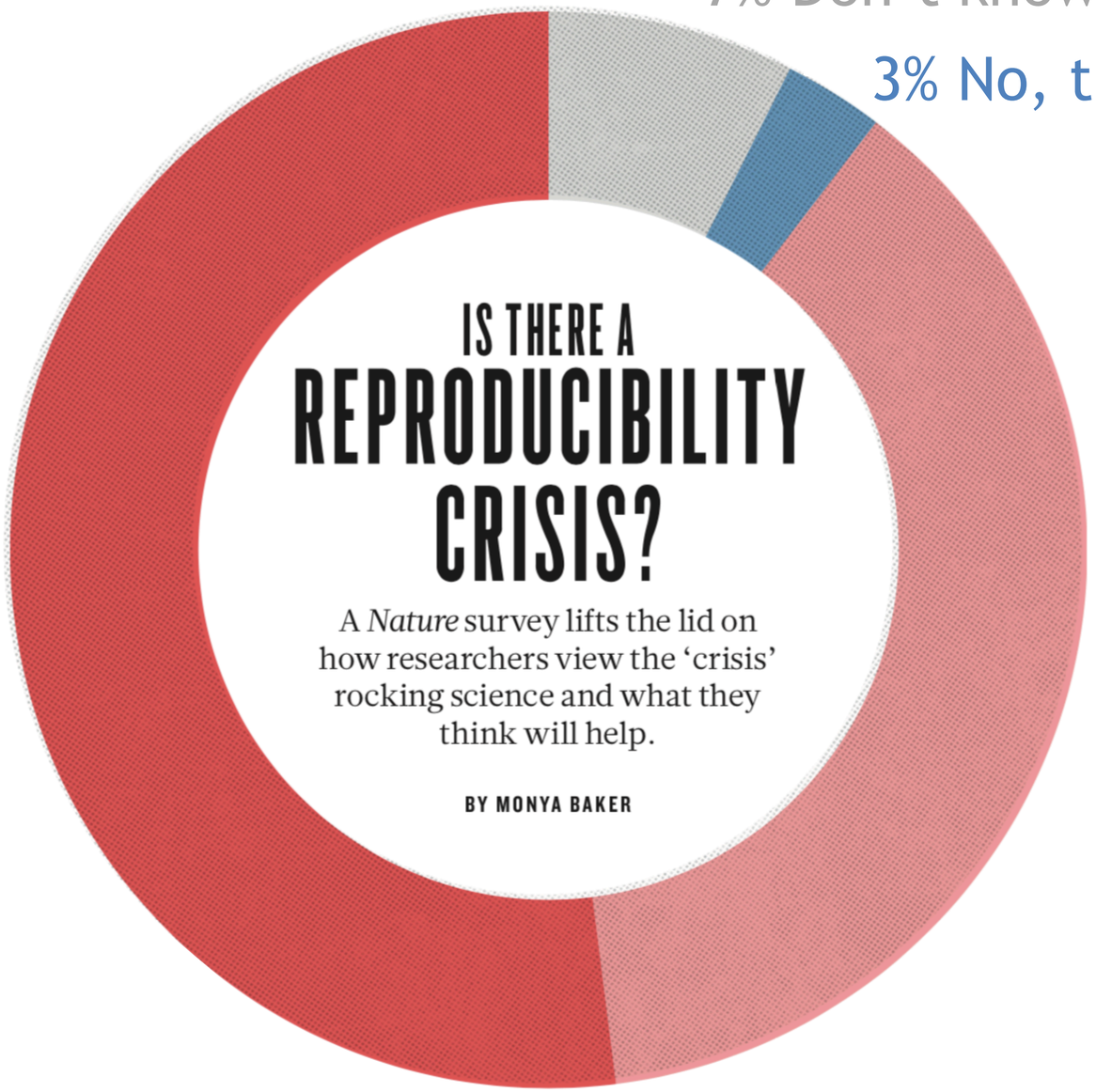
NTNU

7% Don’t know

3% No, there is no crisis

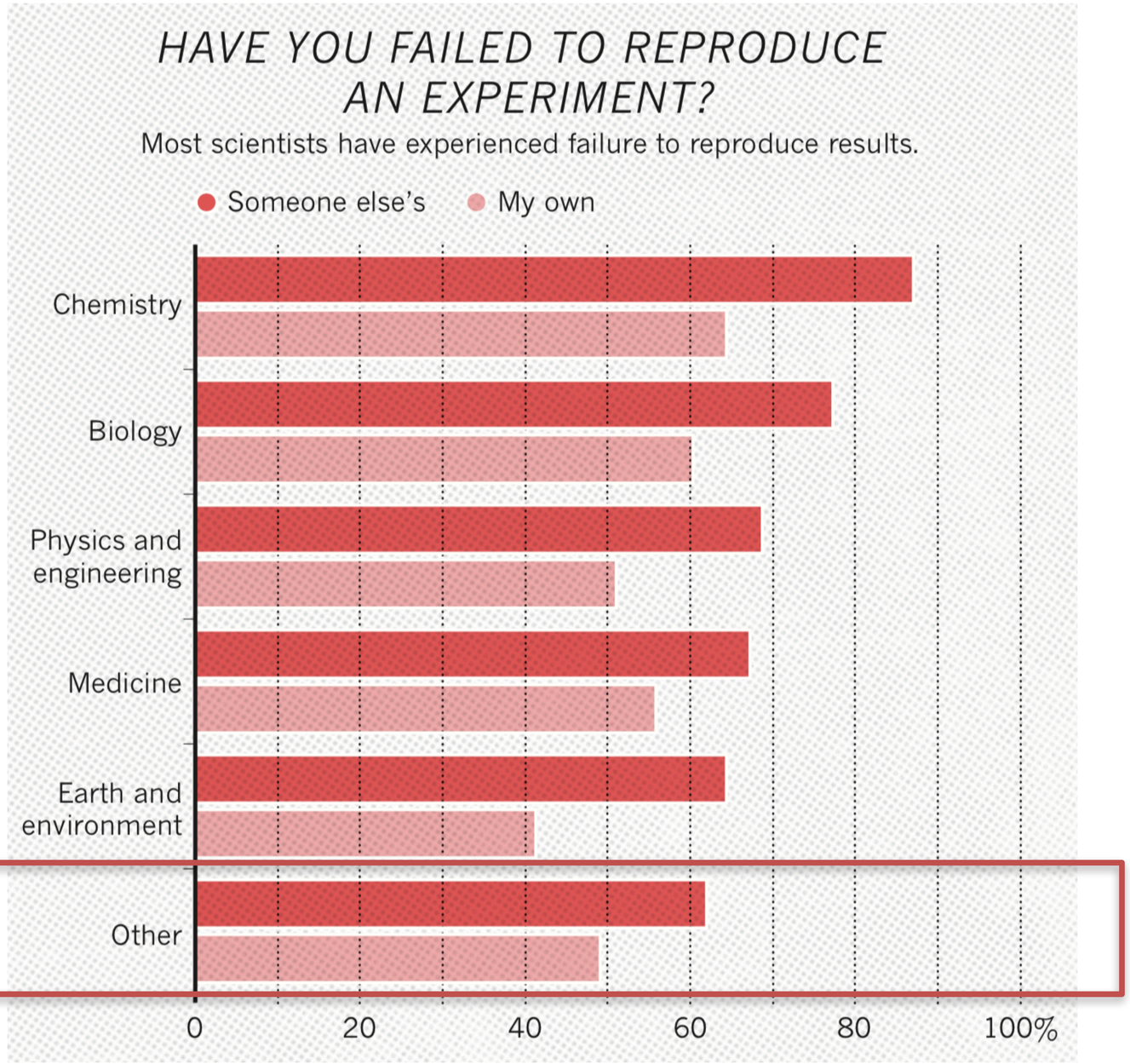
52%
Yes, a
significant
crisis



38%
Yes, a
slight crisis

1,576
RESEARCHERS SURVEYED

(M. Baker, Nature, 2016)

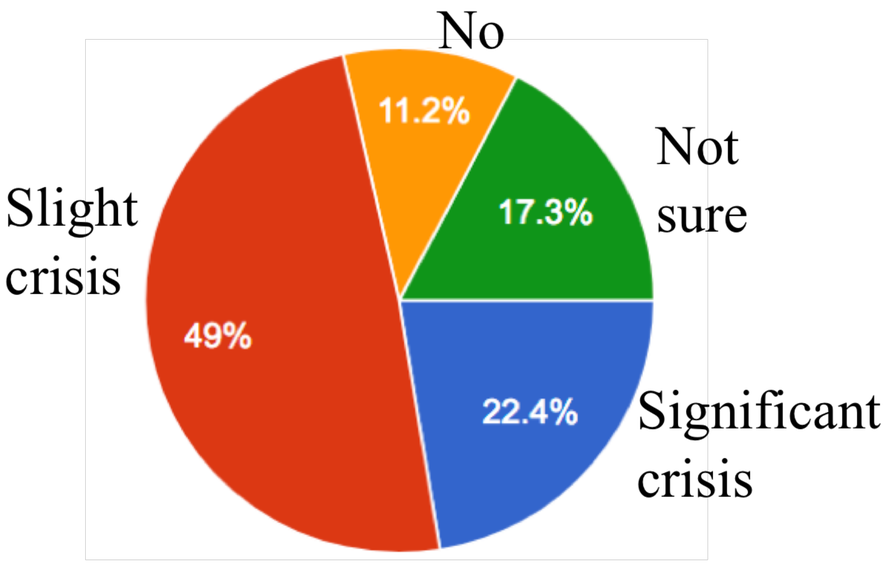


Computer Science

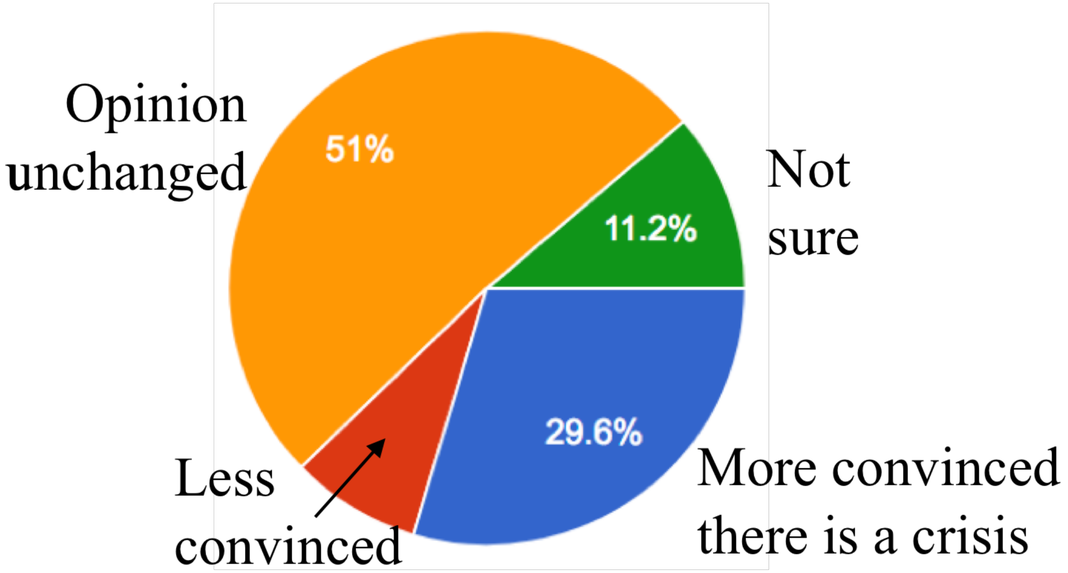
(M. Baker, Nature, 2016)

ICLR 2018 Reproducibility Challenge

Before the challenge (n=98):
"Is there a reproducibility crisis in ML?"



After the challenge (n=98):
"Has your opinion changed?"



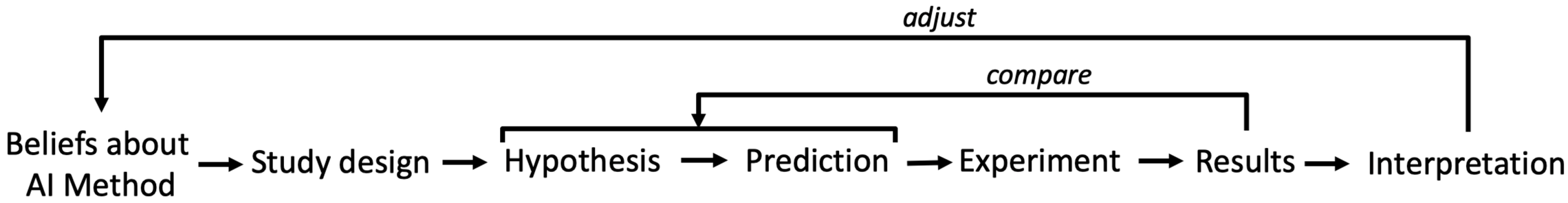
(J. Pineau, ICLR keynote, 2018)

How can we know it is shoulders we stand on?

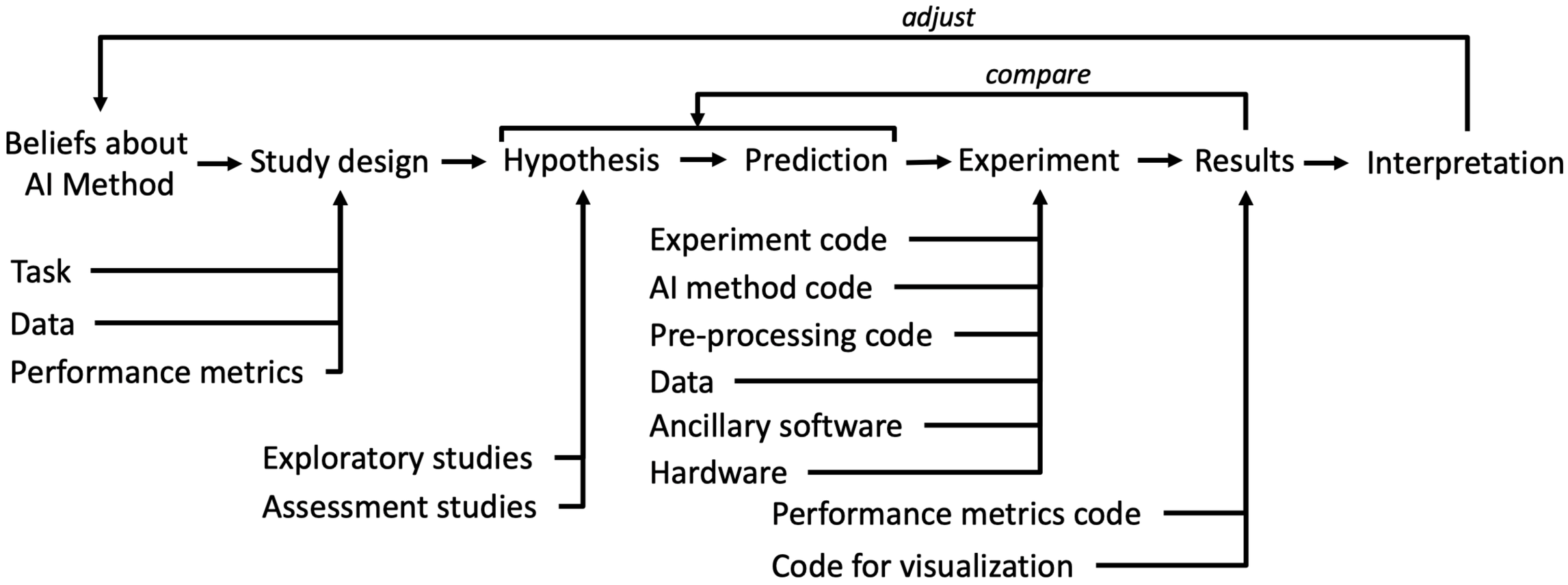
Odd Erik Gundersen, dr. philos.
Chief AI Officer, TrønderEnergi AS
Adjunct Associate Professor, NTNU
odderik@ntnu.no

norwegian open ai lab

The Scientific Method in AI Research

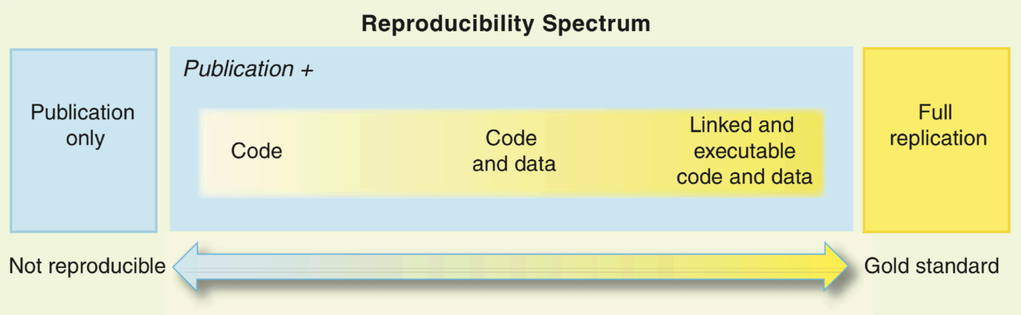


The Scientific Method in AI Research



NTNU

Many Definitions of Reproducibility



(R. D. Peng, Science, 2011)

(V. Stodden, Amstat News, 2011)

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.

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.

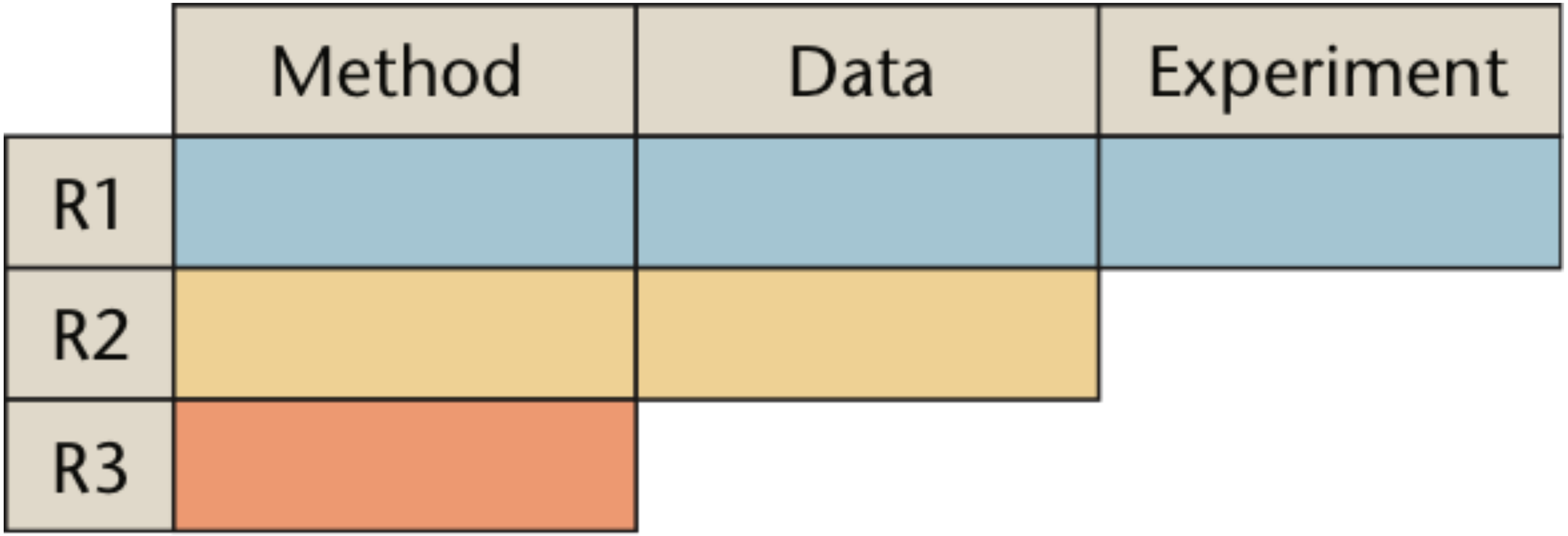
Inferential reproducibility: The drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.

(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.

Degree of Reproducibility



NTNU

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?
	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
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	Method source code	Is the AI system code available open source?
	Experiment source code	Is the experiment code available open source?
	Software dependencies	Are software dependencies specified?
	Hardware	Is the hardware used for conducting the experiment specified?

NTNU

NTNU

Reproducibility Metrics

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

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

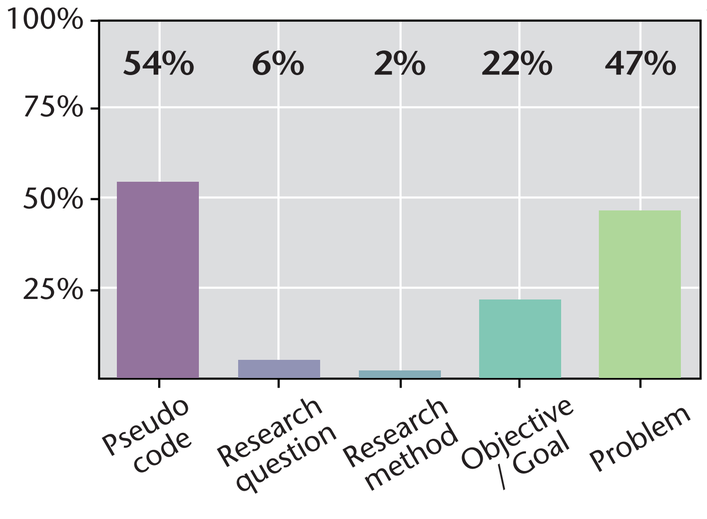
$$R3D(e) = Method(e)$$

NTNU

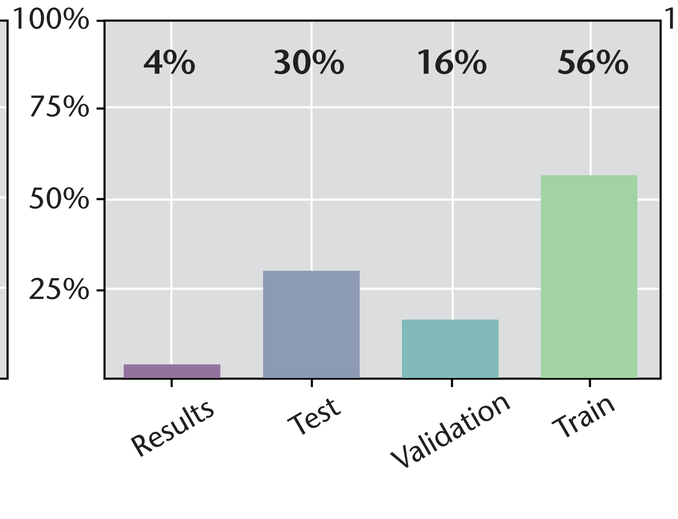
WHAT WE GAIN

We Can Specify How Well Research is Documented

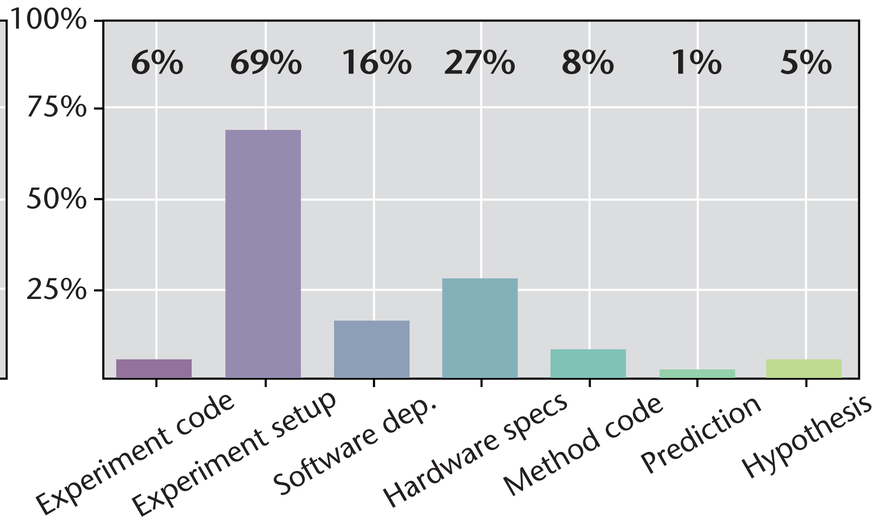
Method



Data

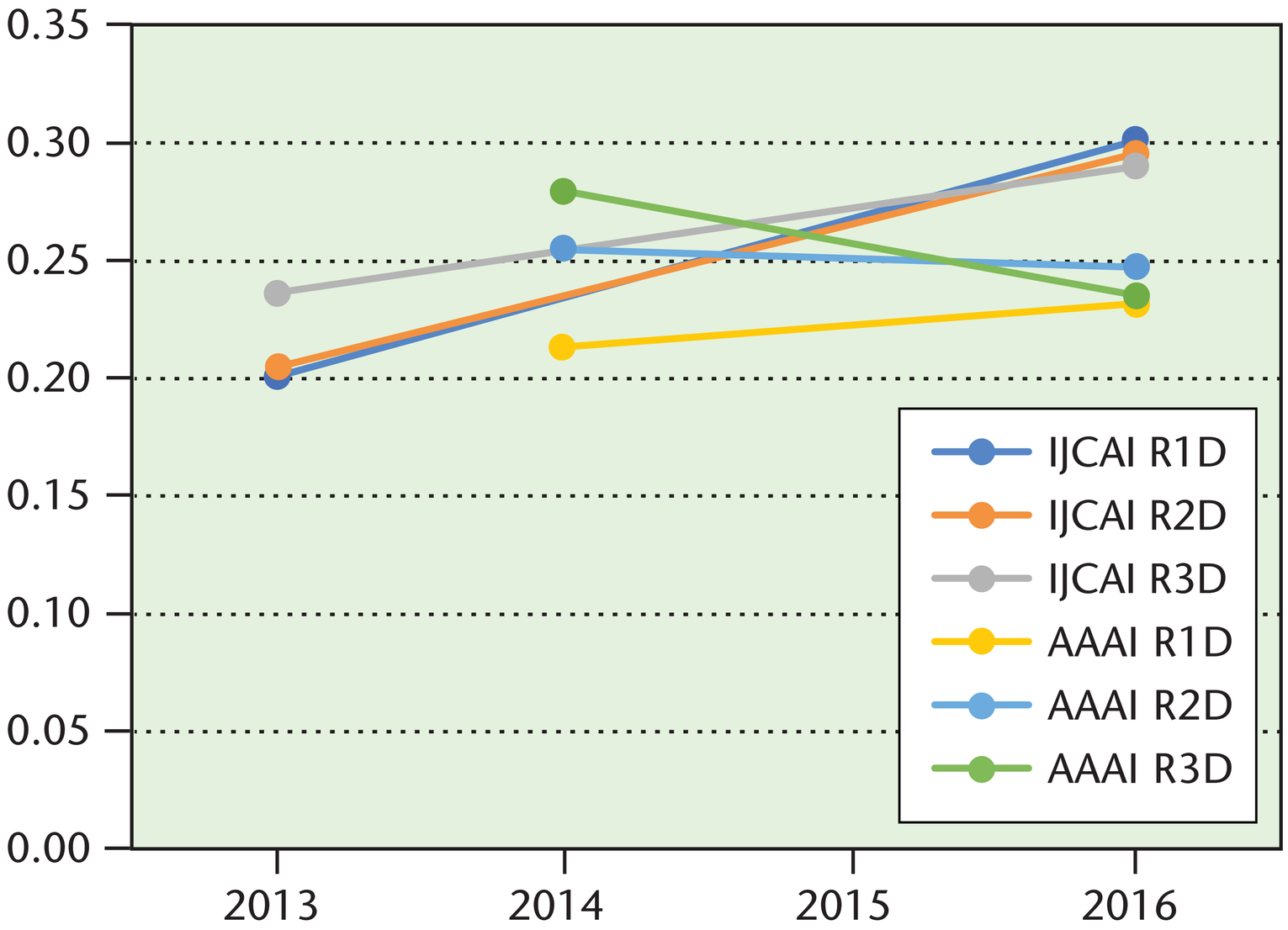


Experiment



(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)

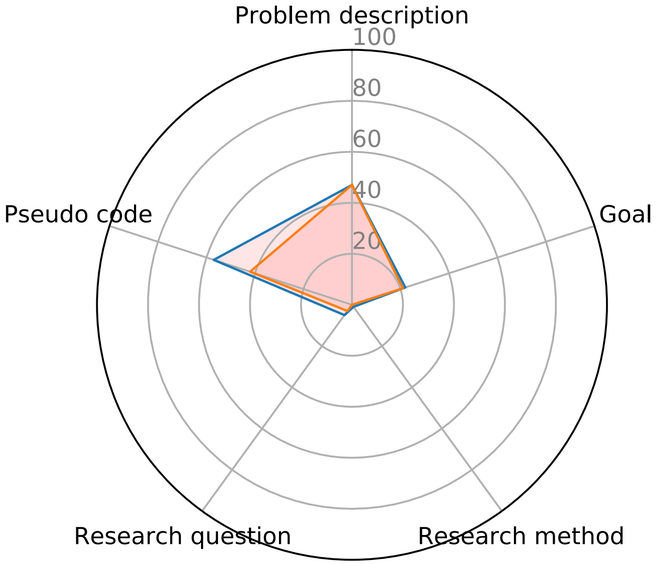
We Can Compare Research: Conferences

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

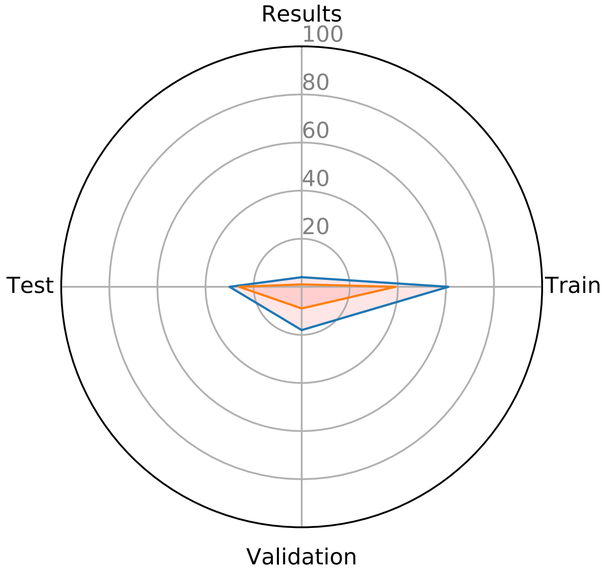
(Gundersen, Kjensmo, AAAI, 2018)

We Can Compare Research: Groups

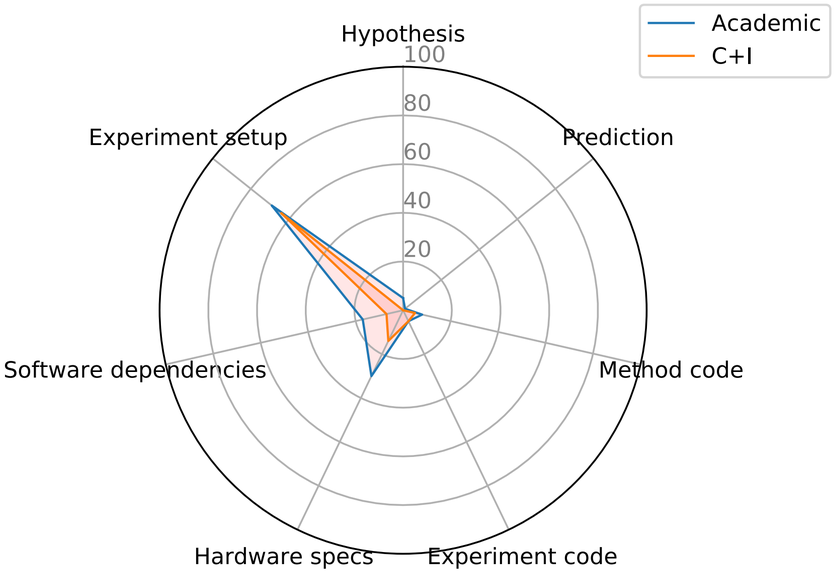
Academia versus Industry



Method



Data



Experiment

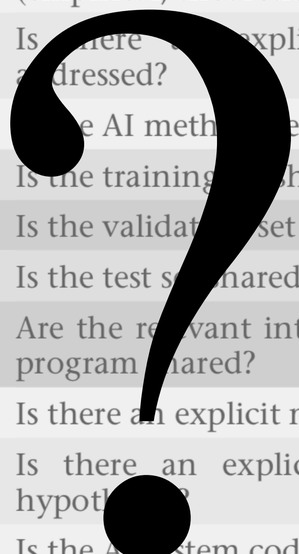
(Gundersen, AI Magazine, forthcoming)

We Can Compare Software Frameworks

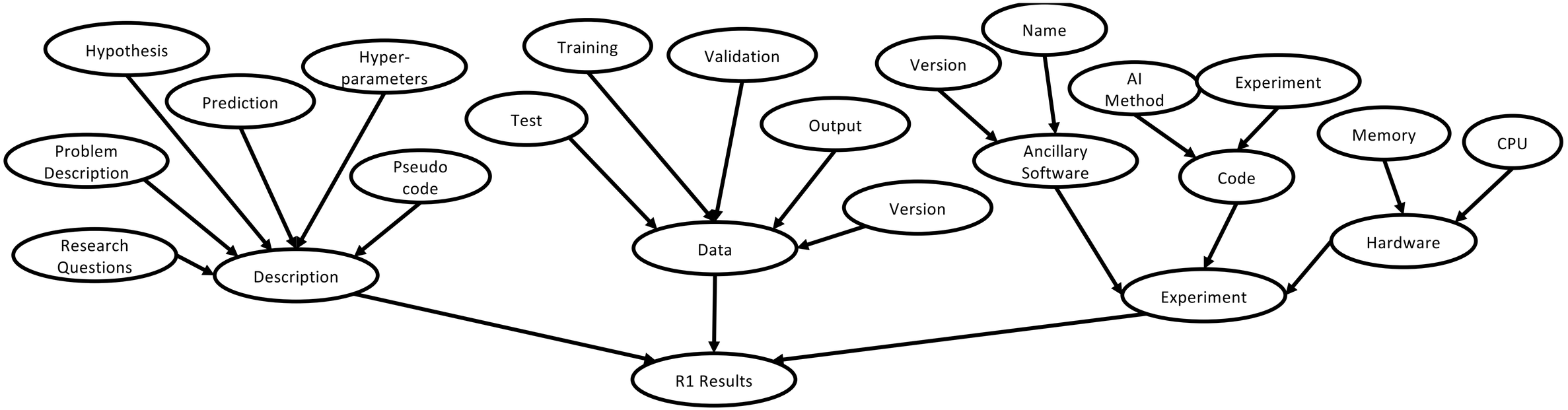
(Isdahl et al, forthcoming)

We Could Empirically Find What Entails Well-Documented Research

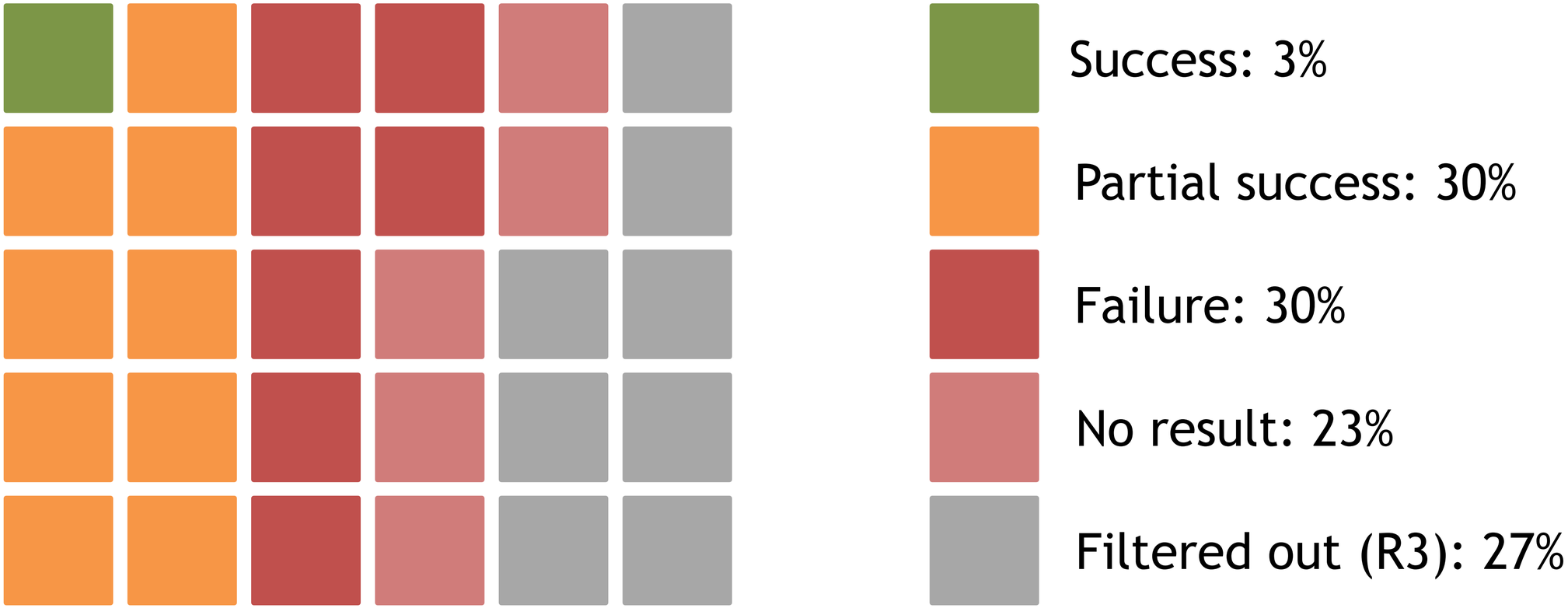
Factor	Variable	Description
Method	Problem	Is there an explicit mention of the problem the research seeks to solve?
	Objective	Is the research objective explicitly mentioned?
	Research method	Is there an explicit mention of the research method used (empirical, theoretical)?
	Research questions	Is there an explicit mention of the research question(s) addressed?
	Pseudocode	Are AI methods 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 hypothesis?
	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

EXPERIMENTS

Experiment I and II

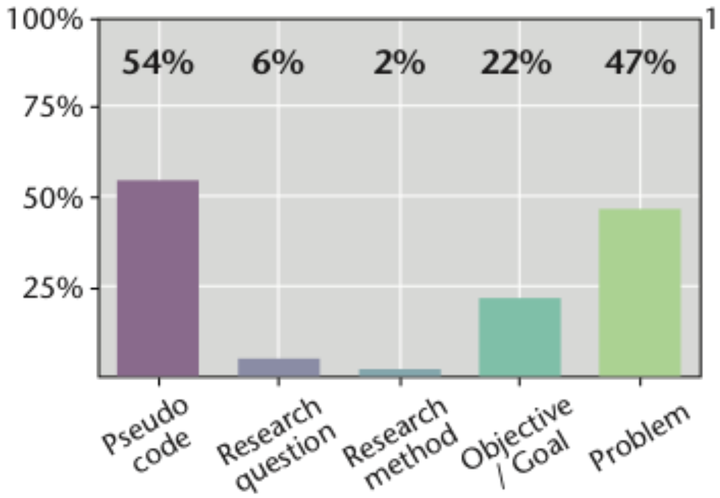
- We surveyed 400 papers.
- 100 papers from each installment of AAAI 2014, AAAI 2016, IJCAI 2013 and IJCAI 2016.
- Six reproducibility metrics proposed for quantifying the reproducibility.

NTNU

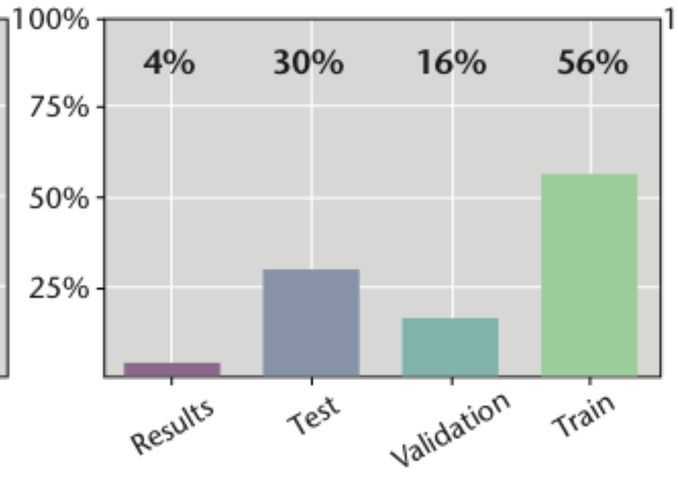
(Gundersen, Kjensmo, AAAI, 2018)

Results I: Factors and Variables

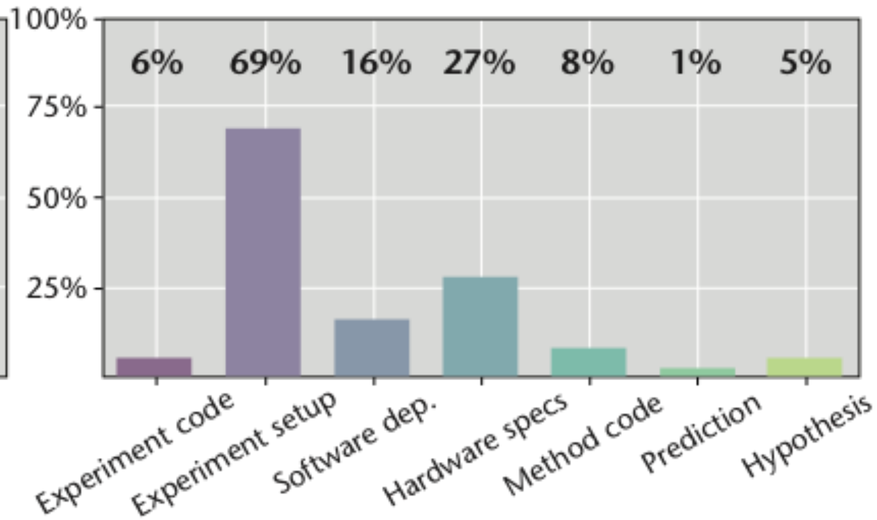
Method



Data

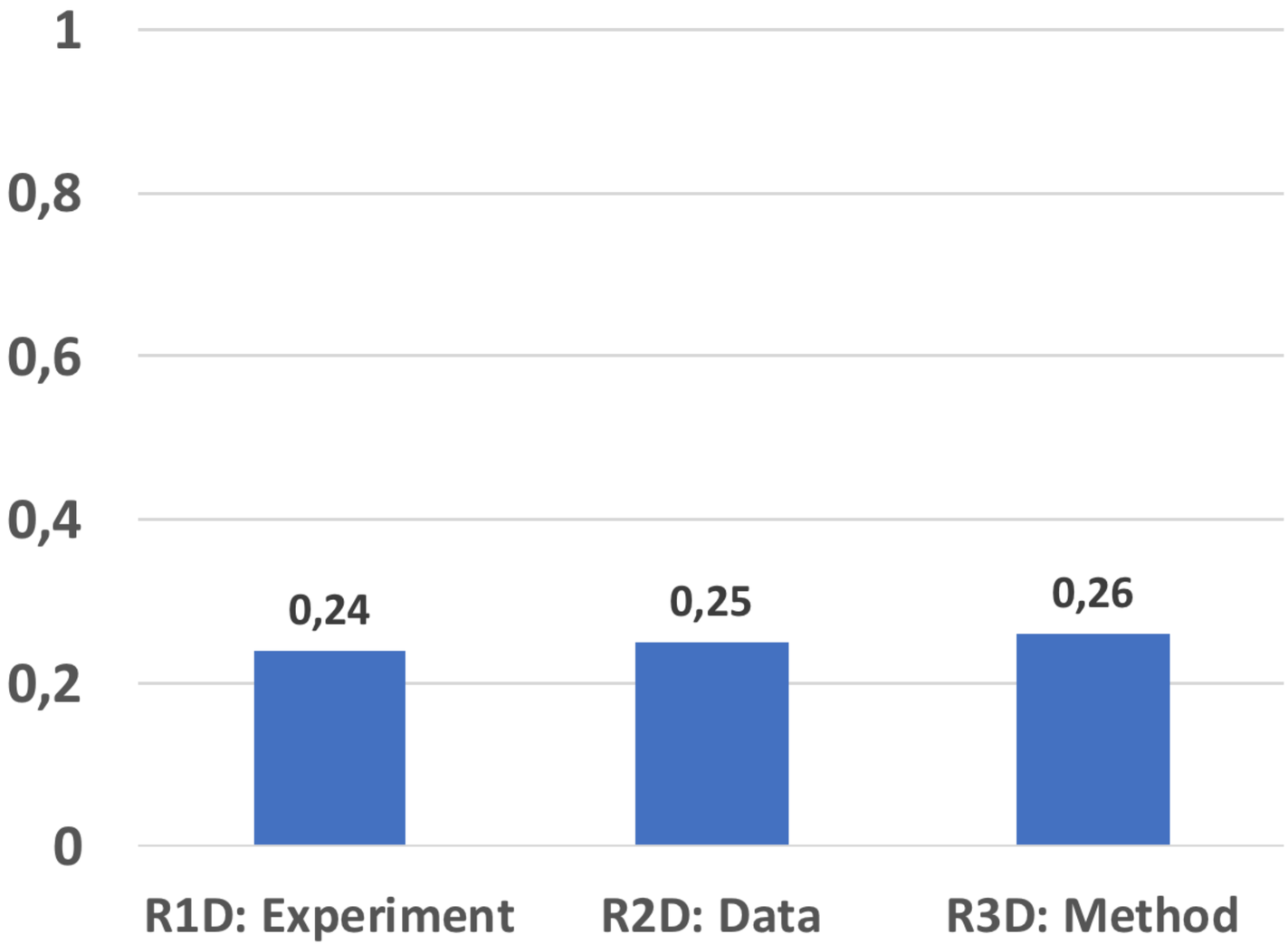


Experiment



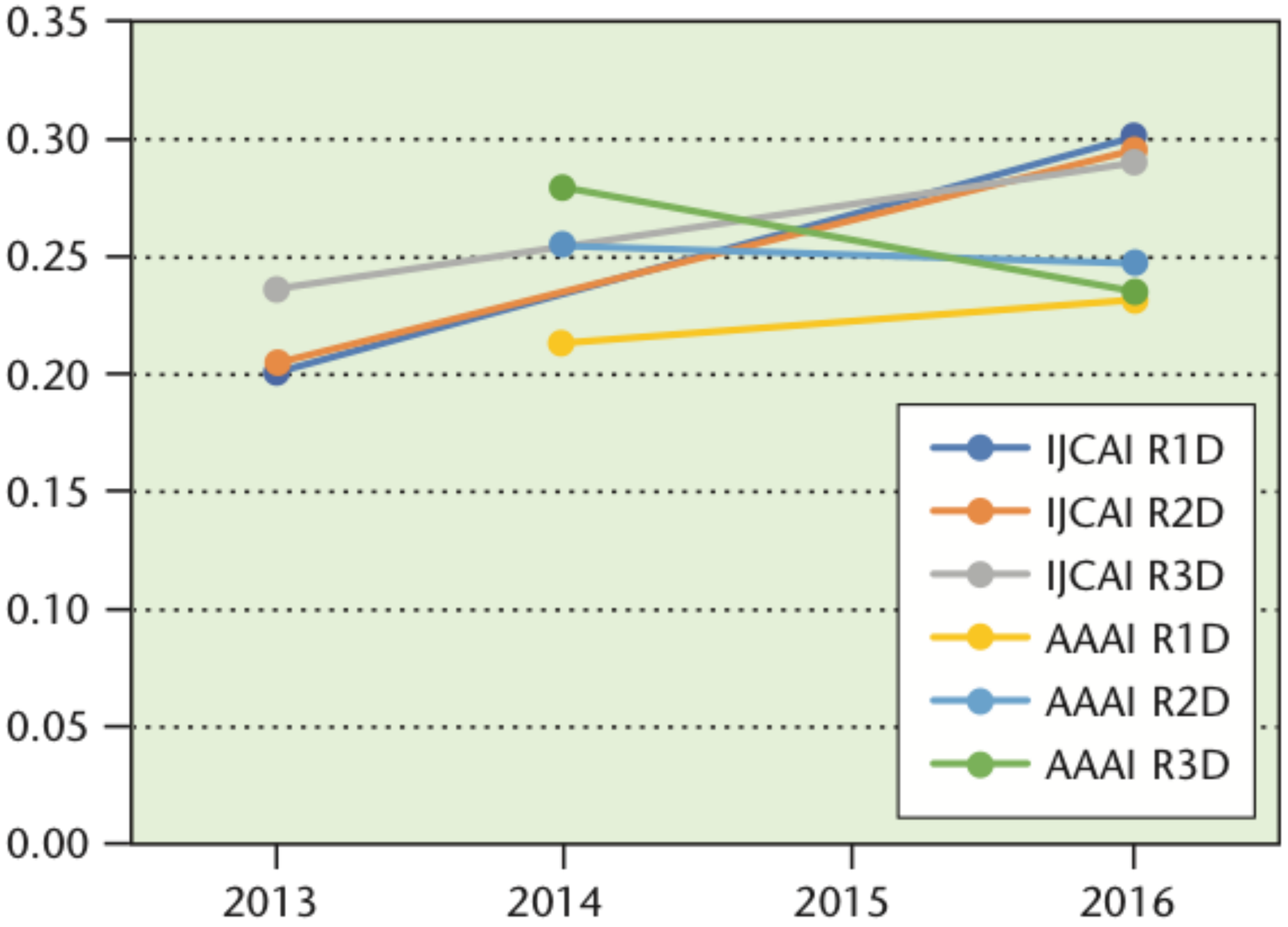
(Gundersen, Kjensmo, AAAI, 2018)

Results II: Reproducibility Degree



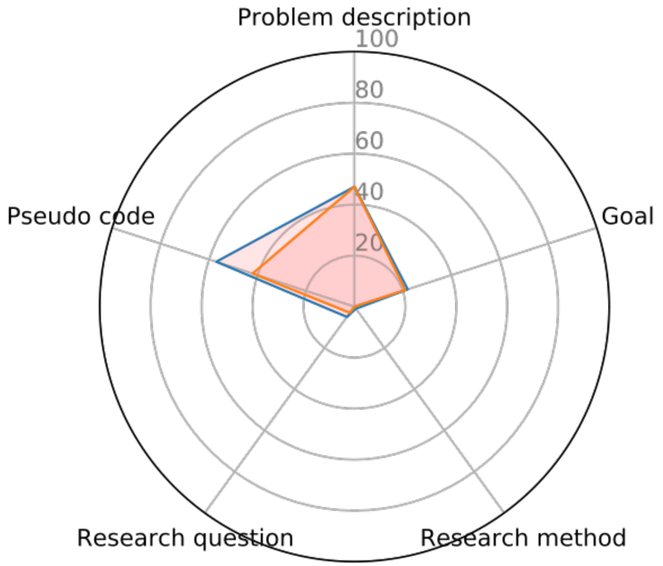
(Gundersen, Kjensmo, AAAI, 2018)

Results III: Change over Time

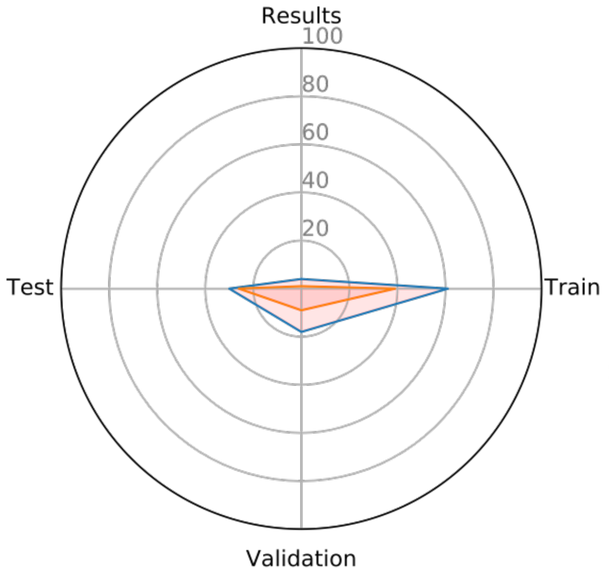


(Gundersen, Kjensmo, AAAI, 2018)

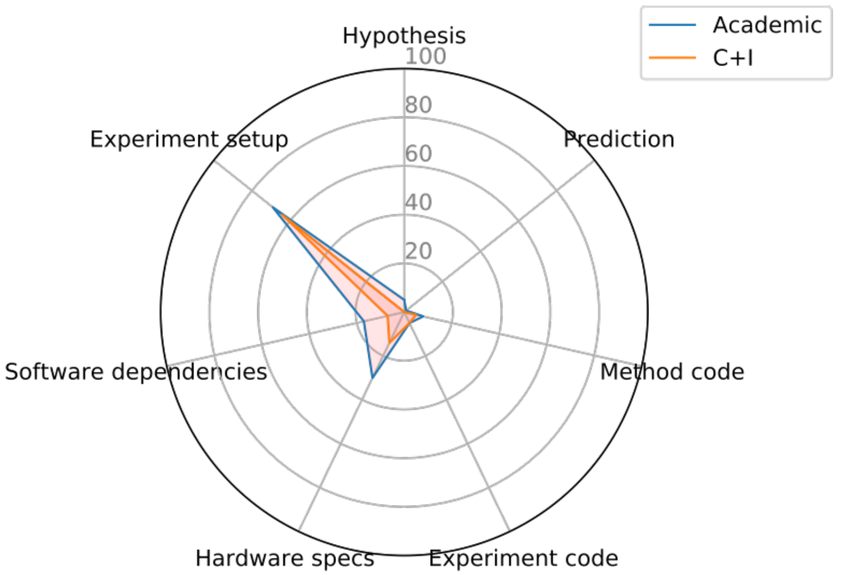
Results IV: Industry vs Academia



Method



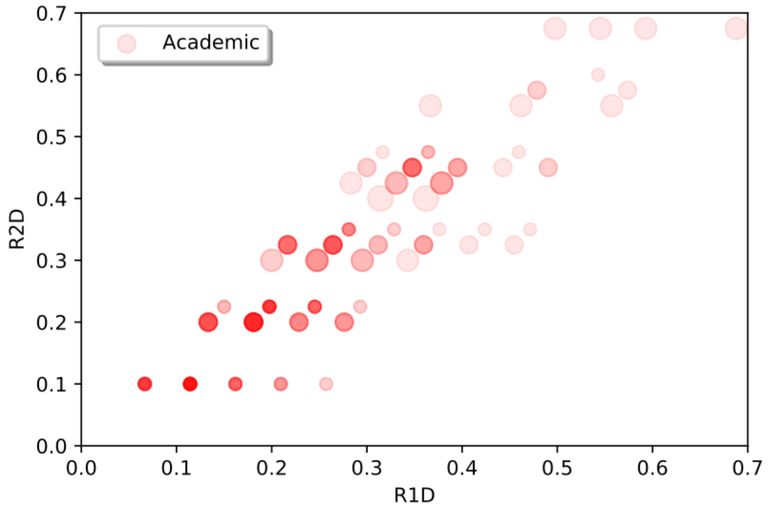
Data

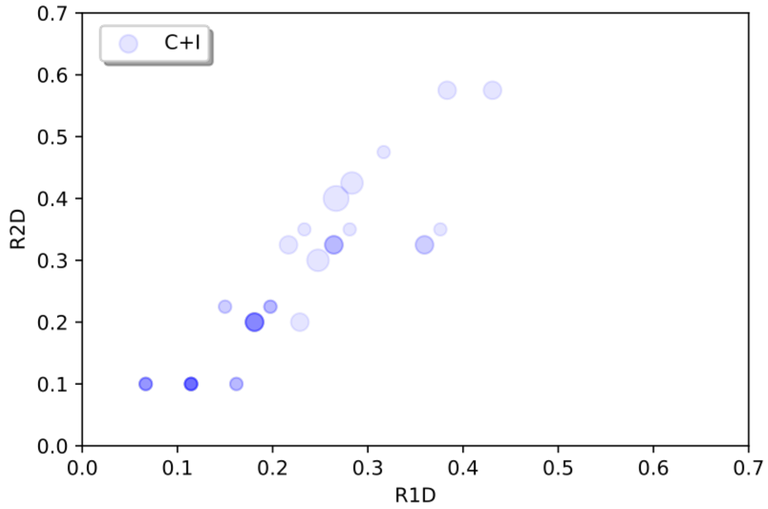


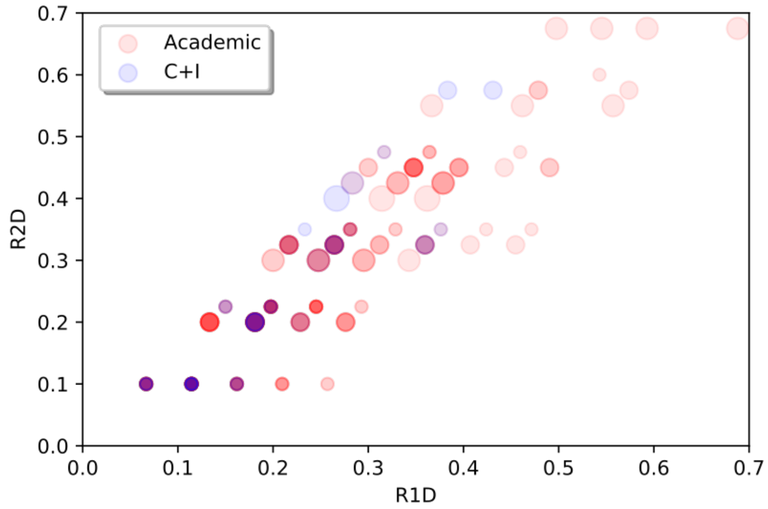
Experiment

(Gundersen, AI Magazine, 2019)

Results V: Industry vs Academia

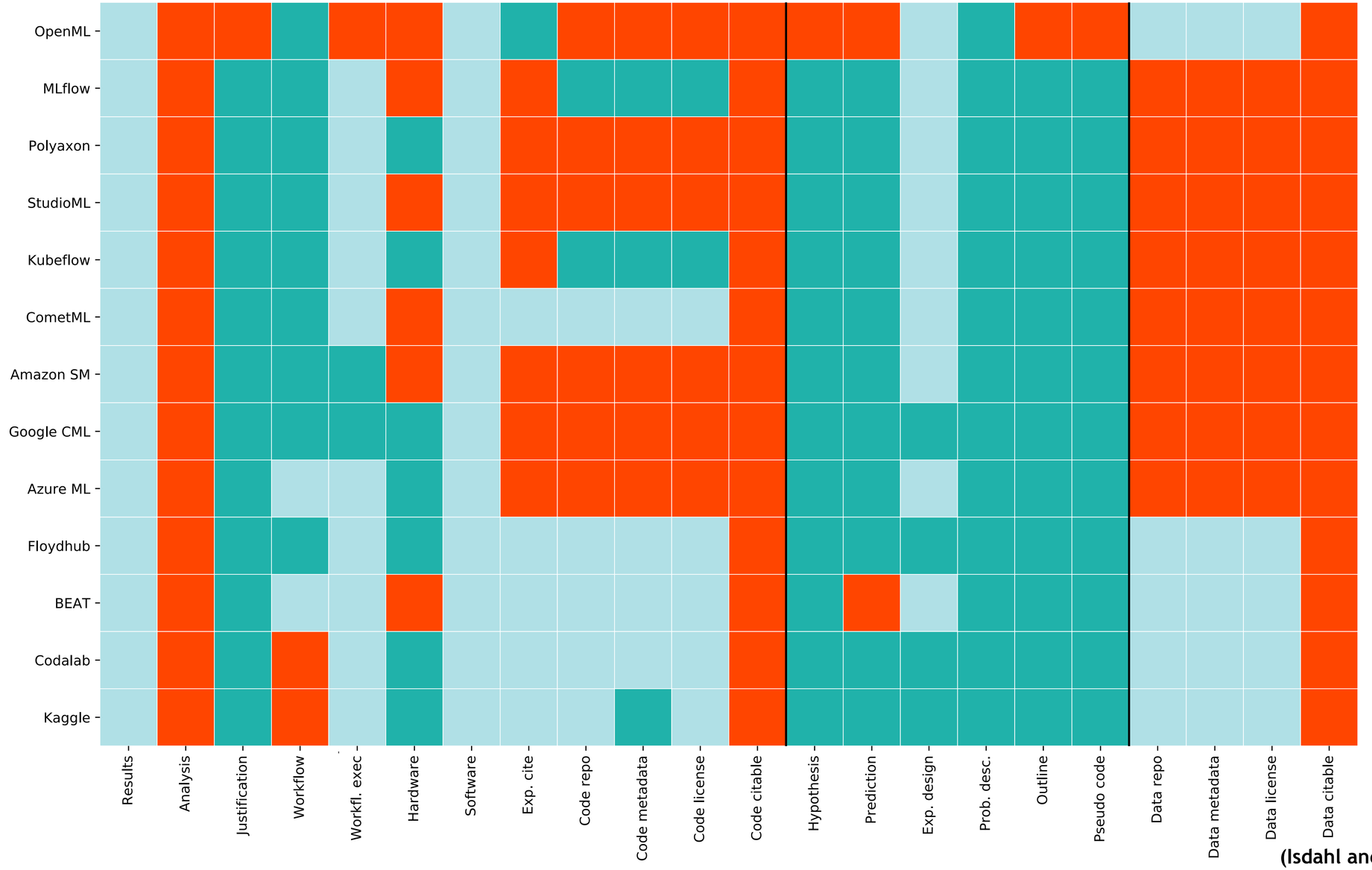




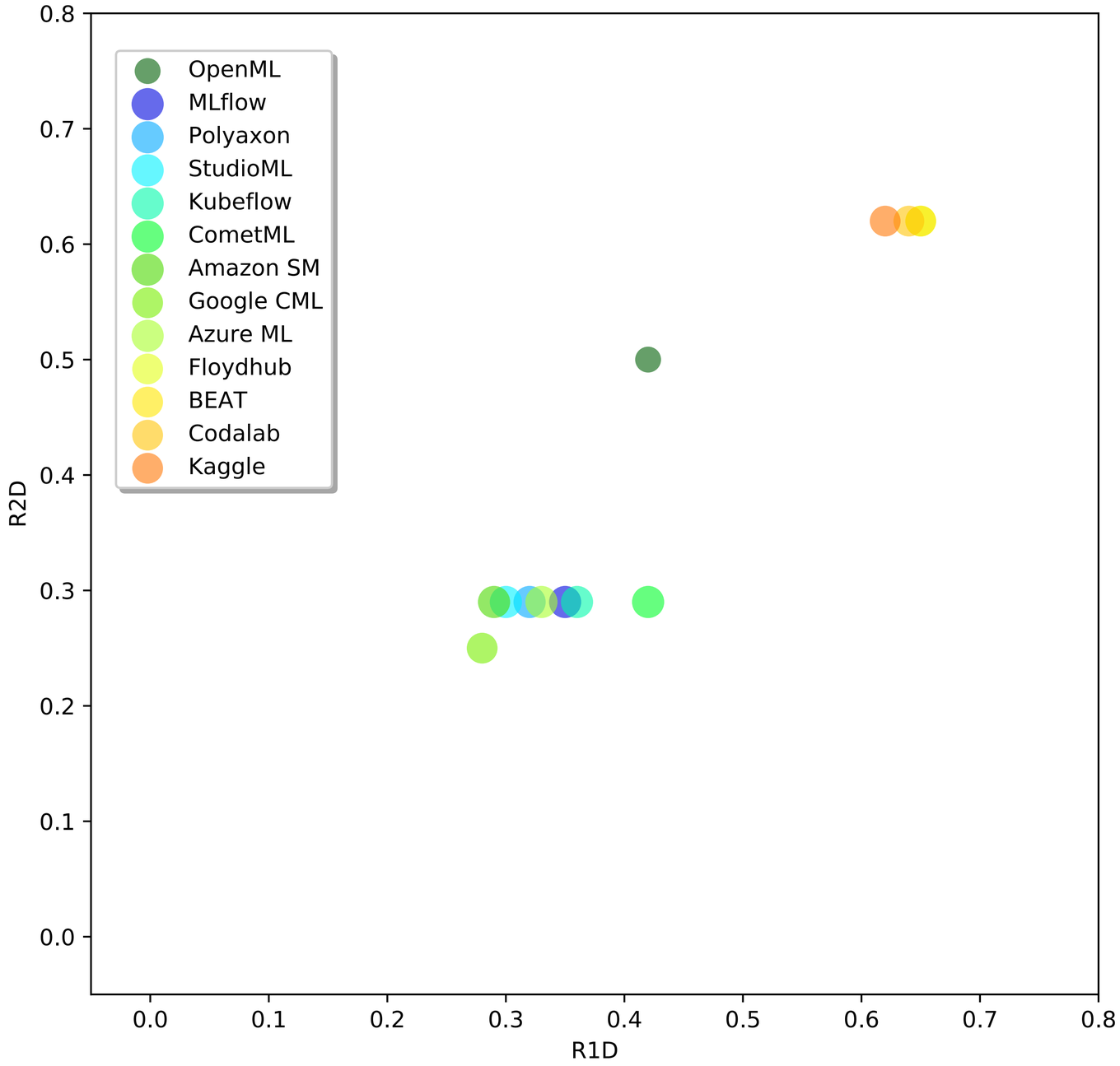


(Gundersen, AI Magazine, 2019)

Experiment III



(Isdahl and Gundersen, eScience, 2019)


OpenML
MLflow
Polyaxon
StudioML
Kubeflow
CometML
Amazon SM
Google CML
Azure ML
Floydhub
BEAT
Codalab
Kaggle
R2D
R1D
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Experiment IV

- We selected 30 papers to reproduce
- Ten most cited AI papers from 2012, 2014 and 2016 based on numbers from Scopus.
- Structured research procedure.

(Gundersen et al, forthcoming)

Research Procedure

- Reproduce research that shared code and data or data (filtered out R3 papers).
- Time-boxed the work put into each research paper to 40 hours effective work time.
- Stopping criteria (computing resources, paywall data sets, only qualitative results presented).

(Gundersen et al, forthcoming)

Results: Outcome per paper

Success: 20%

Partial success: 13%

Failure: 23%

No result: 17%

Filtered out (R3): 27%

(Gundersen et al, forthcoming)

Top Six Causes of Failure

- Aspect of *implementation* not described or ambiguous (R2).
- Aspect of *experiment* not described or ambiguous (R2).
- Not all *hyper-parameters* are specified (R2).
- Mismatch between *data* in paper and available online (R1+R2).
- Method code shared, experiment code not shared (R1).
- Method not described with enough detail (R2).

(Gundersen et al, forthcoming)

NTNU

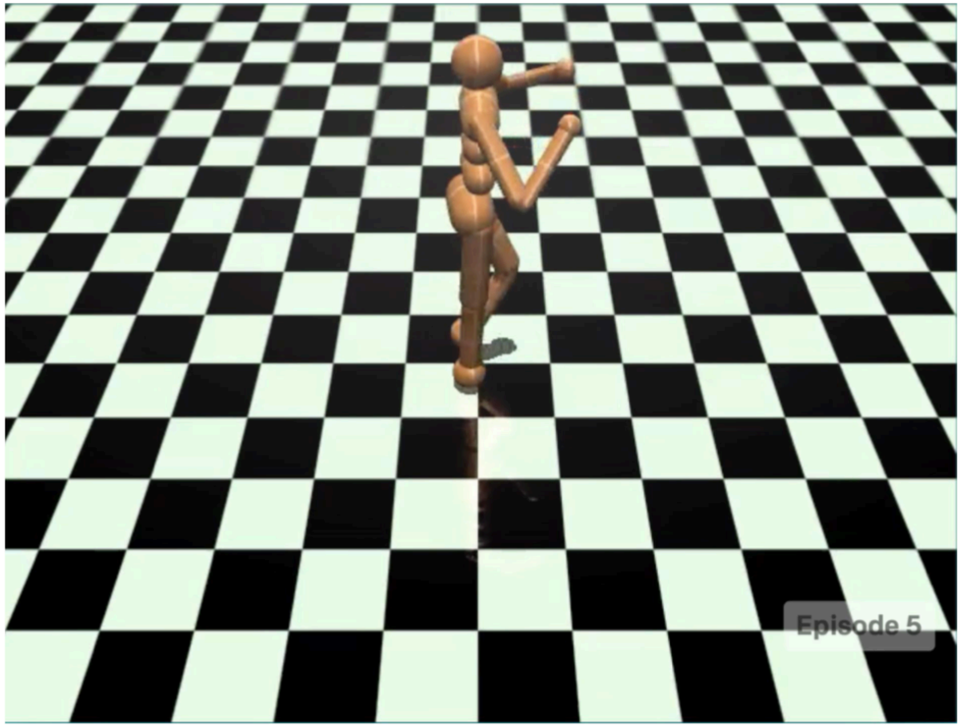
“IT IS MORE LIKE WE ARE STANDING ON EACH OTHERS FEET”

NTNU

EVALUATIONS

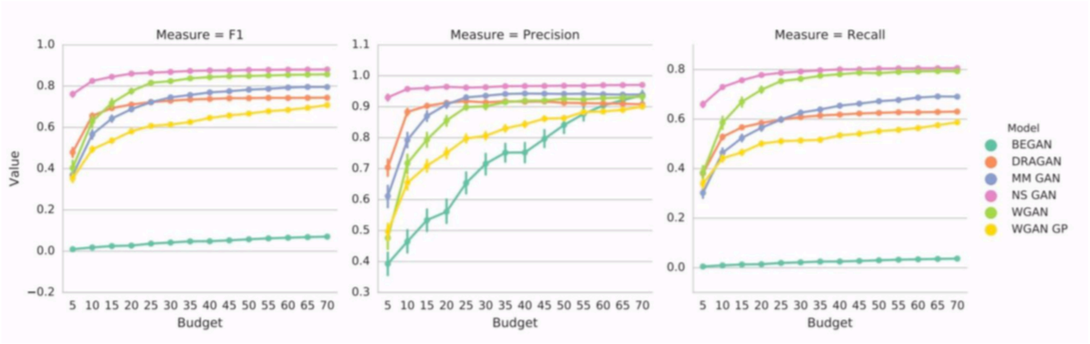
NTNU

Deep Reinforcement Learning that Matters



- Non-determinism in standard benchmark environments and
- Variance intrinsic to the method
- Cause irreproducible results.

Are GANs created equal?



- Study on models and evaluation measures.
- Most models can reach same performance given hyperparameter optimization and random restarts.
- Suggests more systematic and objective evaluation procedures.

NTNU

Software Dependency of Weather Model

TABLE 1. Computing environment including FORTRAN compilers, parallel communication libraries, and optimization levels of the compiler. Identical results are marked by a symbol. Ten ensemble members with different software system are highlighted in boldface.

Name	Machine	FORTRAN compiler	Parallel communication library	Optimization level	Mark
EXP1	**KISTI SUN2**	**INTEL 11.1**	**openmpi 1.4**	**O3**	□
	KISTI SUN2	INTEL 11.1	mvapich2 1.5	O3	□
EXP2	**KISTI SUN2**	**INTEL 11.1**	**mvapich1 1.2**	**O3**	○
	KISTI SUN2	INTEL 11.1	openmpi 1.4	O4	□
EXP3	**KISTI SUN2**	**INTEL 11.1**	**openmpi 1.4**	**O2**	△
EXP4	**KISTI SUN2**	**INTEL 11.1**	**openmpi 1.4**	**O1**	◁
EXP5	**KISTI SUN2**	**INTEL 11.1**	**openmpi 1.4**	**O0**	▷
EXP6	**KISTI SUN2**	**PGI 9.0.4**	**openmpi 1.4**	**O2 (-fastsse)**	■
	KISTI SUN2	PGI 9.0.4	mvapich2 1.5	O2 (-fastsse)	■
	KISTI SUN2	PGI 9.0.4	mvapich1 1.2	O2 (-fastsse)	■
	KISTI SUN2	PGI 8.0.6	mvapich1 1.2	O2 (-fastsse)	■
	YSU Cluster	PGI 10.6	mvapich1 1.2	O2 (-fastsse)	■
	YSU Cluster	PGI 10.6	mvapich1 1.2	O3 (-fastsse)	■
EXP7	**YSU Cluster**	**PGI 10.6**	**mvapich1 1.2**	**O1**	●
EXP8	**YSU Cluster**	**PGI 7.1.6**	**mvapich1 1.2**	**O2 (-fastsse)**	▲
EXP9	**KISTI IBM 1**	**XLF 10.1**	**—**	**O3**	★
	KISTI IBM 2	XLF 12.1	—	O3	★
	KISTI IBM 1	XLF 10.1	—	O4	★
EXP10	**KISTI IBM 1**	**XLF 10.1**	**—**	**O2**	♠
	KISTI IBM 1	XLF 10.1	—	O1	♠

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, Winter 2019.
- **Out-of-the-box Reproducibility: A Survey of Machine Learning Platforms**, R. Isdahl and O. E. Gundersen, eScience 2019.
- **What We Learned When Reproducing the Most Cited AI Research**, O. E. Gundersen, O. Cappelen, N. Grimstad, M. Mølnå, forthcoming.

NTNU Odd Erik Gundersen odderik@ntnu.no