THE SIGNIFICANCE OF EVALUATION IN AI & LAW
A Case Study Re-examining ICAIL Proceedings

Jack G. Conrad
Thomson Reuters Global Resources
Catalyst Lab
Baar, Switzerland 6340

John Zeleznikow
Victoria University
School of Management & Info Systems
Melbourne, Australia 3086

14th Int'l Conf. on Artificial Intelligence & Law
Rome, Italy – 10-14 June 2013

THOMSON REUTERS

OUTLINE

- **Background − Original Study of ICAIL Proceedings**
- **Update − How We've Performed Since**

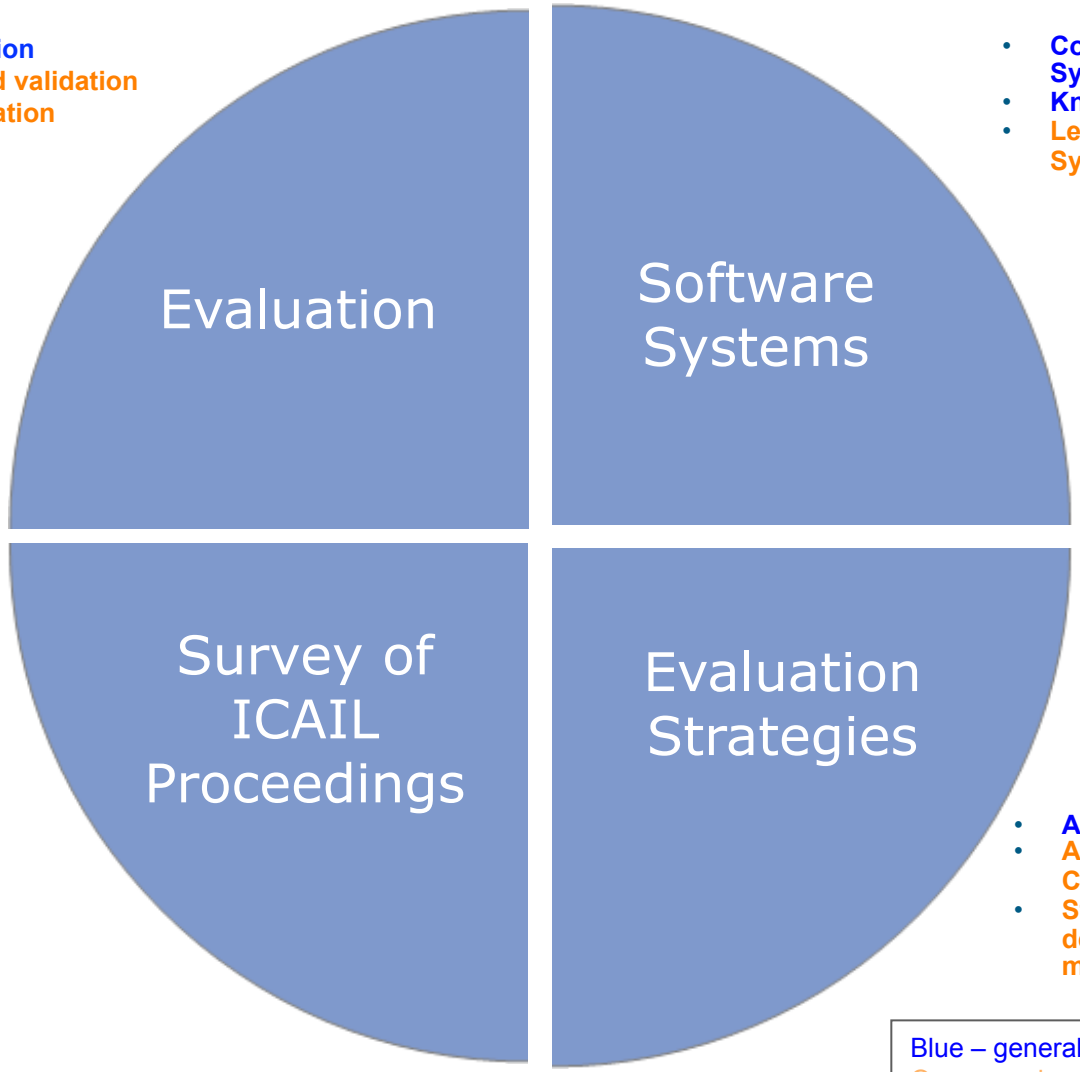
First Study of Evaluation in ICAIL Proceedings

- Marie Jean J. Hall and John Zeleznikow
- **Acknowledging Insufficiency in the Evaluation of Legal Knowledge-based Systems: Strategies Towards a Broad-based Evaluation Model**
- In *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001)* (St. Louis, Missouri), pp. 147-156
- ACM Press, 2001.

Focus: ICAIL Proceedings 1987, 1995, 1997, 1999

“Acknowledging the Insufficiency in the Evaluation of Legal Knowledge-based Systems”

- Verification and validation
- Beyond verification and validation
- Assessment and evaluation



- Conventional Software Systems
- Knowledge-based Systems
- Legal Knowledge-based Systems

- Papers categorized by: “Theoretical”, “Evaluated”,
- “Not Evaluated”
- Focus on last two
- Also examined type of evaluation used

- An evaluation methodology
- An Evaluation Context Checklist
- Strategies beyond development of methodology

Blue – general examination
Orange – domain-specific examination

OUTLINE

- **Background − Original Study of ICAIL Proceedings**
- **Update − How We've Performed Since**

Current Study of Evaluation in ICAIL Proceedings

- A self-reflexive, meta-level study
- Examines the presence of evaluation in works published at ICAIL since 2000 (2001 – 2011)
- Proportion of works that include some form of performance evaluation and their *degree*
- Compares these recent trends with those identified by Hall and Zeleznikow (ICAIL 2001)
- Develops an argument for why evaluation in formal AI and Law reports is significant

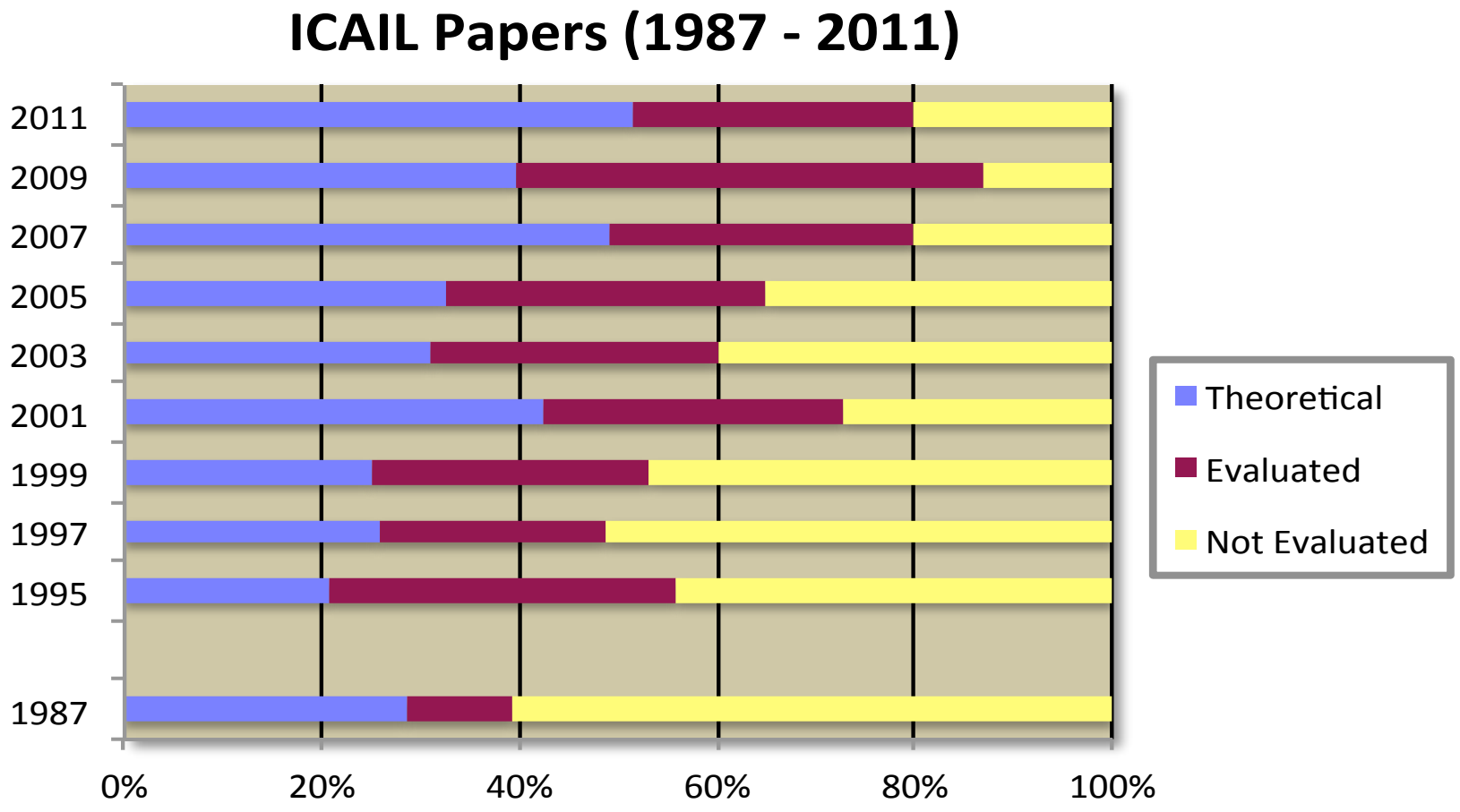
Current Study of Evaluation in ICAIL Proceedings

- **Objective:** Determine how the presence of evaluation at the community level has evolved over time
- **Motivation:** Investigate along one significant dimension if community has matured in use of empirical assessment
- **Proposition:** If fundamental questions unanswered – How good is the system? How reliable is the technique? Does it work? – how can the researcher convince the broader community of the benefits and utility of the work?
- **Definition:** *Evaluation* – systematic determination of subject's merit, worth, significance using criteria governed by a set of standards

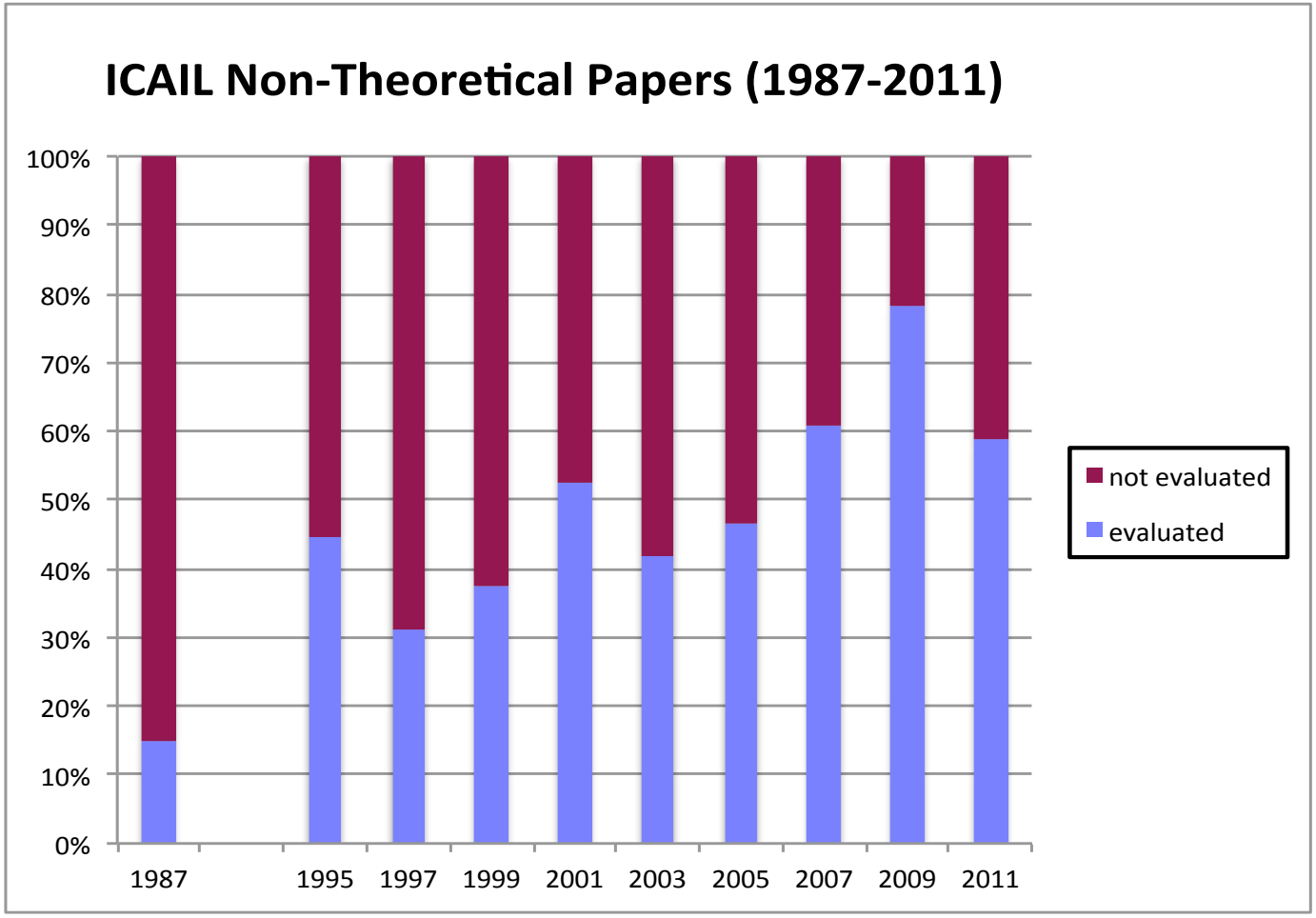
References

- Paul R. Cohen and Adele E. Howe, “How Evaluation Guides AI Research”, *AI Magazine*, 9(4):35-43, Winter, 1988.
- Richard Susskind, “Expert Systems in Law”, In *Proceedings of the 1st International Conference on Artificial Intelligence and Law (ICAIL 1987)* (Boston, MA), pp. 1–8. IAAIL, ACM Press, May 1987.

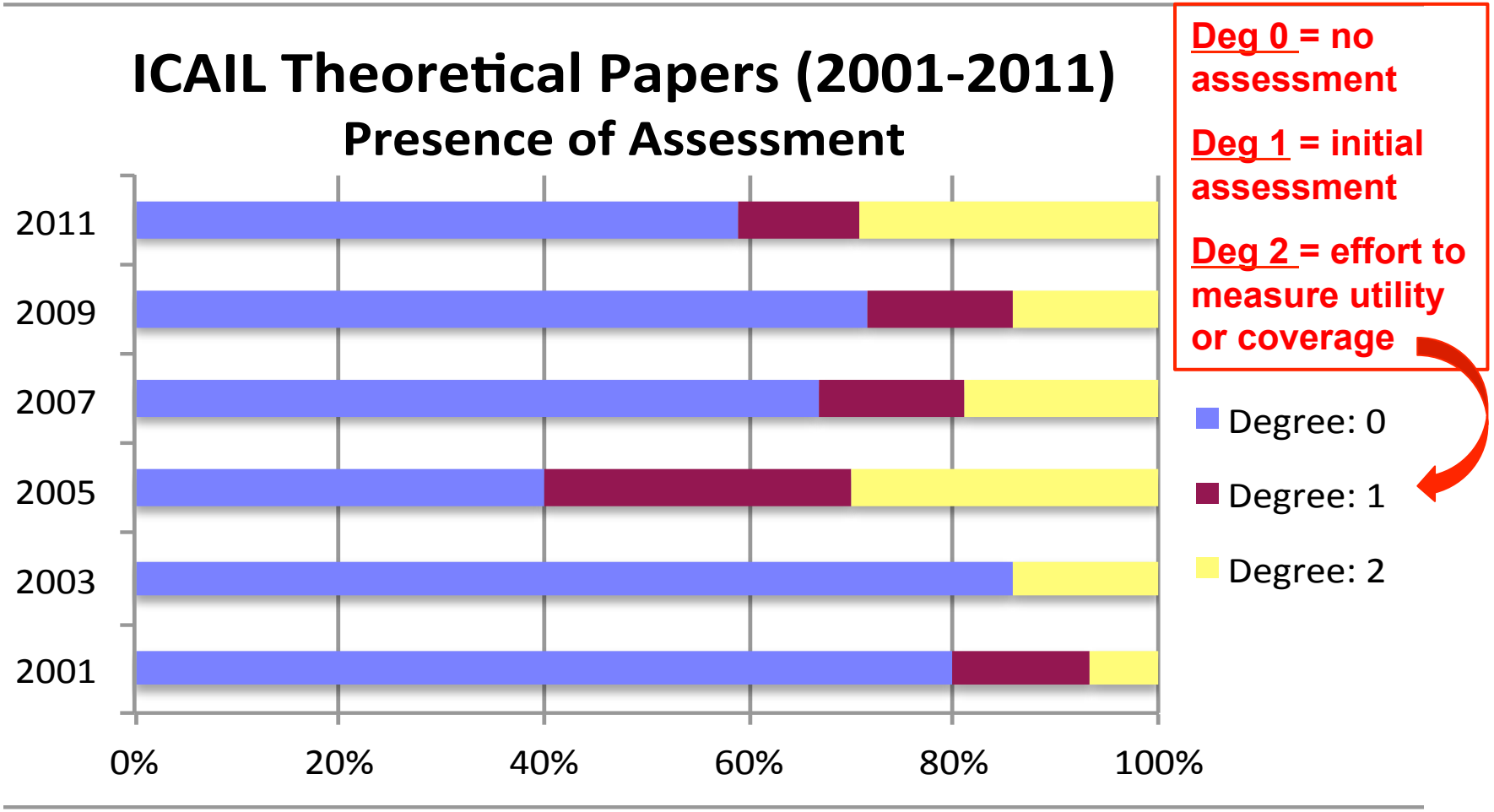
Theoretical vs. Evaluated & Non-Evaluated Works



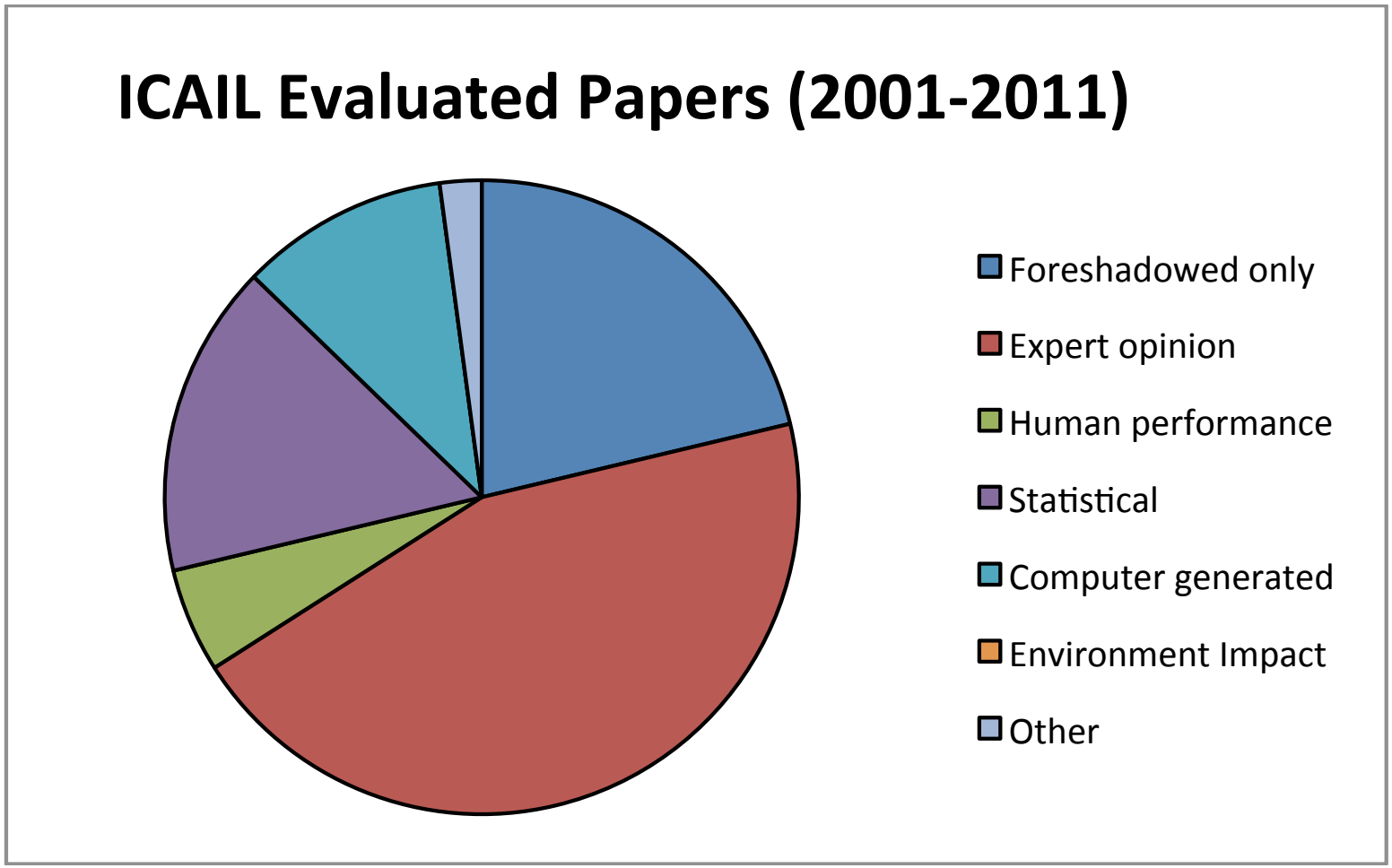
Evaluation in Non-Theoretical Works



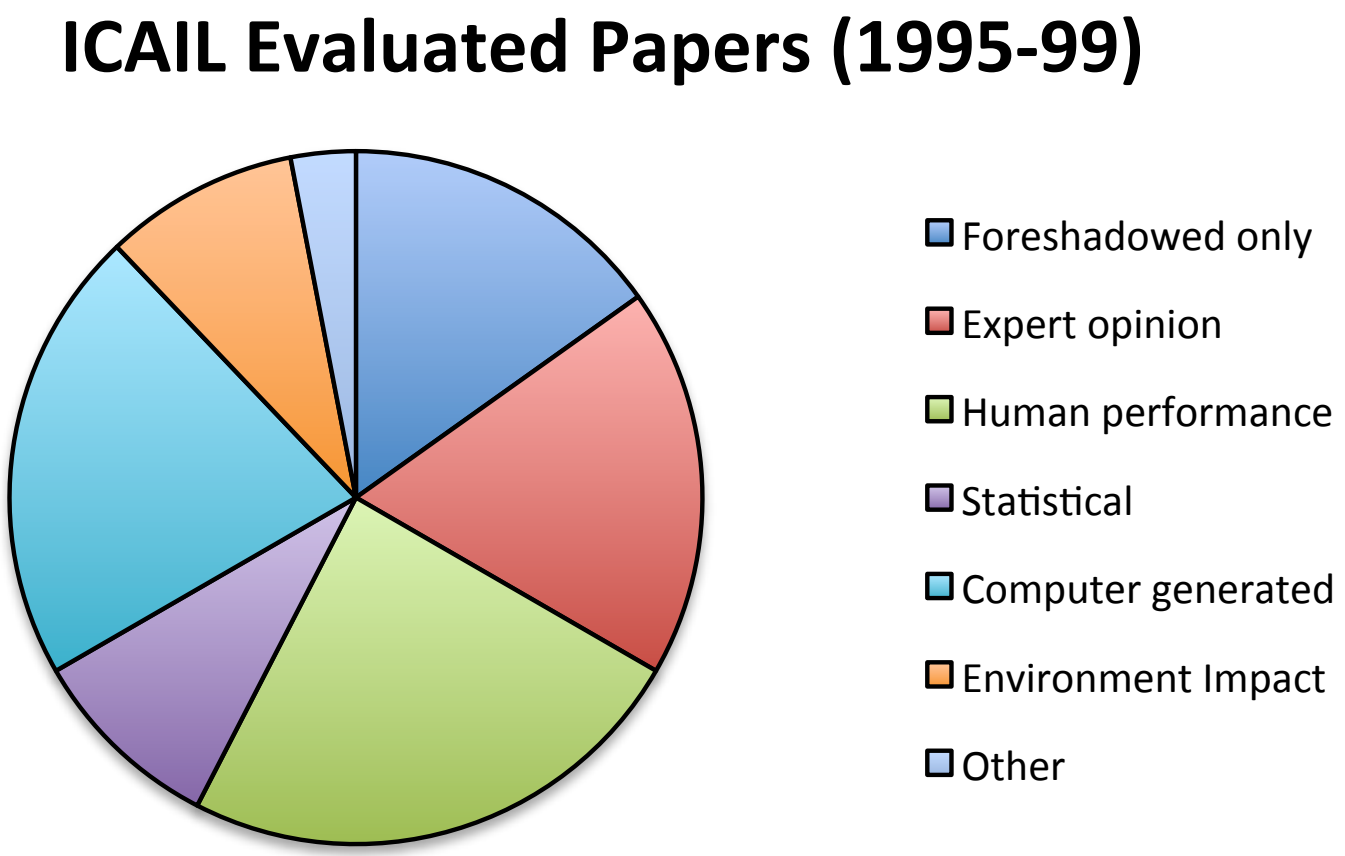
Presence of Assessment in Theoretical Works



Type of Evaluated Works (current study)



Types of Evaluated Works (past study)



Theoretical vs. Evaluated & Non-Evaluated Works

Year	Theoretical	%	Evaluated	%	Not Evaluated	%	Total
1987	8	29%	3	11%	17	61%	28
1995	7	21%	12	35%	15	44%	34
1997	10	26%	9	23%	20	51%	39
1999	8	25%	9	28%	15	47%	32
2001	14	42%	10	30%	9	27%	33
2003	14	31%	13	29%	18	40%	45
2005	14	33%	13	33%	15	14%	42
2007	22	49%	14	31%	9	9%	45
2009	15	39%	18	47%	5	13%	38
2011	18	51%	10	29%	7	20%	35
Total	130	35%	111	30%	130	35%	371

Evaluation Categories

- 0 – **Absent** (F) – no mention of evaluation in any form
- 1 – **Discussion** (D) – discusses how the system or approach could be evaluated
- 2 – **Basic** (C) – preliminary, simply evaluation is performed on a portion of system or data, or evidence of anecdotal assessment
- 3 – **Moderate** (B) – significant evaluation effort is performed on the system or approach
- 4 – **Mature/Comprehensive** (A) – credible degree of evaluation performed, us. multiple assessments
  - E.g., modular vs. end-to-end; vs. baselines; vs. humans

Concluding Remarks

- Current ICAIL evaluation landscape leaves room for improvement
- Short of full-fledged experiments, sketches of how future evaluation should be conducted can be helpful
- Even theoretical works can have extended examples and illustrations of coverage
- To be a mature research community exercising scientific rigor, multi-faceted, in-depth evaluation is required

Current Study of Evaluation in ICAIL Proceedings

- Jack G. Conrad and John Zeleznikow
- **The Significance of Evaluation in AI and Law: A Case Study Re-examining ICAIL Proceedings**
- In *Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL 2013)* (Rome, Italy), pp. 186-191
- ACM Press, 2013.

Focus: ICAIL Proceedings 2001 - 2011

Full-length Version & Recent Paper Classifications

- **http://www.conradweb.org/~jackg/publications.html**

Final Perspectives

- “Evaluation is what we are all about. It is what separates us from other technologists. It is what adds value to our research. We compare what we design with existing baselines to demonstrate that our approach is better, about the same, or worse, but the point is that we investigate the topic from a measurable, highly quantitative and comparable perspective.” -- PEJ

- “One might question whether conferences are the correct place to report evaluation. Conferences might be seen as a place to float initial ideas that report new advances: these are reported so that they can be developed and evaluated subsequently. If everything were required to be evaluated where would such progress come from? Also conference papers are very space limited: often there is too little room to report both the system and detailed experiments.” -- Anonymous Reviewer

THE SIGNIFICANCE OF EVALUATION IN AI & LAW
A Case Study Re-examining ICAIL Proceedings

Jack G. Conrad
Thomson Reuters Global Resources
Catalyst Lab
Baar, Switzerland 6340

John Zeleznikow
Victoria University
School of Management & Info Systems
Melbourne, Australia 3086

Questions & Discussion

THOMSON REUTERS