Using Benchmarking to Advance Research: A Challenge to Software Engineering

Susan Elliot Sim, Steve Easterbrook, Richard Holt

Presenters: Josh Philip and Jan Gorzny
Summary - Benchmarking

- **Definition:** Set of tests to compare performance of different tools/techniques
  - Motivating Comparison, Task Sample, Performance Measures
  - E.g. TPC-A, SPEC CPU2000, TREC Ad Hoc Retrieval

- **Scientific Paradigm Lifecycle:**
  - Prescientific ➔ Normal ➔ Degenerative ➔ Revolution

- **Benchmarks operationalize paradigms** –
  - concretely express problems of interest + solution types sought
  - emerge when technical knowledge and social consensus converge
  - evidence of maturity of discipline

- **Hypothesis:** can be used proactively to accelerate the process of maturity for a discipline
Issues in AI?

- Narrow focus on small set of performance measures at expense of other qualities e.g. Simplicity, elegance, etc.
- No deeper insights into underlying interactions. E.g.
  - **Automated Planning** – FF revolutionized field 15 years ago
    - Complex search algorithm – heuristics, carefully tweaked parameters
    - Excellent results on benchmarks, but not well-understood
  - **Netflix contest** – captivated ML research community
    - $1 million if beat current recommendation system by 10%
    - Stimulated competition & spawned new research & collaborations
    - Winning solution: ensemble of > 100 models
    - MESSY! – who knows/cares why it works!
  - **Deep Learning** – initially rejected for publication
    - Embraced by part of ML community because good results on existing benchmarks
Discussion

- Do current benchmarks encourage high-quality solutions and good practice? If not, intrinsic problem with (mis)use of benchmarks, or are performance measures too simple?
- In CS, which areas could use more benchmarking (HCI?, SE?), and which are too dependent (AI?) on them? More appropriate for some disciplines? Is good mix of empirical methods needed?
- Broadly, how well are benchmarks used in our respective disciplines in CS?
  - Individually: motivation, samples, measures, desired criteria?
  - Collectively: reflect overall research goals? What do they say about priorities of discipline?
Discussion (cont’d)

- Positioned between experiments and case studies – is there naturally a post-positivist stance or are there constructivist elements? Can critical theorists use benchmarks to point out deficiencies in tools/techniques or research goals?
- Does social cohesiveness of community imply it is becoming more narrow/rigid/biased? Is it possible to attain social cohesiveness and still accommodate wide range of views?
- In trying to accelerate the process of maturity, can we determine when community is ready for benchmarks? Or, should we allow creative process to naturally unfold and self-organize into its own structures without imposing benchmarks?
Toward a Framework for Action Research in Information Systems Studies

Francis Lau

Presenters: Jan Gorzny and Josh Philip
Toward a framework for action research in information systems studies

- Definitions:
  - **Action Research**: an iterative process of problem diagnosis, action intervention, and reflective learning
  - **Action Science**: places its emphasis on understanding participants' behaviors as theories-in-use versus their beliefs as espoused theories, and the use of single and double-loop learning for self-improvement
  - **Participatory AR**: a stream of action research that involves practitioners as both subjects and co-researchers
  - **Action Learning**: advocates group participation, programmed instructions, spontaneous questioning, real actions, and experiential learning in different social and organizational contexts.

- Framework: four dimensions
  - Conceptual foundation
  - Study design
  - Research process
  - Role expectations
Toward a framework for action research in information systems studies

<table>
<thead>
<tr>
<th>Dimension and criteria</th>
<th>Classification</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conceptual foundation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aim/question</td>
<td>What is the research aim or question?</td>
<td>Is the research aim or question authentic and practical in addressing an immediate situation?</td>
</tr>
<tr>
<td>Assumptions</td>
<td>Is some form of theory, theme or concept included?</td>
<td>Is the theory, theme or concept authentic?</td>
</tr>
<tr>
<td>Perspective/tradition</td>
<td>Is it interpretive or critical, or one of the community, school, organization development traditions?</td>
<td>Is it authentic according to the paradigm adopted, e.g. are there multiple realities constructed in the study if interpretivism is used?</td>
</tr>
<tr>
<td>Stream</td>
<td>Is it one of the AR, PAR, AS or AL streams?</td>
<td>Is it consistent with the definitions of the respective streams?</td>
</tr>
<tr>
<td><strong>Study design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background</td>
<td>Is there information on the environment being studied?</td>
<td>Does it give sufficient understanding of the total social situation?</td>
</tr>
<tr>
<td>Intended change</td>
<td>What is the nature and extent of planned change?</td>
<td>Is the intended change appropriate and adequate?</td>
</tr>
<tr>
<td>Site</td>
<td>Is it a single or multiple sites? What is type and level of organization involved?</td>
<td>To what extent are the site(s) and the organization involved? And is the level of involvement appropriate and adequate?</td>
</tr>
<tr>
<td>Participants</td>
<td>Who are the participants and what are their background?</td>
<td>Are the participants authentic?</td>
</tr>
<tr>
<td>Data sources</td>
<td>What types of data are collected for the study and how?</td>
<td>Are the data credible, dependable and confirmable?</td>
</tr>
<tr>
<td>Duration</td>
<td>What is the intended length of study in time duration?</td>
<td>Is there adequate time for problem diagnosis, action intervention, and reflective learning to take place?</td>
</tr>
<tr>
<td>Degree of openness</td>
<td>What is the extent of predefined or planned process?</td>
<td>Is the process conducted as planned or would it evolve over time?</td>
</tr>
<tr>
<td>Access/exit</td>
<td>What type, level and extent of access to the organization is intended and is there a defined exit point from the study?</td>
<td>Is the degree of openness appropriate and adequate?</td>
</tr>
<tr>
<td>Presentation</td>
<td>What is the reporting style - case report, ethnographic?</td>
<td>Does reporting provide sufficient information to judge its quality?</td>
</tr>
</tbody>
</table>

(continued)
Toward a framework for action research in information systems studies

- AR: better as a “research method” or “theory of social science”?
- Did Lau miss anything in his framework?
- Why might AR be less common in North American journals compared to European journals?
  - Could this imply anything about the philosophic stances of these regions?
- Is it ever appropriate to not explicitly list interventions taken in such research?
- Can the creation of criteria for assessing action research have the same social implications that a community building a benchmark has?
  - Does it require the same pre-conditions?
Toward a framework for action research in information systems studies

- Why is it important for AR to declare the intent of the study? Or to explicate the perspective?
- Can an iteration be made if there was no/little reflective learning from the last step?
- What bias on roles might a researcher’s philosophical stance have? How could this be avoided?
- Why is it important that AR has an intended change? What happens if AR fails to change anything?