Experimental Evaluation in Computer Science: A Quantitative Study

Paul Lukowicz, Ernst A. Heinz, Lutz Prechelt and Walter F. Tichy

Journal of Systems and Software
January 1995

Outline

• Motivation
• Related Work
• Methodology
• Observations
• Accuracy
• Conclusions
• Future work!

Introduction

• Large part of CS research new designs
  – systems, algorithms, models
• Objective study needs experiments
• Hypothesis
  – Experimental study often neglected in CS
• If accepted, CS inferior to natural sciences, engineering and applied math
• Paper ‘scientifically’ tests hypothesis

Related Work

• 1979 surveys say experiments lacking
  – 1994 say experimental CS under funded
• 1980, Denning defines experimental CS
  – “Measuring an apparatus in order to test a hypothesis.”
  – “If we do not live up to traditional science standards, no one will take us seriously”
• Articles on role of experiments in various CS disciplines
• 1990 experimental CS seen as growing, but 1994
  – “Falls short of science on all levels”
• No systematic attempt to assess research

Methodology

• Select Papers
• Classify
• Results
• Analysis
• Dissemination (this paper)

Select CS Papers

• Sample broad set of CS publications (200 papers)
  – ACM Transactions on Computer Systems (TOCS), volumes 9-11
  – ACM Transactions on Programming Languages and Systems (TOPLAS), volumes 14-15
  – IEEE Transactions on Software Engineering (TSE), volume 19
  – Proceedings of 1993 Conference on Programming Language Design and Implementation
• Random Sample (50 papers)
  – 74 titles by ACM via INSPEC (24 discarded)
    + 30 refereed
Select Comparison Papers

- Neural Computing (72 papers)
  - Neural Computation, volume 5
  - Interdisciplinary: bio, CS, math, medicine ...
  - Neural networks, neural modeling ...
  - Young field (1990) and CS overlap
- Optical Engineering (75 papers)
  - Optical Engineering, volume 33, no 1 and 3
  - Applied optics, opto-mech, image proc.
  - Contributors from: ee, astronomy, optics...
  - Applied, like CS, but longer history

Classify

<table>
<thead>
<tr>
<th></th>
<th>Ernst</th>
<th>Lutz</th>
<th>Paul</th>
<th>Walter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOCS</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Random</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLDI</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOPLAS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSE</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

- Same person read most
- Two read all, save NC

Major Categories

- Formal Theory
  - Formally tractable: theorem's and proofs
- Design and Modeling
  - Systems, techniques, models
  - Cannot be formally proven → require experiments
- Empirical Work
  - Analyze performance of known objects
- Hypothesis Testing
  - Describe hypotheses and test
- Other
  - Ex: surveys

Subclasses of Design and Modeling

- Amount of physical space for experiments
  - Setups, Results, Analysis
  - 0-10%, 11-20%, 21-50%, 51%+
- To shallow? Assumptions:
  - Amount of space proportional to importance by authors and reviewers
  - Amount of space correlated to importance to research
- Also, concerned with those that had no experimental evaluation at all

Assessing Experimental Evaluation

- Look for execution of apparatus, techniques or methods, models validated
- Tables, graphs, section headings…
- No assessment of quality
- But count only ‘true’ experimental work
  - Repeatable
  - Objective (ex: benchmark)
- No demonstrations, no examples
- Some simulations
  - Supplies data for other experiments
  - Trace driven

Outline

- Motivation
- Related Work
- Methodology
- Observations
- Accuracy
- Conclusions
- Future work!
Observation of Major Categories

<table>
<thead>
<tr>
<th></th>
<th>NC</th>
<th>OE</th>
<th>TOCS</th>
<th>Random</th>
<th>PLDI</th>
<th>TOPLAS</th>
<th>TSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Design</td>
<td>40</td>
<td>26</td>
<td>31</td>
<td>15</td>
<td>24</td>
<td>26</td>
<td>47</td>
</tr>
<tr>
<td>Empirical</td>
<td>8</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Emp. + Hyp.</td>
<td>9</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>76</td>
<td>38</td>
<td>50</td>
<td>26</td>
<td>30</td>
<td>87</td>
</tr>
</tbody>
</table>

- Majority is design and modeling
- The CS samples have lower percentage of empirical work than OE and NC
- Hypothesis testing is rare (4 articles out of 403!)

Observation of Design Sub-Classes

- Higher percentage with no evaluation for CS vs. NC+OE (43% vs. 14%)
- Many more NC+OE with 20%+ than in CS
- Software engineering (TSE and TOPLAS) worse than random

Observation of Design Sub-Classes

- Shows percentage that have 20%+ or more to experimental evaluation

Groupwork: How Experimental is WPI CS?

- Take 2 papers: PEDS, SERG, DSRG, ADVIS, REFER, AIRG
- Read abstract, flip through
- Categorize:
  - Formal Theory
  - Design and Modelling
    - Count pages for experiments
  - Empirical
  - Hypothesis Testing
  - Other
- Swap with another group
Outline

• Motivation
• Related Work
• Methodology
• Observations
• Accuracy
• Conclusions
• Future work!

Accuracy of Study

• Deals with humans, so subjective
• Psychology techniques to get objective measure
  – Large number of users
  – Beyond resources (and a lot of work!)
  – Provide papers, so other can provide data
• Systematic errors
  – Classification errors
  – Paper selection bias
• Statistical error

Systematic Error: Classification

<table>
<thead>
<tr>
<th>Theory</th>
<th>Empirical</th>
<th>Hypothesis</th>
<th>Other</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Empirical</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hypothesis</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>2%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>1%</td>
<td>4%</td>
<td>4%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

• Classification differences between 468 article classification pairs

Systematic Error: Paper Selection

• Journals may not be representative of CS
  – PLDI proceedings is a ‘case study’ of conferences
• Random sample may not be “random”
  – Influenced by INSPEC database holdings
  – Further influenced by library holdings
• Statistical error if selection within journals do not represent journals

Overall Accuracy (Maximize Distortion)

No Experimental Evaluation

20% + Space for Experiments
Conclusion

• 40% of CS design articles lack experiments
  – Non-CS around 10%
• 70% of CS have less than 20% space
  – NC and OE around 40%
• CS conferences no worse than journals!
• Youth of CS is not to blame
• Experiment difficulty not to blame
  – Harder in physics
  – Psychology methods can help
• Field as a whole neglects importance

Guidelines

• Higher standards for design papers
• Recognize empirical as first class science
• Need more publicly available benchmarks
• Need rules for how to conduct repeatable experiments
• Tenure committees and funding orgs need to recognize work involved in experimental CS
• Look in the mirror

Future Work

• Experiment in 1994 ... how is CS today?
• 30 people in class
• 200 articles
• Each categorized by 2 people
• About 15 articles each
  ⇒ Publish the results!
• (Send me email if interested)