Tutorial@IJCAI19

What I talk about when I talk about 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

TrønderEnergi®

norwegian
open ai lab

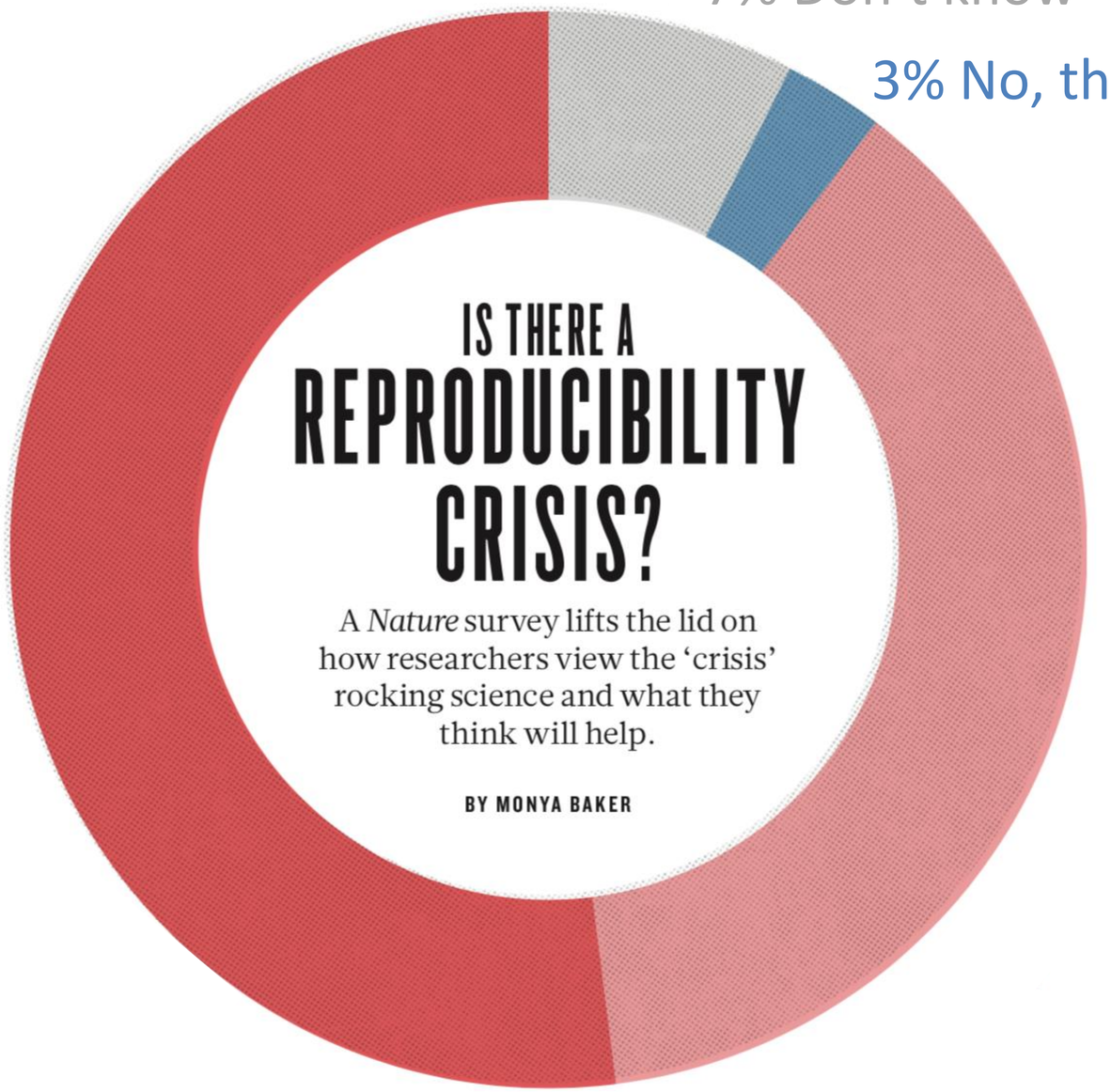
NTNU

INTRODUCTION

7% Don’t know

3% No, there is no crisis

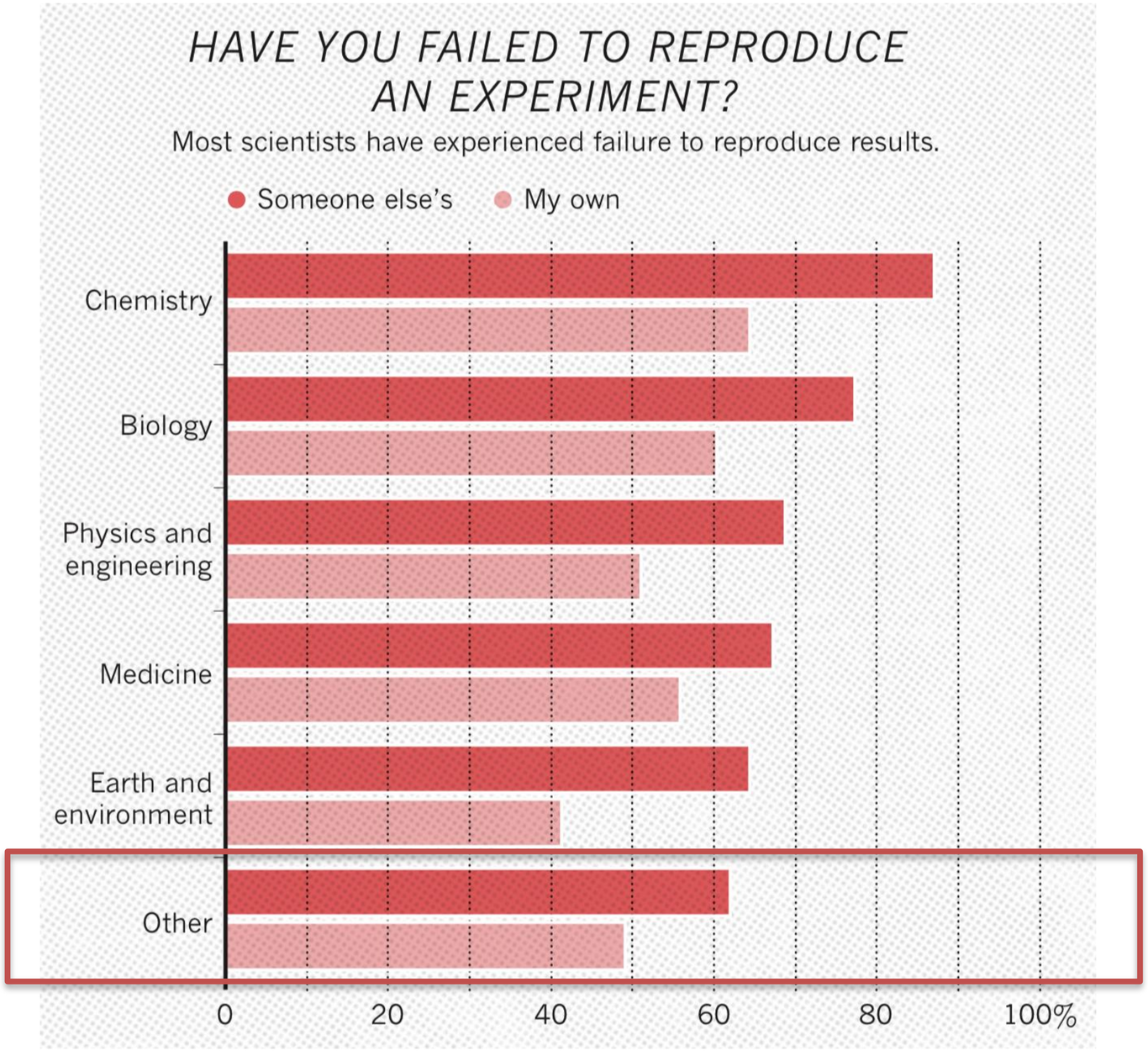
52%
Yes, a significant crisis



38%
Yes, a slight crisis

1,576
RESEARCHERS SURVEYED

(M. Baker, Nature, 2016)

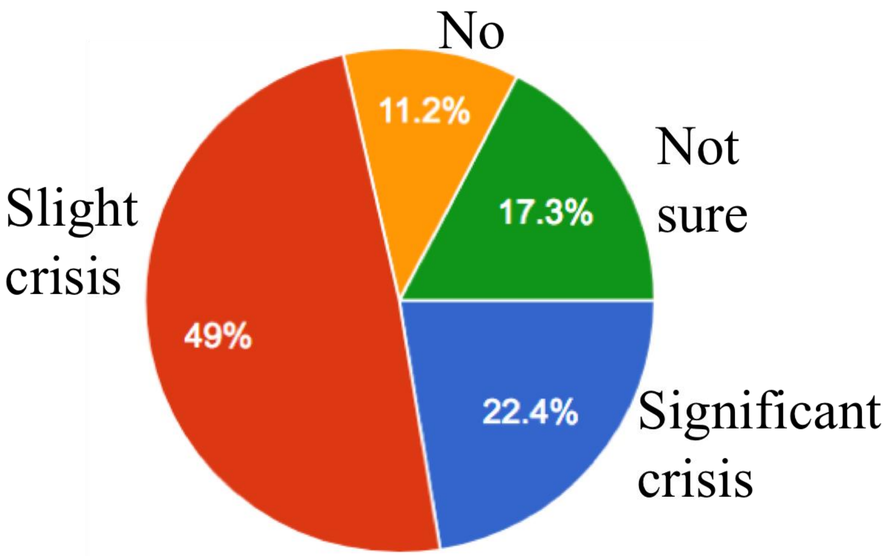


Computer Science

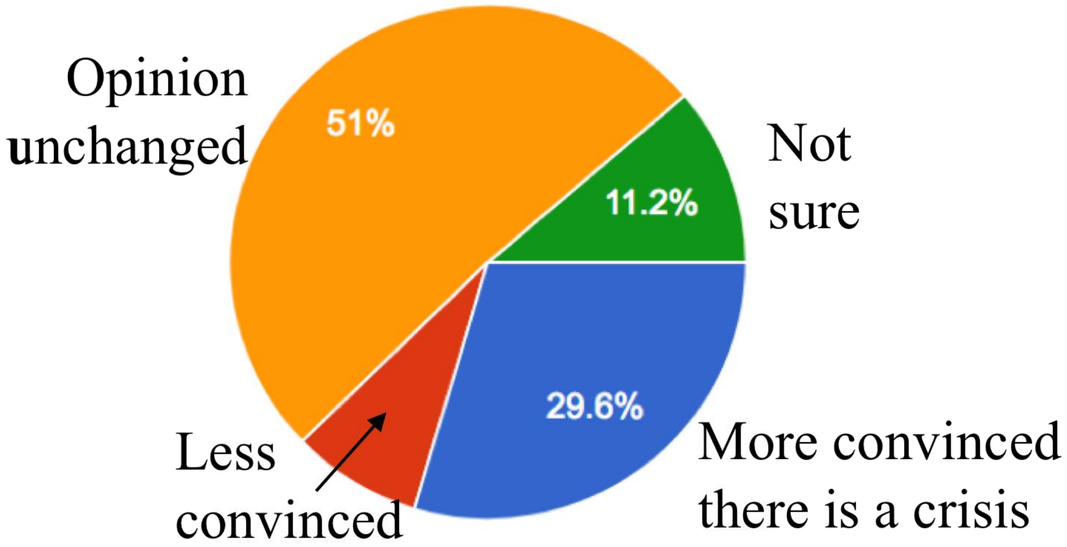
(M. Baker, Nature, 2016)

ICLR 2018 Reproducibility Challenge

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

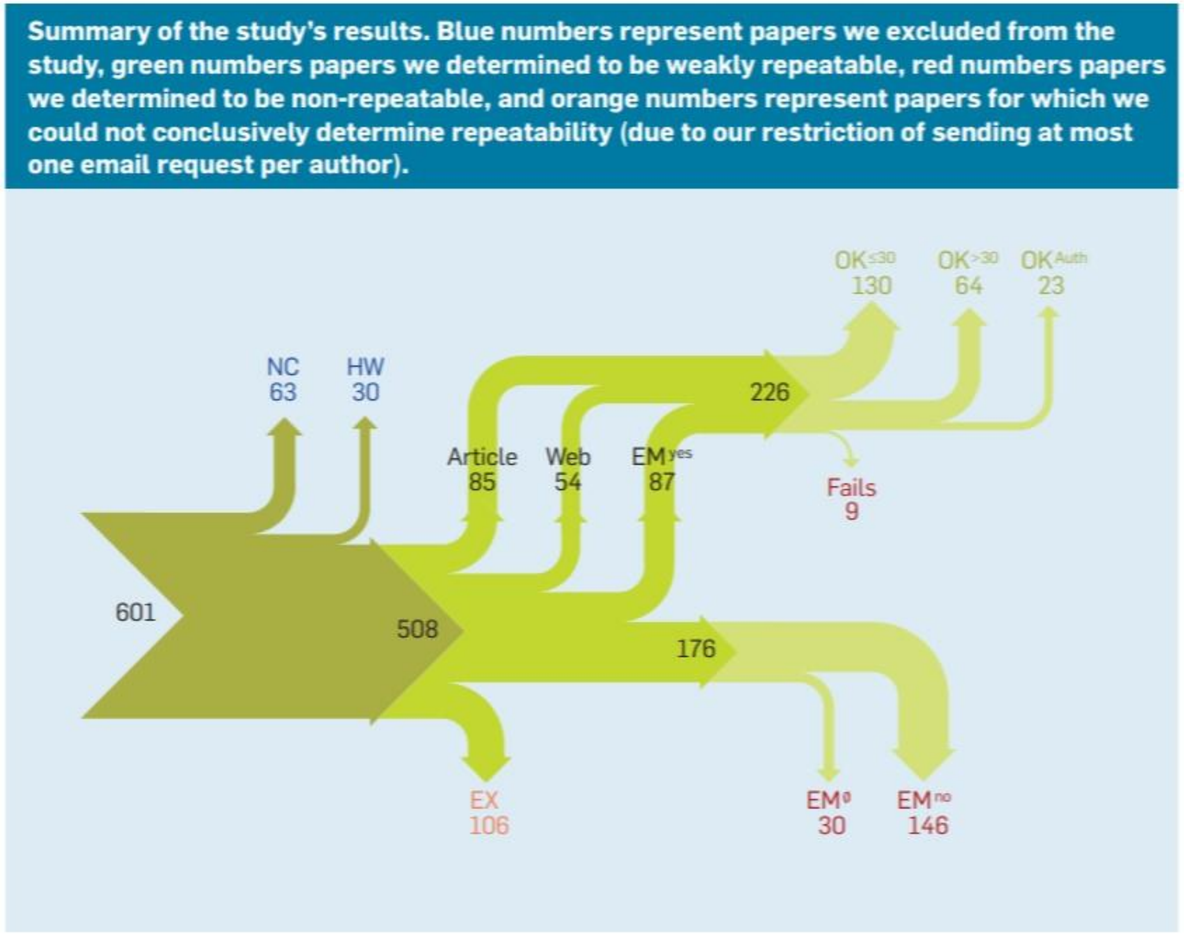


After the challenge (n=98):
"Has your opinion changed?"



(J. Pineau, ICLR keynote, 2018)

Repeatability in Comp. Sys. Research



- Analyzed 601 ACM papers.
- Out of these
- Locate and build source code.
- Able to for 32.3% w.o. communicating with authors.
- Increase to 48.3% with communication.

(C. Collberg and T. A. Proebsting, Communications of the ACM, 2016)

Journal Policy Effectiveness Analysis

Table 1. Responses to emailed requests (n = 180)

Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	20	11
Asked for reasons	20	11
Refusal to share	12	7
Directed back to supplement	6	3
Unfulfilled promise to follow up	5	3
Impossible to share	3	2
Shared data and code	65	36
Email bounced	3	2
No response	46	26

- Science policy to include code and data.
- Requested from 204 papers from Science.
- Obtained artifacts from 44%.
- Able to reproduce 26%.

(C. Collberg and T. A. Proebsting, Communications of the ACM, 2016)

NTNU

PART 1

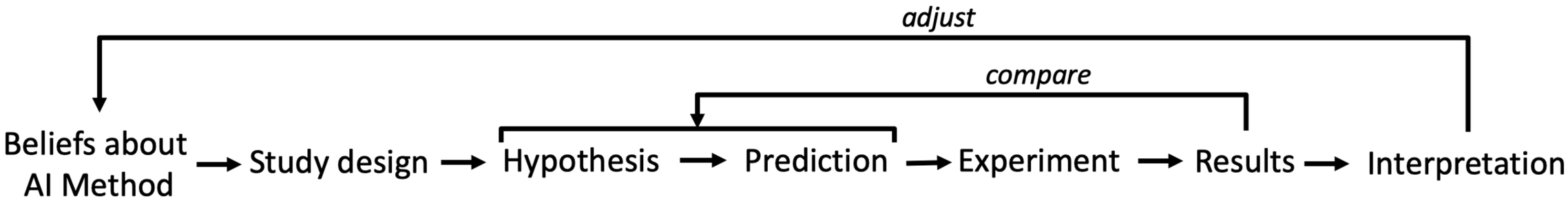
UNDERSTANDING REPRODUCIBILITY

NTNU

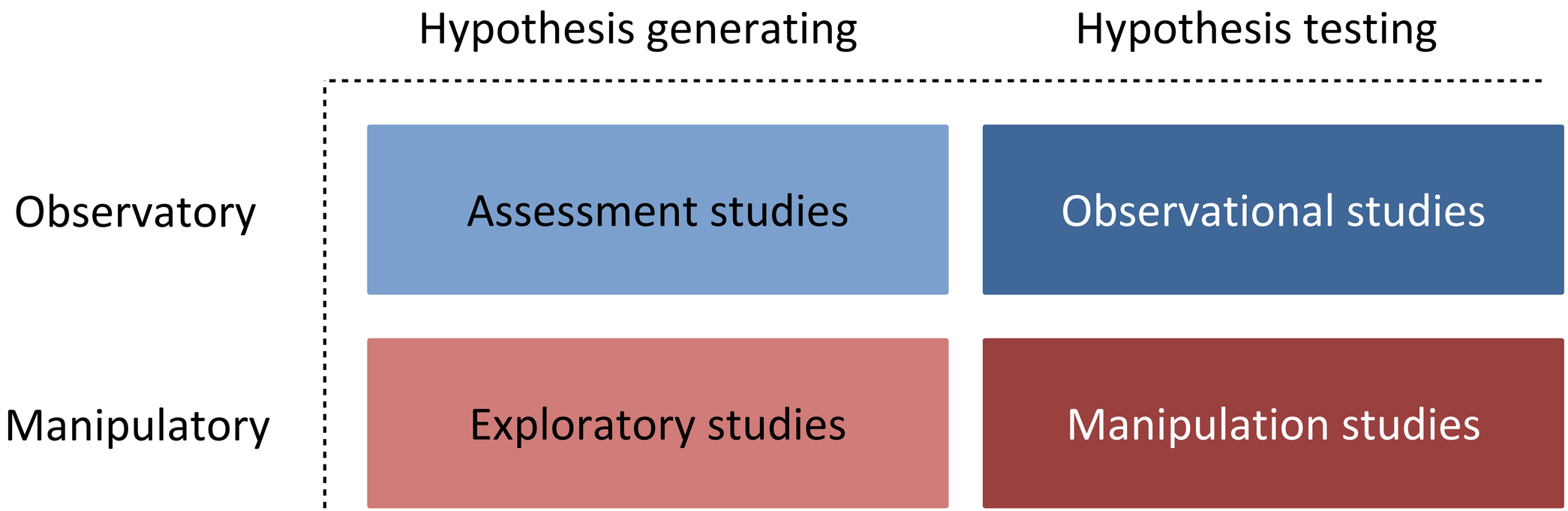
Reproducibility

SO WHAT IS IT?

The Scientific Method in Empirical AI Research

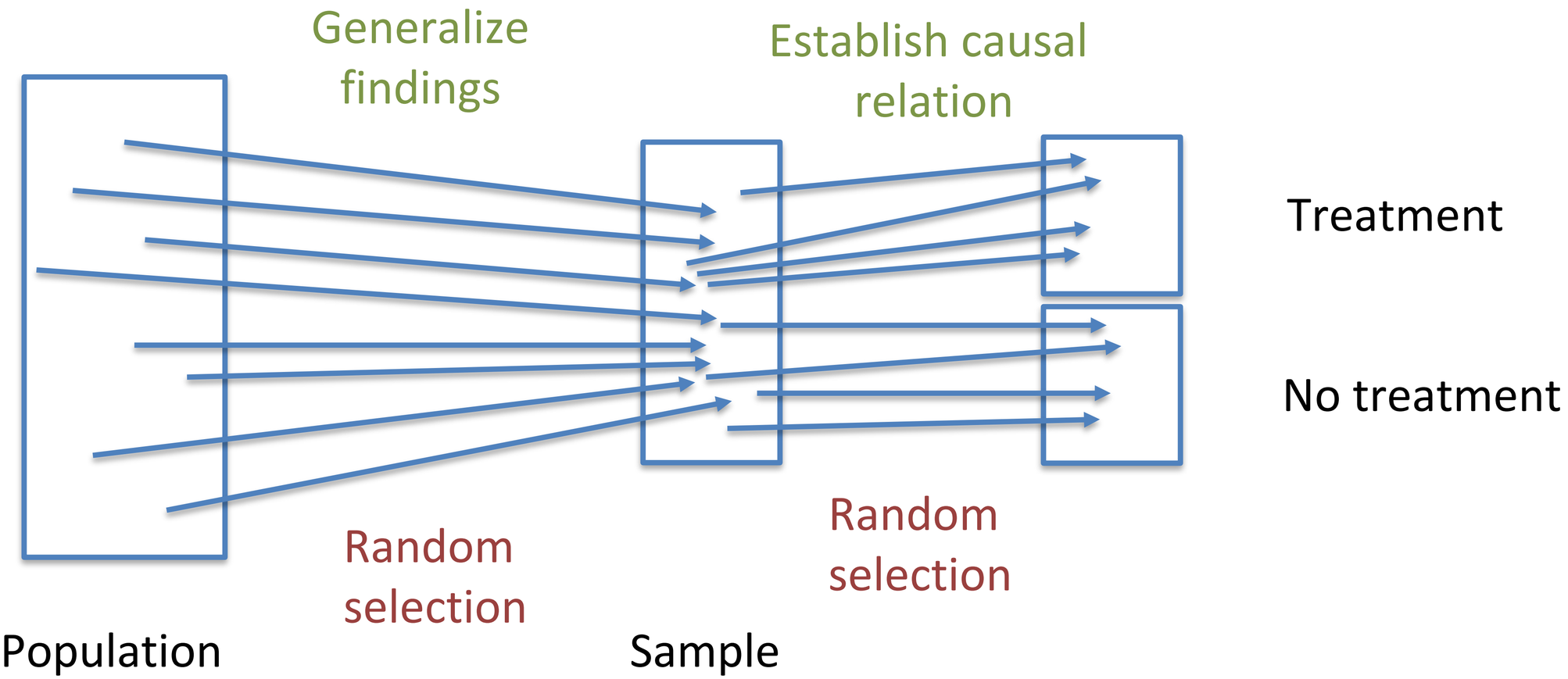


Types of studies

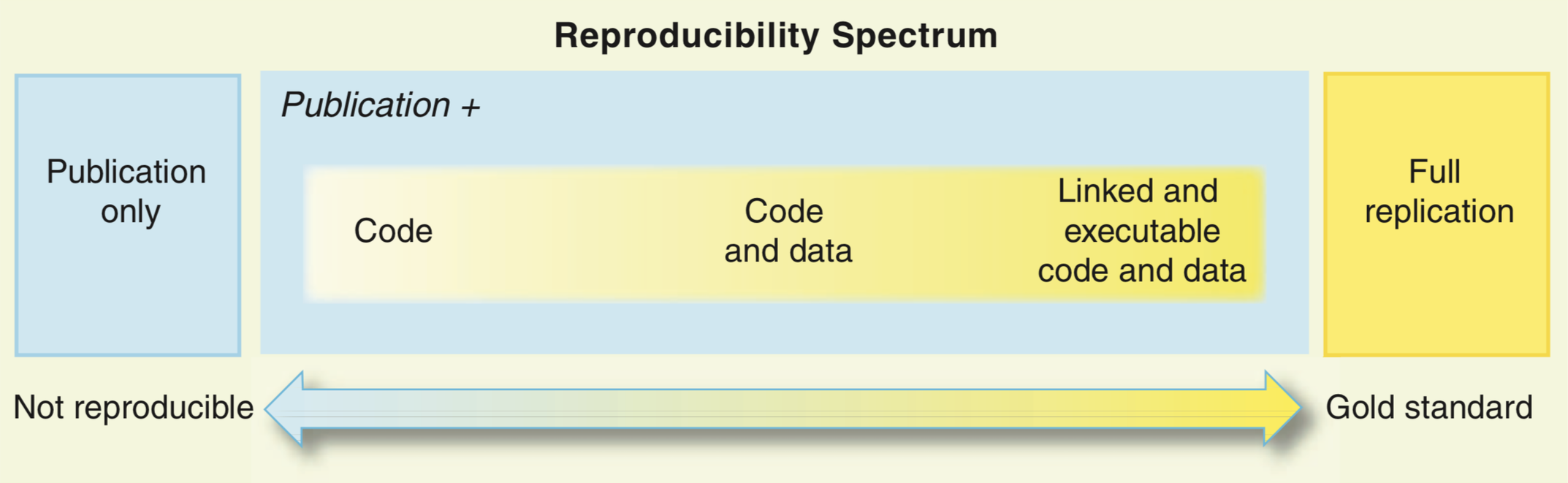


(P. R. Cohen, MIT Press, 1995)

Which conclusions can we draw?



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

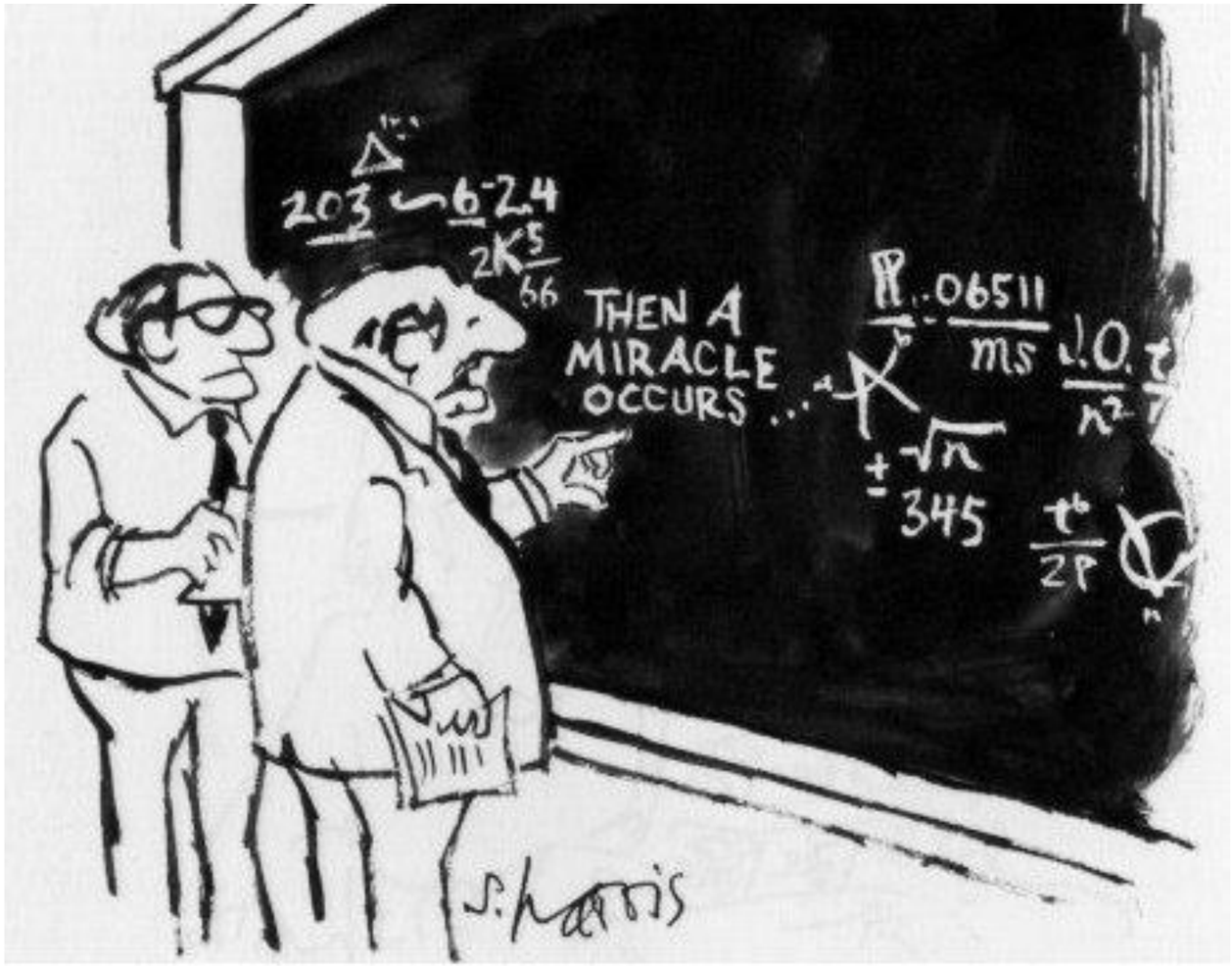
Reproducibility in empirical AI research is the ability of an **independent** research team to produce the same **results** using the same AI method based on the **documentation** made by the original research team.

NTNU

In order to reproduce results

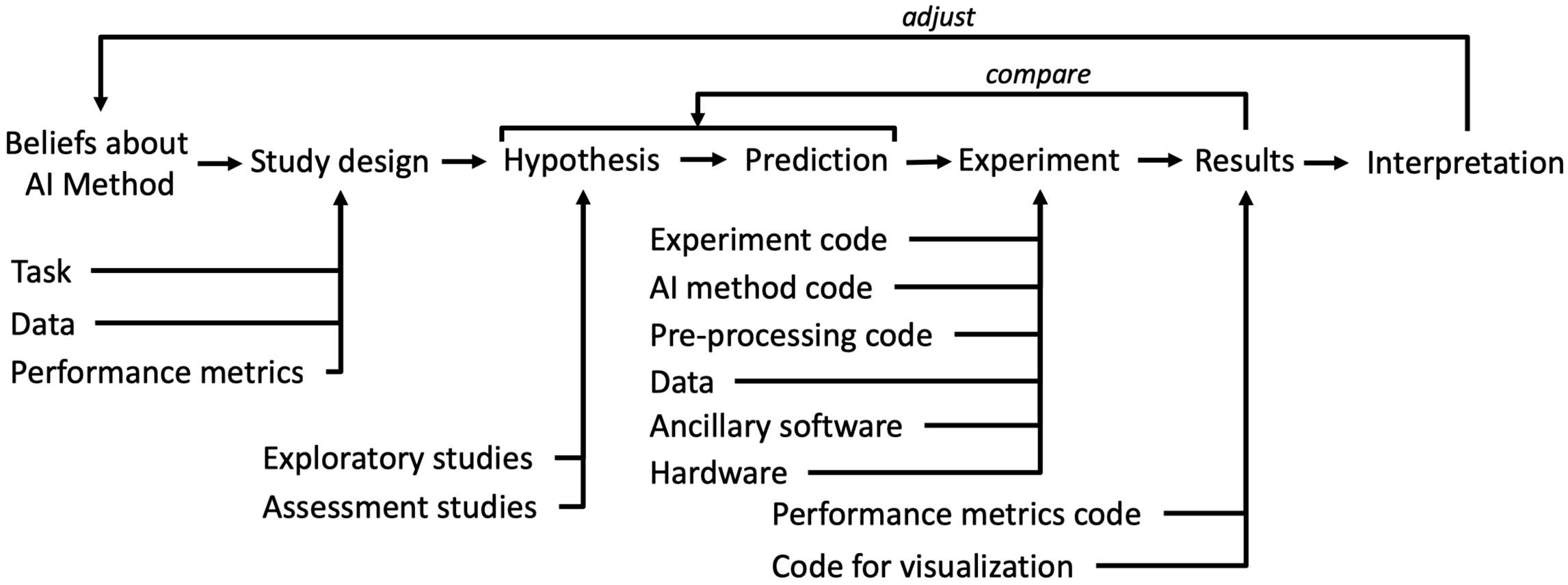
WHAT MUST BE DOCUMENTED?

Reproducibility crisis



"I think you should be more explicit here in step two."

The Scientific Method in Empirical AI Research



NTNU

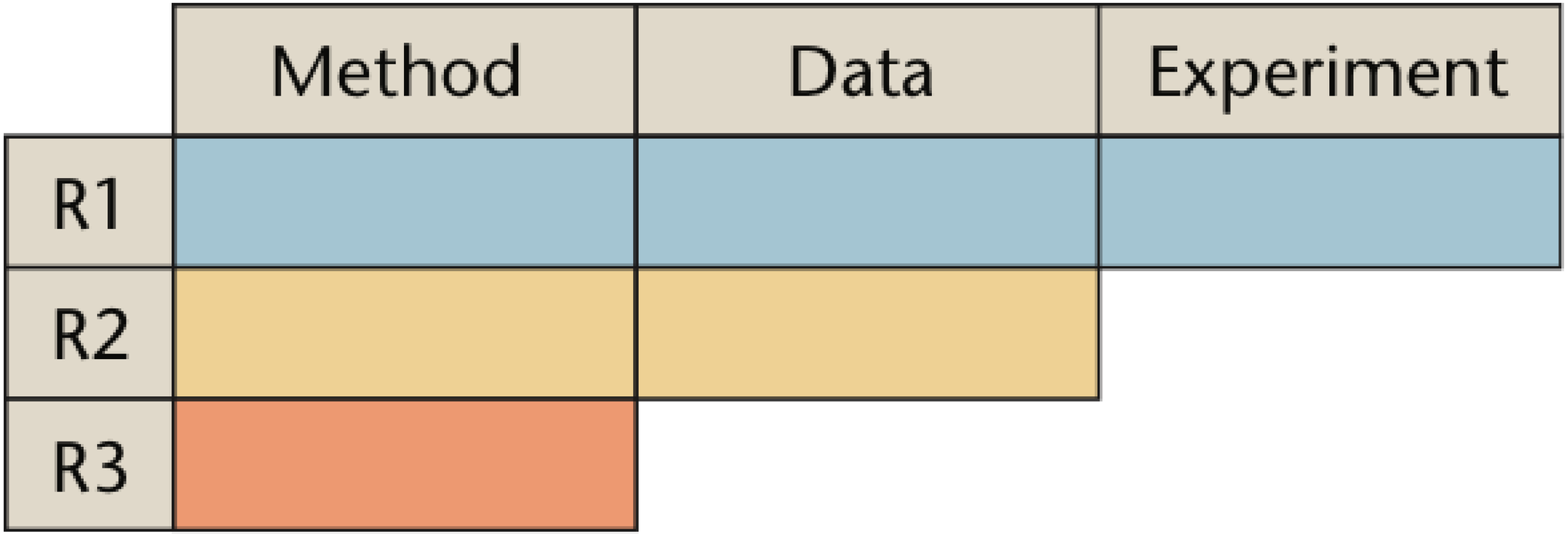
Documentation

- **Method (text):** Description of AI method (system/algorithm), study design, experiment description - human to human, abstract concepts.

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

- **Experiment (software):** AI method code + experiment code + experiment setup + HW + SW + analysis code

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

QUANTIFYING REPRODUCIBILITY

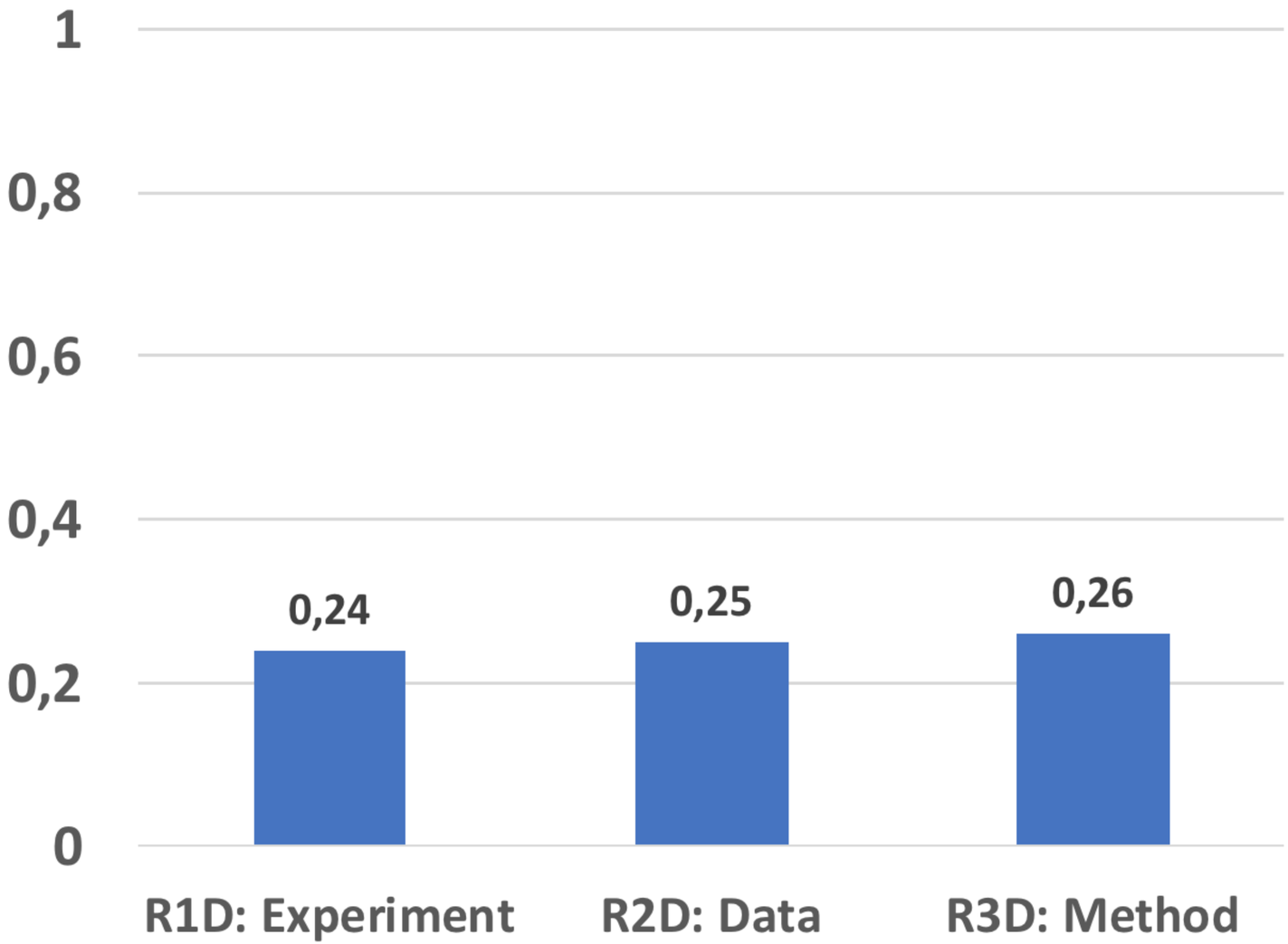
Quantifying Reproducibility

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

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

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

A Normalized Metric



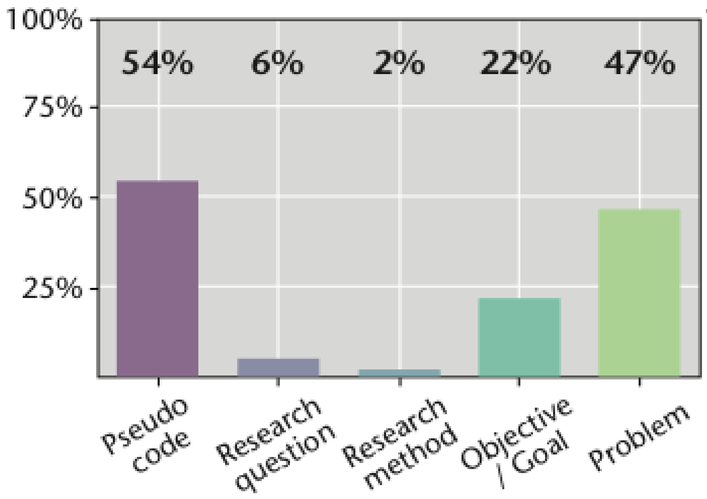
(Gundersen, Kjensmo, AAAI, 2018)

NTNU

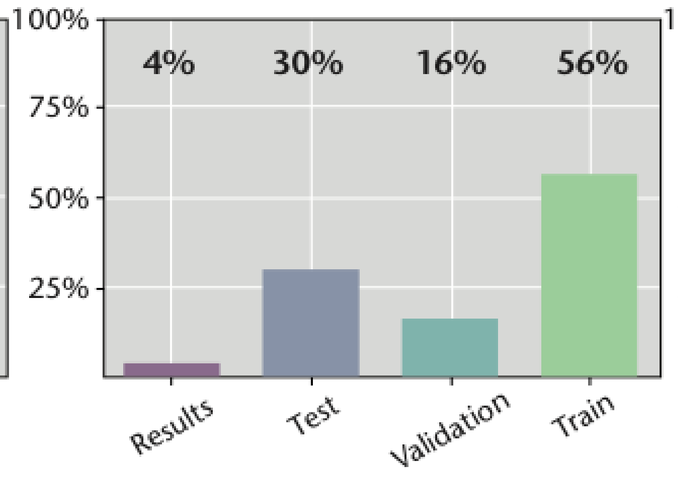
WHAT WE GAIN

We Can Specify How Well Research is Documented

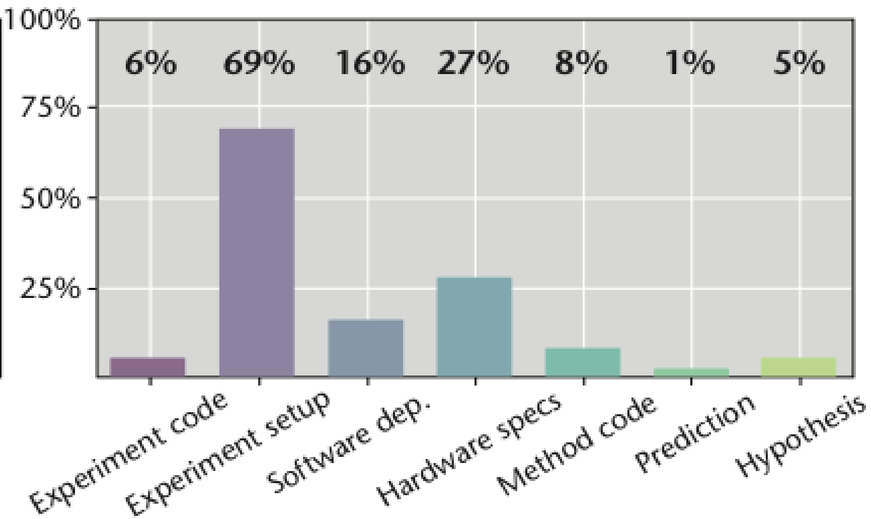
Method



Data

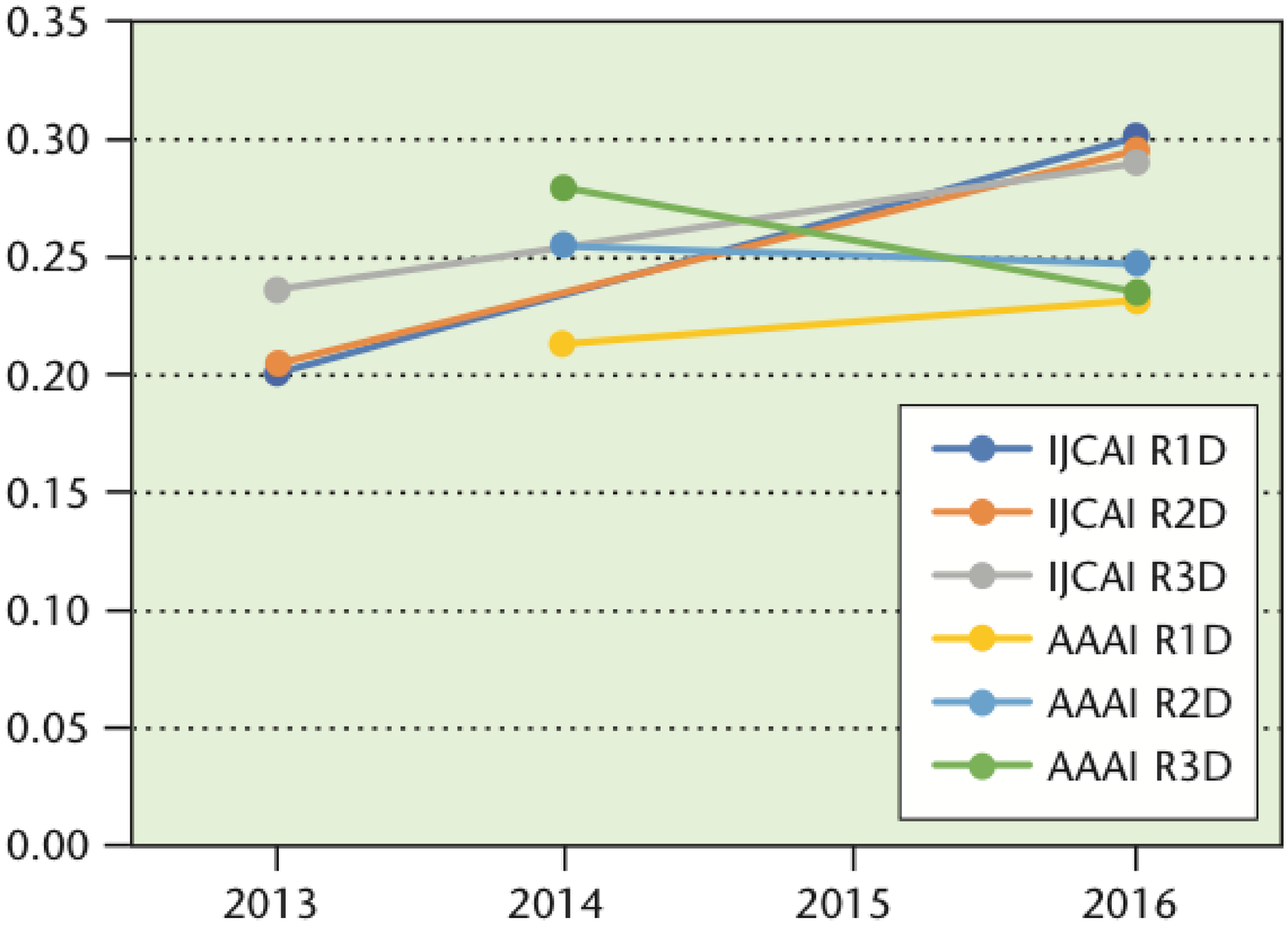


Experiment



(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)

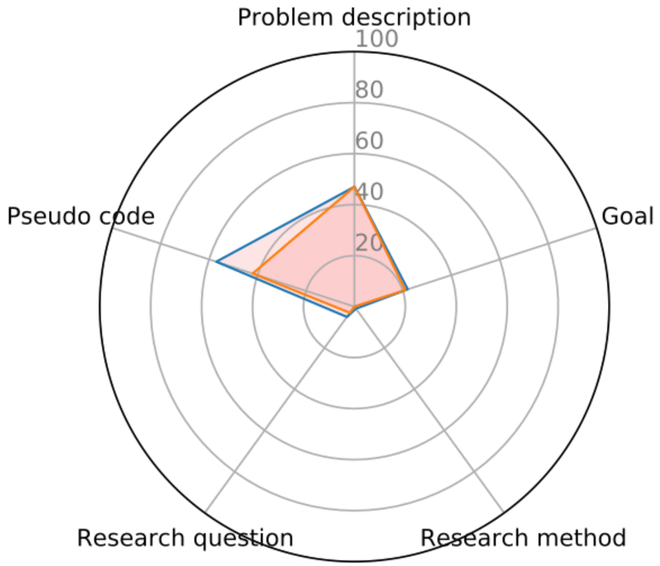
We Can Compare Research: Conferences

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

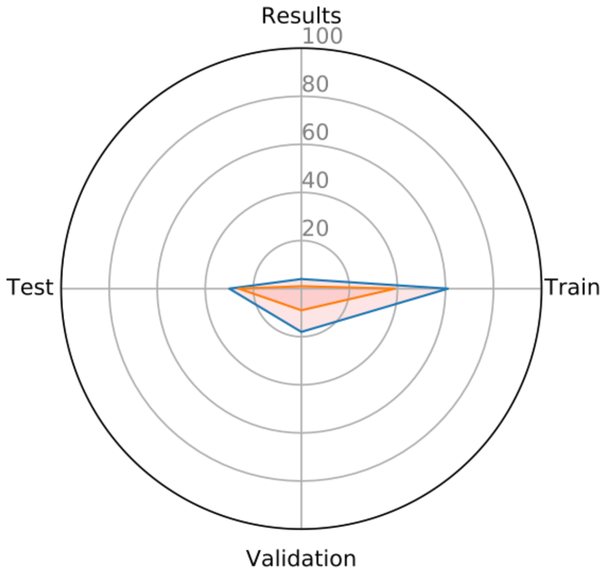
(Gundersen, Kjensmo, AAAI, 2018)

We Can Compare Research: Groups

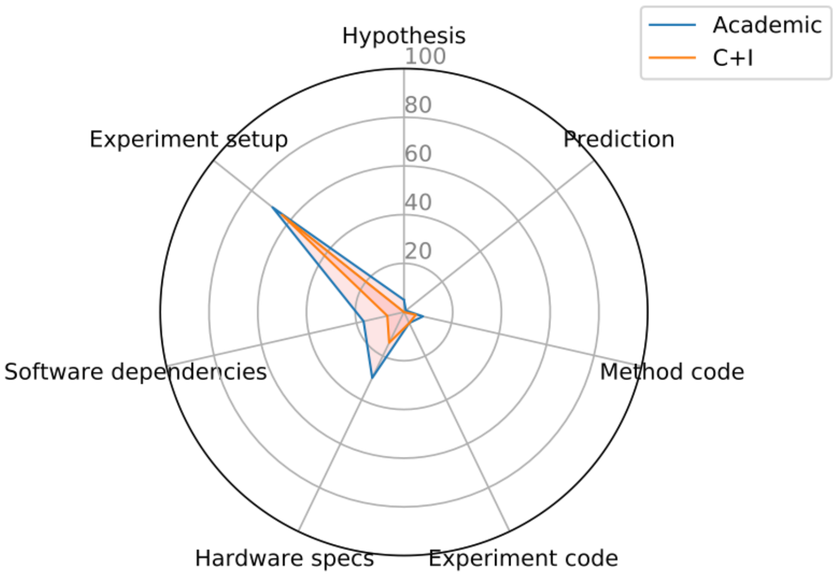
Academia versus Industry



Method



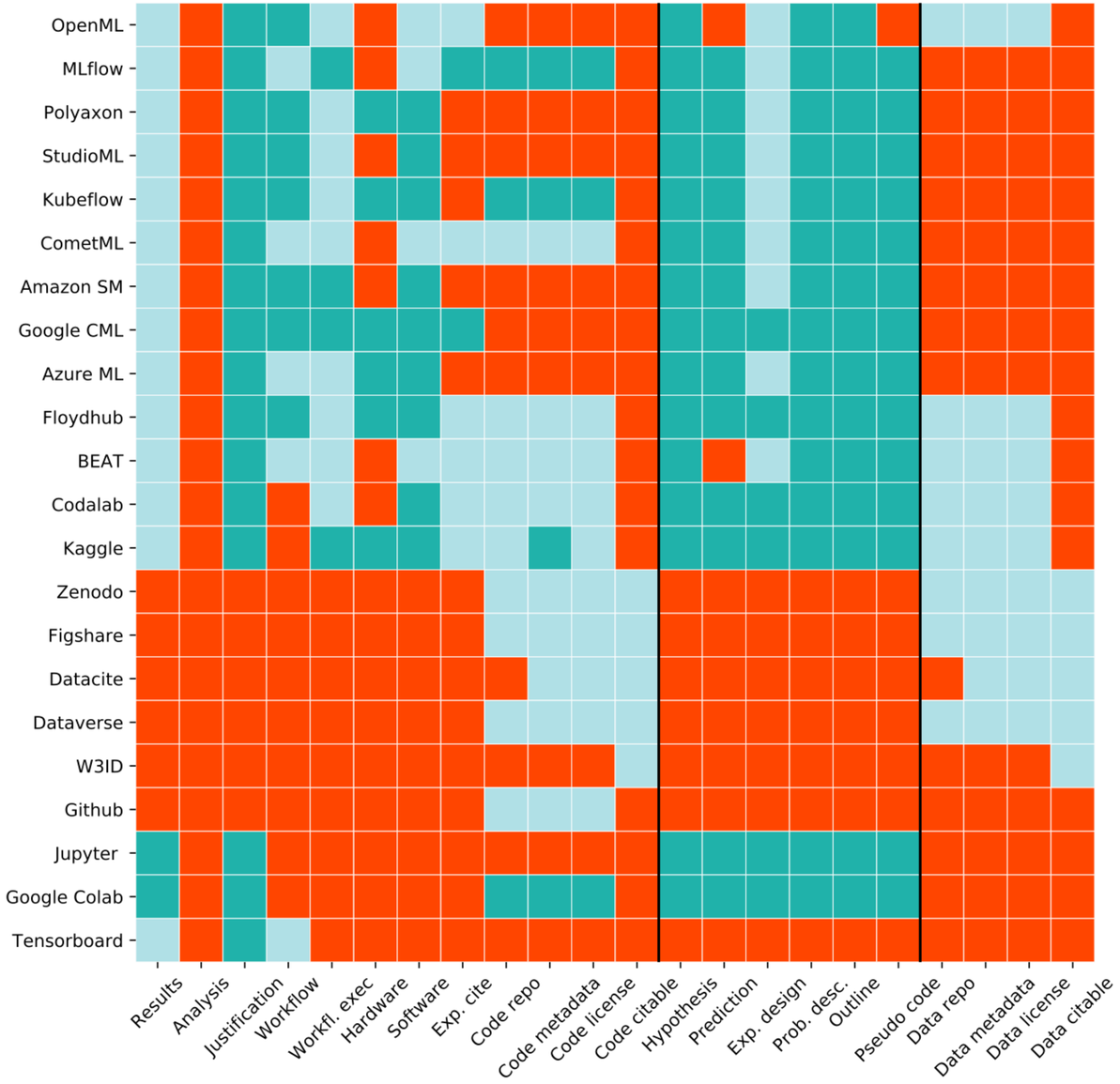
Data



Experiment

(Gundersen, AI Magazine, forthcoming)

We Can Compare Software Frameworks

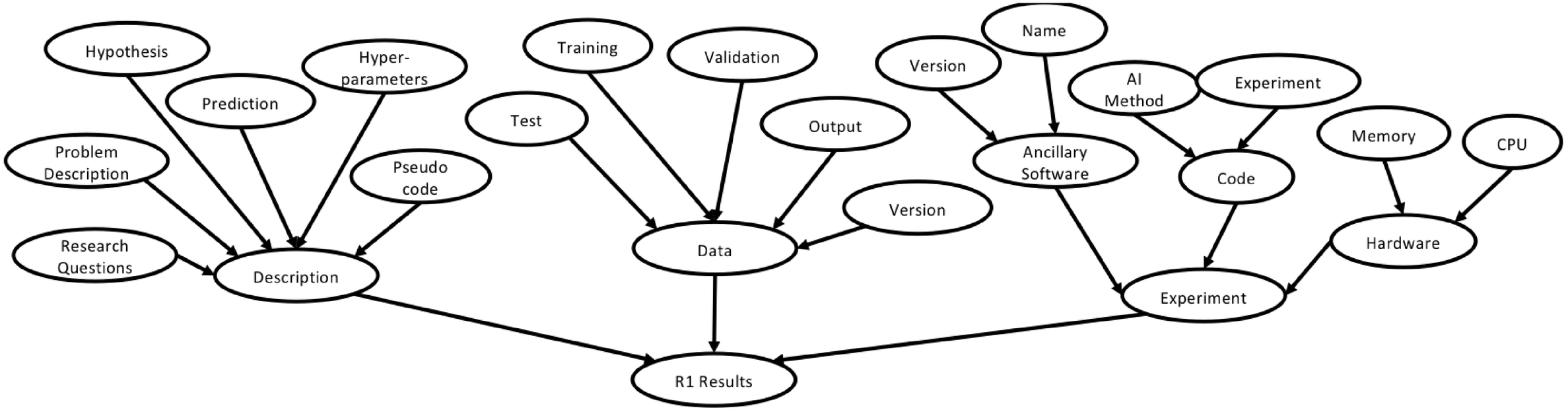


(Isdahl and Gundersen, forthcoming)

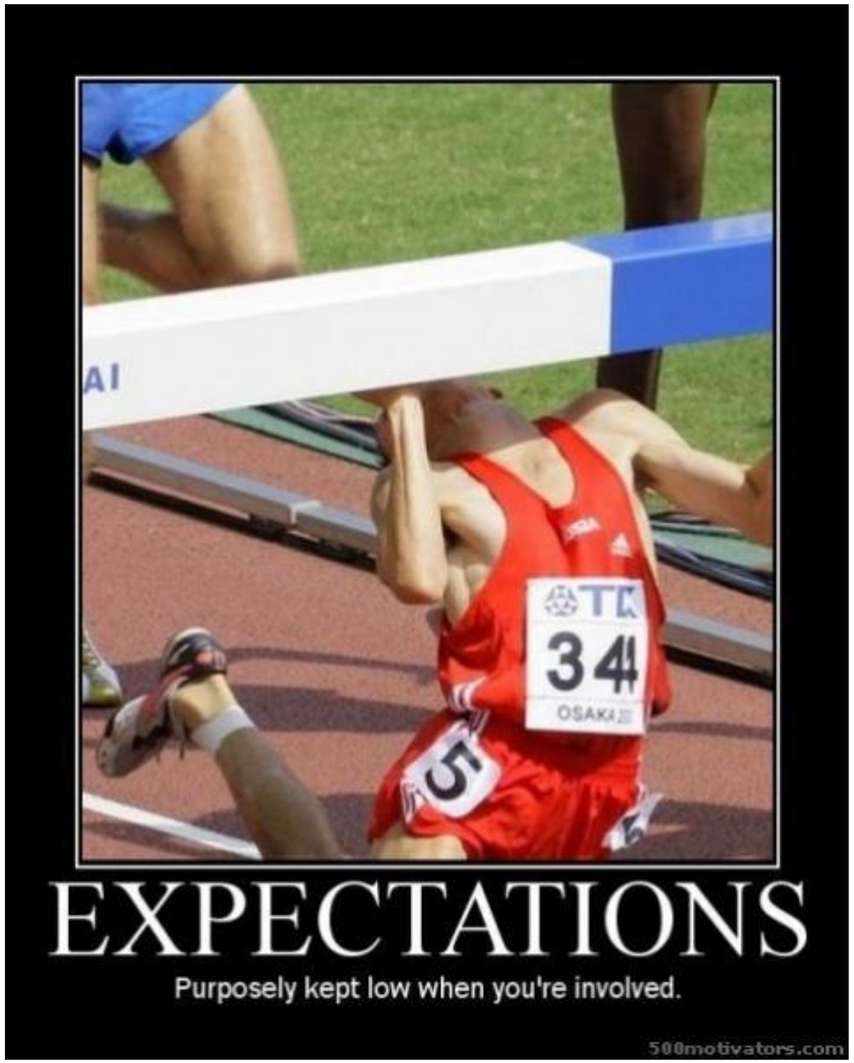
We Could Empirically Find What Entails Well-Documented Research

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 [illegible] explicit mention of the research question(s) addressed?
	Pseudocode	[illegible]he AI meth[illegible] described using pseudocode?
Data	Training data	[illegible]he training [illegible] shared?
	Validation data	Is the valida[illegible] set shared?
	Test data	Is the test [illegible] shared?
	Results	Are the [illegible]evant 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 hypot[illegible]s?
	Method source code	Is the [illegible] stem 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 Can Set the Bar Based on What We Want to Achieve



NTNU

CASE STUDY

HOW WELL IS AI RESEARCH DOCUMENTED?

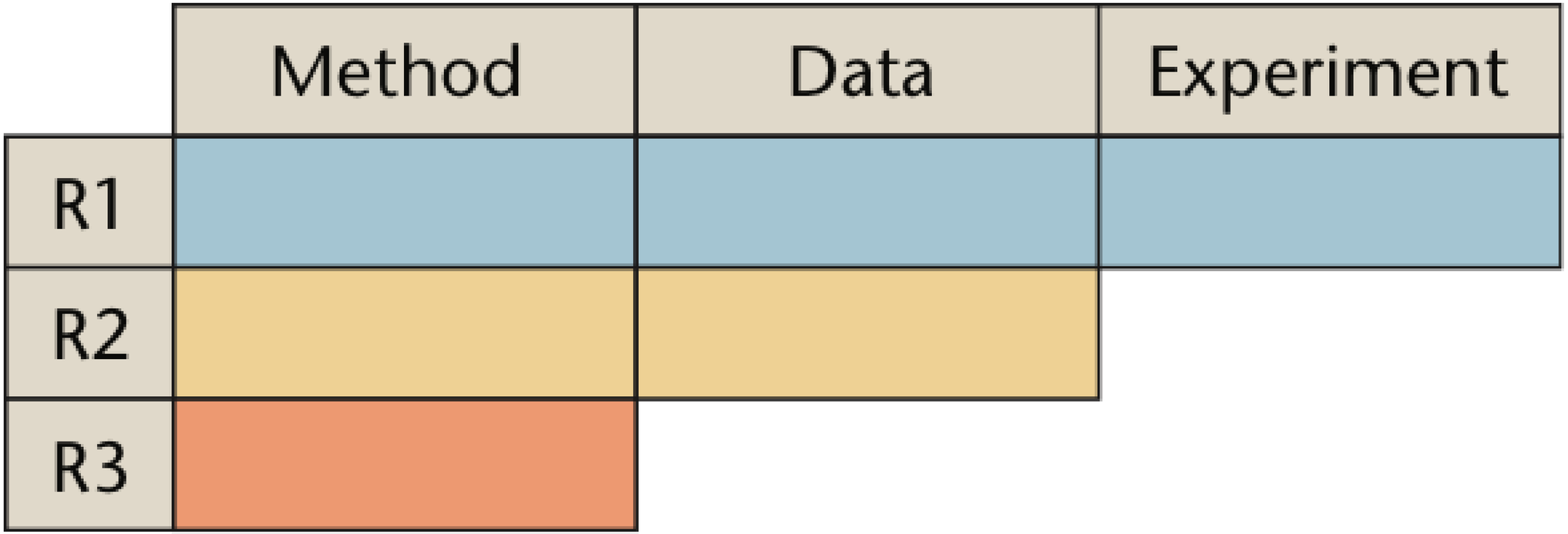
Experiment I

- 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)

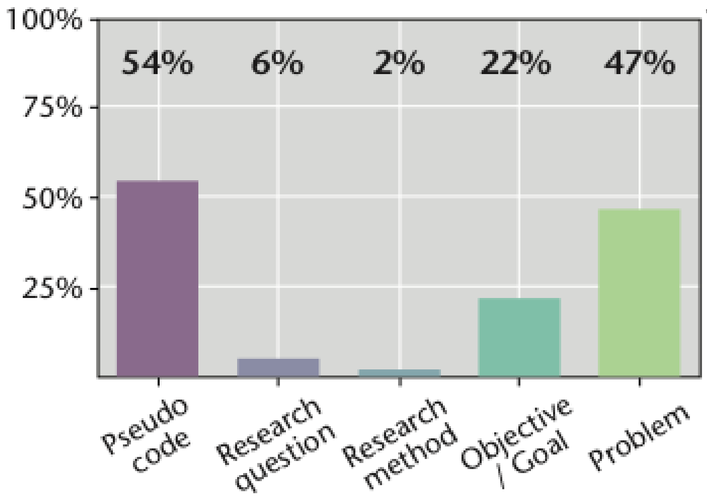
Degree of Reproducibility



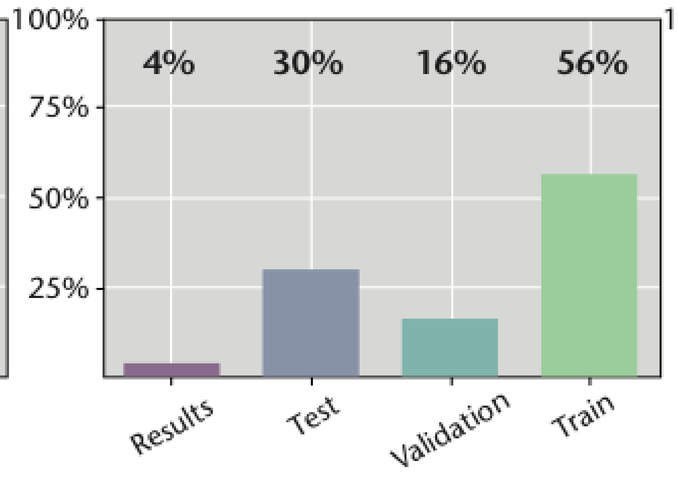
NTNU

Results I: Factors and Variables

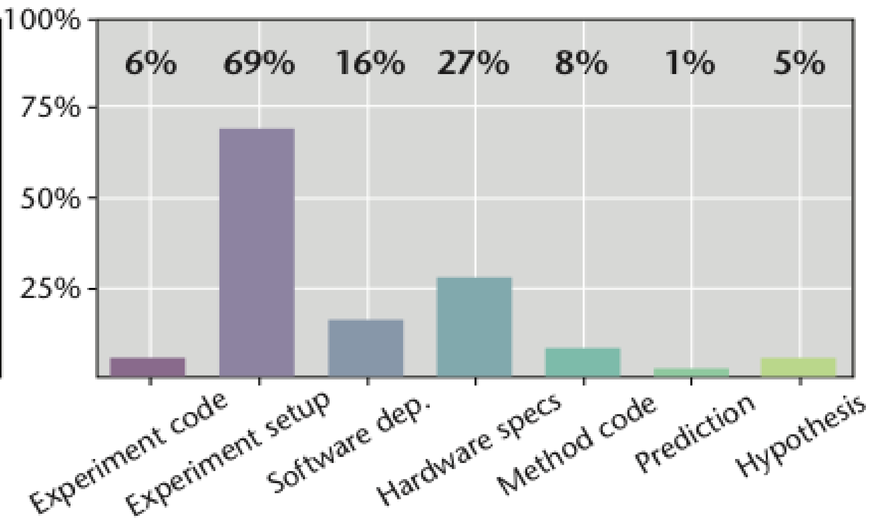
Method



Data

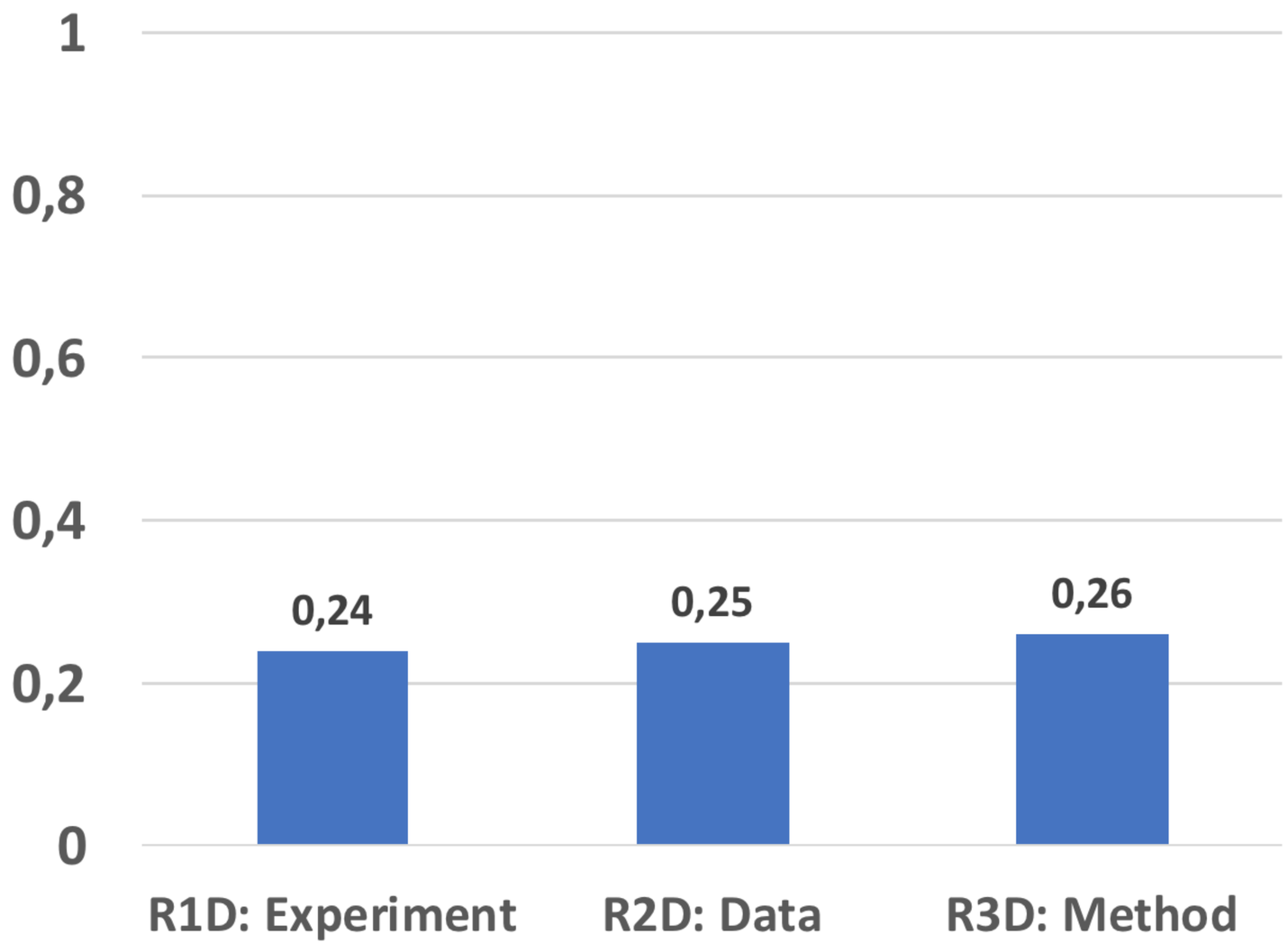


Experiment



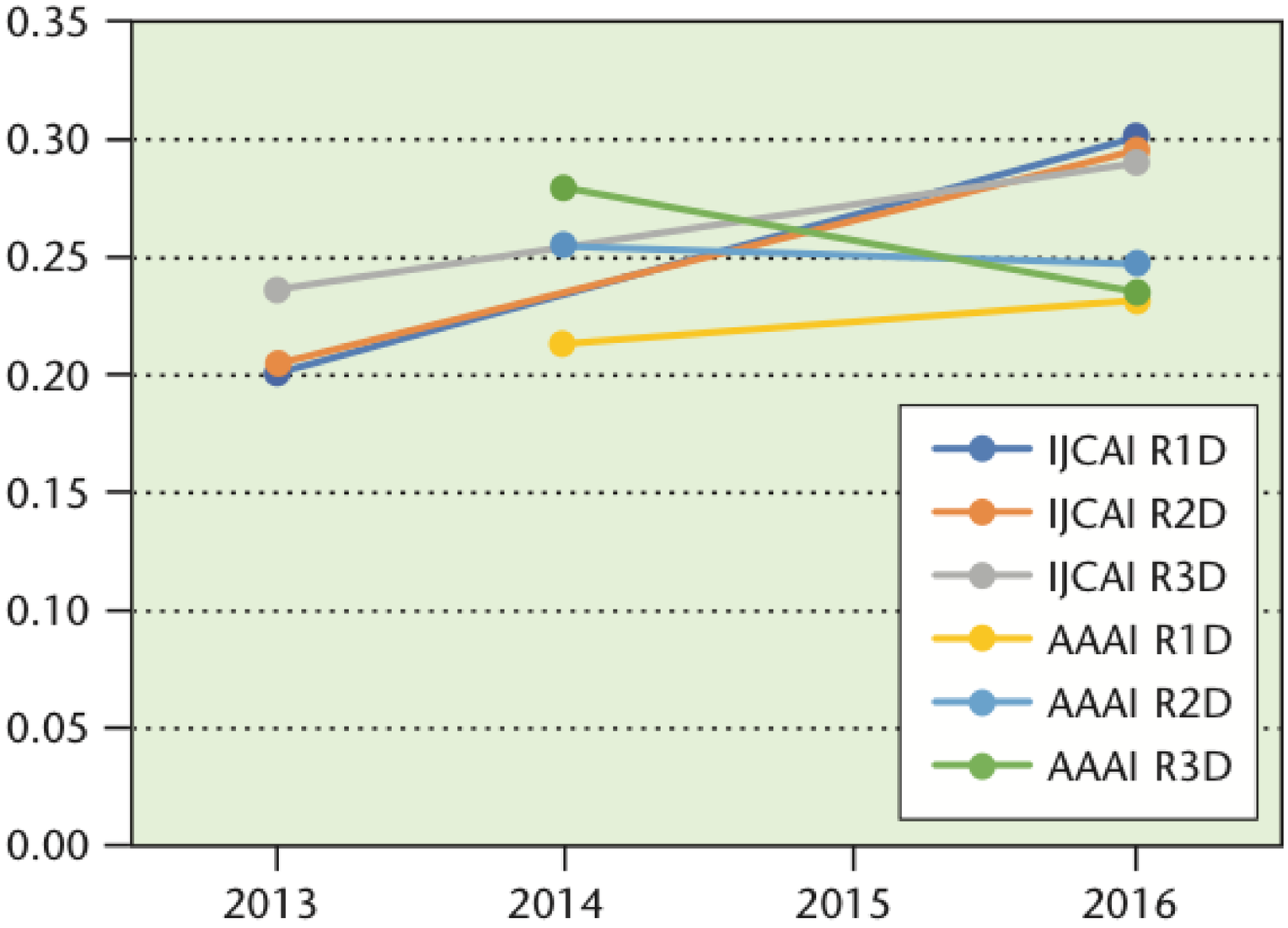
(Gundersen, Kjensmo, AAAI, 2018)

Results II: Reproducibility Degree



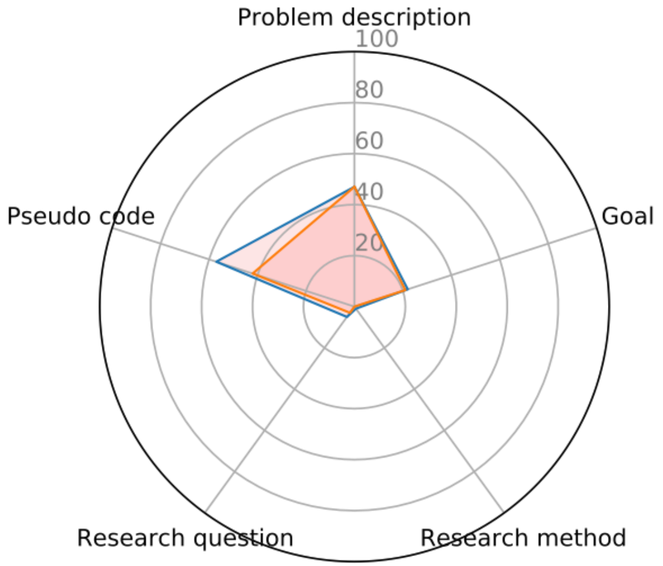
(Gundersen, Kjensmo, AAAI, 2018)

Results III: Change over Time

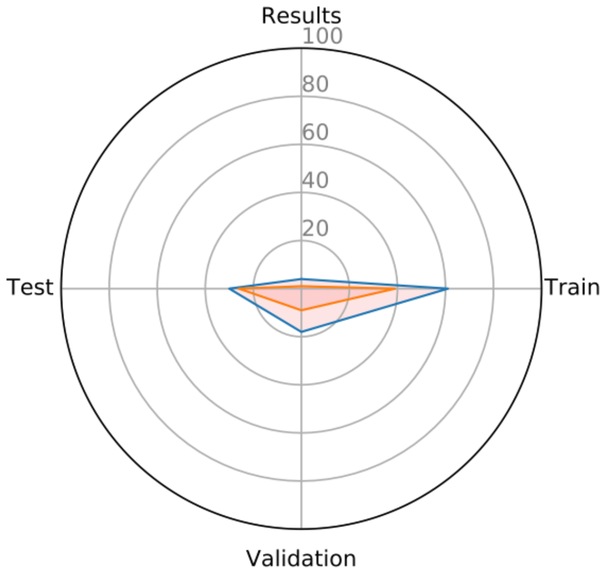


(Gundersen, Kjensmo, AAAI, 2018)

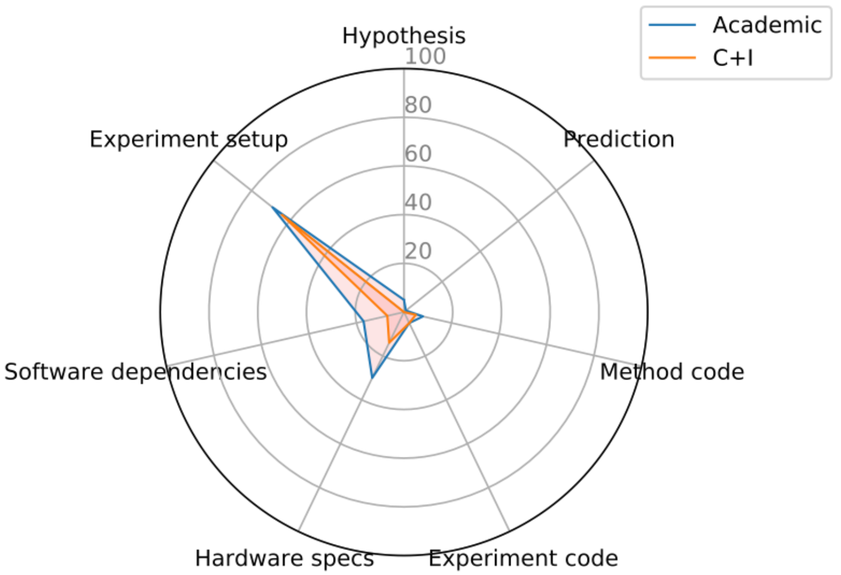
Results IV: Industry vs Academia



Method



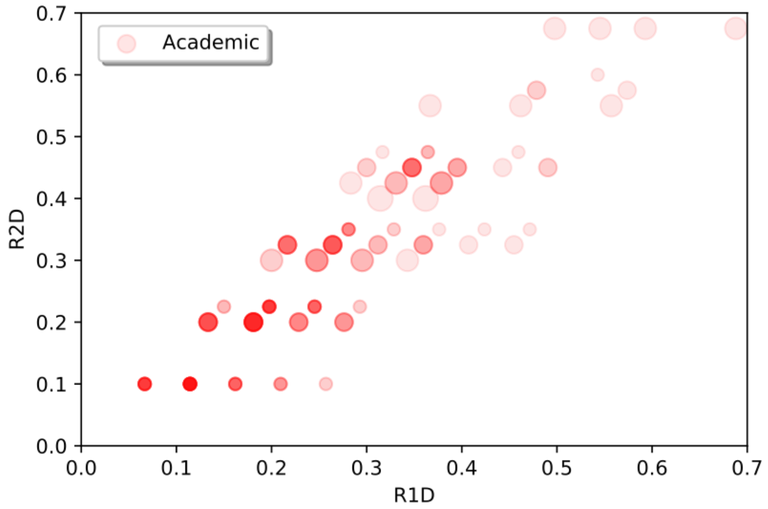
Data

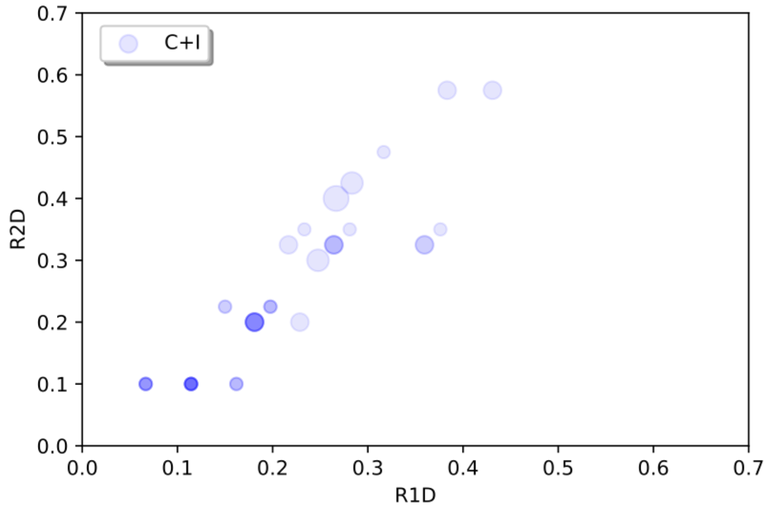


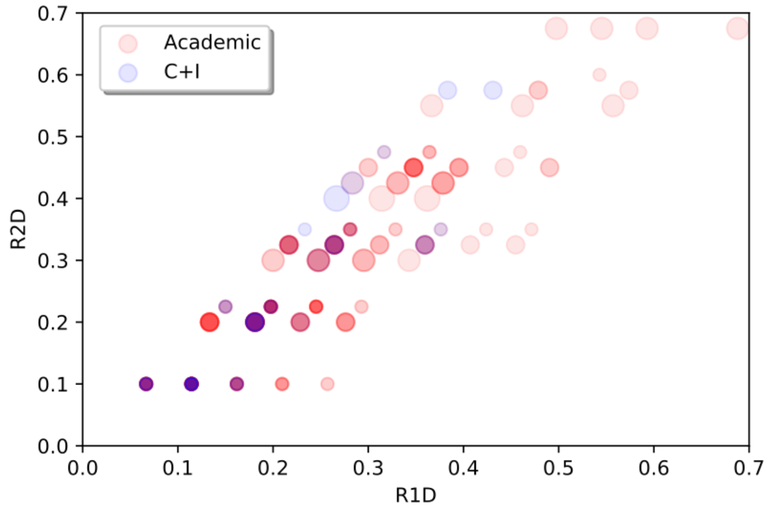
Experiment

(Gundersen, AI Magazine, forthcoming)

Results V: Industry vs Academia





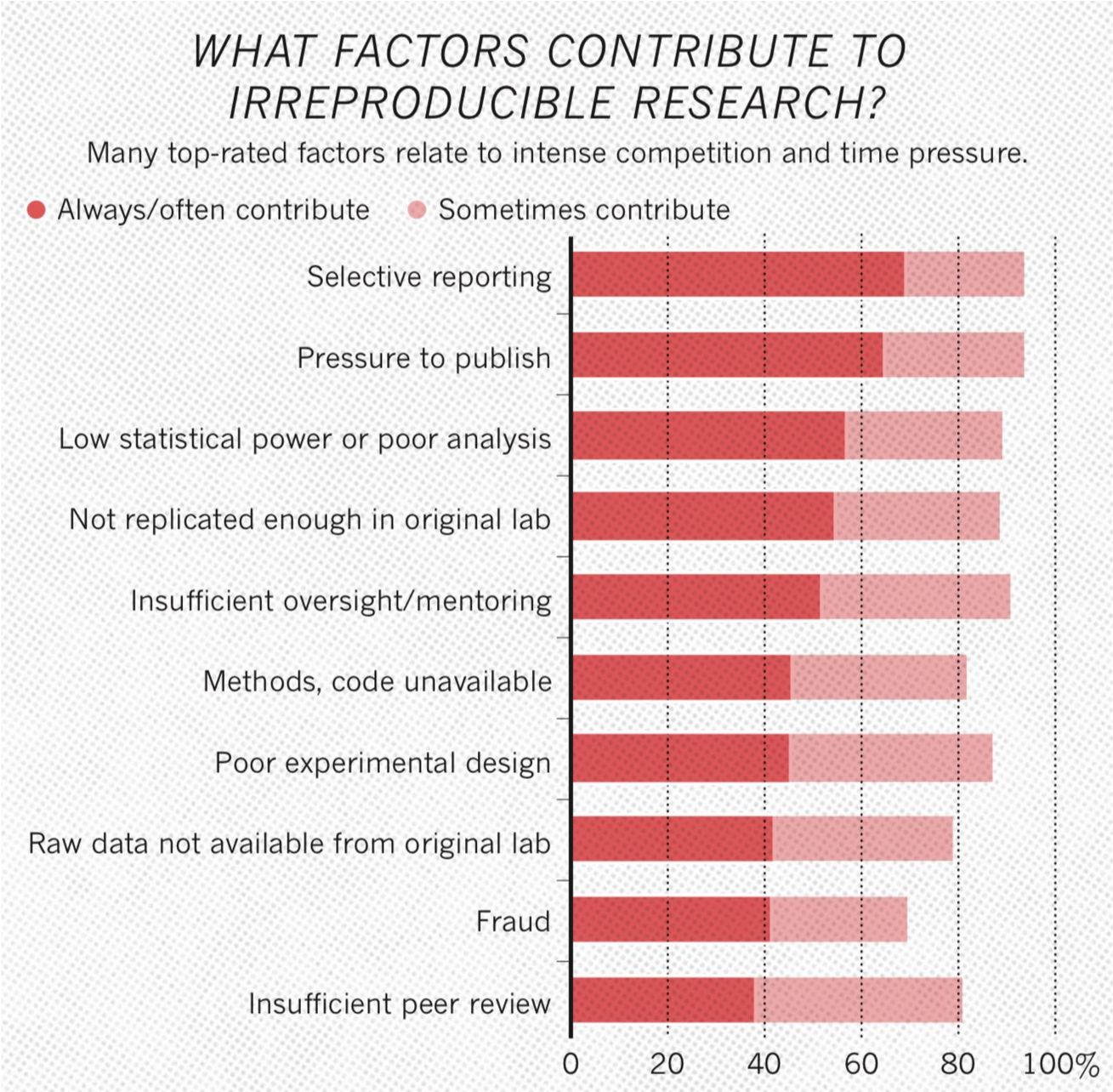


(Gundersen, AI Magazine, forthcoming)

NTNU

PART 2

CAUSES OF IRREPRODUCIBILITY

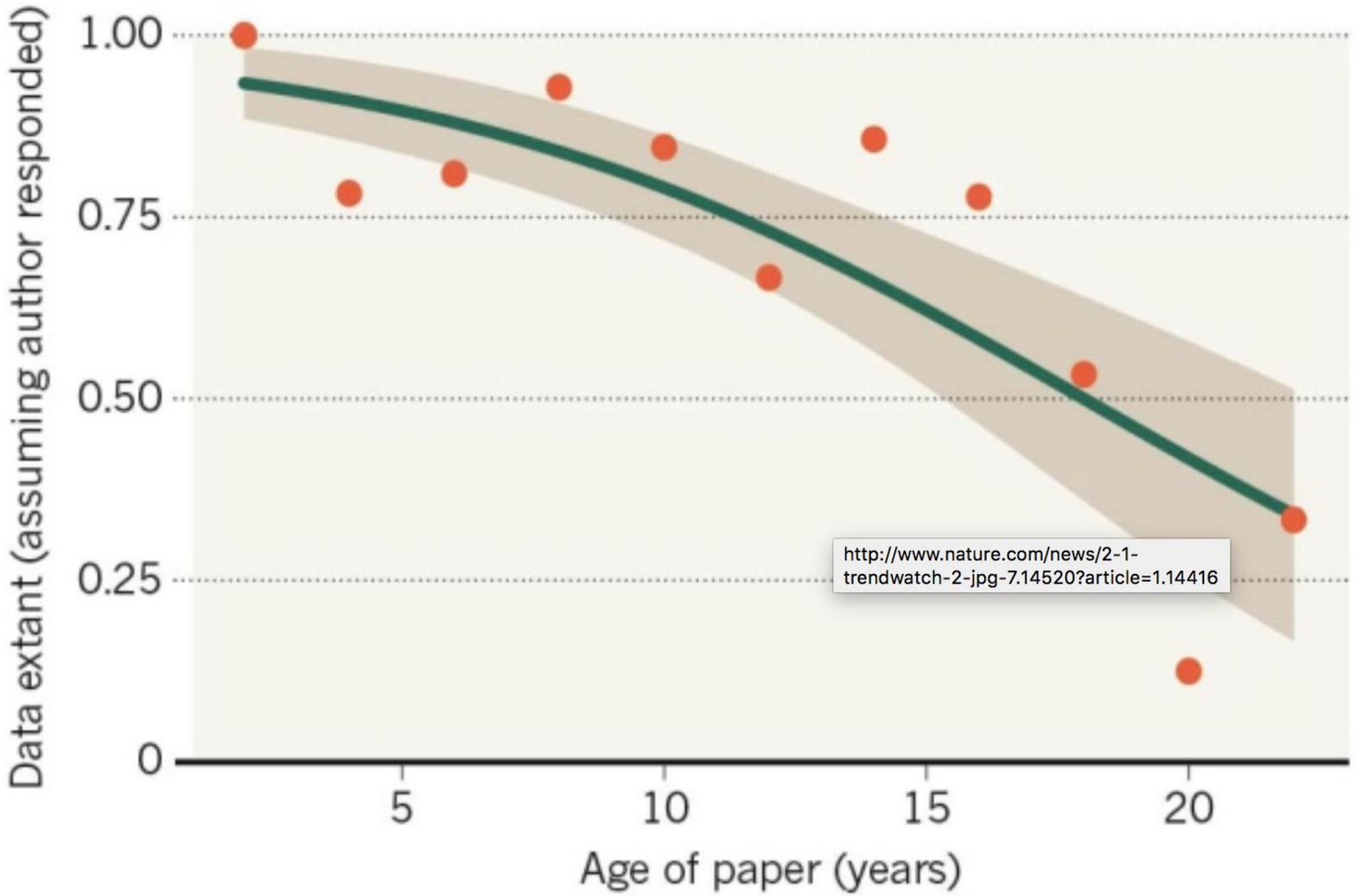


(M. Baker, Nature, 2016)

NTNU

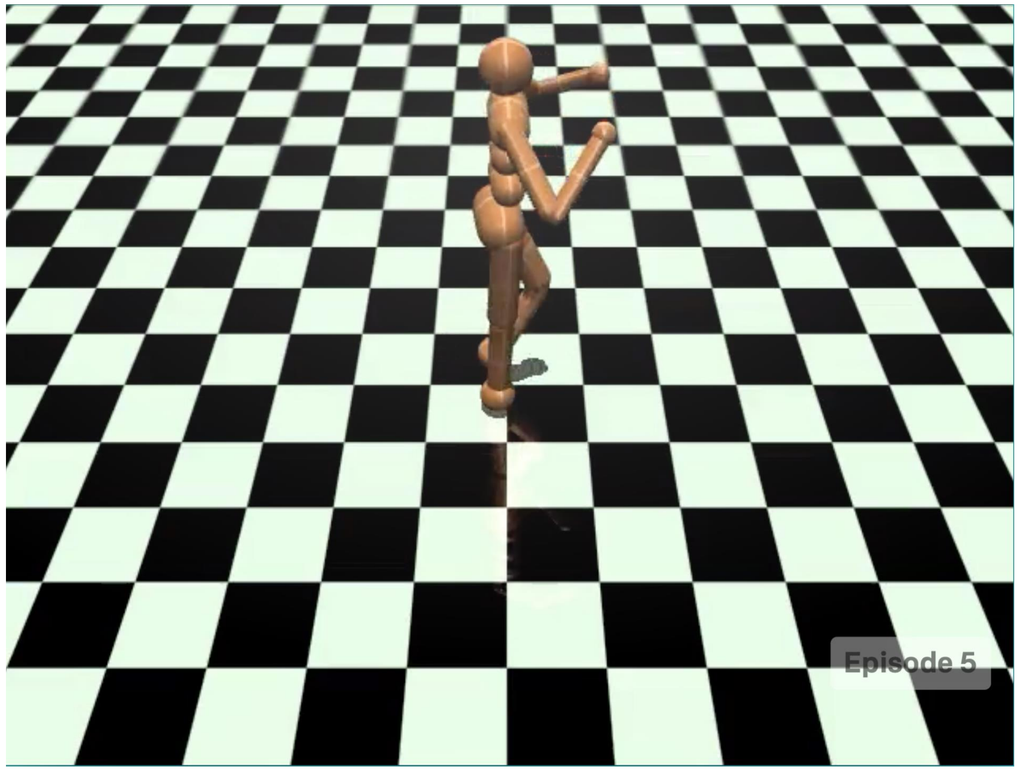
MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



(M. Baker, Nature, 2016)

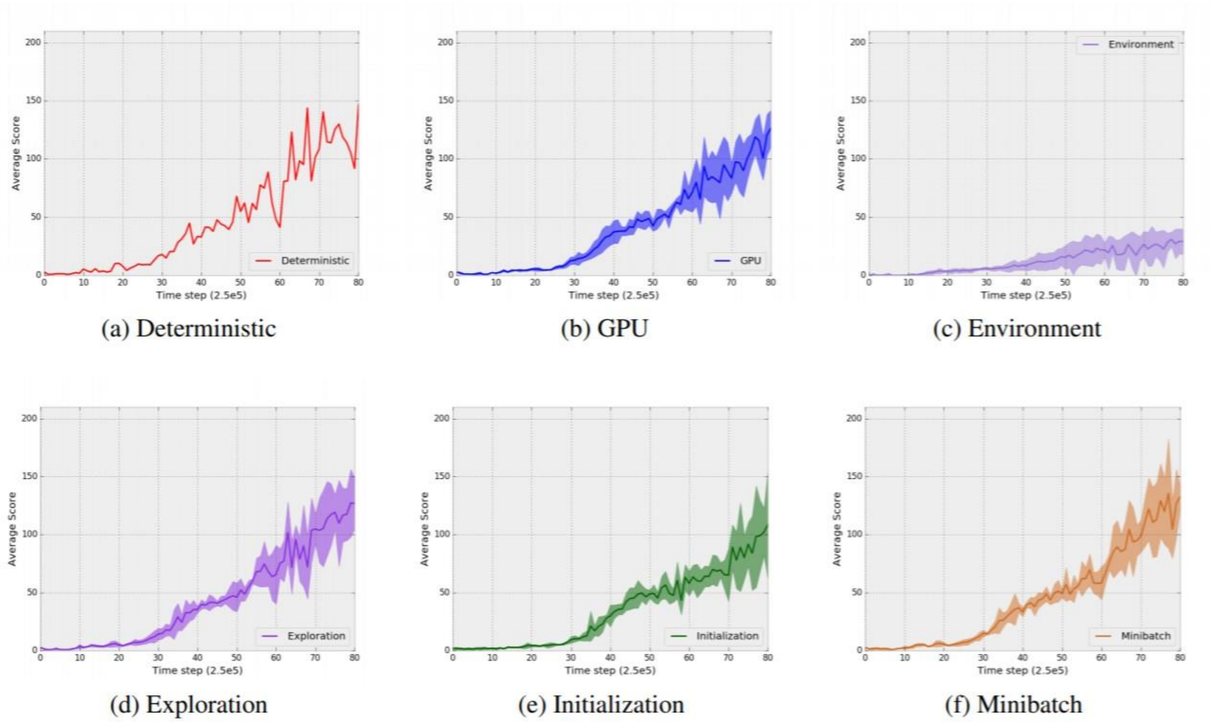
Deep Reinforcement Learning that Matters



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

(Henderson et al, AAAI 2018, 2018)

Deterministic Implementations for Reproducibility in DRL



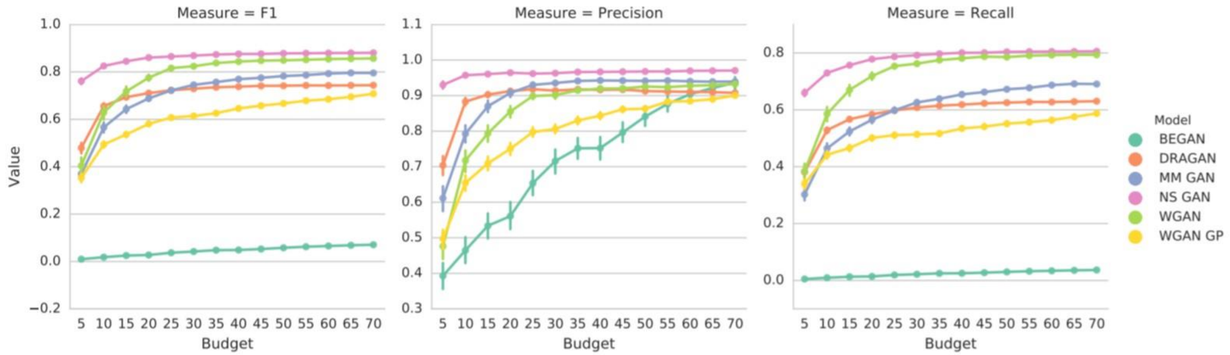
(a) Deterministic (b) GPU (c) Environment

(d) Exploration (e) Initialization (f) Minibatch

- Non-determinism in training process.
- Deterministic implementation of Q-learning.
- Measure impact of different sources of nondeterminism.
- Different sources have huge impact on performance.

(P. Nagarajan, G. Warnell, P. Stone, AAAI RAI 2019)

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.

(Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O., NIPS 2017)

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	♠

(S. HONG, M. KOO, and J. JANG, 2013)

NTNU

Case study

REPRODUCING THE MOST CITED AI RESEARCH

Experiment

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

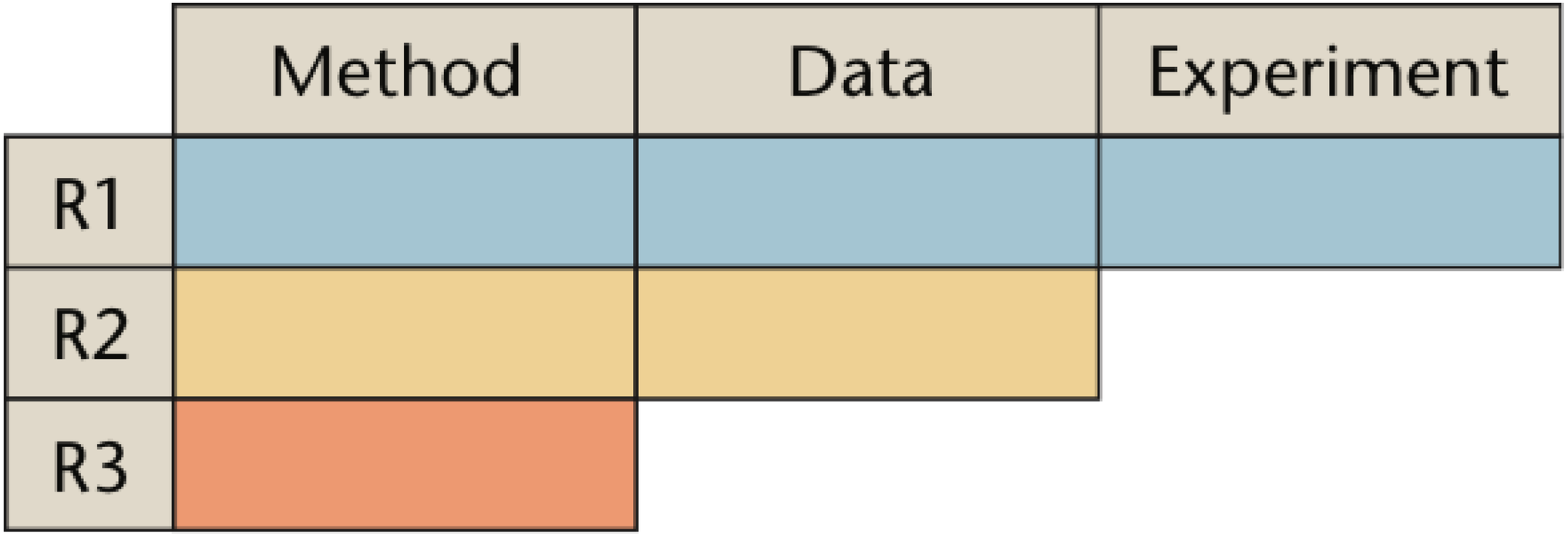
(Gundersen et al, forthcoming)

Research Procedure

- Reproduce research classified as R1 and R2 reproducible.
- 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).

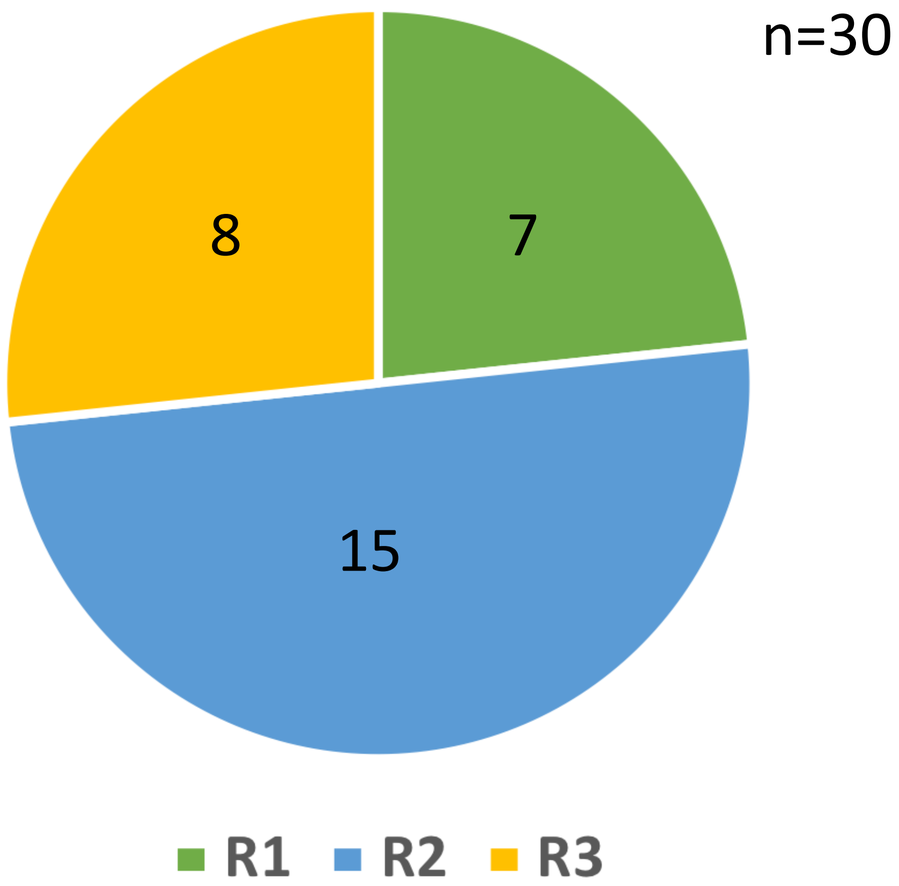
(Gundersen et al, forthcoming)

Degree of Reproducibility



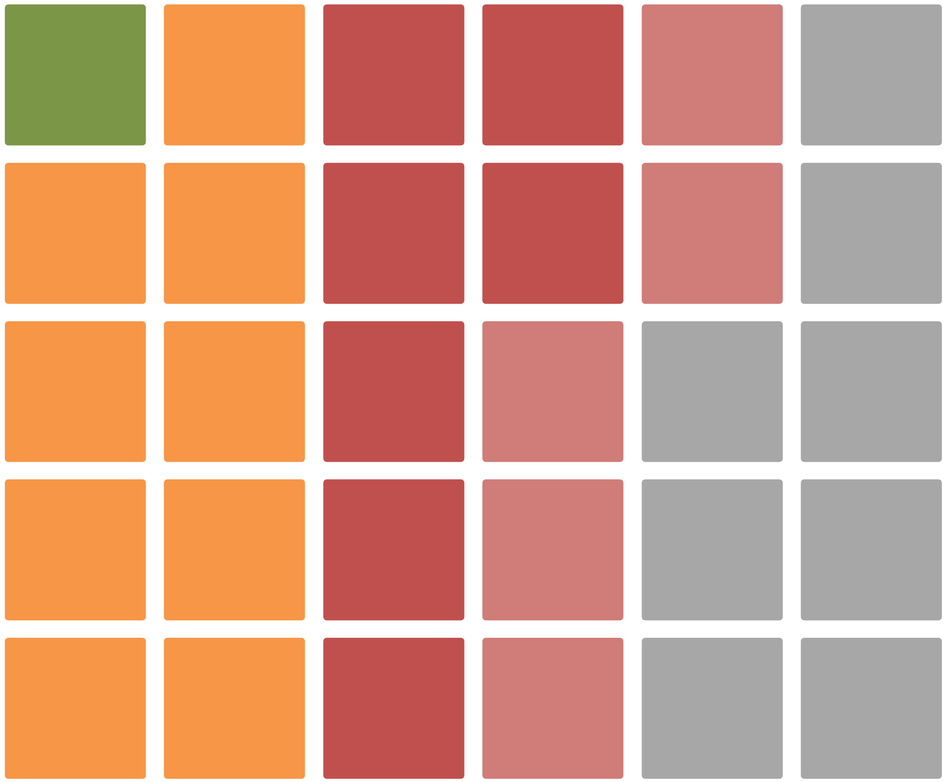
NTNU

Results: Reproducibility Degree



(Gundersen et al, forthcoming)

Results: Outcome per paper



Success: 3%

Partial success: 30%

Failure: 30%

No result: 23%

Filtered out (R3): 27%

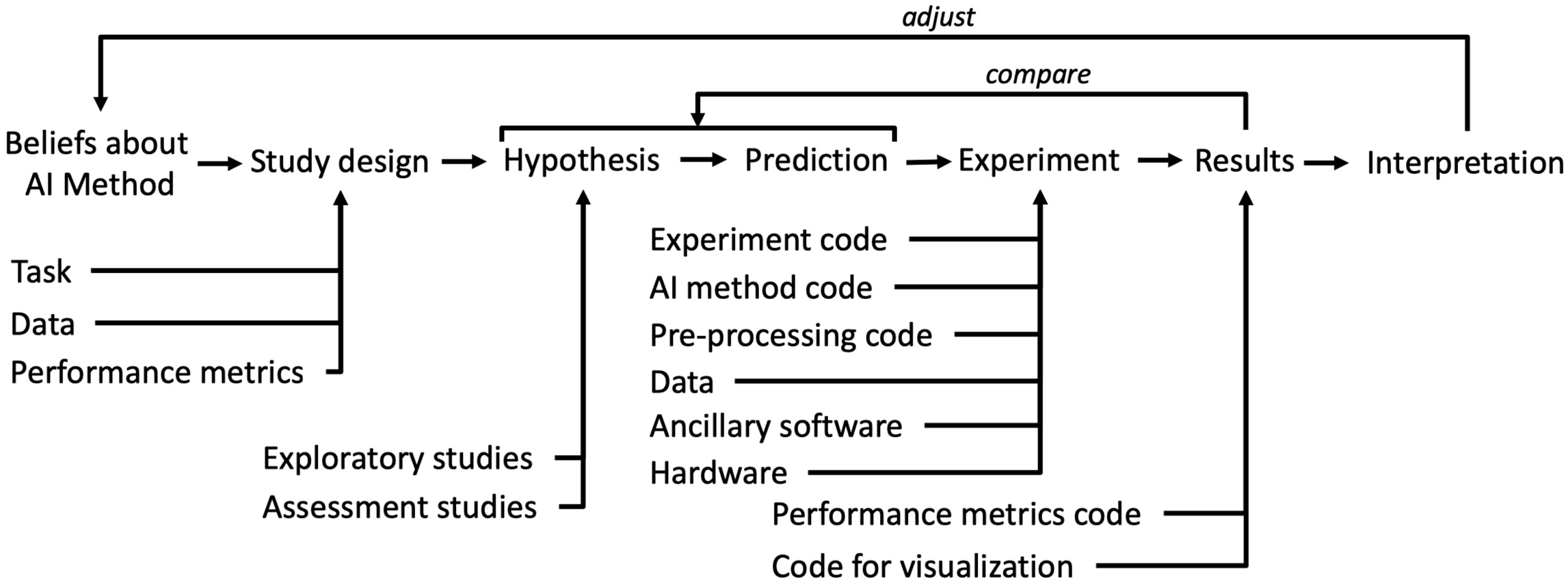
(Gundersen et al, forthcoming)

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)

The Scientific Method in Empirical AI Research



NTNU

Part 3

RECOMMENDATIONS

Recommendations in AI Magazine

- Author checklist of 24 practical recommendations

- Four groups:
  - Data
  - Source code
  - AI Methods
  - Experiments

- Summary:
  - Open Science - share data, code and procedures.
  - Build digital scholarship
  - Version code and data!

(Gundersen, Gil and Aha, AI Magazine, Fall 2018)

Data

Recommendations	*Data mentioned in a publication should:*
1.	Be available in a shared community repository, so anyone can access it
2.	Include basic metadata, so others can search and understand its contents
3.	Have a license, so anyone can understand the conditions for reuse of the data
4.	Have an associated digital object identifier (DOI) or persistent URL (PURL) so that the data is available permanently
5.	Be cited properly in the prose and listed accurately among the references, so readers can identify the datasets unequivocally and data creators can receive credit for their work

Source code

Recommendations	*Source code used for implementing an AI method and executing an experiment should:*
6.	Be available in a shared community repository, so anyone can access it
7.	Include basic metadata, so others can search and understand its contents
8.	Include a license, so anyone can understand the conditions for use and extension of the software
9.	Have an associated digital object identifier (DOI) or persistent URL (PURL) for the version used in the associated publication so that the source code is permanently available
10.	Be cited and referenced properly in the publication so that readers can identify the version unequivocally and its creators can receive credit for their work

AI Methods

Recommendations	*AI methods used in a publication should be:*
11.	Presented in the context of a problem description that clearly identifies what problem they are intended to solve
12.	Outlined conceptually so that anyone can understand their foundational concepts
13	Described in pseudocode so that others can understand the details of how they work

Experiments

Recommendations	*Descriptions of experiments in a publication should:*
14.	Explicitly present the hypotheses to be assessed, before other details concerning the empirical study are presented
15.	Present the predicted outcome of the experiment, based on beliefs about the AI method and its application
16.	Include the experiment design (parameters and the conditions to be tested) and its motivation, such as why a specific number of tests or data points are used based on the desired statistical significance of results and the availability of data
17.	Identify and describe the measure and metrics
18.	Provide the evaluation protocol
19.	Share the results
20.	Describe the results and the analysis
21.	Be described as a workflow that summarizes how the experiment is executed and configured
22.	Include documentation on workflow executions or execution traces that provide parameter settings and initial, intermediate, and final data
23.	Specify the hardware used to run the experiments
24	Be cited and published separately when complex, so that others can unequivocally refer to the individual portions of the method that they reuse or extend

The ML Reproducibility Checklist

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.
- ❑ A link to a downloadable source code, with specification of all dependencies, including external libraries.

For any **theoretical claim**, check if you include:

- ❑ A statement of the result.
- ❑ A clear explanation of any assumptions.
- ❑ A complete proof of the claim.

(J. Pineau, https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf)

The ML Reproducibility Checklist

For all **figures** and **tables** that present empirical results, check if you include:

- ❑ A complete description of the data collection process, including sample size.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ An explanation of any data that were excluded, description of any pre-processing step.
- ❑ An explanation of how samples were allocated for training / validation / testing.
- ❑ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ❑ The exact number of evaluation runs.
- ❑ A description of how experiments were run.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ Clearly defined error bars.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. stddev).
- ❑ A description of the computing infrastructure used.

(J. Pineau, https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf)

Reproducibility in Computational and Experimental Mathematics 2012

Implementation Criteria I:

- *A precise statement of assertions to be made in the paper.*
- *Full statement (or valid summary) of experimental results.*
- *Salient details of data reduction & statistical analysis methods.*
- Necessary run parameters were given.
- *A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.*
- Complete statements of, or references to, every algorithm used, and salient details of auxiliary software (both research and commercial software) used in the computation.

(ICERM workshop report, https://icerm.brown.edu/topical_workshops/tw12-5-rcem/icerm_report.pdf)

Reproducibility in Computational and Experimental Mathematics 2012

Implementation Criteria II:

- Discussion of the adequacy of parameters such as precision level and grid resolution.
- Proper citation of all code and data used, including that generated by the authors.
- Availability of computer code, input and output data, with some reasonable level of documentation.
- *Avenues of exploration examined throughout development, including information about negative findings.*
- *Instructions for repeating computational experiments described in the article.*
- Precise functions were given, with settings.
- Salient details of the test environment, including hardware, system software, and number of processors used.

(ICERM workshop report, https://icerm.brown.edu/topical_workshops/tw12-5-rcem/icerm_report.pdf)

Reproducibility in Computational and Experimental Mathematics 2012

Archiving Criteria:

- Data documented to clearly explain what each part represents.
- *Data archived with significant longevity expected.*
- *Data location provided in the acknowledgements.*
- Authors have documented use and licensing rights.
- Software documented well enough to run it and what it ought to do.
- *The code is publicly available with no download requirements.*
- There was some method to track changes/to the software, as well as some certainty that the code is securely archived.

(ICERM workshop report, https://icerm.brown.edu/topical_workshops/tw12-5-rcem/icerm_report.pdf)

NTNU

PART 4

CHALLENGES

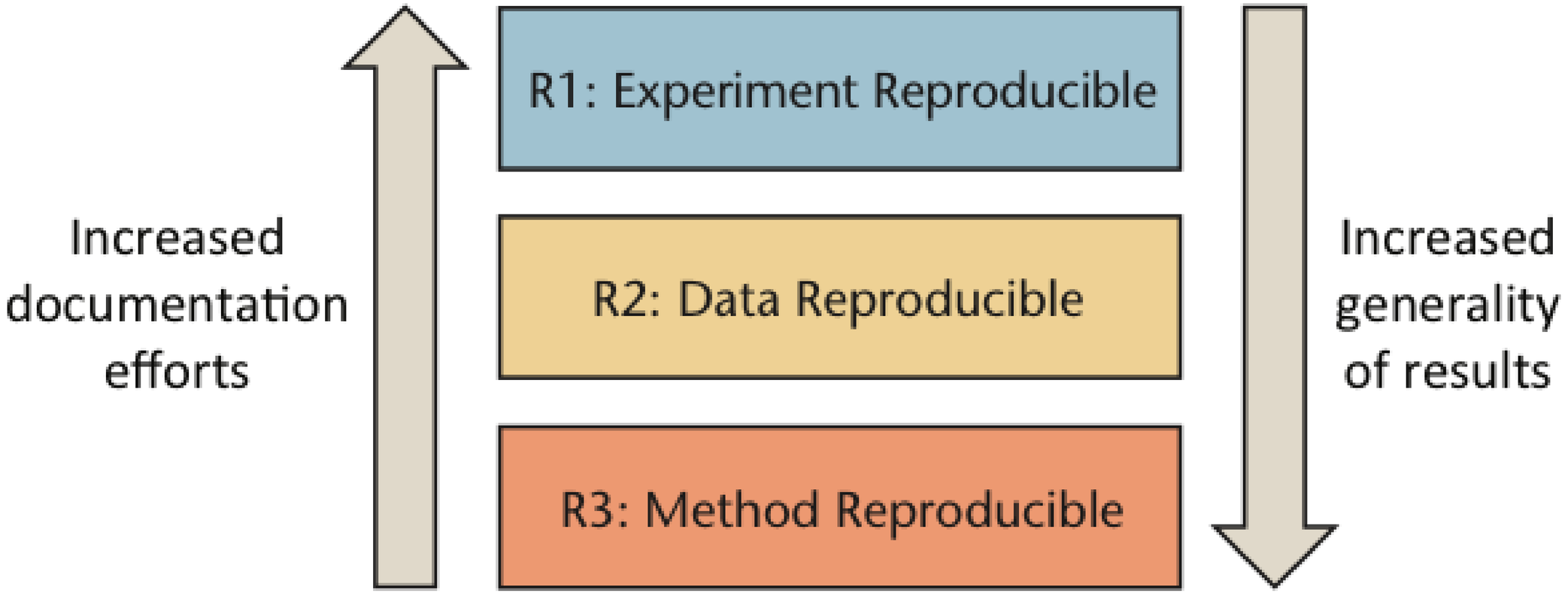
Open Questions for Reproducibility

- Can we agree on a definition of reproducibility?
- Will it rely on documenation or result or both?

- What do we mean by the same result?
- Does this change between reproducibility degrees?

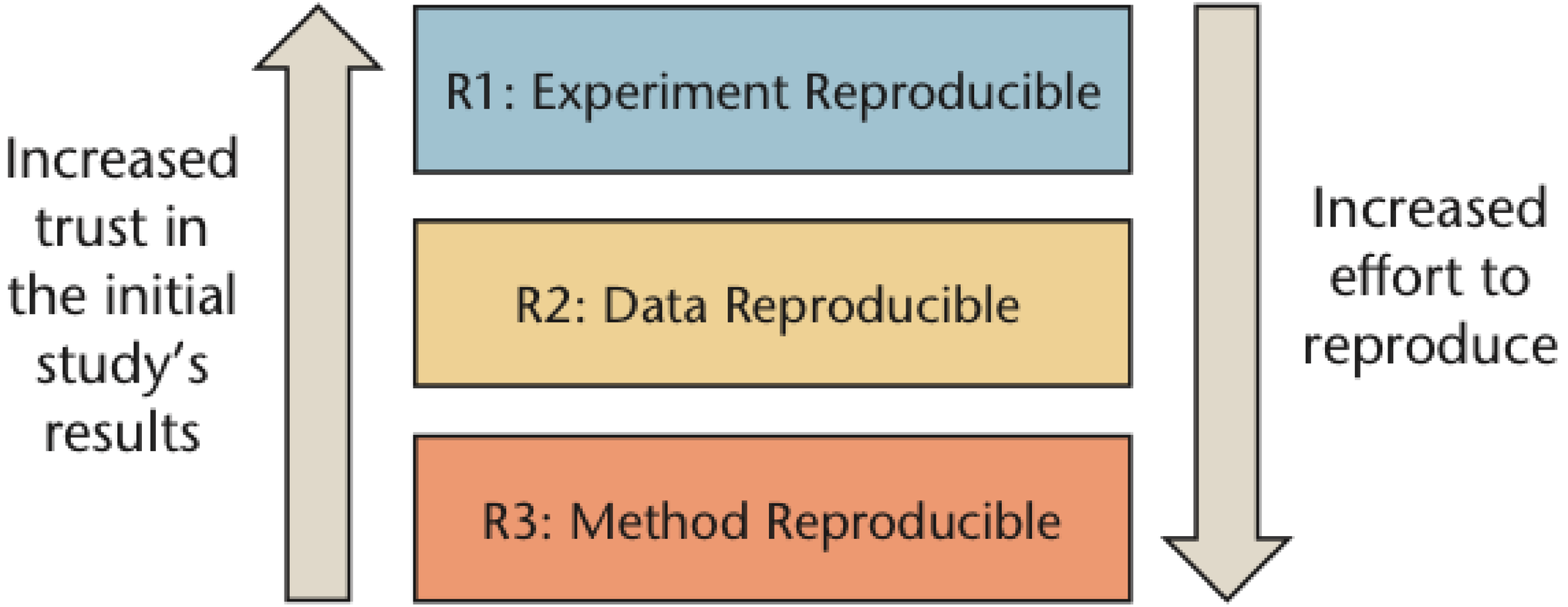
- Are the levels properly defined?
- What exactly must be documented?
- For each level?

PoV of Original Researchers



(Gundersen, Gil and Aha, AI Magazine, forthcoming 2018)

PoV of Independent Researchers



(Gundersen, Gil and Aha, AI Magazine, Fall 2018)

Barriers to reproducibility

- **Time consuming:** Proper documentation, questions from external researchers, maintenance cost.
- **No incentives:** Not required by publishers, grant makers, evaluating committees for research positions.
- **Risk future work:** Sharing of data, code and detailed experiment procedures will enable independent researchers to quickly build on the published research, and jeopardize possible new publications.

(Gundersen, AI Magazine, forthcoming)

NTNU

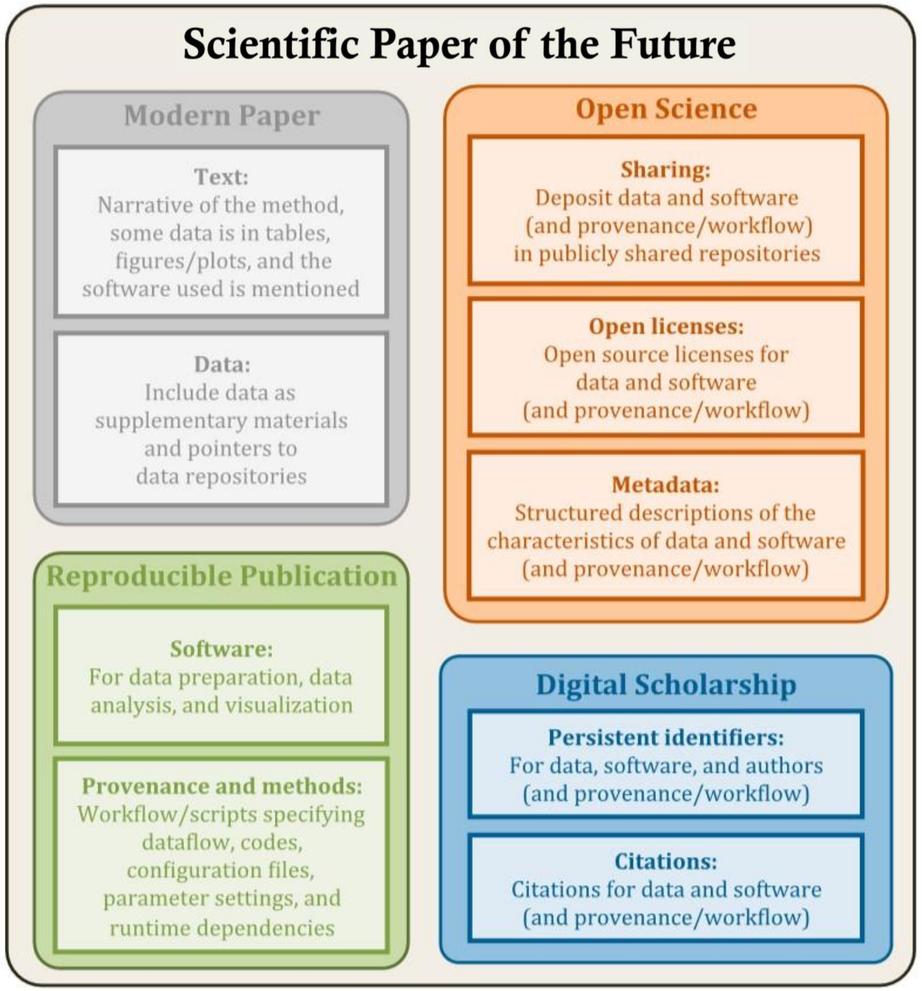
PART 5

FUTURE

Removing Barriers

- **Build infrastructure:** Reduce the effort for individuals.
- **Provide infrastructure:** Publishers, academic institutions, and grant makers could provide the infrastructure.
- **Eligibility requirements:** Make reproducibility a requirement for academic positions.
- **Emphasize quality:** Not quantity when evaluating researchers for positions.
- **Reward sharing:** Not only review how many papers have been published, but also how many data sets and code repositories are shared when reviewing candidates.
- **Reward reproducibility:** Have others reproduced the research? Does it represent scientific knowledge?

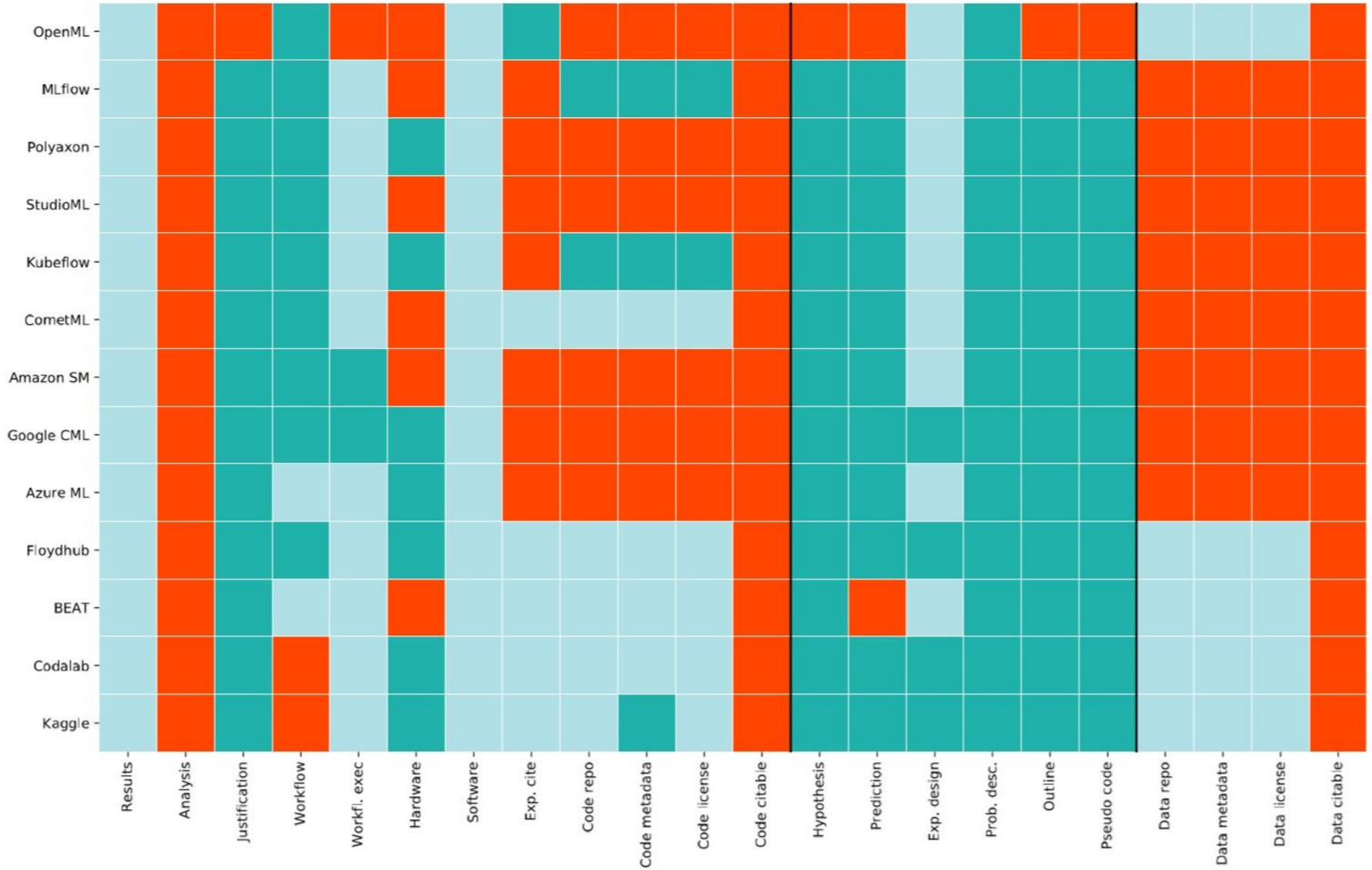
Scientific Paper of the Future



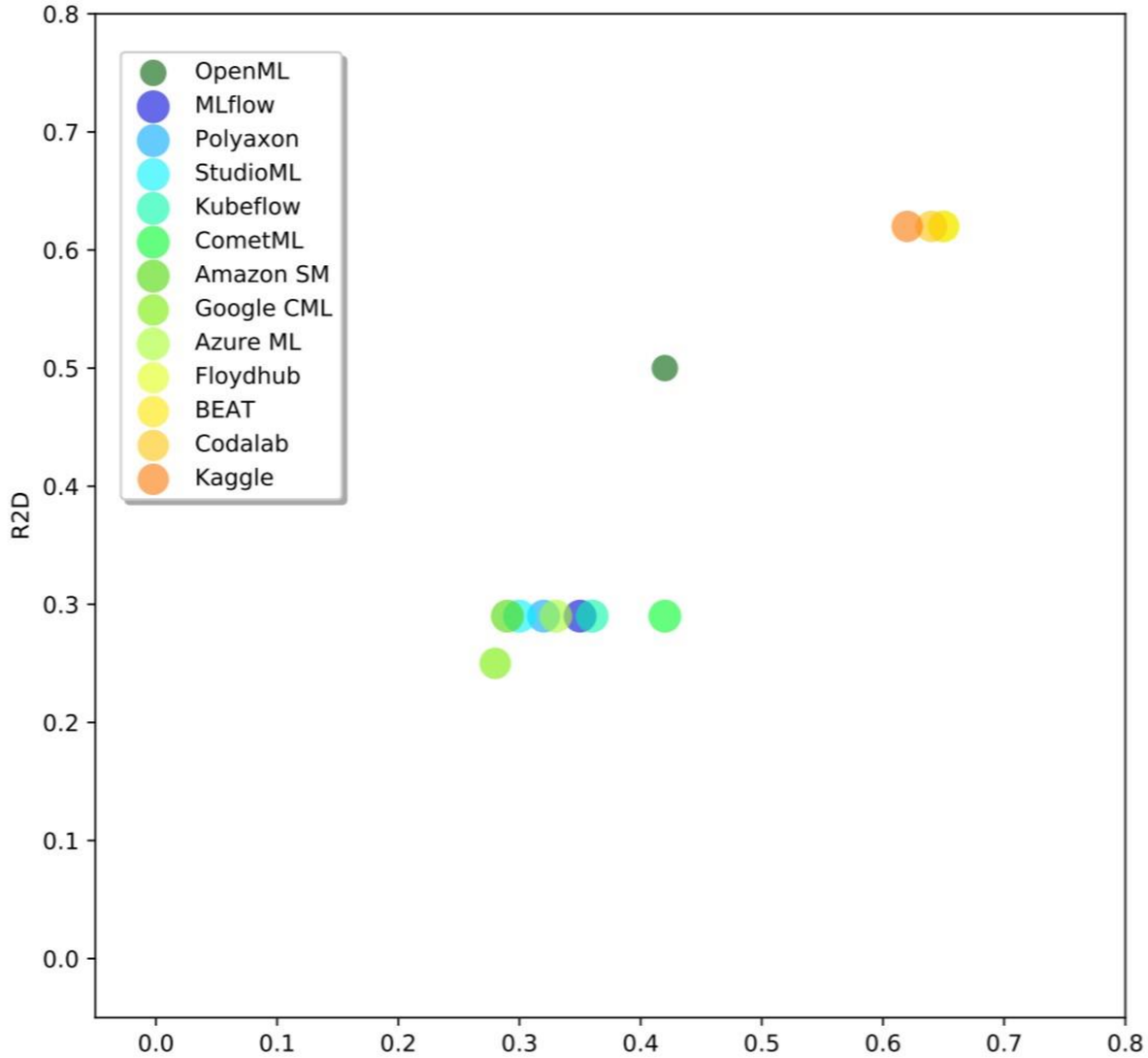
- Tutorial @ AAAI 2017
- Pratical advise on how to make your research reproducible.
- Complete tutorial available at website.

(Y. Gil, D. Garijo, G. Peretsman-Clement, AAAI 2017 Tutorial)

Machine Learning Platforms



Machine Learning Platforms



In the Future, Who Are Responsible for What?

- **Academic institutions:** Evaluate whether open positions will be filled by scientists whos research is reproducible.
- **Publishers:** Provide infrastructure, such as code and data repositories, and guidelines on how to publish research including code and data.
- **Grant makers:** Require funded research to be reproducible and that data and code are shared.
- **Scientists:** Ensure that proper science is conducted.

What if We Cannot Share?

- Many valid reasons for not sharing.
  - Privacy, cannot share private data.
  - Data set is too large.
  - Company IP.
  - Commercial software.

Open Questions

- Should public money fund research for which results are not shared with the public?
- Should such a requirement apply to academia only?
- What if academia collaborate with industry?
- Is it not better to publish papers that describe ideas and do not share code and data, than not publish the ideas?
- How can we ensure that industry continue to publish?
- Should we be more explicit; should we label our research with the reproducibility degree?

SUMMARY

NTNU |Odd Erik Gundersen|odderik@ntnu.no

Current State

- Most research is not reproducible.
- Research is so poorly documented that it is hard if even possible to reproduce the results.
- According to Ioannidis, most research findings are false.

(J. P. Ioannidis, PLoS Medicine, 2005)

Scenario I - Dark Future

- We continue in the same vain.
- AI research loses credibility, as it is do not follow scientific method.

(J. P. Ioannidis, PLoS Medicine, 2005)

Scenarion II - Bright Future

- Many good tools exist that support reproducible experiments. We start using them.
- The amount of research that share data and code increases.
- We see that small changes make a huge impact and we improve further.
- Fewer dead-ends are visited and knowledge improves faster.
- Virtuous circle.

Important Notes

- Perfectly executed research need not be reproducible!
- Claims made by reproducible research might be wrong!
- Reproducibility is about transparency and enabling others to validate results.

NTNU

Reproducibility is a core part of science

Are your experiments reproducible?

References

- **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.
- **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, M. Mølnå and N. Grimstad, forthcoming.

References

- **Deterministic Implementations for Reproducibility in Deep Reinforcement Learning**, P. Nagarajan, G. Warnell, P. Stone, AAAI 2019 workshop on Reproducible AI, 2019
- **The Machine Learning Reproducibility Checklist,** J. Pineau, 2019
- **Setting the Default to Reproducible**, ICERM workshop report, https://icerm.brown.edu/topical_workshops/tw12-5-rcem/icerm_report.pdf, 2012
- **An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility**, Stodden, Seiler and Ma, PNAS, 2018.
- **An Evaluation of the Software System Dependency of a Global Atmospheric Model**, S. HONG, M. KOO, and J. JANG, Monthly Weather Review, 2013.
- **Are GANs Created Equal? A Large-scale Study** Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O., *Advances in neural information processing systems*. 2018.
- **1,500 scientists lift the lid on reproducibility**, M. Baker, *Nature News*, *533*(7604), 452, 2016.

References

- **Unreproducible Research is Reproducible**, X. Bouthillier, C. Laurent, and P. Vincent, *International Conference on Machine Learning*. 2019.
- **Repeatability in computer systems research**, C. Collberg and T. A. Proebsting, Communications of the ACM, 2016
- **Learn to Write a Scientific Paper of the Future: Reproducible Research, Open Science, and Digital Scholarship,** Y. Gil, D. Garijo, G. Peretsman-Clement, AAAI 2017 Tutorial, http://scientificpaperofthefuture.org/
- **Why Most Published Research Findings Are False**, J. P. Ioannidis, *PLoS medicine*, *2*(8), e124, 2005