//C
RETURN $b, $d, $c
```

- Variables `$b`, `$c` not bound in `return` clause.
- Have you ever run this through *any* language processor?

Be Inventive *Before* Entering the Experimental Section

```
for $p in (bib.xml)/paper,
    $t=$p/title,
    $y=$p/year,
    $c=$p/confer,
    $a=$p/authors/author,
    $f=$a/first_name,
    $l=$a/last_name
where $t/text()="XML" and $c/text()=      and
      $y/text()="2007"
return <author> {$f} {$l} </author>
```

Be Inventive *Before* Entering the Experimental Section

```
for $p in (bib.xml)/paper,
    $t=$p/title,
    $y=$p/year,
    $c=$p/confer,
    $a=$p/authors/author,
    $f=$a/first_name,
    $l=$a/last_name
where $t/text()="XML" and $c/text()=        and
      $y/text()="2007"
return <author> {$f} {$l} </author>
```

- Does "=" mean XQuery's `for` or `let` here?
- You never ran this. What **did** you run then?

Be Inventive *Before* Entering the Experimental Section

```
for $p in (bib.xml)/paper,
    $t=$p/title,
    $y=$p/year,
    $c=$p/confer,
    $a=$p/authors/author,
    $f=$a/first_name,
    $l=$a/last_name
where $t/text()="XML" and $c/text()=        and
      $y/text()="2007"
return <author> {$f} {$l} </author>
```

- Does "=" mean XQuery's `for` or `let` here?
- You never ran this. What **did** you run then?
- You included performance numbers, but you measured **something else**.

Brevity is a Virtue!

7 Experimental Evaluations

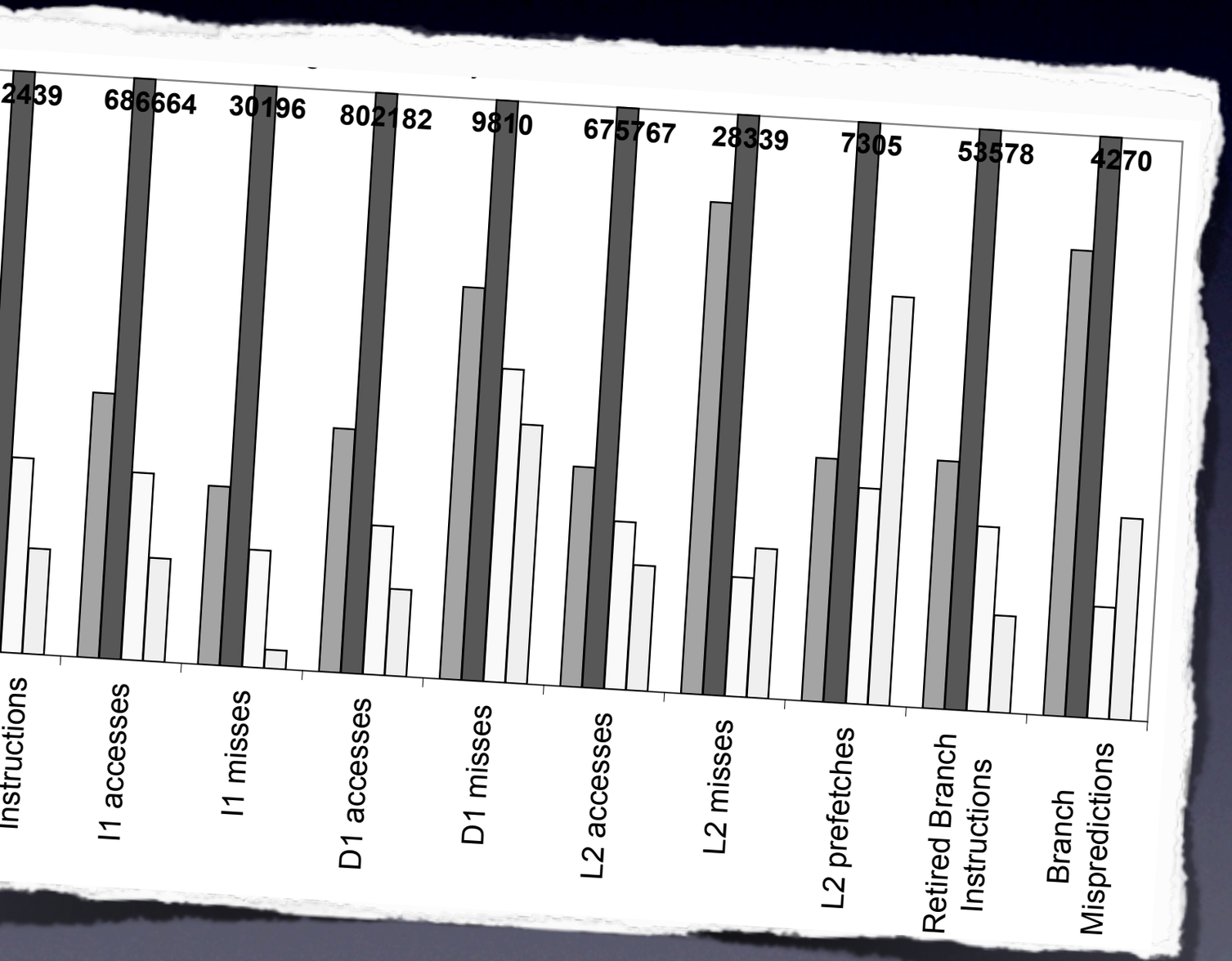
In this section, we present an experimental study to verify the effectiveness of our proposed techniques. All experiments were run on a machine with 3.4GHZ. The experiments were run in warm memory. The proposed techniques were implemented in C++. The synthetic graphs were gen-

Brevity is a Virtue!

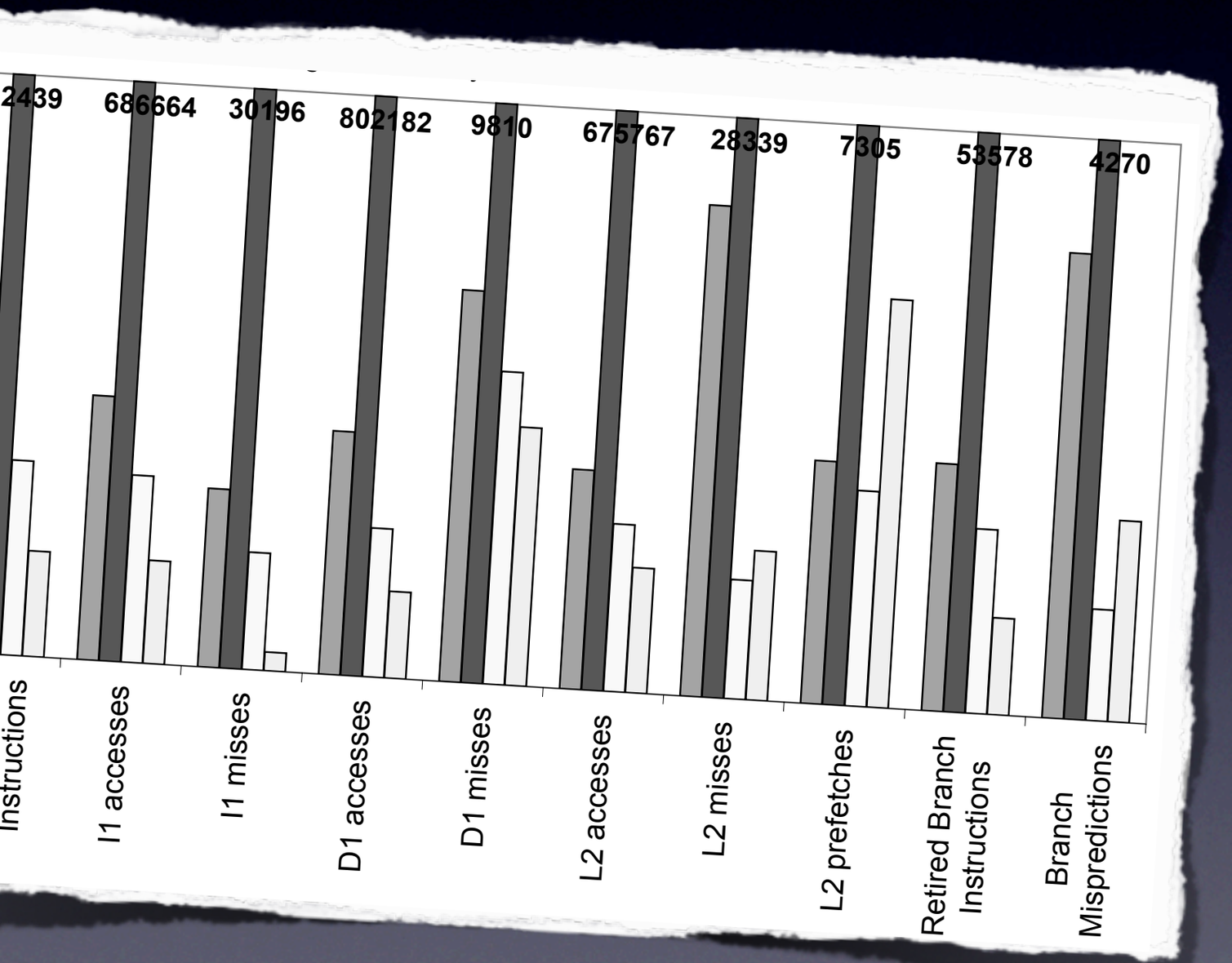
7 Experimental Evaluations

In this section, we present an experimental study to verify the effectiveness of our proposed techniques. All experiments were run on a machine with 3.4GHZ. The experiments were run in warm memory. The proposed techniques were implemented in C++. The synthetic graphs were gen-

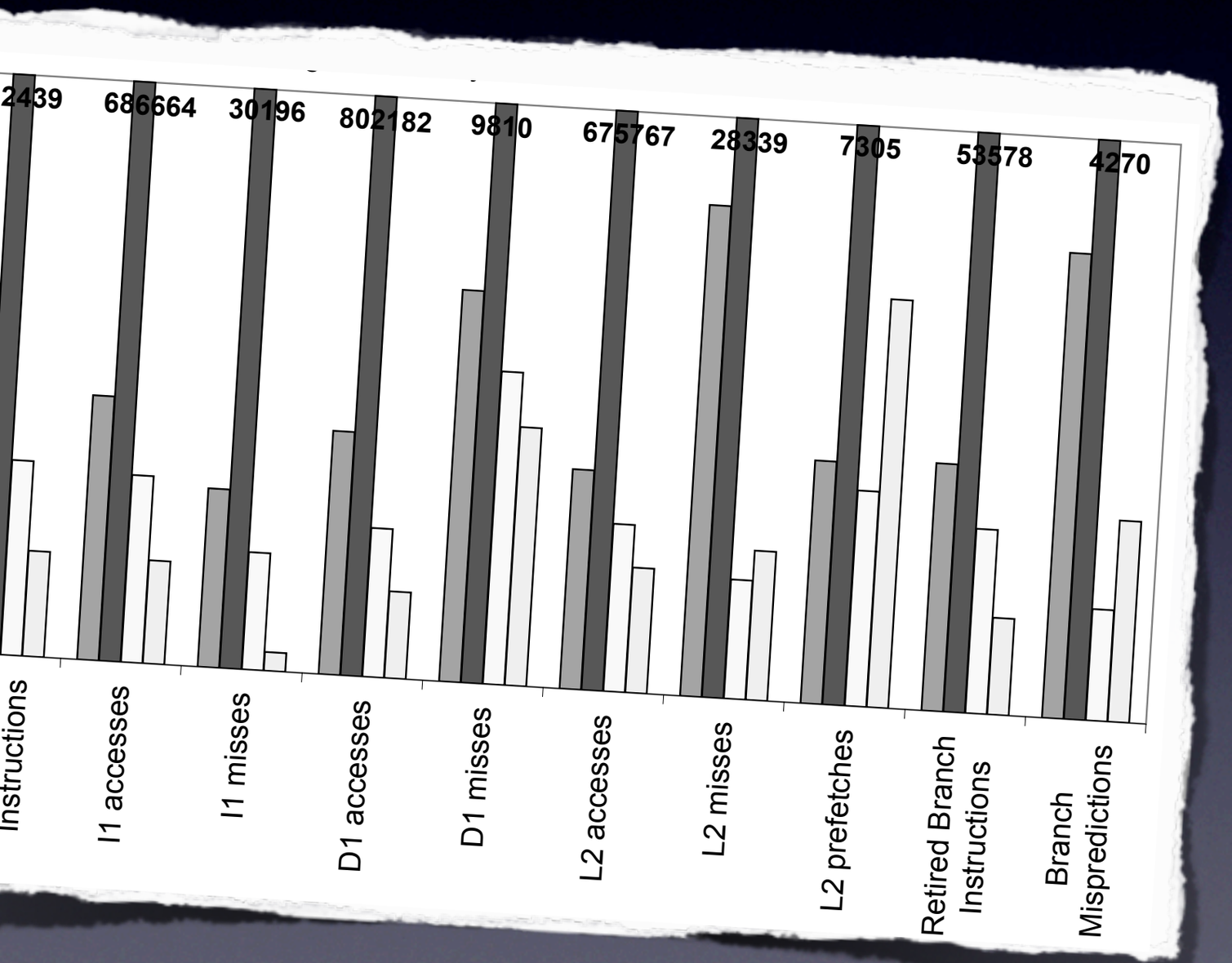
- "We studied it — but you will never be able to experience it yourselves."


2439
686664
30196
802182
9810
675767
28339
7305
53578
4270
Instructions
I1 accesses
I1 misses
D1 accesses
D1 misses
L2 accesses
L2 misses
L2 prefetches
Retired Branch Instructions
Branch Mispredictions

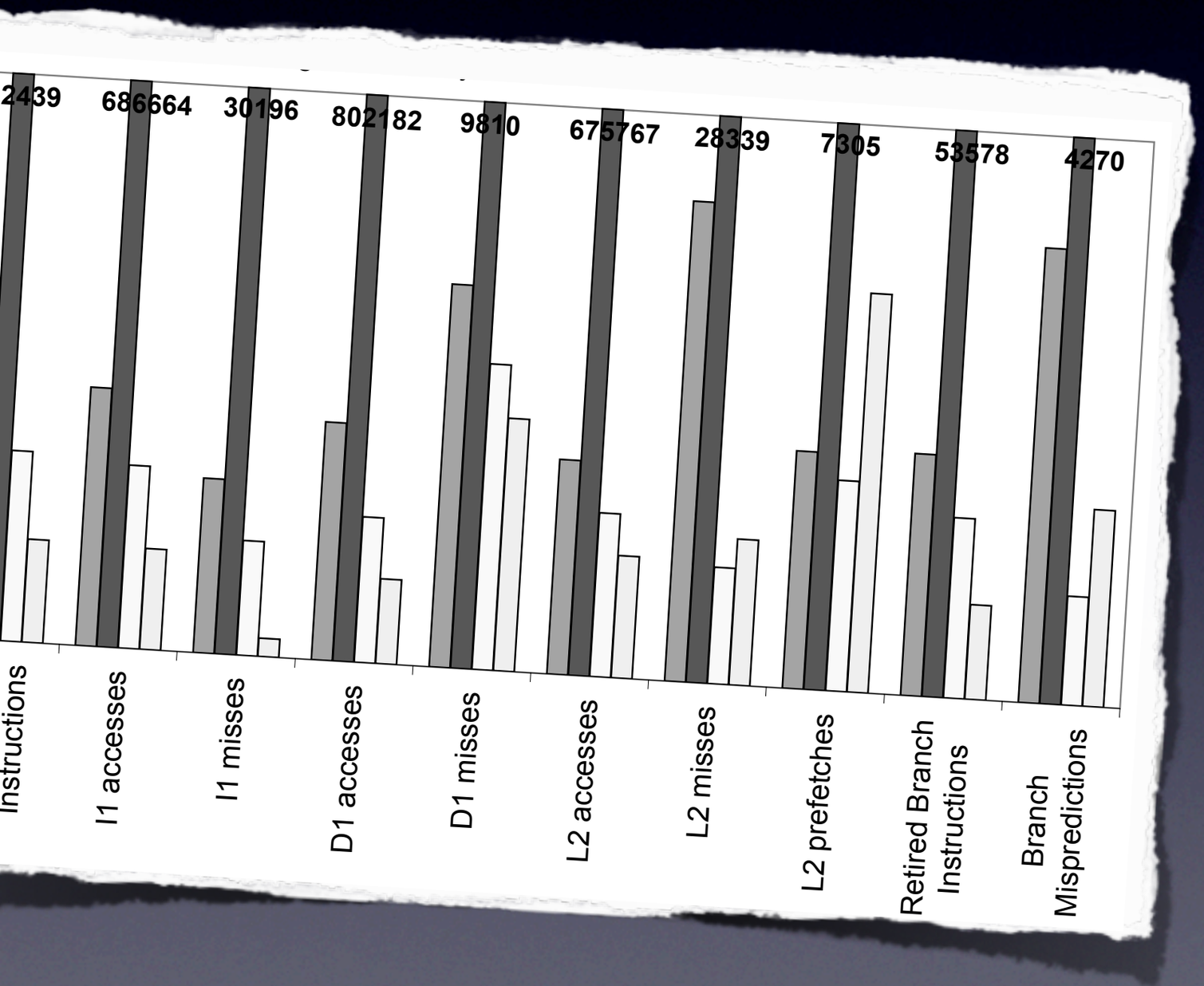
“We consistently outperform



“We consistently outperform … this hopeless case.”

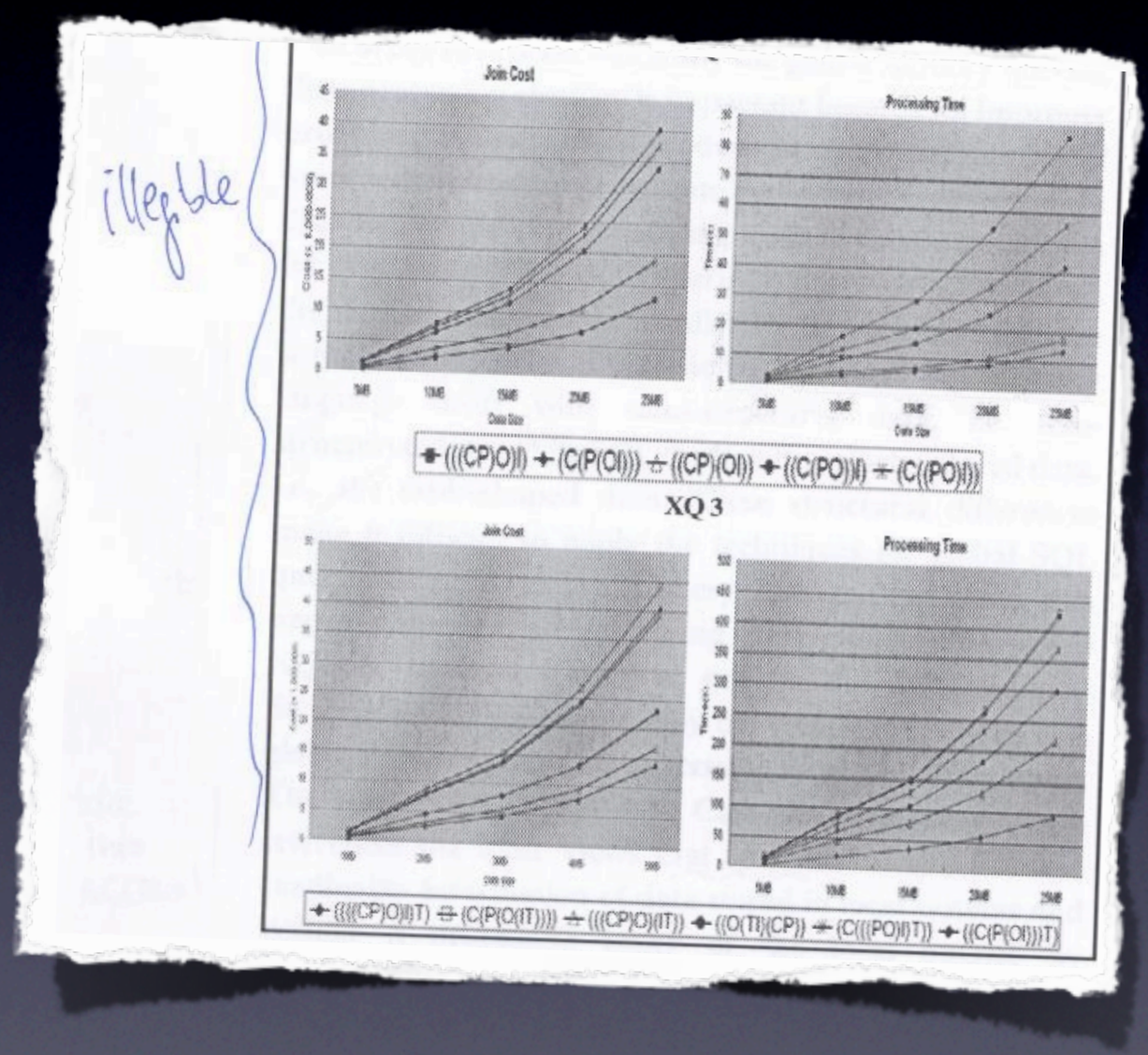


“We consistently outperform … this hopeless case.”

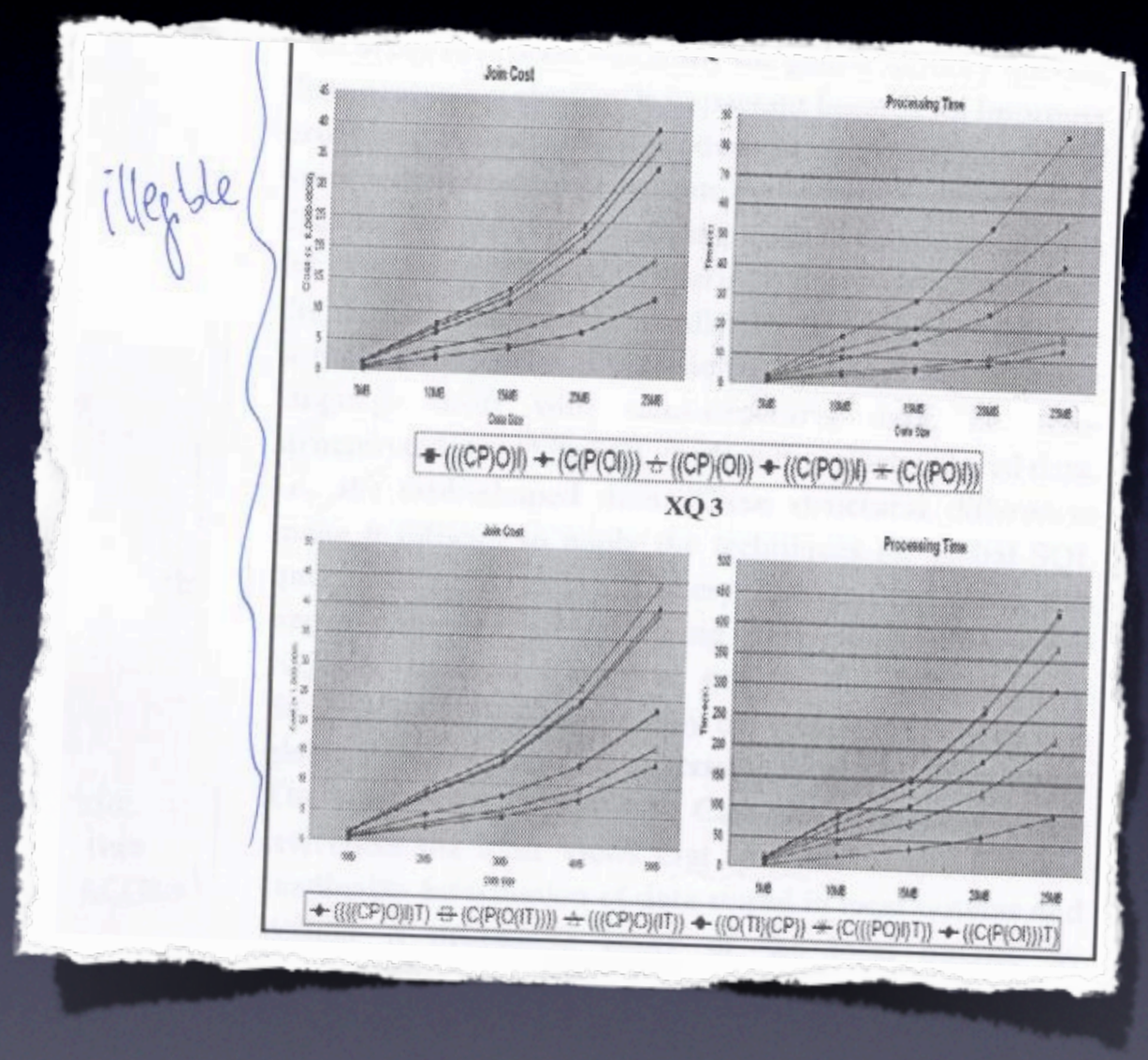


- Compare against the **real** competition.
- Makes for a more interesting analysis, too.

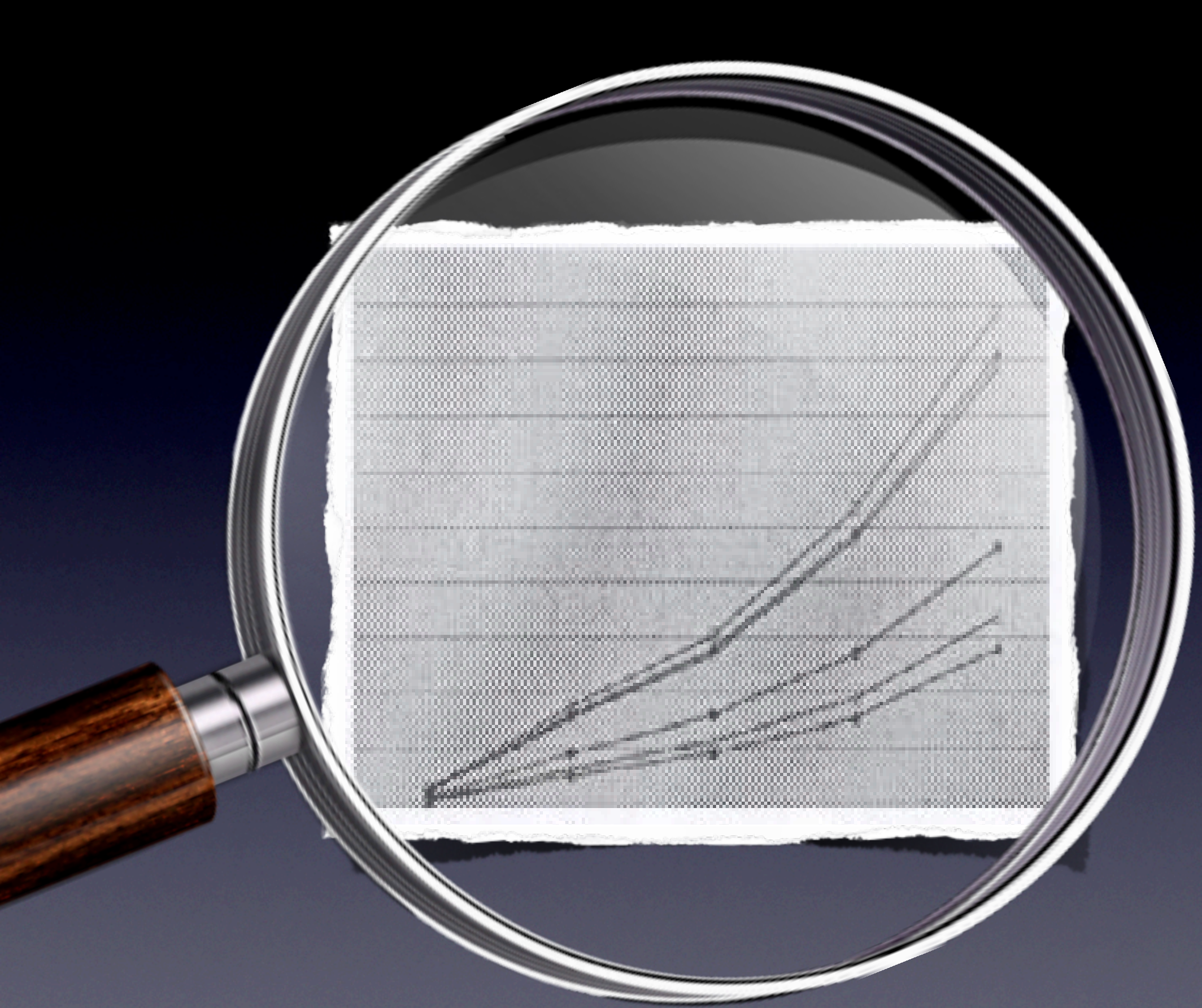
“If it was hard to measure, why should it be easy to read?”



“If it was hard to measure, why should it be easy to read?”



- These graphs contain the **core message** of the work.



Bedtime stories: Experimental validation

Martin Kersten

Experimental validation

- What are the requirements for a credible experimental assessment ?
- Are there classes of papers that do not need experimental validation?

Experiment Metrics

- Platform accessibility
  1. Off-the-shelf
  2. Accessible to scientist
  3. For rich only
- Software accessibility
  1. Open-source,
  2. Built your self
  3. Proprietary
- Parameter space
  1. Space exploration
  2. Public points
  3. Private point
- Address a desire
  1. Real-life
  2. Simulation
  3. Theory
- Metric Monsters
  1. Colleagues
  2. Compiler
  3. Clock

Experiment Metrics

- Platform accessibility
  1. Off-the-shelf
  2. Accessible to scientist
  3. **For rich only**
- Software accessibility
  1. Open-source,
  2. Built your self
  3. **Proprietary**
- Parameter space
  1. Space exploration
  2. Public points
  3. **Private point**
- Address a desire
  1. Real-life
  2. Simulation
  3. **Theory**
- Metric Monsters
  1. Colleagues
  2. Compiler
  3. **Clock**

Experiment Metrics

- Platform accessibility
  1. Off-the-shelf
  2. Accessible to scientist
  3. For rich only

- Software accessibi
  1. Open-sourc
  2. Built v
  3. P

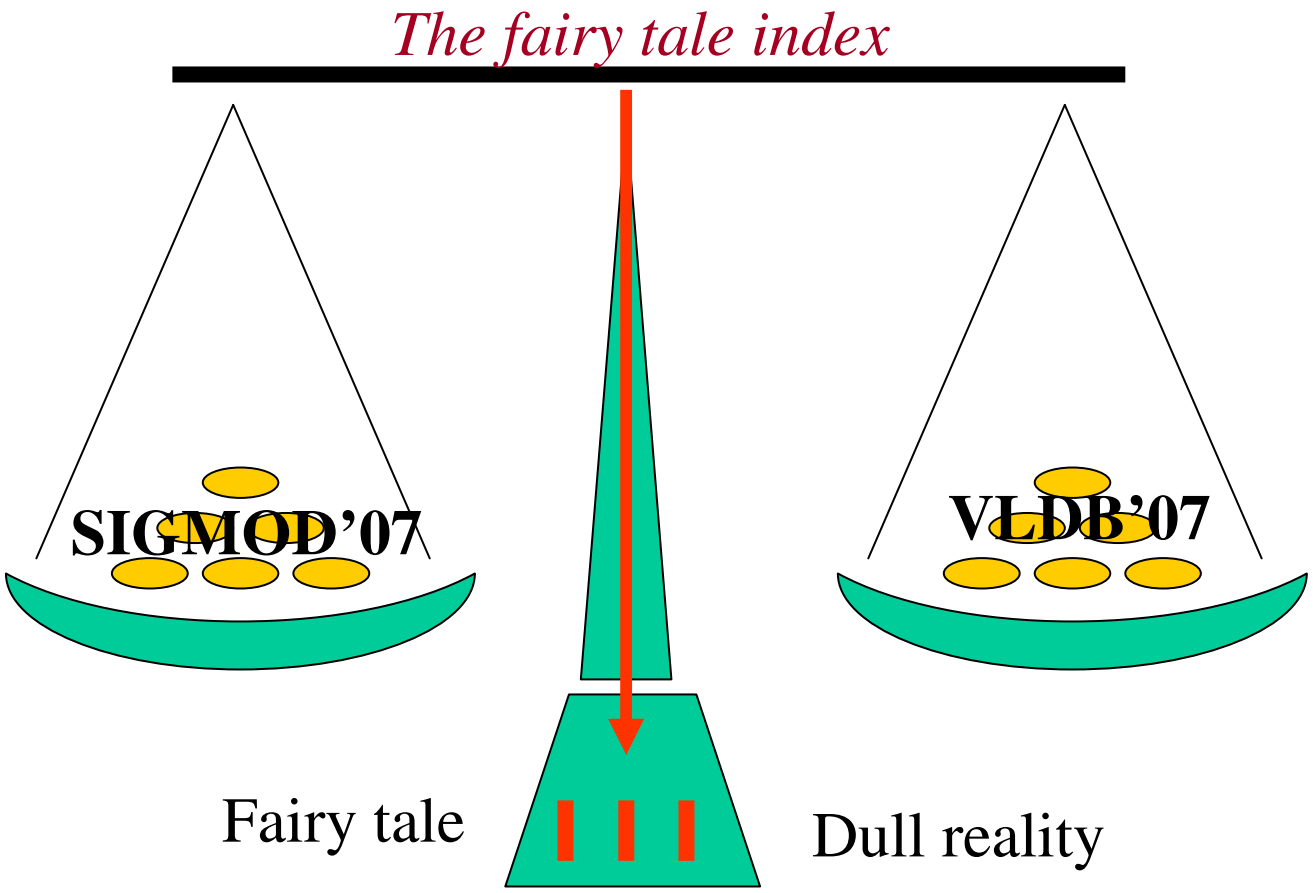
- Address a
  1. Re

ic Monsters
1. Colleagues
2. Compiler
3. Clock

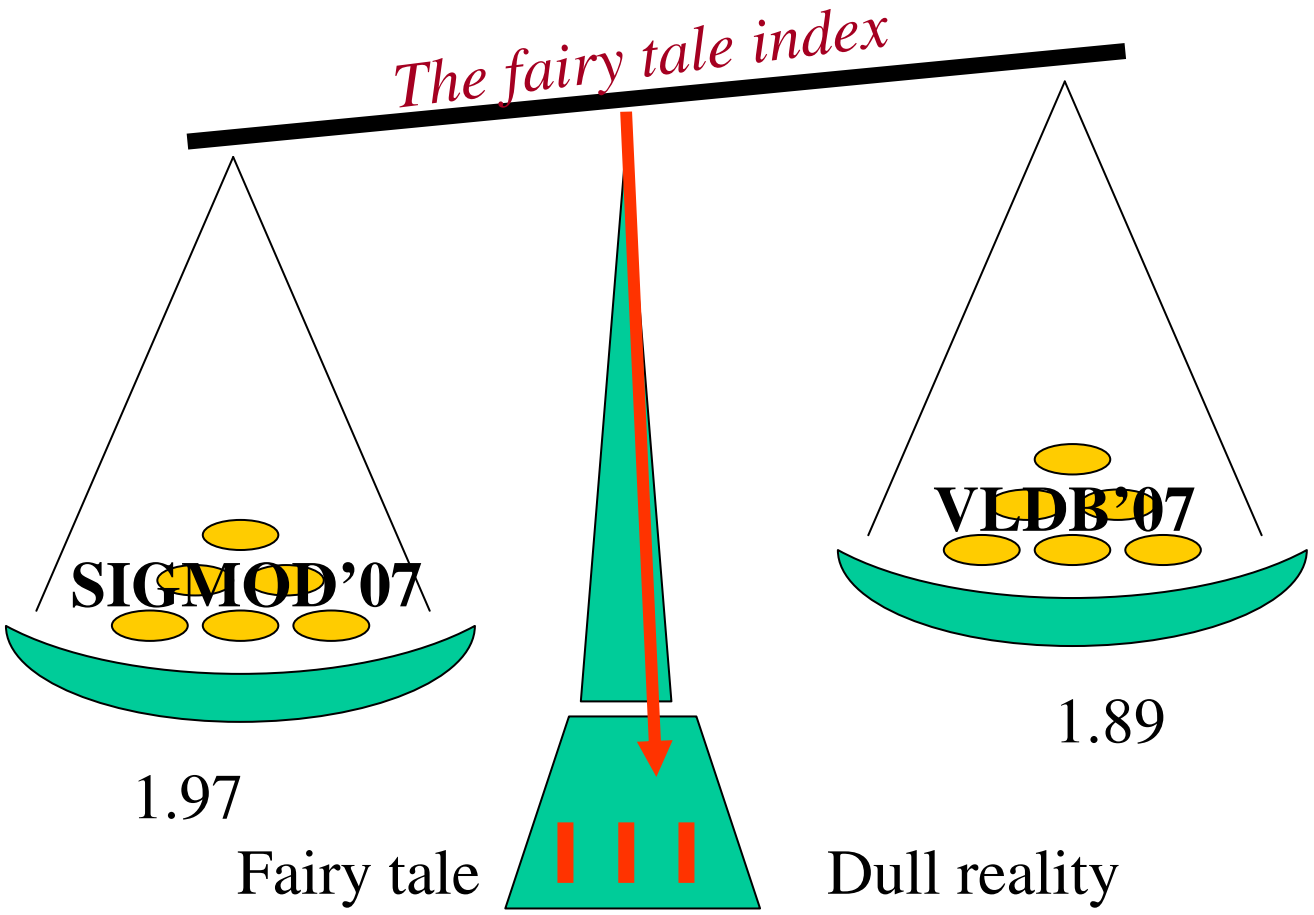
pace
exploration
ublic points
3. Private point

The Fairy Tale Index

The Verdict



The Verdict



Performance Evaluation and Experimental Assessment

Paul Larson
Microsoft Research

Typical experimental evaluations of limited value

- Database systems used for lots of different purposes
  - Different databases, workloads, hardware
- Easy to find a case where your idea improves performance by X %
- Lots of work to find out
  - Does the improvement hold up in different contexts?
  - How does it interact with other features?
  - What's the effects on other quality measures?
- Solid experimental performance evaluation is difficult and takes a lot of work

Benchmarks and performance comparisons

- Good benchmarks are hard to design but very useful
- Thorough experimental evaluations and comparisons are extremely valuable
- So why do we have so few papers on new benchmarks or comparing performance?
  - few submitted or few accepted?

Sigmod experimental repeatability requirement

- Experiments verified by a committee
- Submit code and data sets
- Doubt we have a big problem with fraudulent results?
- Impractical - Lots of work for what benefit?
- Industrial labs not able to participate
  - Can't distribute code without license
  - Can't distribute experimental code

Performance Evaluation and Experimental Assessment

Guido Moerkotte

- Should experimental assessment and performance evaluation be considered part of research or rather part of engineering?
- Who cares.

- What are the requirments for a credible assessement?
- Answer doesn’t fit into 5 minutes.

- Are current experimental benchmarks up to the task?
- Not necessarily.

- Are there classes of papers that do not need experimental validation?
- Yes: PODS papers.
- Yes: those with time/space complexity analysis

- Are there other metrics than performance that could/should be assessed empirically?
- Yes, but not in databases.

- Would a requirement list or even template help to ensure standardized and complete representation?
- Yes, see TPC. But for universities this is too heavy.
- And: standardized benchmarks only exist for old problems.

- Is comparison with commercial systems possible?
- Yes, it reveals deficiencies and potentially proposes remedies.

- What are the minimal requirements on experimental validations?
- plausibility. [no cheating!]
- completeness: e.g.: index: time to load, query, update, query. plus space
- approximate reproducability.

- Should we modify the reviewing process to solicit more disclosure of data and code?
- Who is going to read the code anyway?

- Answers are only valid, if you don’t want a paper to be accepted.

Karl Popper, anonymity, the “12 pages” and repeatability

Let's hear it from the Viennese

Falsifiability is the demarcation
between science and non-science

Karl Popper

Thanks wikipedia!
I can now (pretend I have) read Popper's works.

The easy way to do it

- Detail/document the experimental procedure
  - Data set, algorithm
  - Archive; SIGMOD?
- or provide an (online) system

…and while talking about online systems…

Shameless (yet can be repeatable) Advertising Section

Check out **app2you.org:** Create custom, interactive, database-driven web applications in minutes!

for classroom management, graduate admissions, hiring, event planning, and all sorts of collaborative processes you need

The easy way became hard by conference paper regulations

- “12 pages” do not fit all
  - Allow pointers to web sites having data sets, detailed descriptions, online demos
- Clashes with anonymity

Repeatability proposal

- Will discourage some flagrant cases
- … but onerous and “offensive” [per member of my thesis committee]
- Anonymity-complete bruhaha

Conclusion

- Strongly promote “repeatability” aspects
- Remove regulations that collide with them
- Measure the effect, feedback