NTNU

How Can We Know It Is Shoulders We Stand On?

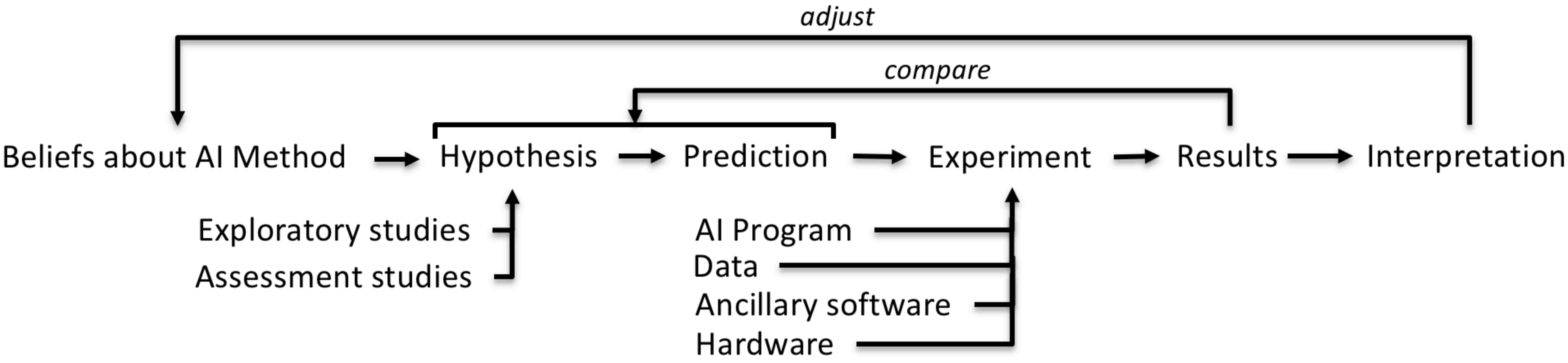
Measuring reproducibility

Odd Erik Gundersen, dr. philos.
Chief AI Officer, TrønderEnergi AS
Adjunct Associate Professor, NTNU
odderik@ntnu.no

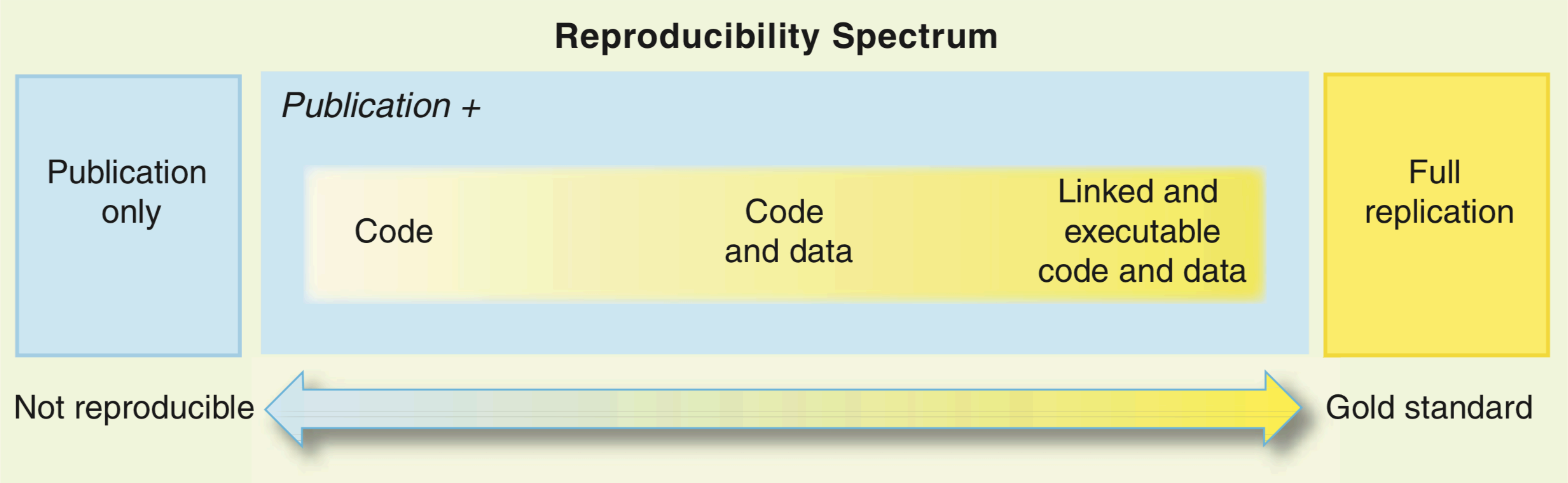
NTNU
Norwegian University of Science and Technology

norwegian open ai lab

The Scientific Method in Empirical AI Research



Defining Reproducibility I



(R. D. Peng, Science, 2011)

Defining Reproducibility II

Replication is to re-run the experiment with code and data provided by the author.

Reproduction implies both replication and the regeneration of findings with at least some independence from the [original] code and/or data.

(V. Stodden, Amstat News, 2011)

Defining Reproducibility III

Methods reproducibility: The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.

Results reproducibility: The production of corroborating results in a new study, having used the same experimental methods.

Inferential reproducibility: The drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.

(S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, Science Translational Medicine, 2016)

Definition of Reproducibility

Reproducibility in empirical AI research is the ability of an **independent** research team to produce the same **results** using the same AI method based on the **documentation** made by the original research team.

Documentation

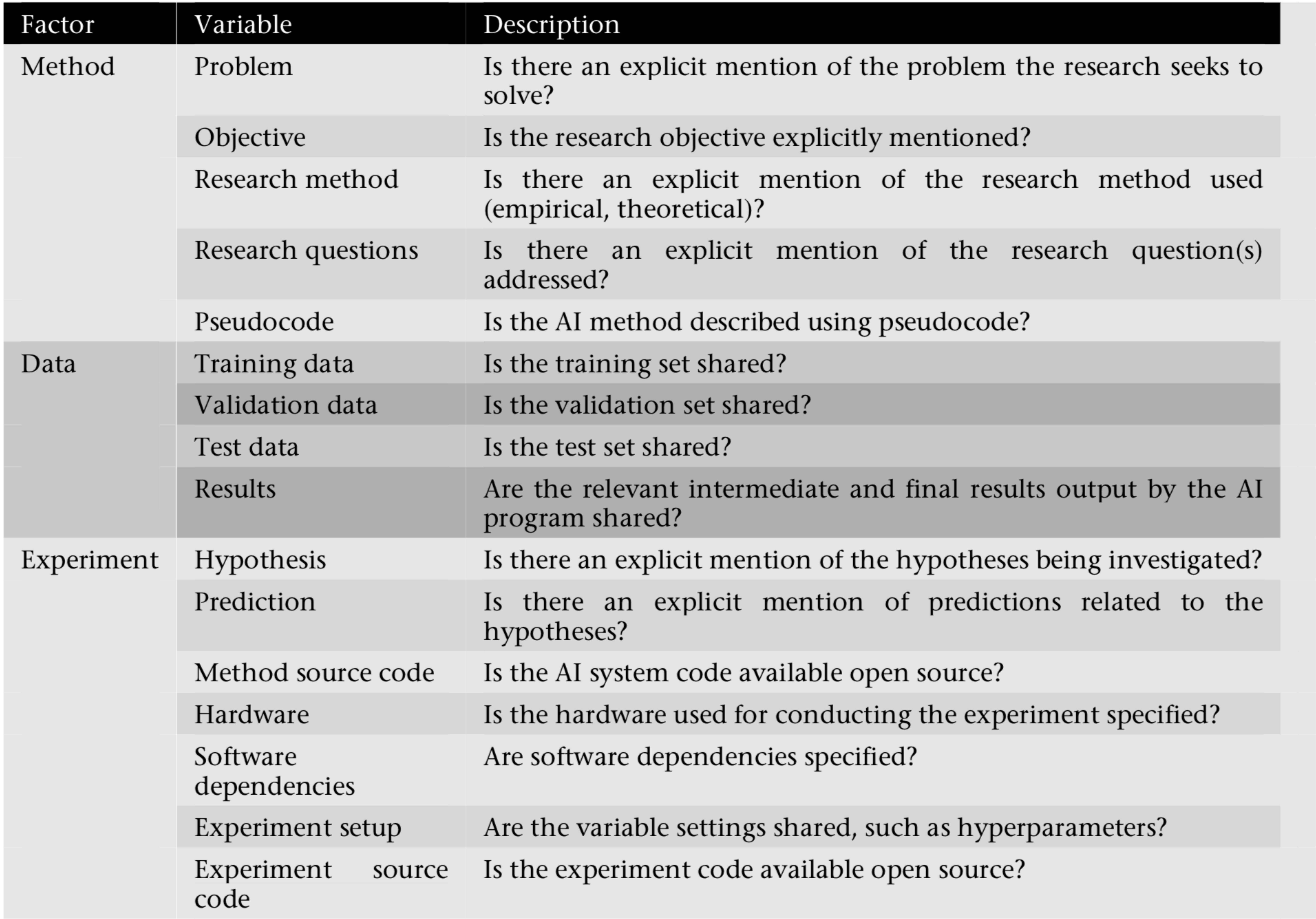
- **Method:** Report, the textual description of method (system/algorithm/experiment) - human to human - abstract concepts.

- **Data:** Represents the world the AI method operates in. Used for testing hypotheses.

- **Experiment Setup:** Code (AI method implementation + experiment code) + hardware

Degree of Reproducibility

	Method	Data	Experiment
R1	🟦	🟦	🟦
R2	🟨	🟨	
R3	🟧		



Factor	Variable	Description
Method	Problem	Is there an explicit mention of the problem the research seeks to solve?
	Objective	Is the research objective explicitly mentioned?
	Research method	Is there an explicit mention of the research method used (empirical, theoretical)?
	Research questions	Is there an explicit mention of the research question(s) addressed?
	Pseudocode	Is the AI method described using pseudocode?
Data	Training data	Is the training set shared?
	Validation data	Is the validation set shared?
	Test data	Is the test set shared?
	Results	Are the relevant intermediate and final results output by the AI program shared?
Experiment	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Method source code	Is the AI system code available open source?
	Hardware	Is the hardware used for conducting the experiment specified?
	Software dependencies	Are software dependencies specified?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
	Experiment source code	Is the experiment code available open source?

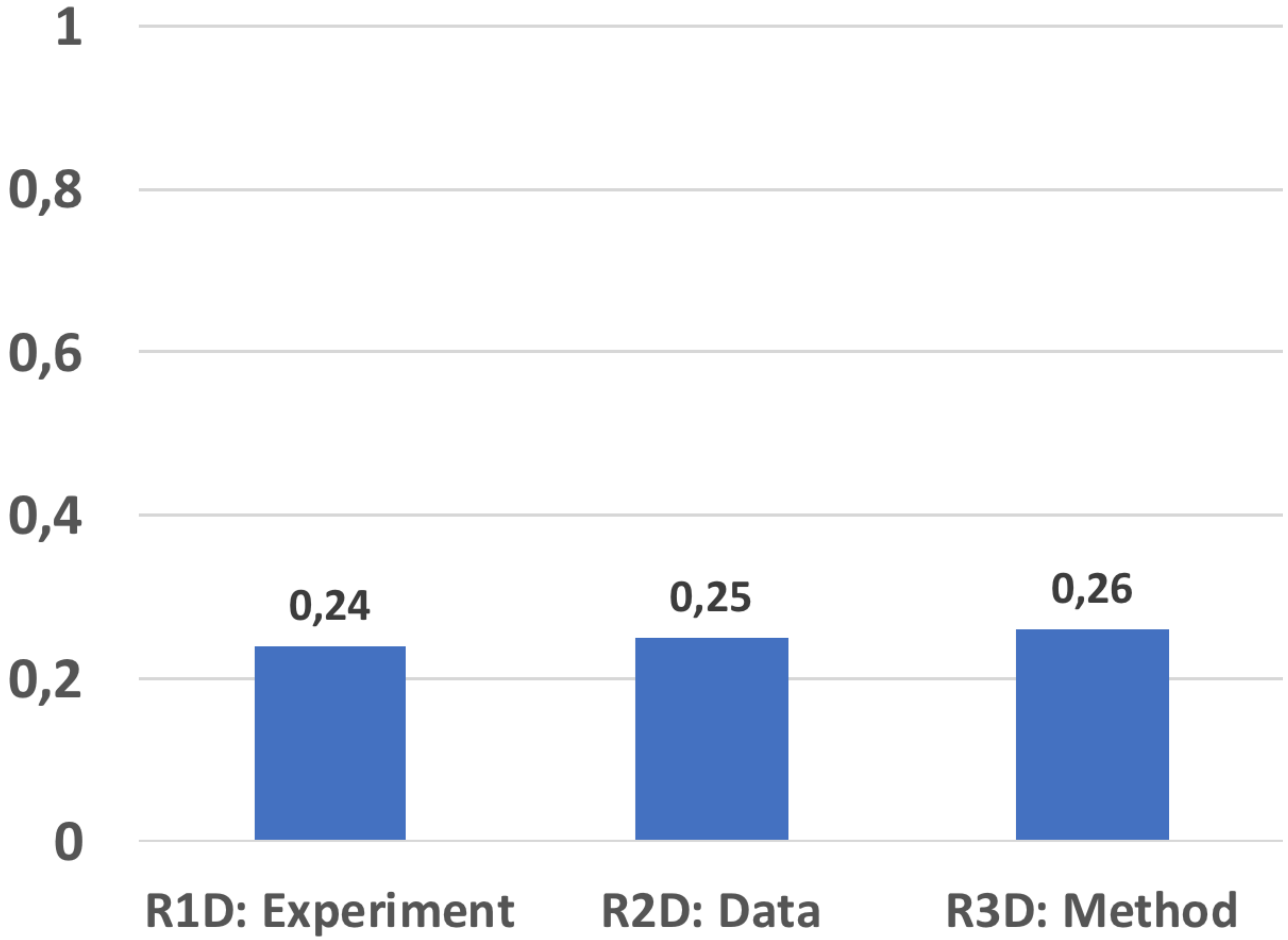
Quantifying Reproducibility

$$R1D(e) = \frac{\delta_1 Method(e) + \delta_2 Data(e) + \delta_3 Exp(e)}{\delta_1 + \delta_2 + \delta_3}$$

$$R2D(e) = \frac{\delta_1 Method(e) + \delta_2 Data(e)}{\delta_1 + \delta_2}$$

$$R3D(e) = Method(e)$$

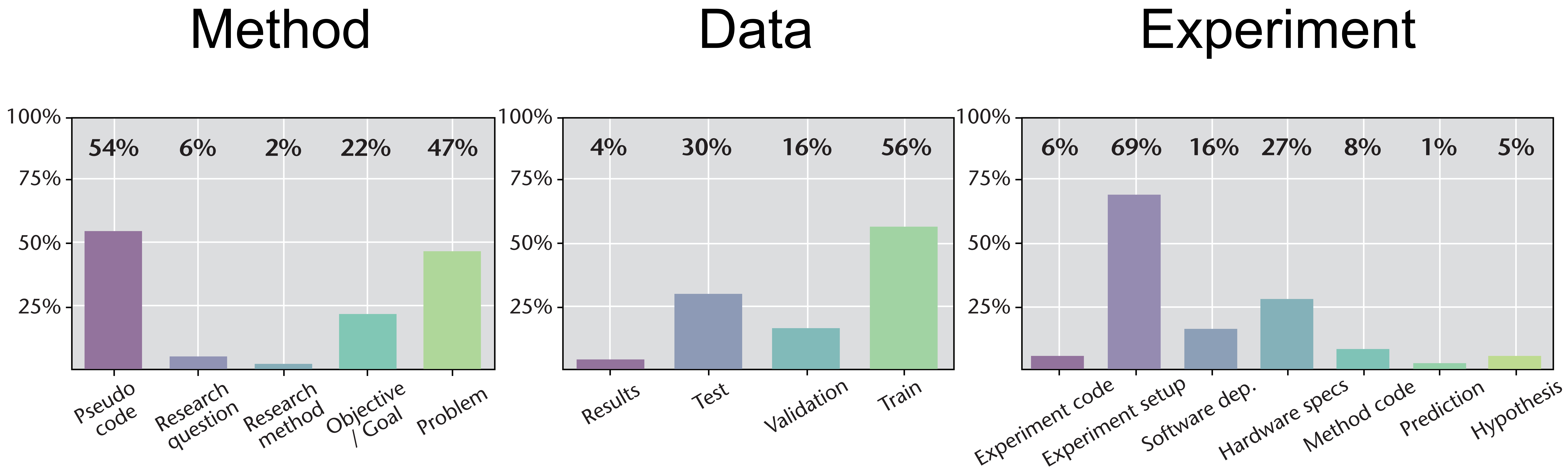
A Normalized Metric



(Gundersen, Kjensmo, AAAI, 2018)

WHAT WE GAIN

We Can Specify How Well Research is Documented



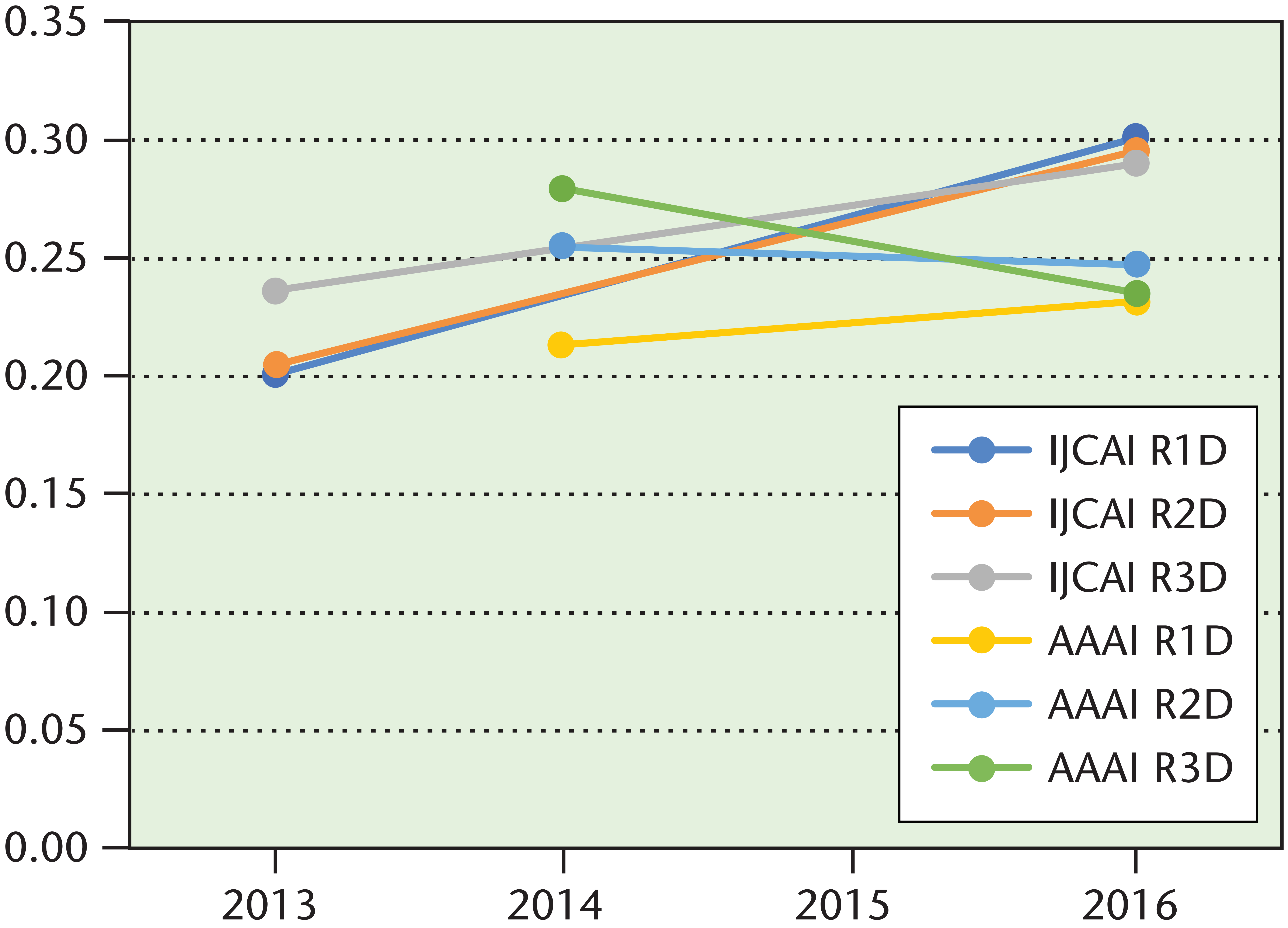
Method

Data

Experiment

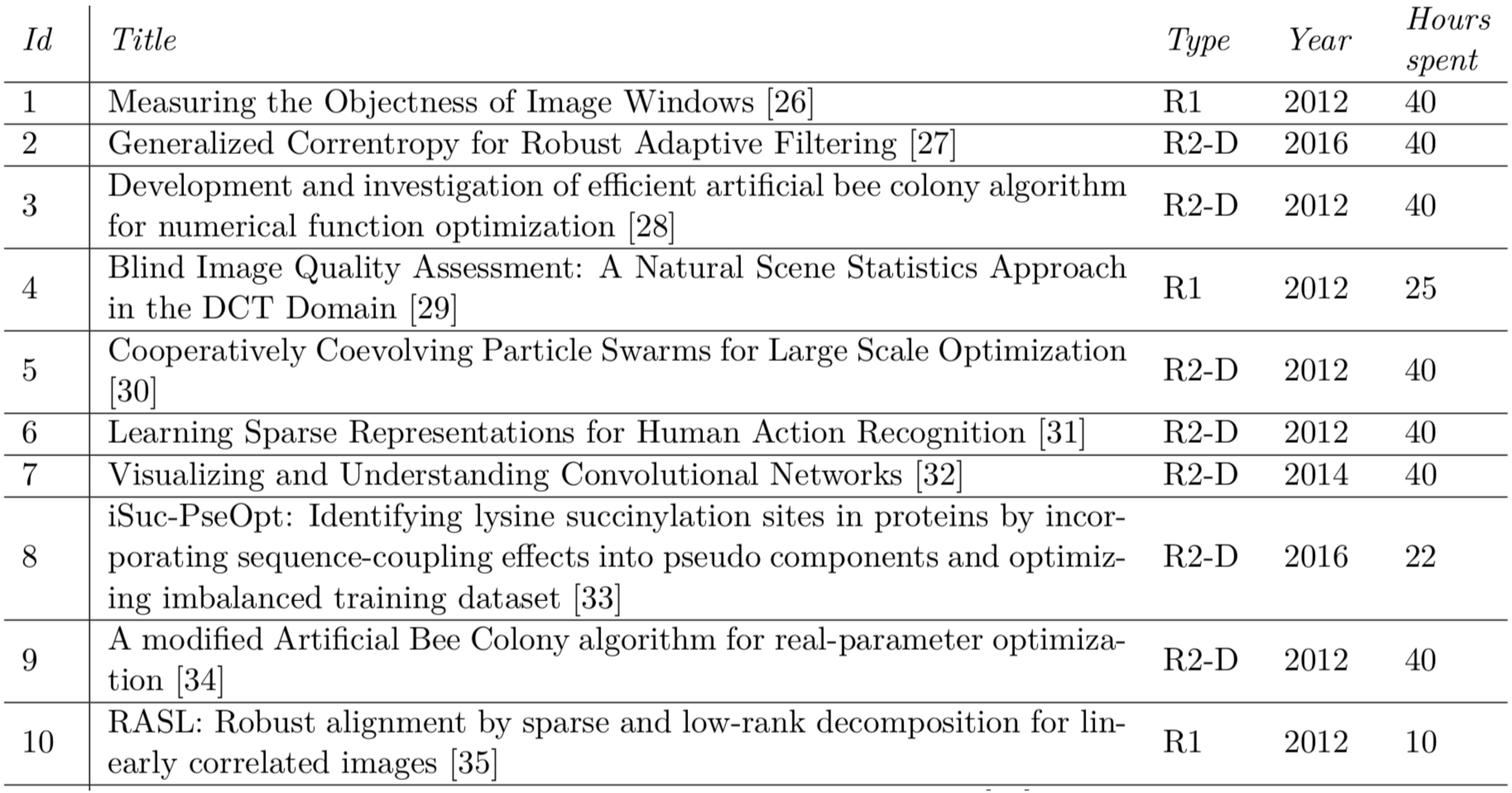
(Gundersen, Kjensmo, AAAI, 2018)

We Can Measure Improvement



(Gundersen, Kjensmo, AAAI, 2018)

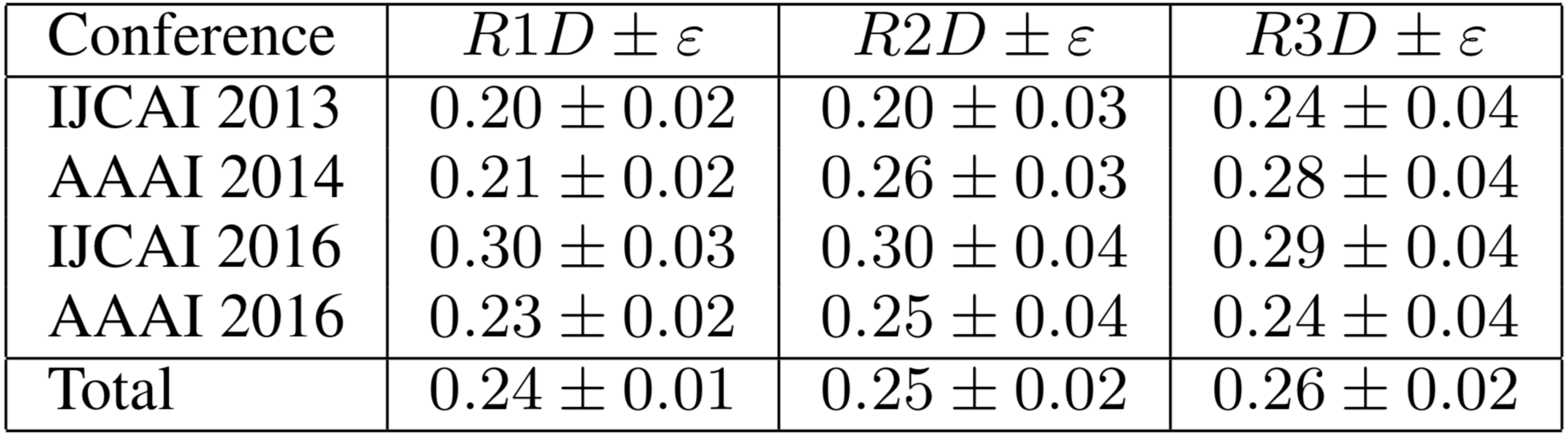
We Can Compare Research: Papers



Id	*Title*	*Type*	*Year*	*Hours spent*
1	Measuring the Objectness of Image Windows [26]	R1	2012	40
2	Generalized Correntropy for Robust Adaptive Filtering [27]	R2-D	2016	40
3	Development and investigation of efficient artificial bee colony algorithm for numerical function optimization [28]	R2-D	2012	40
4	Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain [29]	R1	2012	25
5	Cooperatively Coevolving Particle Swarms for Large Scale Optimization [30]	R2-D	2012	40
6	Learning Sparse Representations for Human Action Recognition [31]	R2-D	2012	40
7	Visualizing and Understanding Convolutional Networks [32]	R2-D	2014	40
8	iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset [33]	R2-D	2016	22
9	A modified Artificial Bee Colony algorithm for real-parameter optimization [34]	R2-D	2012	40
10	RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images [35]	R1	2012	10

(Gundersen et al, forthcoming)

We Can Compare Research: Conferences

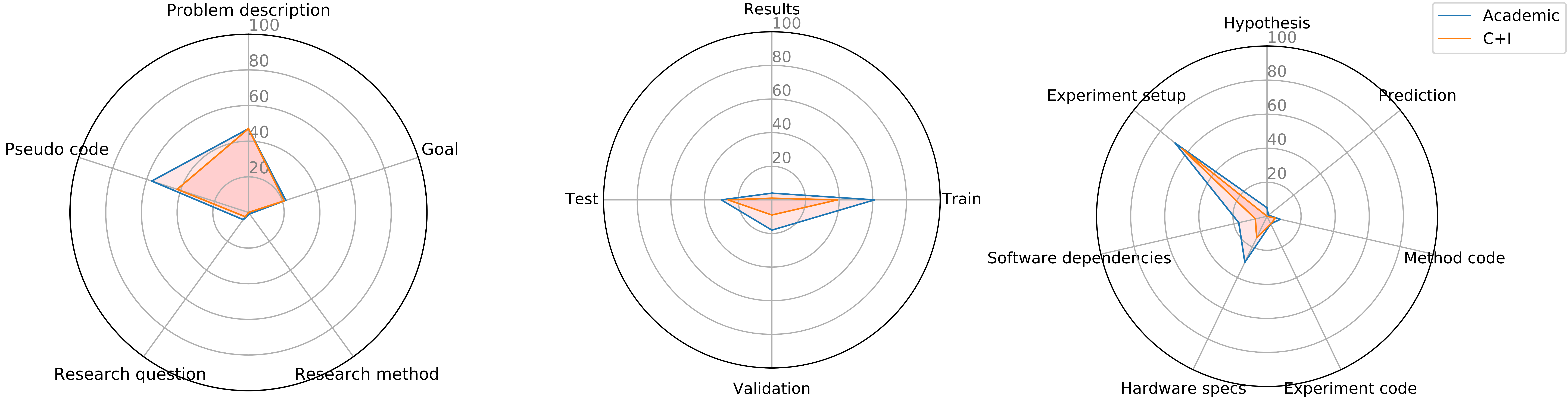


Conference	$R1D \pm \varepsilon$	$R2D \pm \varepsilon$	$R3D \pm \varepsilon$
IJCAI 2013	0.20 ± 0.02	0.20 ± 0.03	0.24 ± 0.04
AAAI 2014	0.21 ± 0.02	0.26 ± 0.03	0.28 ± 0.04
IJCAI 2016	0.30 ± 0.03	0.30 ± 0.04	0.29 ± 0.04
AAAI 2016	0.23 ± 0.02	0.25 ± 0.04	0.24 ± 0.04
Total	0.24 ± 0.01	0.25 ± 0.02	0.26 ± 0.02

(Gundersen, Kjensmo, AAAI, 2018)

We Can Compare Research: Groups

Academia versus Industry



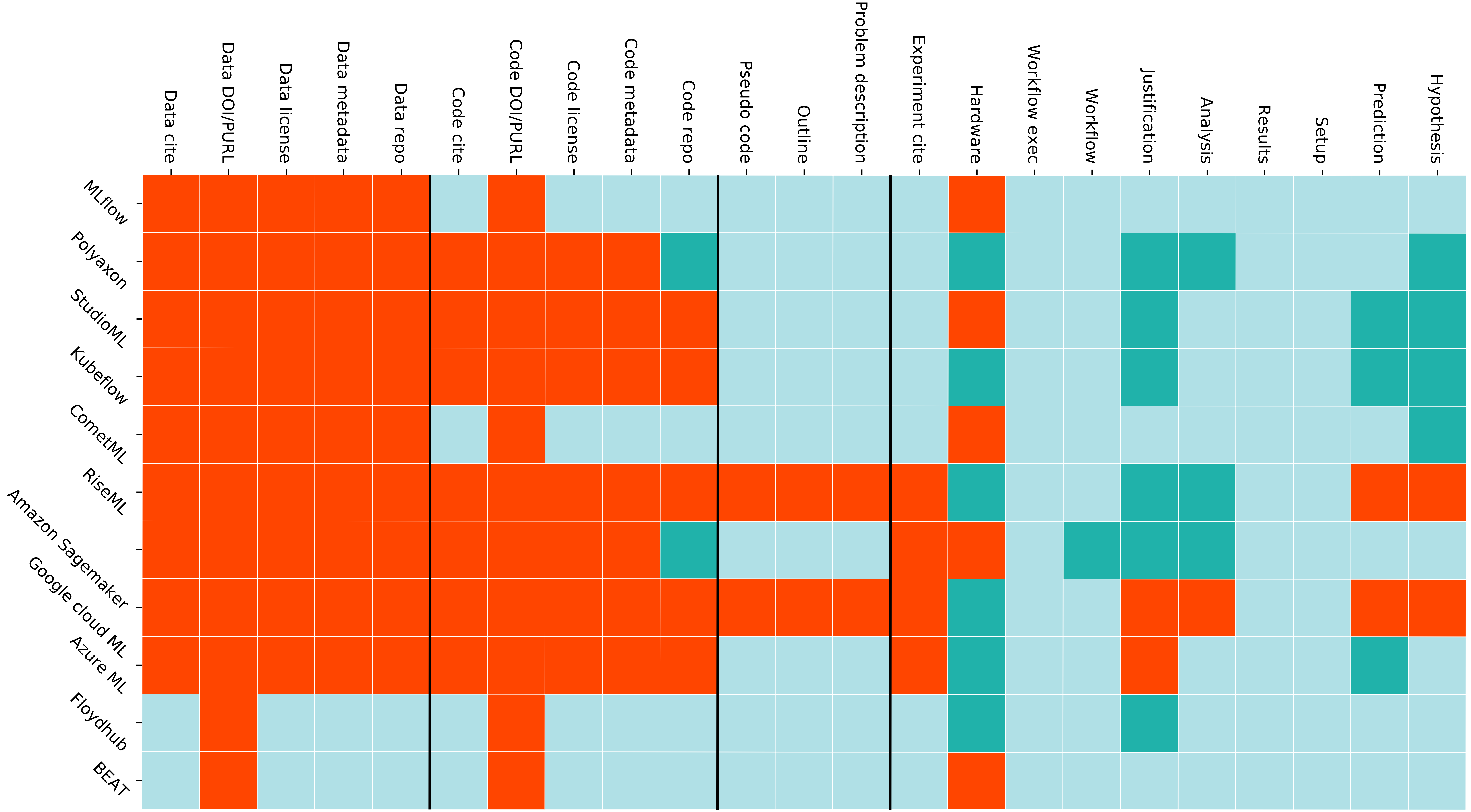
Method

Data

Experiment

(Gundersen, AI Magazine, forthcoming)

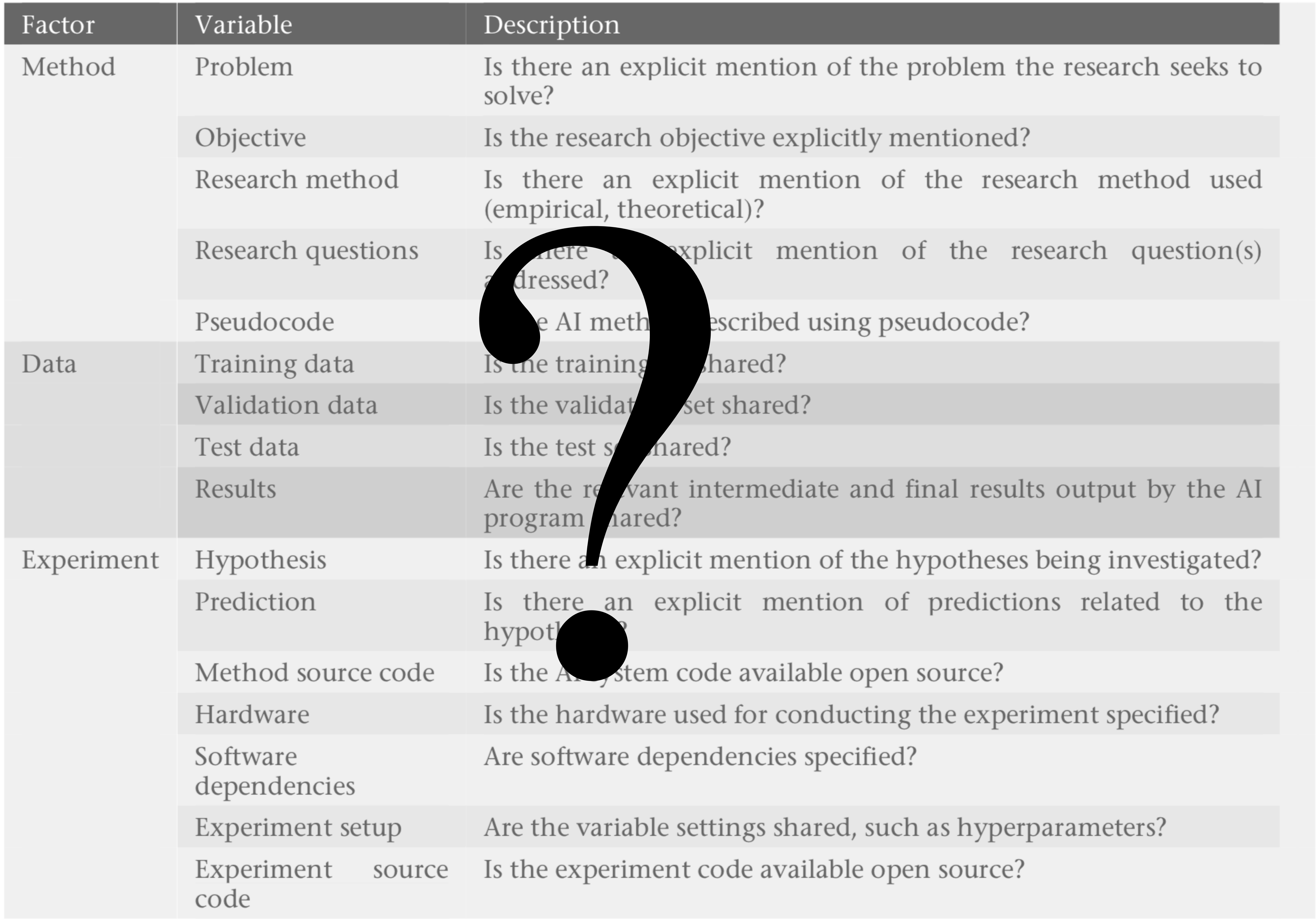
We Can Compare Software Frameworks



	Data cite	Data DOI/PURL	Data license	Data metadata	Data repo	Code cite	Code DOI/PURL	Code license	Code metadata	Code repo	Pseudo code	Outline	Problem description	Experiment cite	Hardware	Workflow exec	Workflow	Justification	Analysis	Results	Setup	Prediction	Hypothesis
MLflow	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Polyaxon	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟦	🟦	🟦	🟦	🟩	🟦	🟦	🟩	🟩	🟦	🟦	🟦	🟩
StudioML	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟩	🟦	🟦	🟦	🟩	🟩
Kubeflow	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟦	🟦	🟦	🟩	🟦	🟦	🟩	🟦	🟦	🟦	🟩	🟩
CometML	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟩
RiseML	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟦	🟦	🟩	🟩	🟦	🟦	🟧	🟧
Amazon Sagemaker	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟦	🟦	🟦	🟧	🟧	🟦	🟩	🟩	🟩	🟦	🟦	🟦	🟦
Google cloud ML	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟦	🟦	🟧	🟧	🟦	🟦	🟧	🟧
Azure ML	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟦	🟦	🟧	🟩	🟦	🟦	🟧	🟦	🟦	🟦	🟩	🟦
Floydhub	🟦	🟧	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟩	🟦	🟦	🟩	🟦	🟦	🟦	🟦	🟦
BEAT	🟦	🟧	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

(Isdahl et al, forthcoming)

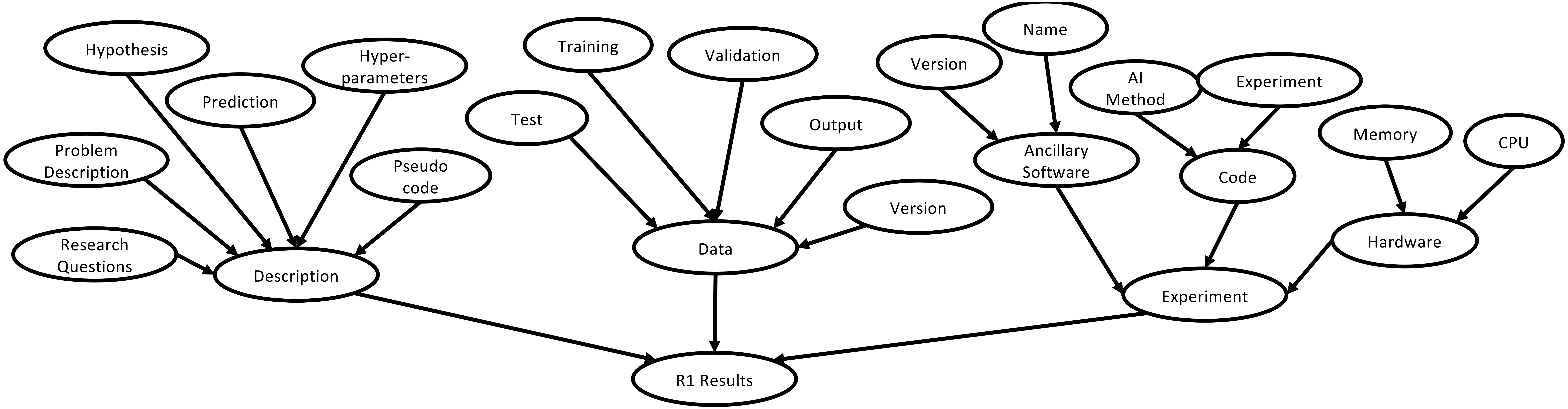
We Could Empirically Find What Entails Well-Documented Research



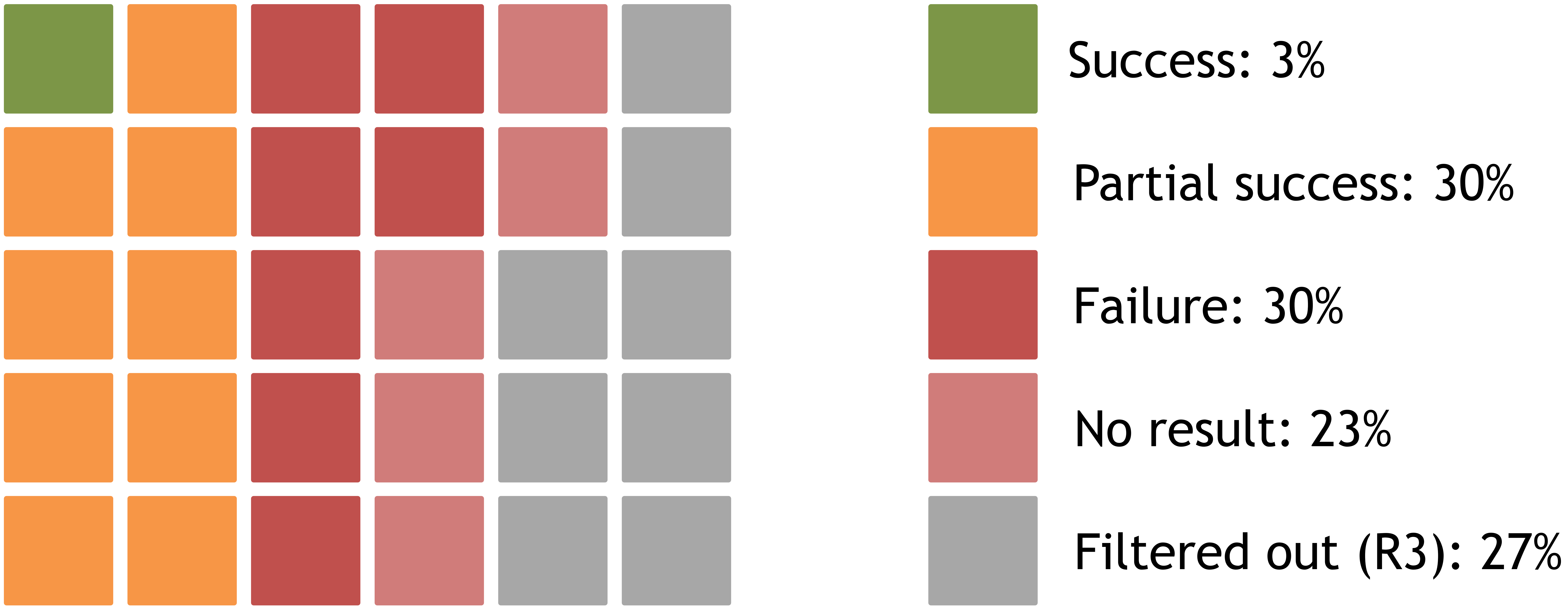
Factor	Variable	Description
Method	Problem	Is there an explicit mention of the problem the research seeks to solve?
	Objective	Is the research objective explicitly mentioned?
	Research method	Is there an explicit mention of the research method used (empirical, theoretical)?
	Research questions	Is there an explicit mention of the research question(s) addressed?
	Pseudocode	Is the AI method described using pseudocode?
Data	Training data	Is the training set shared?
	Validation data	Is the validation set shared?
	Test data	Is the test set shared?
	Results	Are the relevant intermediate and final results output by the AI program shared?
Experiment	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Method source code	Is the AI system code available open source?
	Hardware	Is the hardware used for conducting the experiment specified?
	Software dependencies	Are software dependencies specified?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
	Experiment source code	Is the experiment code available open source?

?

Compute the Likelihood of Success?

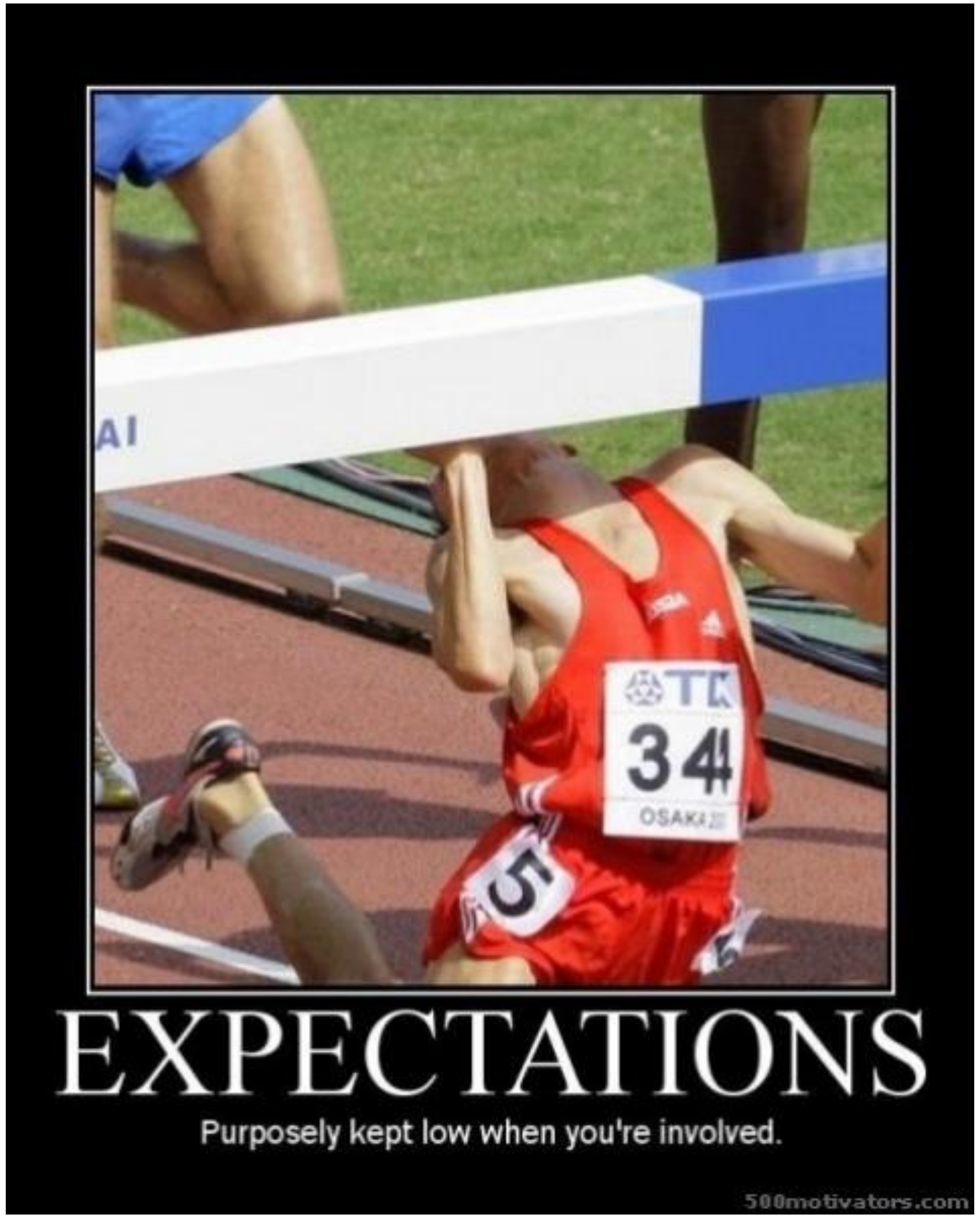


We Should Be Able to Measure Success



(Gundersen et al, forthcoming)

We Can Set the Bar Based on What We Want to Achieve



Research

- **State of the Art: Reproducibility in Artificial Intelligence** O. E. Gundersen and S. Kjensmo, AAAI 2018
- **On Reproducible AI** O. E. Gundersen, Y. Gil and D. W. Aha, AI Magazine, Fall 2018.
- **Standing on the Feet of Giants** O. E. Gundersen, AI Magazine, forthcoming 2019.
- **Supporting Reproducible Experiments - A Survey**, R. Isdahl and O. E. Gundersen, forthcoming 2019.
- **What We Learned When Reproducing the Most Cited AI Research**, O. E. Gundersen, O. Cappelen, N. Grimstad, M. Mølnå, forthcoming 2019.

NTNU Odd Erik Gundersen odderik@ntnu.no