MAKING YOUR RESEARCH REPRODUCIBLE

Odd Erik Gundersen, dr. philos.

Chief AI Officer, TrønderEnergi AS

Adjunct Associate Professor, NTNU

NTNU

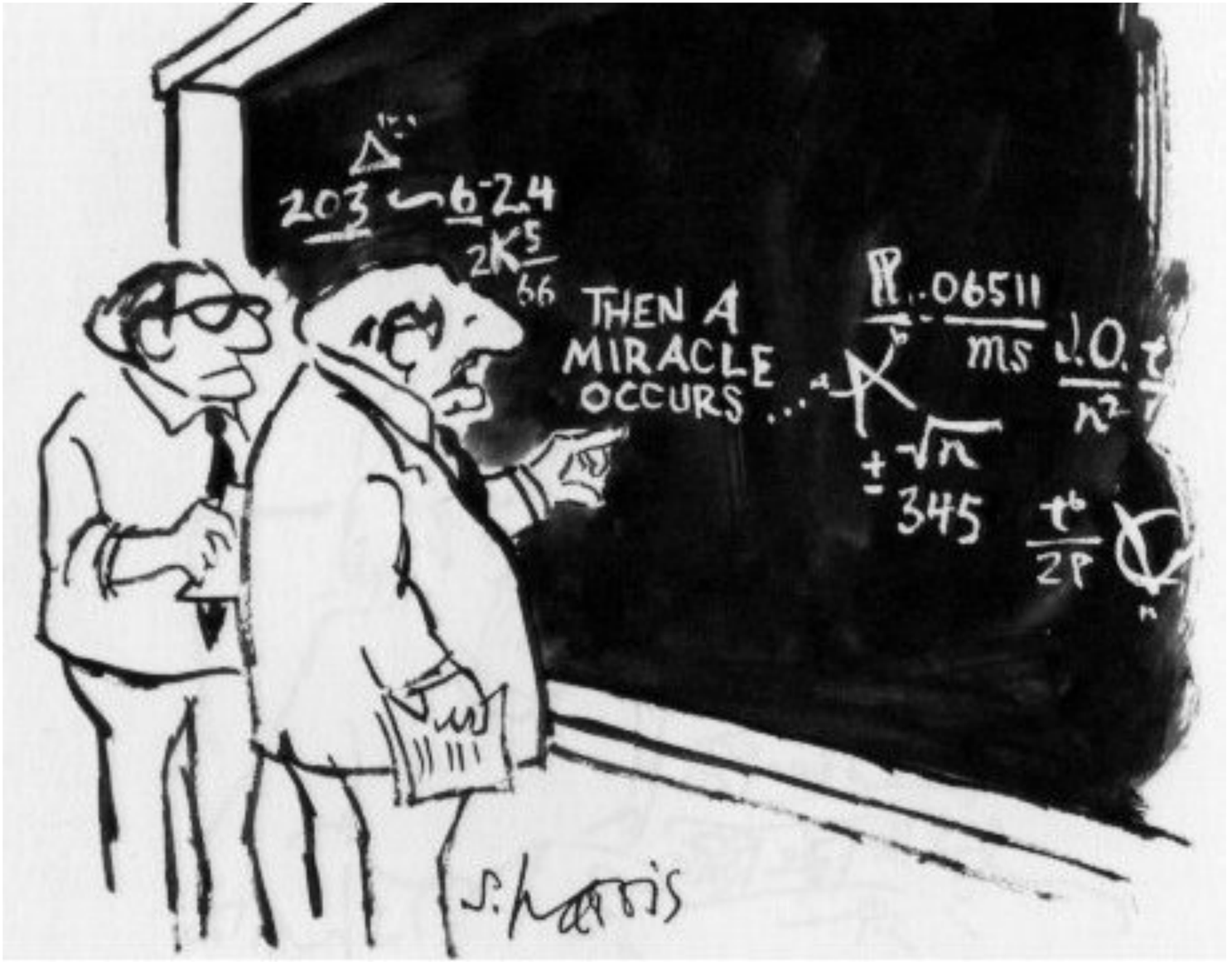
TrønderEnergi®

Norwegian Open AI Lab

AlphaGo

“Impressive results. No code. No model.”

(Silver et al, 2016)



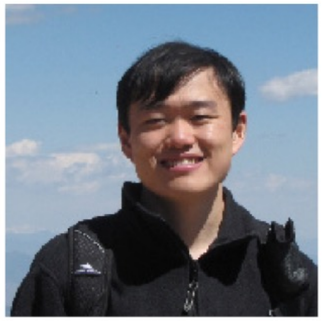
"I think you should be more explicit here in step two."

NTNU

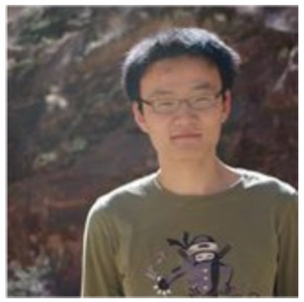
Reproducing AlphaZero with ELF: What we learned

Yuandong Tian

Facebook AI Research

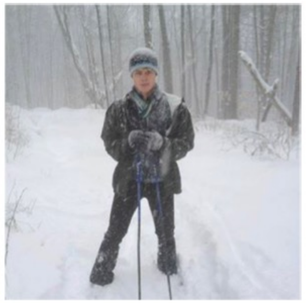

Yuandong Tian


Jerry Ma*


Qucheng Gong*


Shubho Sengupta*

Zhuoyuan Chen


James Pinkerton


Larry Zitnick

NTNU

Reproducing AlphaZero with Elf

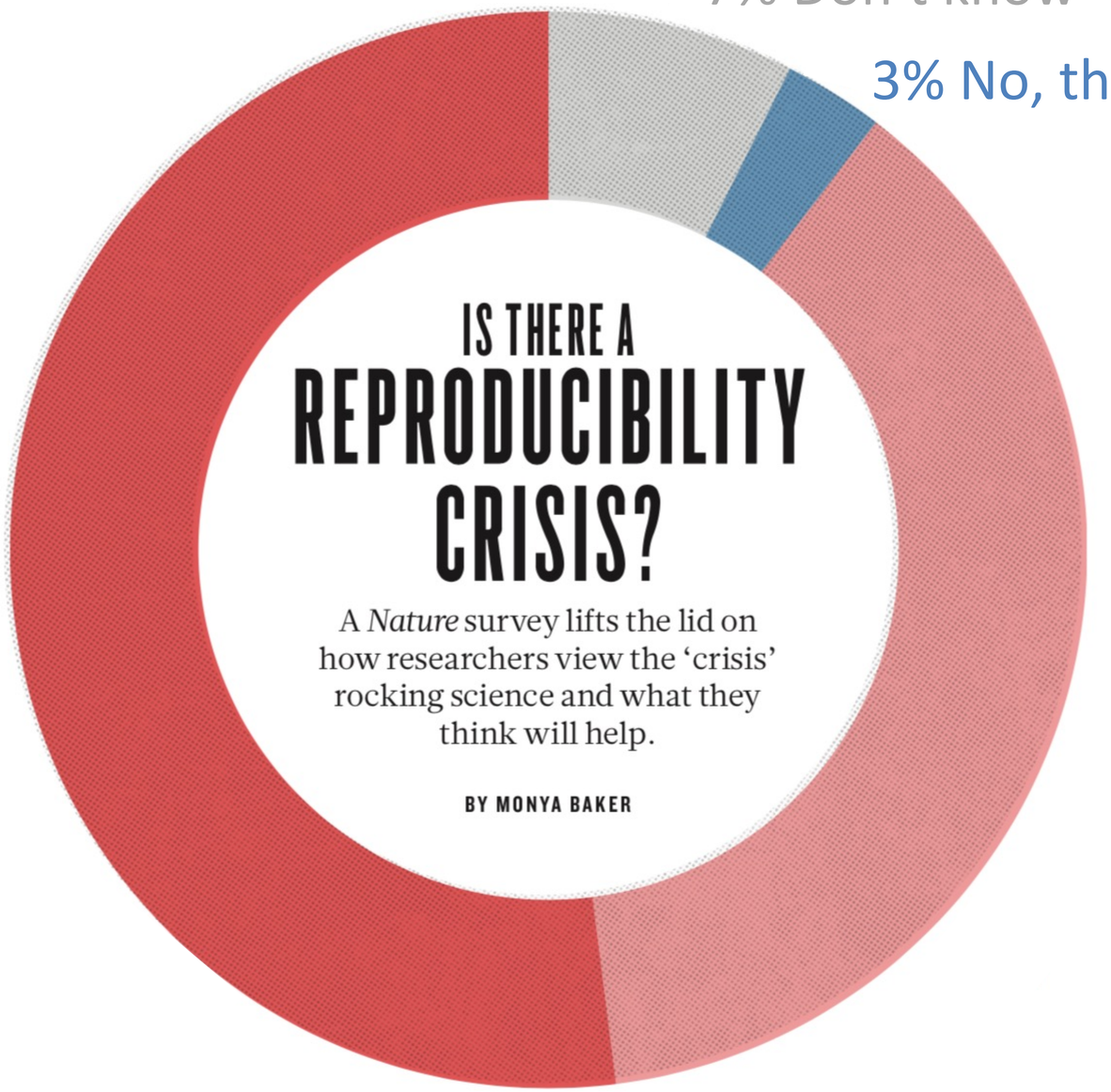
- Hard to reproduce
  - **Details are missing in the paper**
  - **Huge computational cost (15.5 years to generate 4.9M selfplays with 1 GPU)**
  - **Sophisticated (distributed) systems.**
- Lack of ablation analysis
  - What factor is critical for the performance?
  - Is the algorithm robust to random initialization and changes of hyper parameters?
  - How the ladder issue is solved?
- Lots of mysteries
  - Is the proposed algorithm really universal?
  - Is the bot almighty?
  - Is there any weakness in the trained bot?

ELF OpenGo: An Analysis and Open Reimplementation of AlphaZero, Tian et al, ICML 2019.

7% Don’t know

3% No, there is no crisis

52%
Yes, a significant crisis

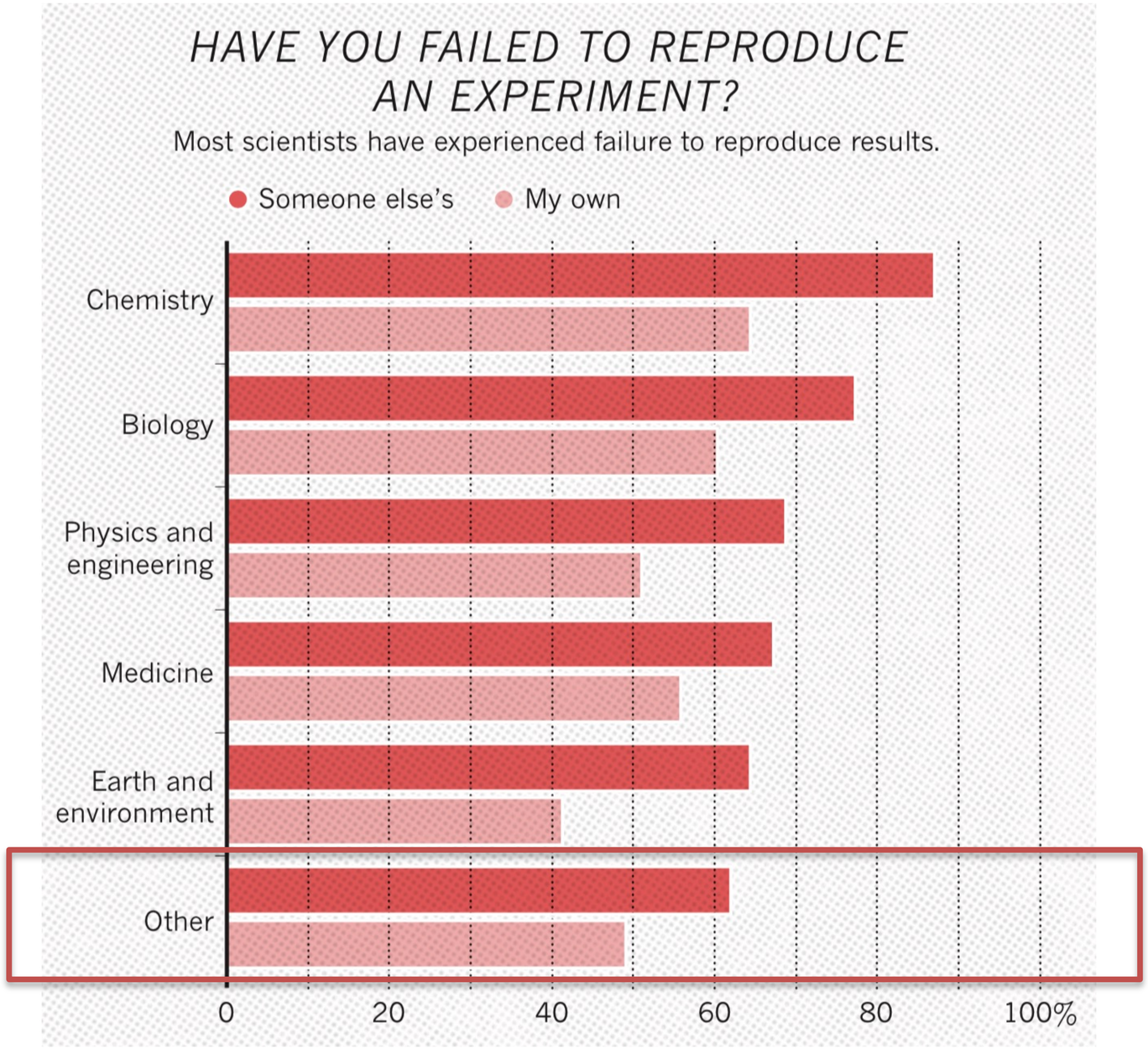


38%
Yes, a slight crisis

1,576
RESEARCHERS SURVEYED

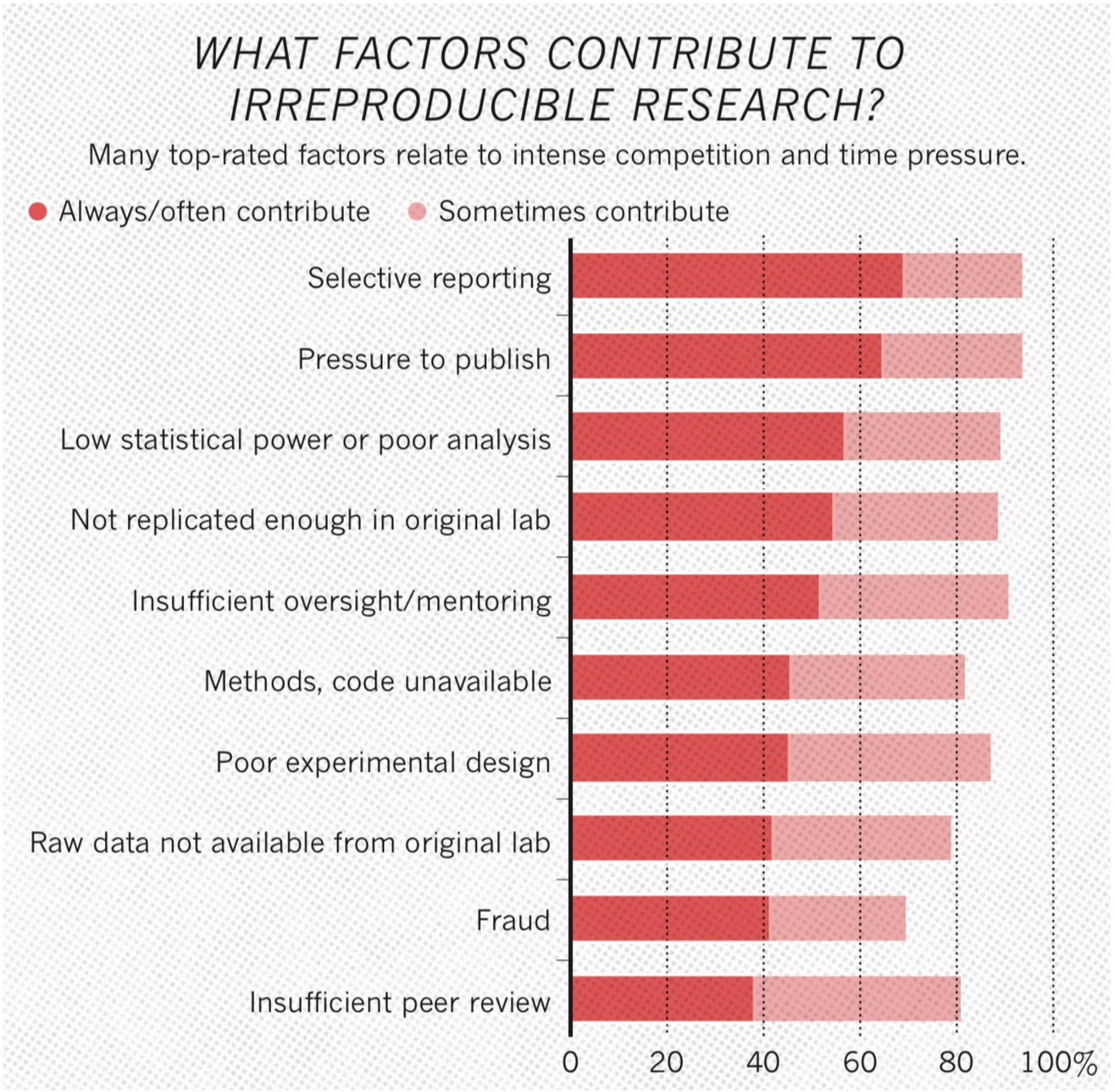
(M. Baker, Nature, 2016)

(Gundersen , 2020)



Computer Science

(M. Baker, Nature, 2016)

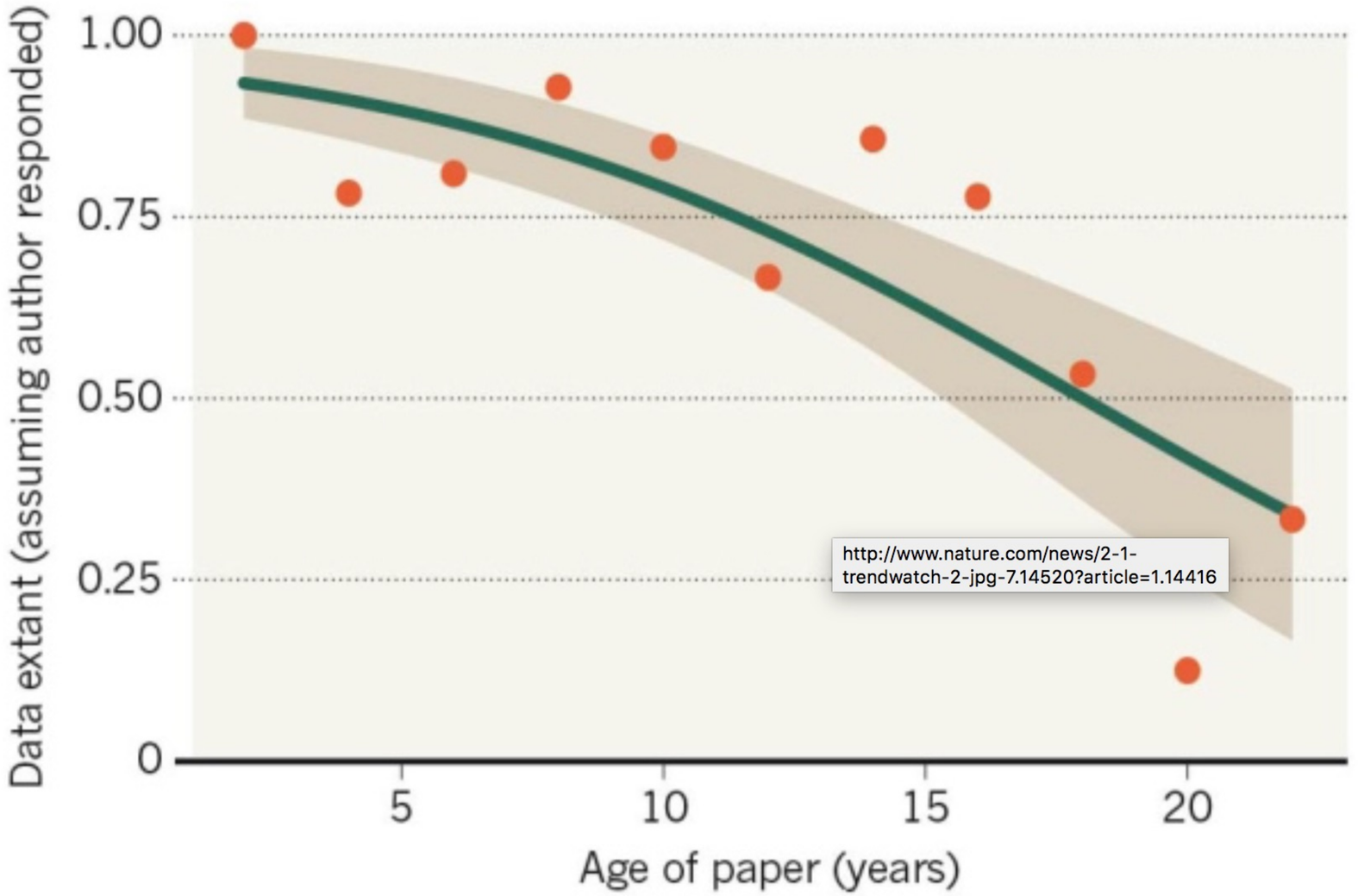


(M. Baker, Nature, 2016)

NTNU

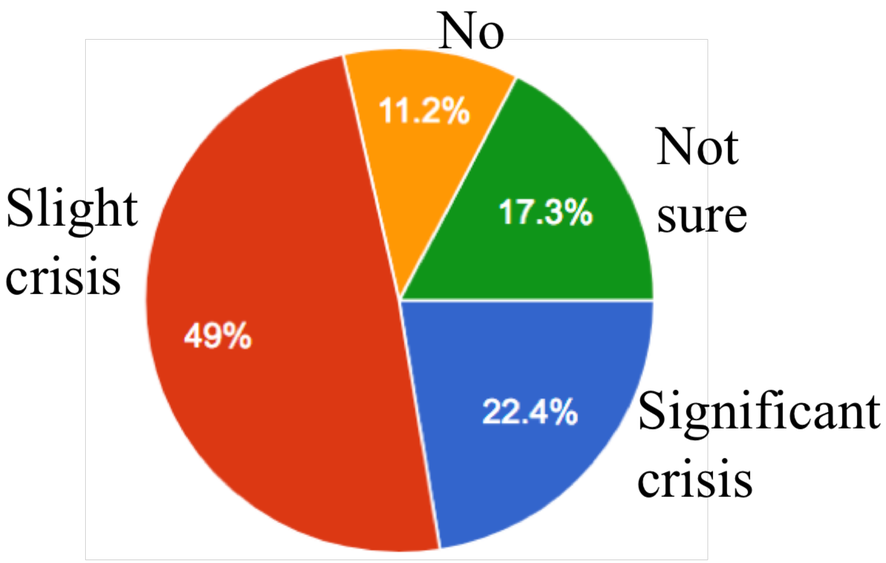
MISSING DATA

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

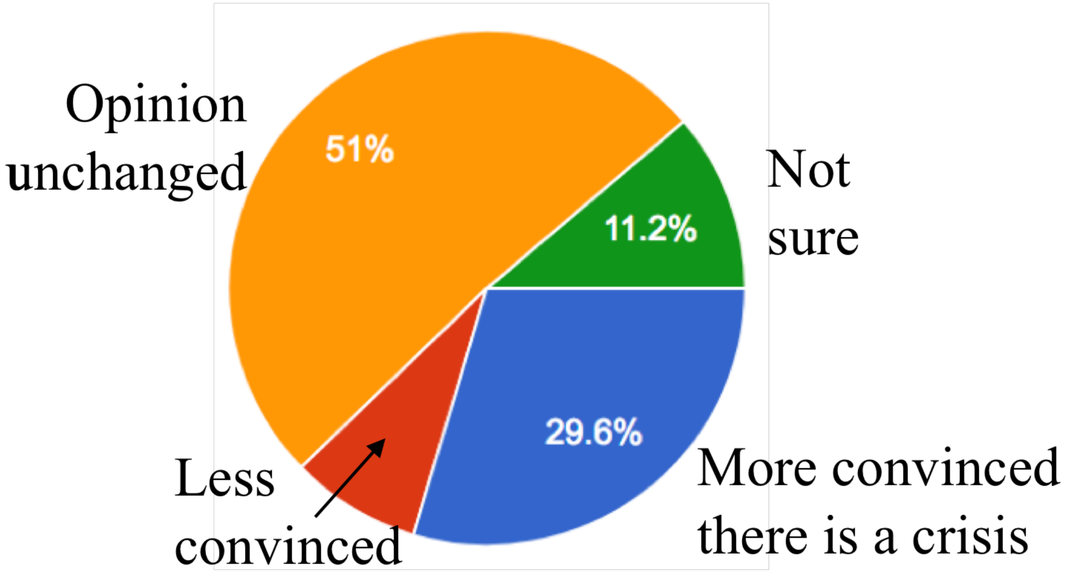


ICLR 2018 Reproducibility Challenge

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



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



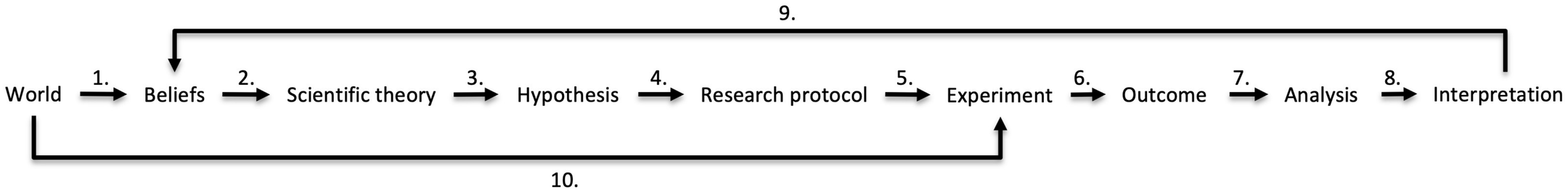
(J. Pineau, ICLR keynote, 2018)

NTNU

REPRODUCIBILITY

PART II

The Scientific Method - Process



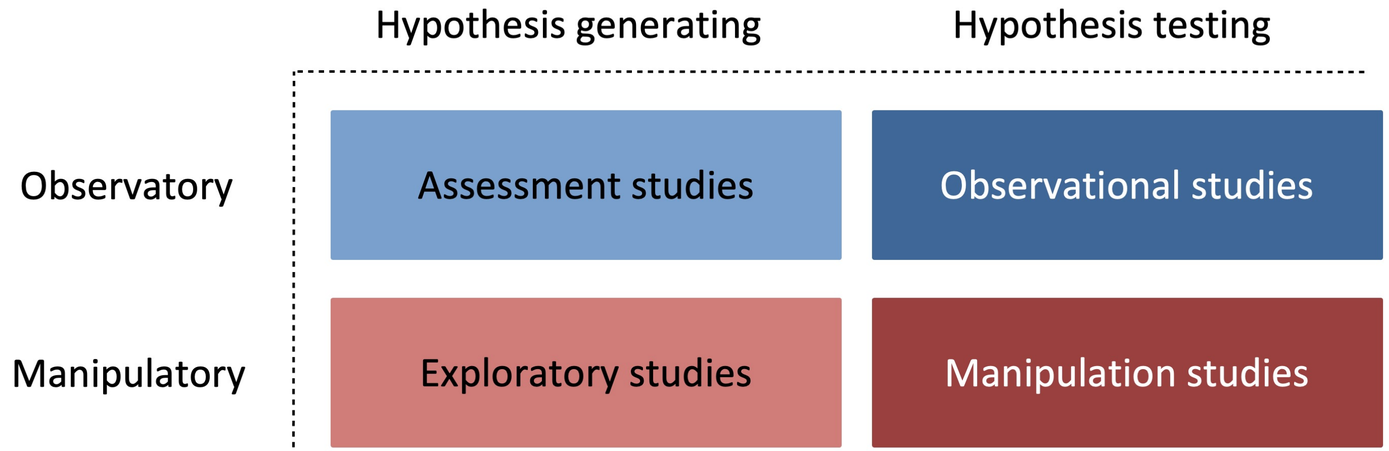
(Gundersen 2021)

The Scientific Method - Steps

1. Observe the world and form beliefs about it
2. Explain causes and effects by forming a scientific theory
3. Formulate a genuine test of the scientific theory as a hypothesis
4. Design an experiment to test the hypothesis and document it in a research protocol
5. Implement the experiment so that it is ready to be conducted
6. Conduct the experiment to produce results
7. Analyze the results to make an analysis
8. Interpret the findings
9. Update beliefs according to the interpretation
10. Observe the world in a structured manner

(Gundersen 2021)

Types of Empirical Studies



Hypothesis generating - identify and suggest possible hypotheses.

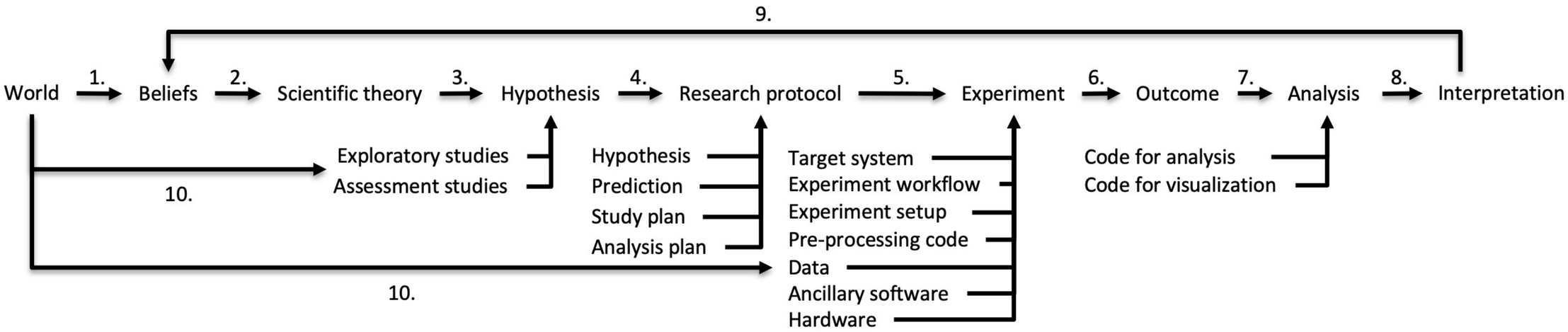
- ***Exploratory*** Yields casual hypotheses by collecting data and analyzing it in many ways.
- ***Assessment*** Establish baselines and ranges as well as other behaviors of system or environment.

Hypothesis testing - test explicit and precise hypotheses

- ***Observation*** Collect data in a way that does not directly interfere with how the data arise, establish an association.
- ***Manipulation*** Test hypotheses about causal influences of factors by manipulating them and and noting effects on measure variables.

(Cohen 1995)

The Scientific Method in ML



(Gundersen 2021)

Example of Experiment

Multi-column Deep Neural Networks for Image Classification

Dan Cireşan, Ueli Meier and Jürgen Schmidhuber
IDSIA-USI-SUPSI
Galleria 2, 6928 Manno-Lugano, Switzerland
{dan,ueli,juergen}@idsia.ch

Abstract

Traditional methods of computer vision and machine learning cannot match human performance on tasks such as the recognition of handwritten digits or traffic signs. Our biologically plausible, wide and deep artificial neural network architectures can. Small (often minimal) receptive fields of convolutional winner-take-all neurons yield large network depth, resulting in roughly as many sparsely connected neural layers as found in mammals between retina and visual cortex. Only winner neurons are trained. Several deep neural columns become experts on inputs preprocessed in different ways; their predictions are averaged. Graphics cards allow for fast training. On the very competitive MNIST handwriting benchmark, our method is the first to achieve near-human performance. On a traffic sign recognition benchmark it outperforms humans by a factor of two. We also improve the state-of-the-art on a plethora of common image classification benchmarks.

1. Introduction

Recent publications suggest that unsupervised pretraining of deep, hierarchical neural networks improves supervised pattern classification [2, 10]. Here we train such nets by simple online back-propagation, setting new, greatly improved records on MNIST [19], Latin letters [13], Chinese characters [22], traffic signs [33], NORB (jittered, cluttered) [20] and CIFAR10 [17] benchmarks.

We focus on deep convolutional neural networks (DNN), introduced by [11], improved by [19], refined and simplified by [1, 32, 7]. Lately, DNN proved their mettle on data sets ranging from handwritten digits (MNIST) [5, 7], handwritten characters [6] to 3D toys (NORB) and faces [34]. DNNs fully unfold their potential when they are wide (many maps per layer) and deep (many layers) [7]. But training them requires weeks, months, even years on CPUs. High data transfer latency prevents multi-threading and multi-CPU code from saving the situation. In recent years, however, fast parallel neural net code for graphics cards (GPUs) has overcome this problem. Carefully designed GPU code for image classification can be up to two orders of magnitude faster than its CPU counterpart [35, 34]. Hence, to train huge DNN in hours or days, we implement them on GPU, building upon the work of [5, 7]. The training algorithm is fully online, i.e. weight updates occur after each error back-propagation step. We will show that properly trained wide and deep DNNs can outperform all previous methods, and demonstrate that unsupervised initialization/pretraining is not necessary (although we don't deny that it might help sometimes, especially for datasets with few samples per class). We also show how combining several DNN columns into a Multi-column DNN (MCDNN) further decreases the error rate by 30-40%.

2. Architecture

The initially random weights of the DNN are iteratively trained to minimize the classification error on a set of labeled training images; generalization performance is then tested on a separate set of test images. Our architecture does this by combining several techniques in a novel way:

(1) Unlike the small NN used in many applications, which were either shallow [32] or had few maps per layer (LeNet7, [20]), ours are deep and have hundreds of maps per layer, inspired by the Neocognitron [11], with many (6-10) layers of non-linear neurons stacked on top of each other, comparable to the number of layers found between retina and visual cortex of macaque monkeys [3].

(2) It was shown [14] that such multi-layered DNN are hard to train by standard gradient descent [36, 18, 28], the method of choice from a mathematical/algorithmic point of view. Today's computers, however, are fast enough for this, more than 60000 times faster than those of the early 90s[1]. Carefully designed code for massively parallel graphics processing units (GPUs normally used for video games) allows for gaining an additional speedup factor of 50-100 over serial code for standard computers. Given enough labeled data, our networks do not need additional heuristics

[1] 1991 486DX-33 MHz, 2011 i7-990X 3.46 GHz

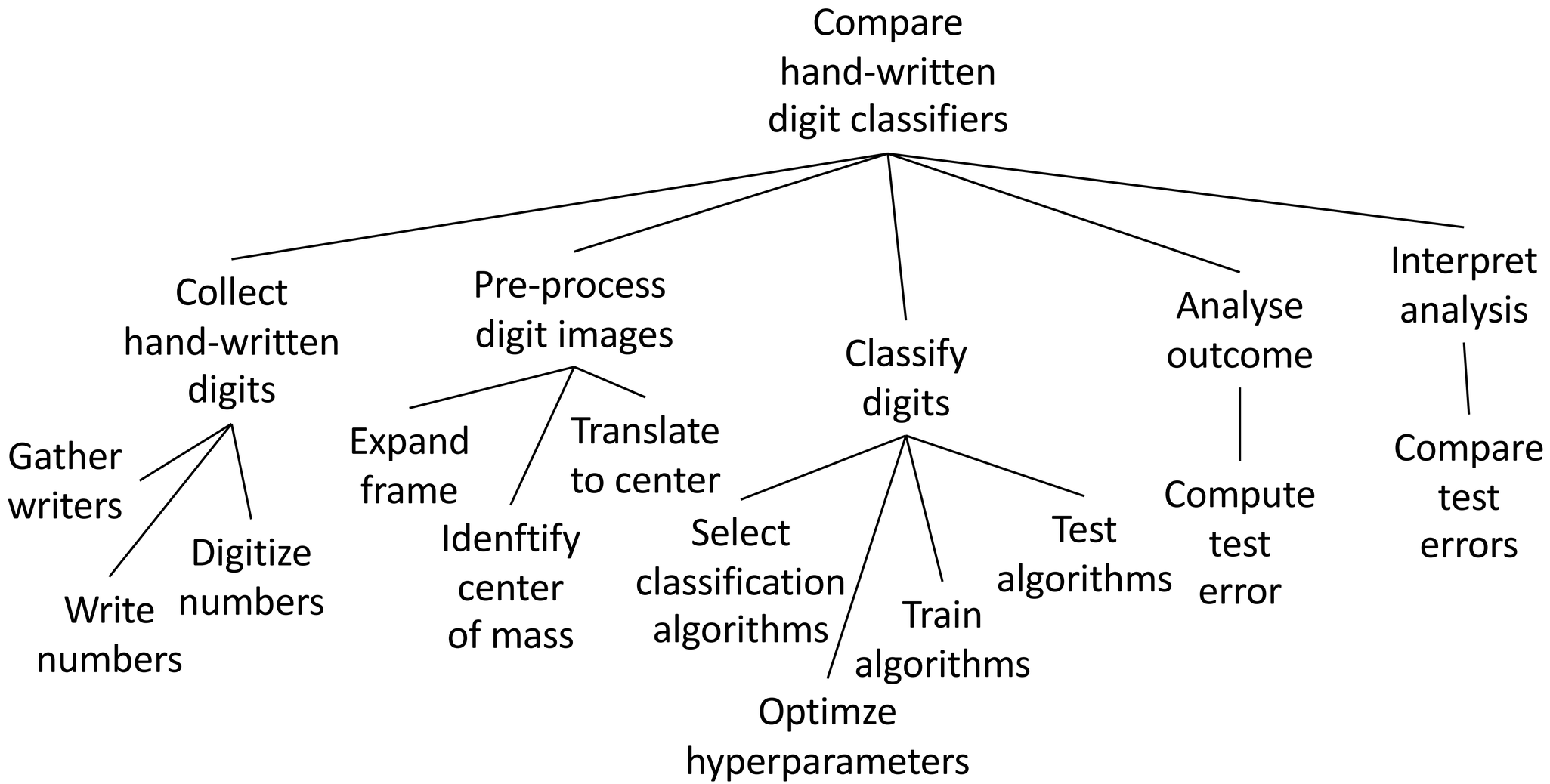
978-1-4673-1228-8/12/$31.00 ©2012 IEEE

Scientific theory: Deep neural networks are models of the brain, although simple ones, and as such intelligence could emerge from them.

Hypothesis: The performance of biological inspired deep convolutional neural networks is competitive with human performance on computer vision benchmark tasks.

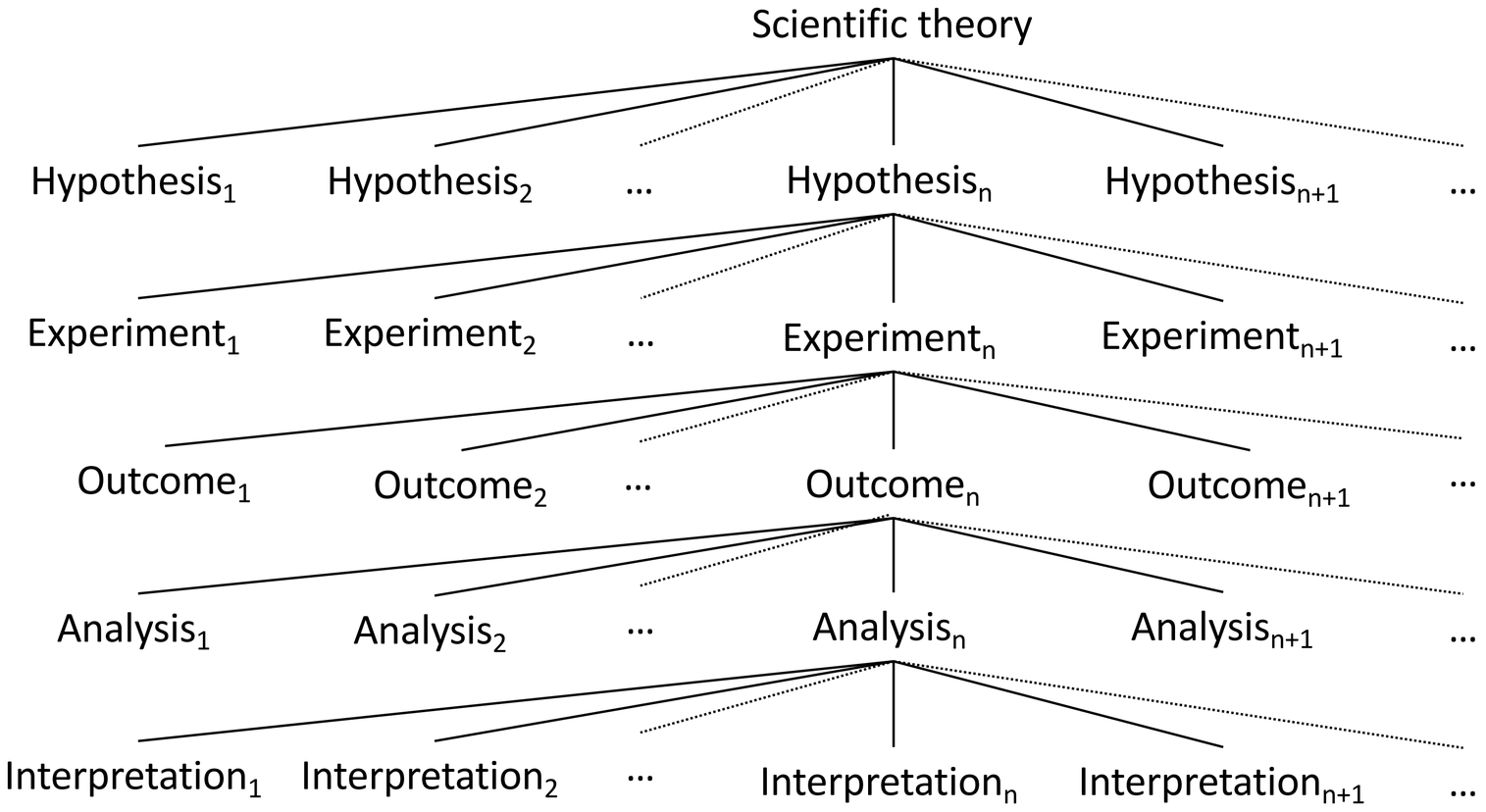
(Gundersen 2021)

Experiment: DNN for Image Classification



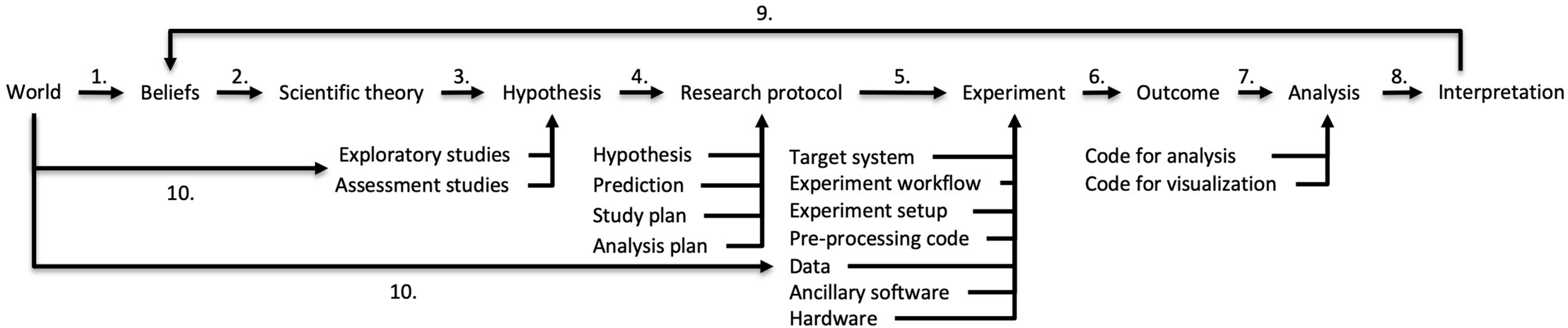
(Gundersen 2021)

The Scientific Method - Steps



(Gundersen 2021)

The Scientific Method in ML



(Gundersen 2021)

Definition of Reproducibility

Reproducibility is the ability of ***independent investigators*** to draw the ***same conclusions*** from an experiment by following the ***documentation*** shared by the original investigators.

(Gundersen 2021)

The Three Types of Reproducibility

Outcome reproducible *The outcome of the reproducibility experiment is the same as the outcome produced by the original experiment.*

Analysis reproducible *Outcome might differ, but same analysis and interpretation on different outcome leads to same conclusion.*

Interpretation reproducible *Neither the outcome nor the analysis need to be the same if the interpretation leads to the same conclusion.*

(Gundersen 2021)

The Three Types of Documentation

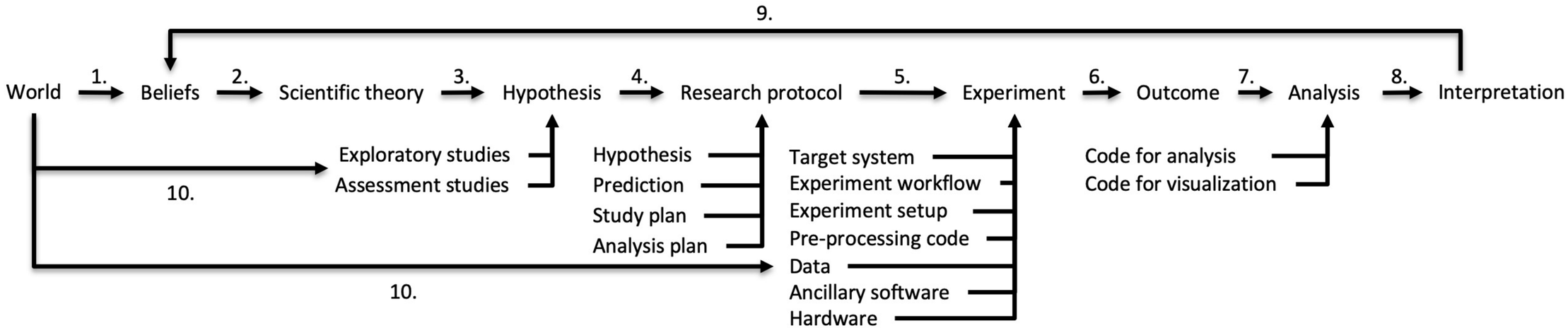
Description *Description of the AI method implemented by the AI program, the experiment being conducted and the analysis of the results as well as the hardware and ancillary software used for conducting the experiment.*

Code *AI Program code, code for setup and configuration, code controling workflow, code for analysis of results and visualization.*

Data All data used for conducting the experiment. Are the samples used for *training, validation and test specified? What about the results?*

(Gundersen 2021)

The Scientific Method in ML



Documentation → Experiment → Result

Description

Code

Data

Outcome

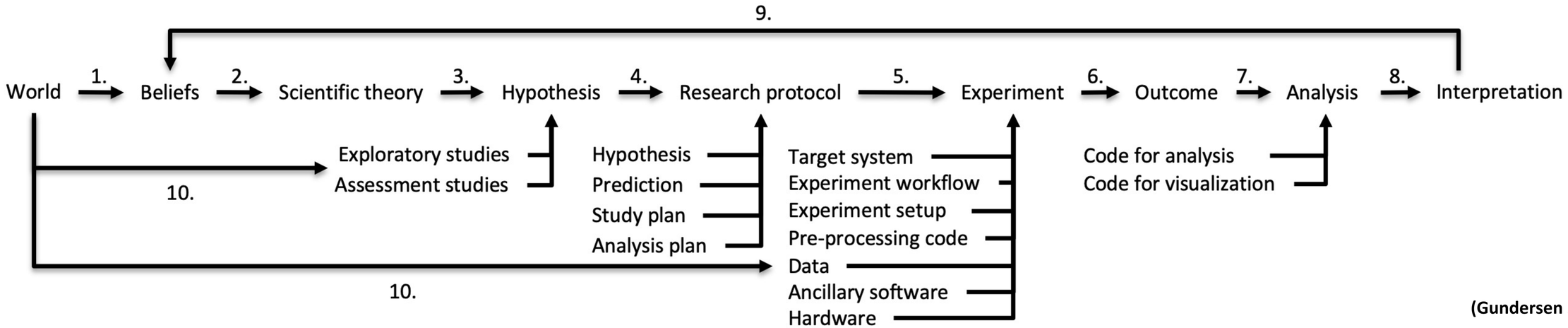
Analysis

Interpretation

(Gundersen 2021)

Degrees of Reproducibility

	Outcome	Analysis	Interpretation
R1 Outcome	**Same**	**Same**	**Same**
R2 Analysis	**Different**	**Same**	**Same**
R3 Interpretation	**Different**	**Different**	**Same**

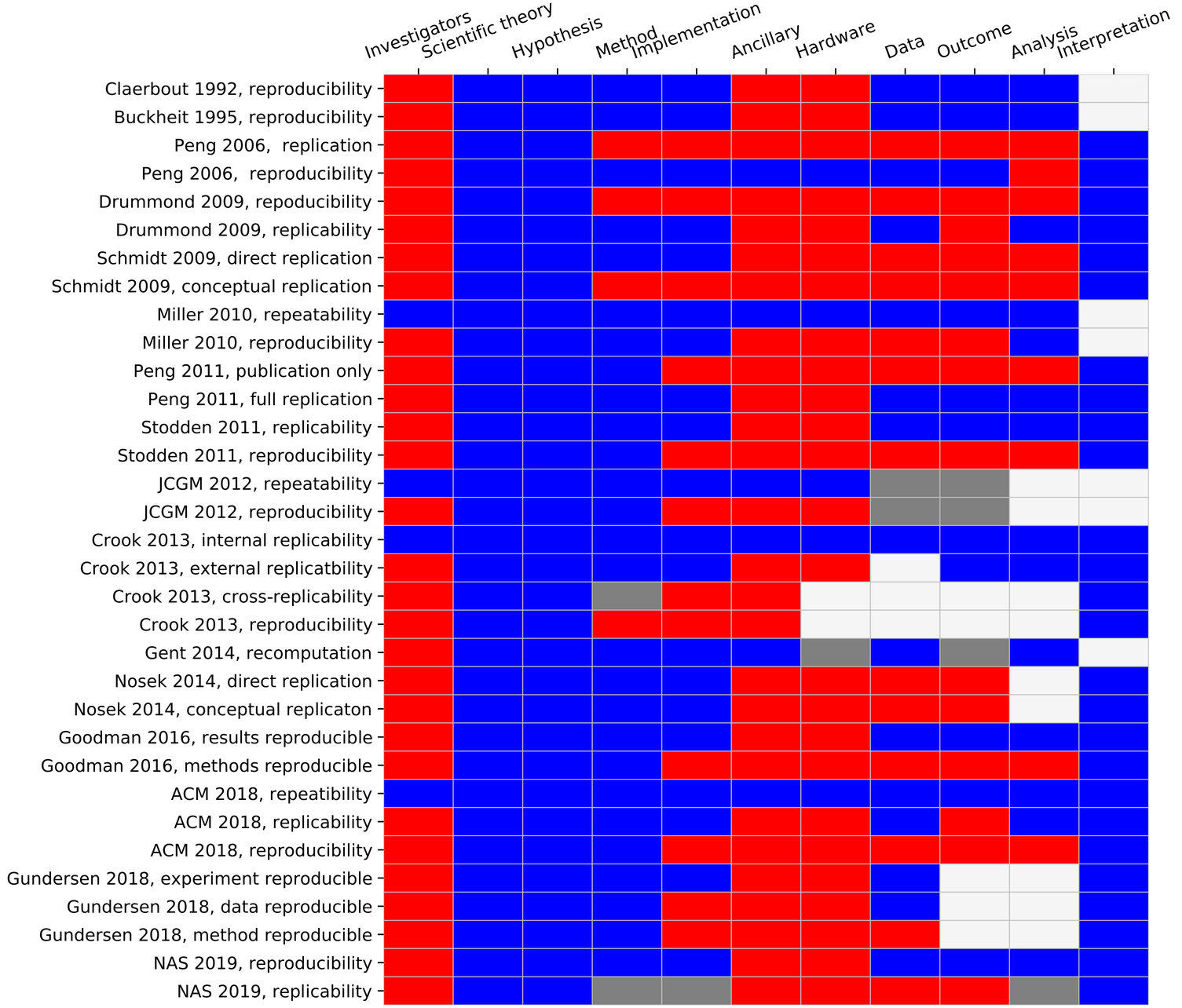


(Gundersen 2021)

Four Types of Reproducibility

	Text	Code	Data
Description	■		
Code	■	■	
Data	■		■
Experiment	■	■	■

(Gundersen 2021)



(Gundersen 2021)

NTNU

THE AAAI REPRODUCIBILITY CHECKLIST

PART III

General Reproducibility Guidelines for AI Research

Source: https://folk.idi.ntnu.no/odderik/reproducibility_guidelines.pdf

Version: 1.3 June 25, 2020

Authors: Odd Erik Gundersen, Yolanda Gil, Mausam

For each experiment, check that the following is described:

- How the experimental design rigorously tests the claims.
- The evaluation metrics and the motivation for choosing these metrics.
- All (hyper-)parameters for each model/algorithm, number and range of values tried per parameter, and the criterion for selecting best parameter setting.
- The final parameters for each model/algorithm.
- The computing infrastructure used for running the experiment (hardware and software), such as which software and version (libraries, frameworks, operating system etc), processing units (GPU/CPU), memory and more.
- For each reported result, the number of algorithm-runs it is averaged over and its variance.

For data used in the paper, check the following:

- For closed datasets, describe the dataset.
- For a new dataset, deposit it to a public repository with a description and metadata.
- For a new dataset, release it with a license that allows free usage for research purposes.
- All open datasets are cited.

For all code, check the following:

- All source code required for conducting the experiment is shared and cited.
- The version of the code used for conducting the experiments is specified.
- A license is added with the source code to allow free usage for research purposes.

For the paper, check the following:

- Claims being investigated are stated clearly.
- For theoretical papers, complete proofs are provided (for example in the appendix).
- Assumptions and limitations are identified.
- A conceptual outline and pseudo code describing the AI method is given.
- Statements about how the results substantiate the claims.

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

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

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

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

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

AAAI Reproducibility Checklist

Four sections:

1. The paper
2. Theoretical contributions
3. Data sets
4. Computational experiments

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.

This paper:

- clearly states what claims are being investigated (yes/partial/no)
- explains how the results substantiate the claims (yes/partial/no)
- explicitly identifies limitations or technical assumptions (yes/partial/no)
- includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA)

Does this paper make theoretical contributions? (yes/no)

If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
- Proofs of all novel claims are included. (yes/partial/no)
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
- Appropriate citations to theoretical tools used are given. (yes/partial/no)

Does this paper rely on one or more data sets? (yes/no)

If yes, please complete the list below.

- All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA)
- All datasets that are not publicly available are described in detail (yes/partial/no/NA)

Does this paper include computational experiments? (yes/no)

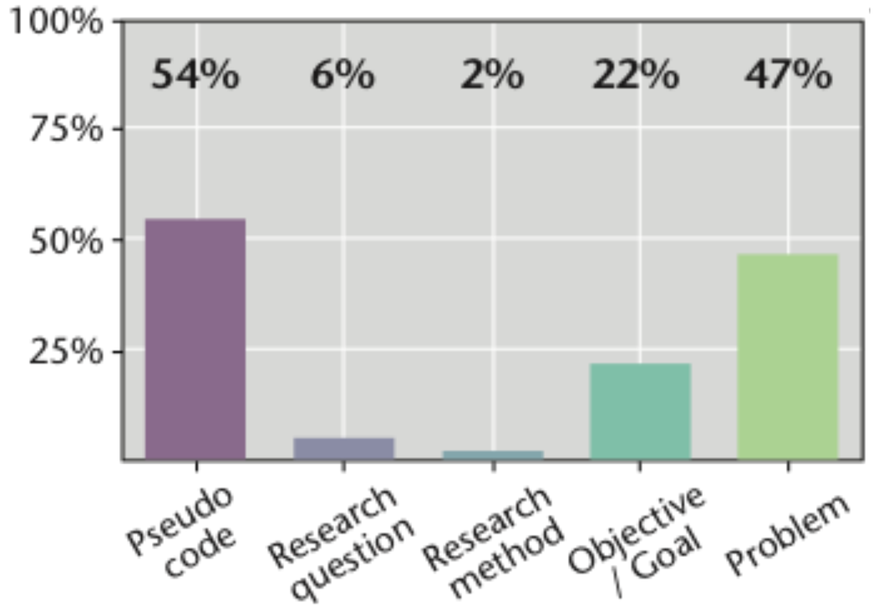
If yes, please complete the list below.

- All source code required for conducting experiments is included in a code appendix (yes/partial/no).
- All source code required for conducting experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no)
- This paper states the number of algorithm runs used to compute each reported result (yes/no)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA)
- This paper states the number and range of values tried per (hyper-)parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA)

Source: https://aaai.org/Conferences/AAAI-22/reproducibility-checklist/

The paper

- Claims are clearly stated.
- Explain how the results substantiate the claims.
- Explicitly identify limitations and or technical assumptions.
- Include conceptual outline/pseudocode of AI methods introduced.



(Gundersen and Kjensmo, 2018)

Research Protocol

Project Description

Short, high-level, overview

Research Questions

What is being investigated? What are the main research questions you are asking?

Why is the problem important, has anyone else said so? Briefly review previous research on each research question.

What is your contribution? How is your research topic different from what has been done before?

Methodology

What do you intend doing? Briefly describe the methods that you will use to answer your research questions.

Why is this strategy being adopted? Why is this necessary for your study?

Work Detail

Decide on the stages of the project and the dependencies between them. Compile a project plan.

- Risks (e.g., delays in obtaining key resources) and Risk Management Strategies.
- Timeline, including Gantt chart. Use specific dates so that you finish on time.
- Resources required (equipment, people, special software etc)
- Deliverables
- Milestones (which should refer to the Timeline)

Evaluation of Research Questions

You should have a plan for testing your system when it is complete. Work this out now; everything will be wasted if you finish your implementation but cannot evaluate your "advance" convincingly.

Indicate the interpretation and conclusions that you will place upon the results. What difference will they make? Indicate the implications of your research for current theory and practice.

Anticipated Outcomes

- What might we expect the outcomes of your project to be? What do you expect to find? The aim here is not to anticipate your study but rather to give an outline of what you envisage.
- Major software artefacts to be produced, their key features and major design challenges.
- Key success factors – how will you judge whether the project has succeeded or not?

Bibliography and Previous Systems

List the main sources on which your research will be based. In the proposal we want a preliminary outline of the key works. All work must be properly cited.

As your work progresses you have to show that you have read the relevant papers and books and understand the field. You should show that you know which important contributions are and how they are related and may be grouped. You should know where the concepts you use were first described.

Research Questions Clearly state what you are investigating?

Methodology How do you conduxt your investigation? Describe your experiment.

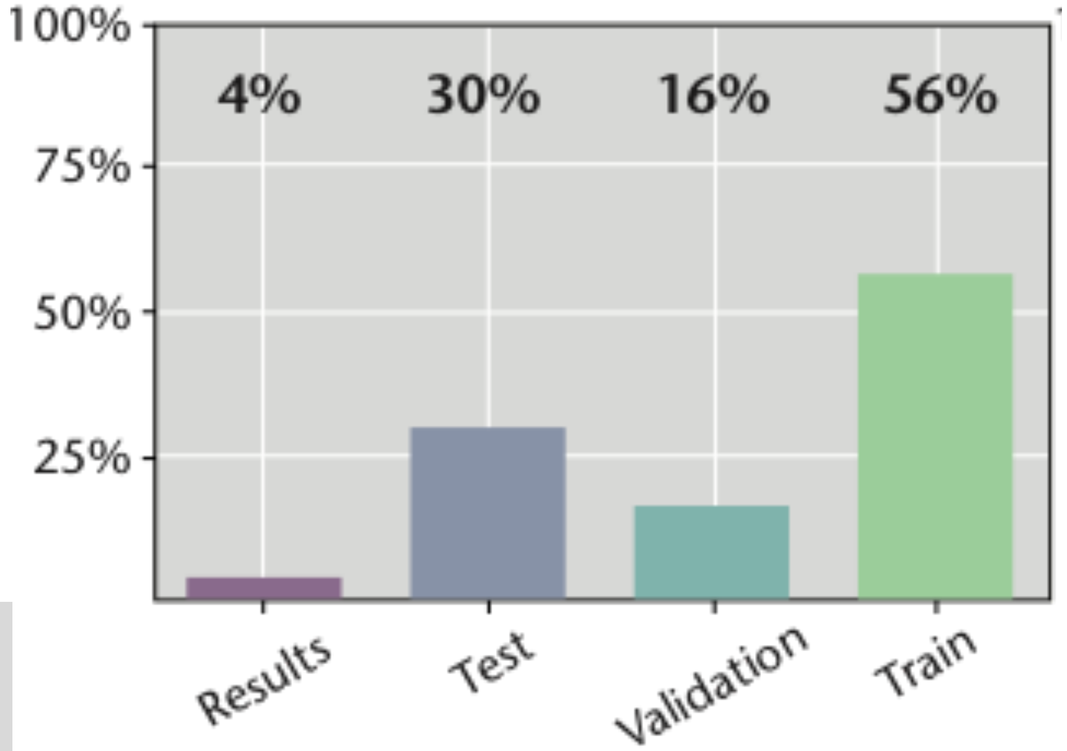
Evaluation of Research Questions What is the best way to evaluate the outcome of the experiment? Explain it.

Anticipate Outcomes Make a prediction. What do you expect and why?

Source: https://www.cs.uct.ac.za/teaching/forms/researchProposalGuide_2007.pdf

How well is data documented?

- We know we should not train and test on the same data.
- Is *Outcome Reproducible* an option if we do not know which samples were used for what?
- Can only check if *Outcome Reproducible* if results are shared.



The order a machine learning algorithm is fed training samples can affect the performance.

(Gundersen and Kjensmo, 2018)

An Unbiased Look at Dataset Bias

Selection bias *Does the dataset represent a fair sampling of the world?*

Capture Bias *Are the samples represented fairly (centered object, handle direction of mugs?)*

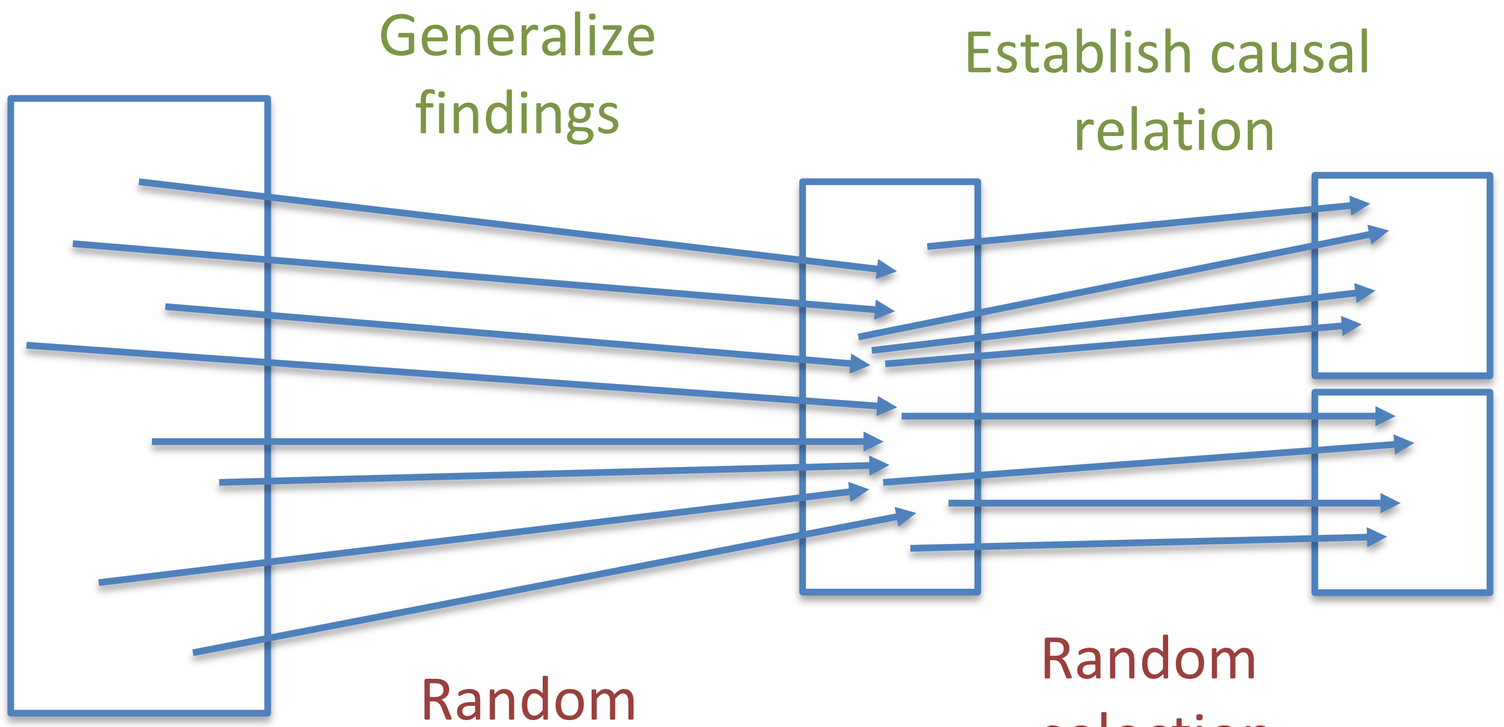
Negative bias *Does the data set contain negative examples as well?*

(Torralba and Efros 2011

Other issues

- Data version:
  - Are there different versions of the same dataset?
  - Some software libraries provide standard datasets as well i.e. seaborn and GluonTS.
  - Sometimes these differ from the original ones. Cite the correct version.
  - Sometimes the reported data is not the same as the published data (different number of samples).
- Large dataset:
  - Webscale datasets might not be stored after analysis. Outcome reproducibility not possible.
- Concept drift:
  - The real changes and datasets are static.
  - What was true one day is not true the next.
  - If the dataset is not shared it is impossible to know whether any differences are caused by concept drift or other issues related to the quality of the research.

Which Conclusions Can Be Drawn?



Hardware and Ancillary Software

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	♠

(Hong et al, 2013)

Code Version

Menu

[95] Groundhog: Addressing The Threat That R Poses To Reproducible Research

Posted on January 5, 2021 by Uri Simonsohn

R, the free and open source program for statistical computing, poses a substantial threat to the reproducibility of published research. This post explains the problem and introduces a solution.

The Problem: Packages

R itself has some reproducibility problems (see example in this footnote [1]), but the big problem is its packages: the addon scripts that users install to enable R to do things like run meta-analyses, scrape the web, cluster standard errors, format numbers, etc. The problem is that packages are constantly being updated, and sometimes those updates are not backwards compatible. This means that the R code that you write and run today may no longer work in the (near or far) future because one of the packages your code relies on has been updated. But worse, R packages depend on other packages. Your code could break after a package you don't know you are using updates a function you have never even used.

What data does R keep if you run *distinct(data, Subject)*?

Depends. When did you last update {dplyr} ?

Subject	dv	condition	mediator
11543	70	treatment	5
11543	70	treatment	5
555	3	control	6
555	3	control	6
47888	110	placebo	3
47888	110	placebo	3

Before June 24 2016

Subject	dv	condition	mediator
11543	70	treatment	5
555	3	control	6
47888	110	placebo	3

After June 24 2016

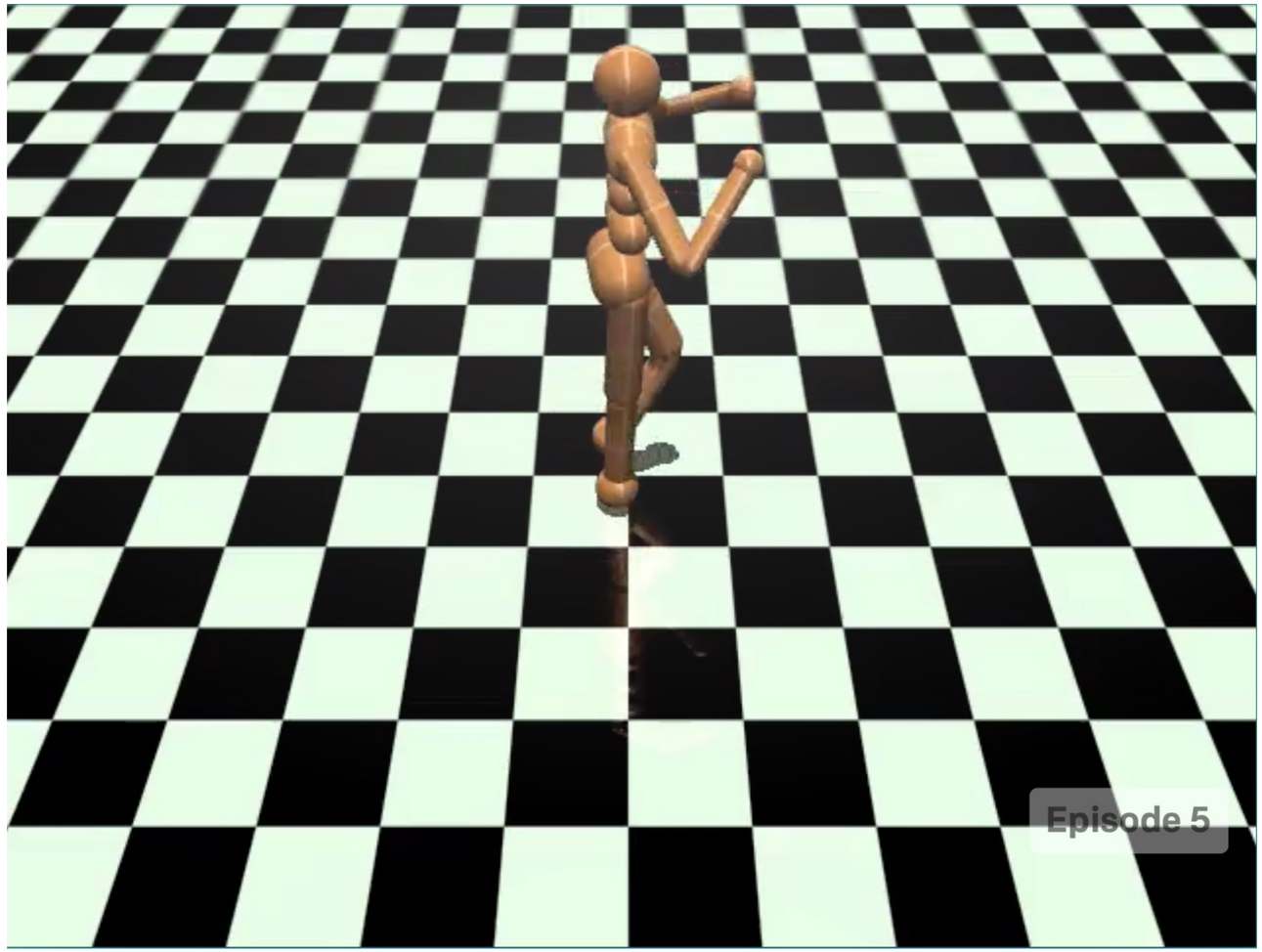
Subject
11543
555
47888

When you run the code later, you might get different results!

Computational experiments II

- This paper formally describes **evaluation metrics** used and explains the motivation for choosing these metrics.
- This paper states the number of **algorithm runs** used to compute each reported result.
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include **measures of variation**, confidence, or other distributional information
- This paper lists all final **(hyper-)parameters** used for each model/algorithm in the paper's experiments.
- This paper states the **number and range** of values tried per (hyper-)parameter during development of the paper, along with

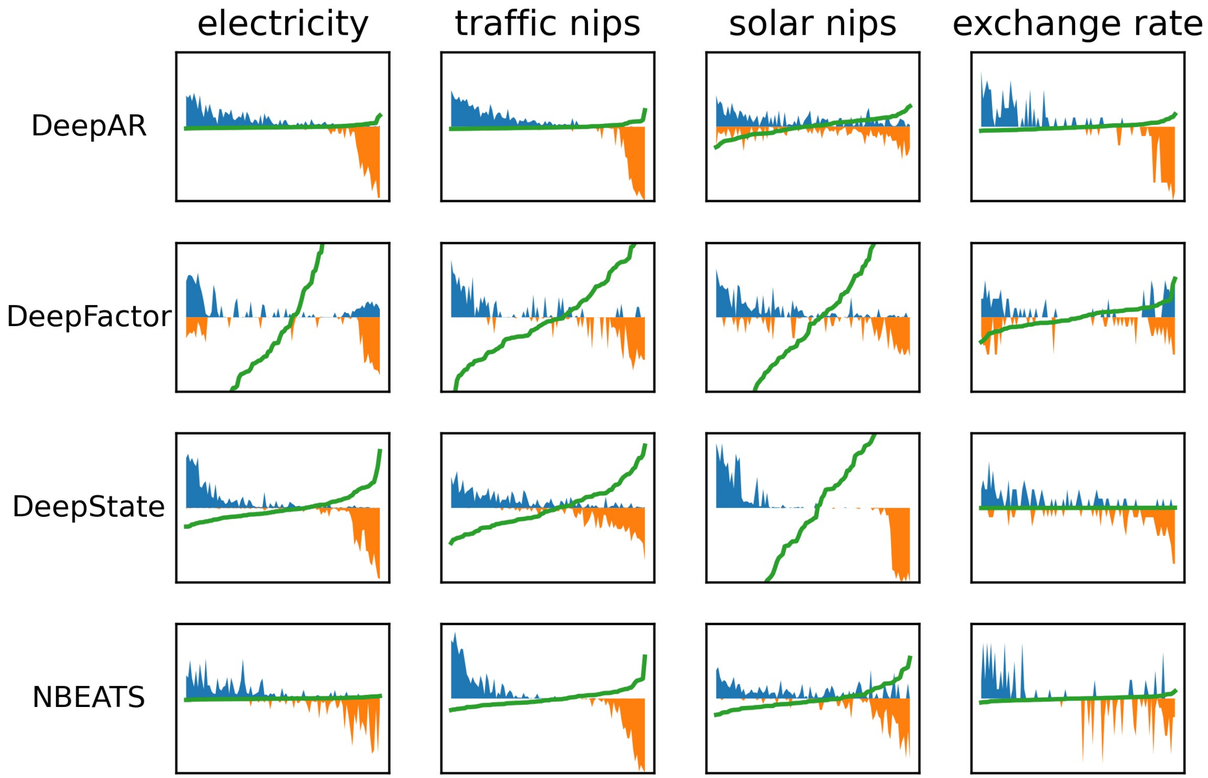
Deep Learning that Matters



- **Hyperparameter search** will have a huge effect on results. Ranges rarely documented properly.
- Simple changes in **network architecture** can have make large changes to result.
- **Different implementations** of same baseline algorithm can yield very different results.

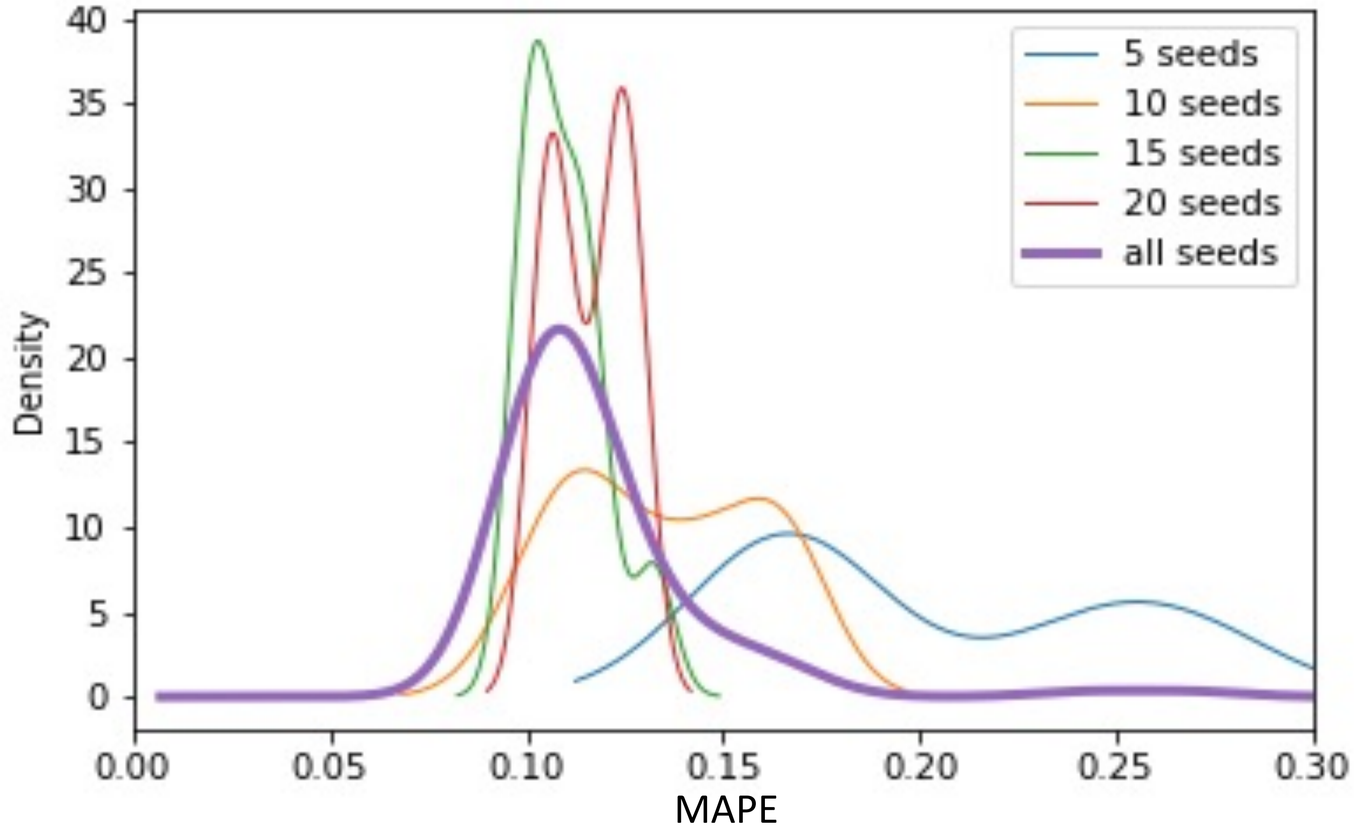
(Henderson et al, AAAI 2018)

Algorithm Runs and Variation I



Ran the same experiment 100 times. Only difference was which seeds we used to initialize the pseudorandom number generator

Algorithm Runs and Variation II



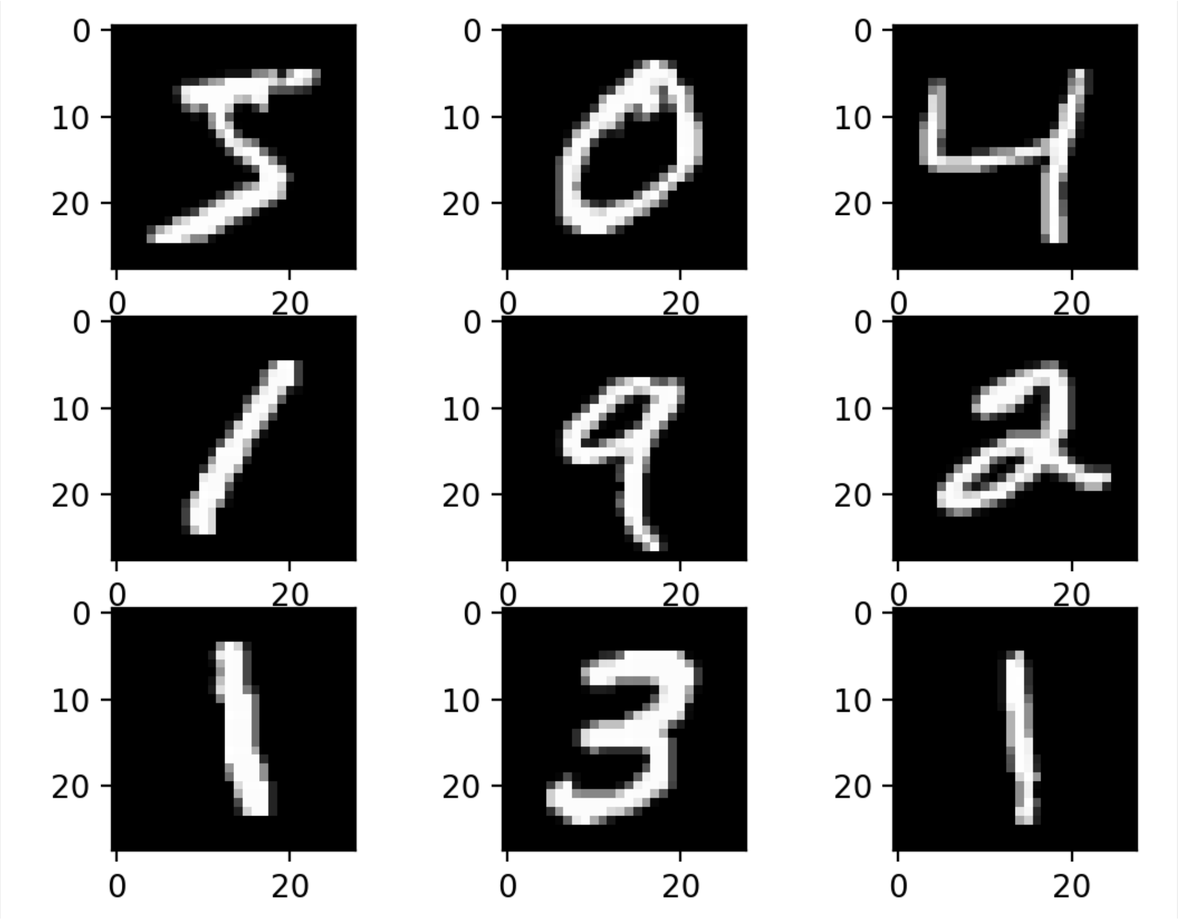
KDE used to smooth out the variance of a selection of seeds.

See how different the average MAPE scores for those seeds will be.

Assuming a similar distribution for our baseline, we can manipulate results by selecting the best set of 5 seeds for our algorithm and the 5 worst seeds for our baseline.

Experiment: MNIST Classification I

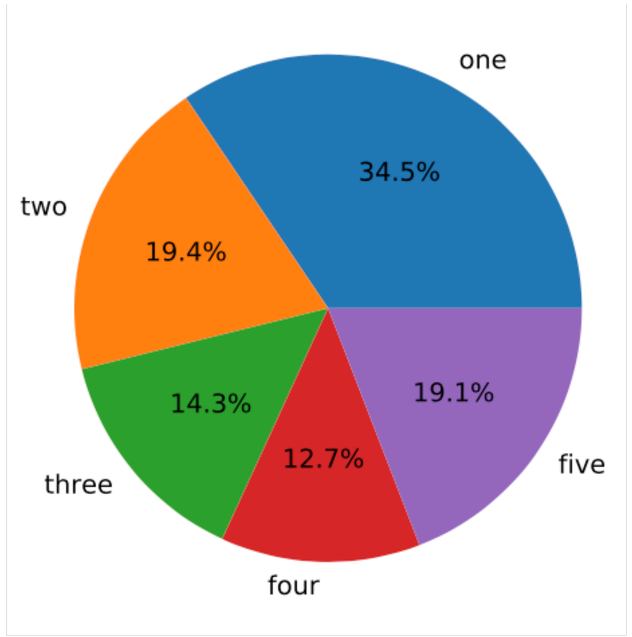
- Same experiment conducted 20 times on four different machine learning platforms.
- Code = *same*
- *Data* = *same*
- HW = ***!same***
- Ancillary SW = ***!same***



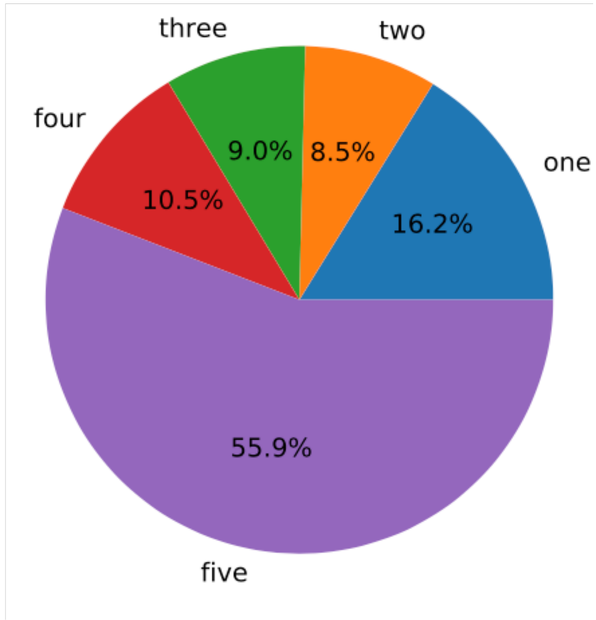
(Gundersen, Shamsaliei and Isdahl, forthcoming)

Experiment: MNIST Classification II

When models are wrong, how many are wrong?



CPU, random seed ***not*** fixed

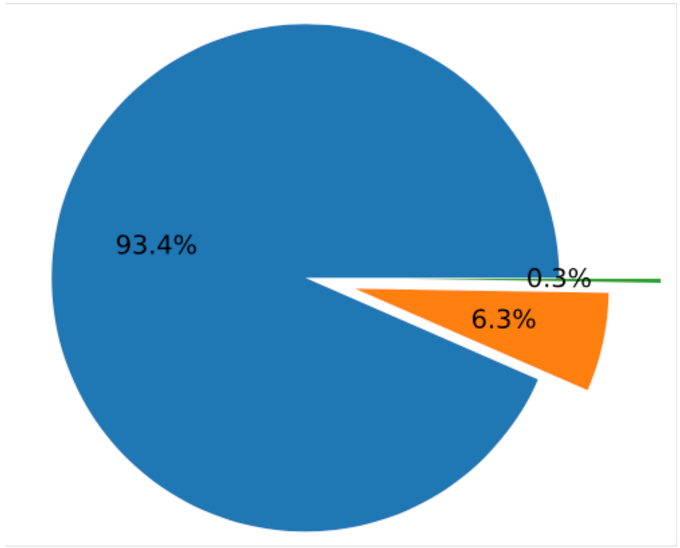


CPU, random seed fixed

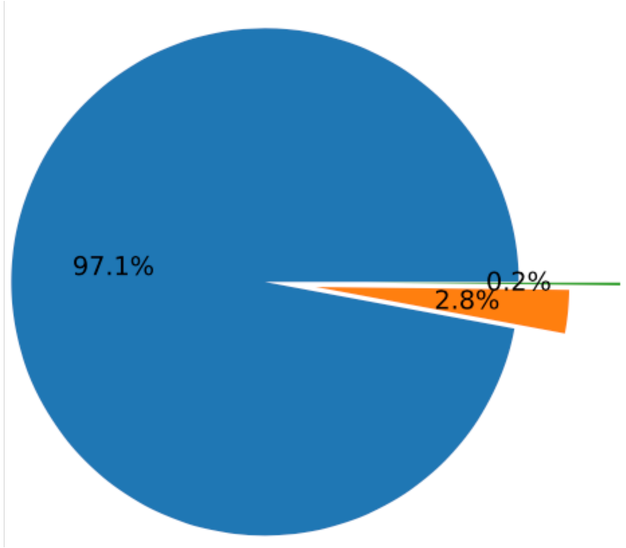
(Gundersen, Shamsaliei and Isdahl, forthcoming)

Experiment: MNIST Classification III

When models are wrong, how many different classes do they see?



CPU, random seed ***not*** fixed



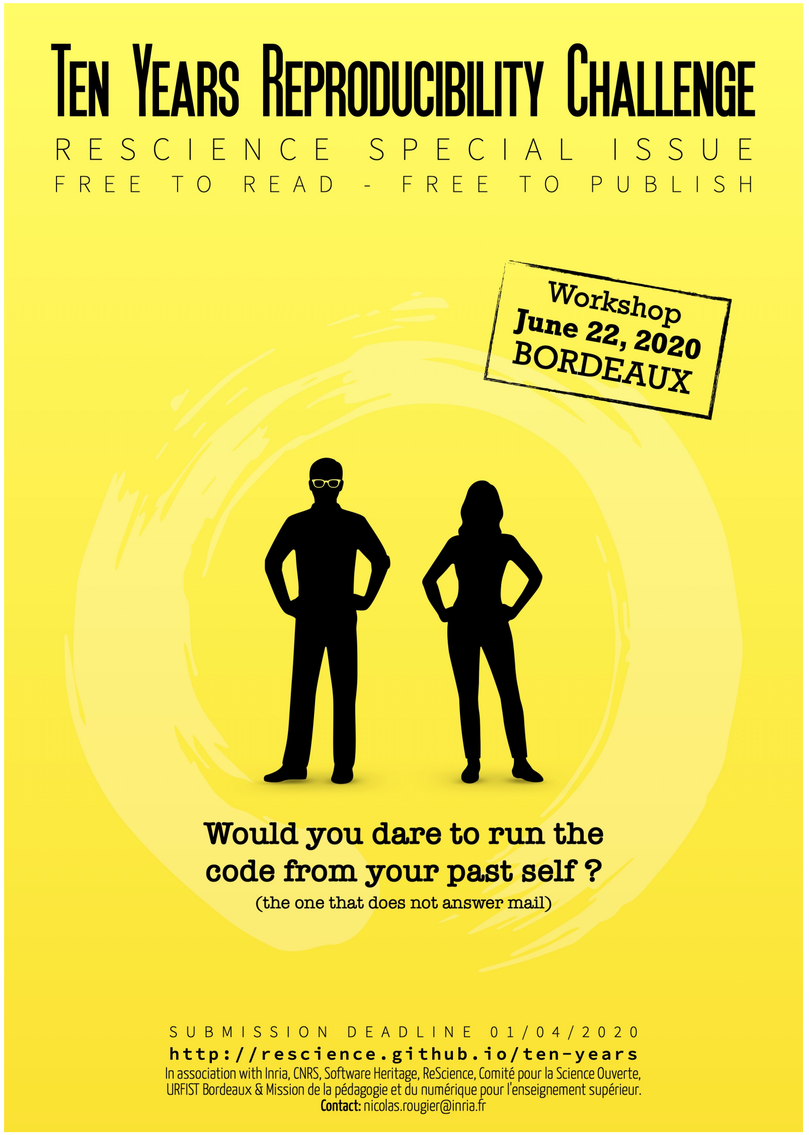
CPU, random seed fixed

(Gundersen, Shamsaliei and Isdahl, forthcoming)

THE VALUE AND CHALLENGES OF TRANSPARENT RESEARCH

PART IV

The Ten Years Reproducibility Challenge



"Programming languages evolve, as do the computing environments in which they run, and code that works flawlessly one day can fail the next."

- Nicolas Rougier, Nature, 2020

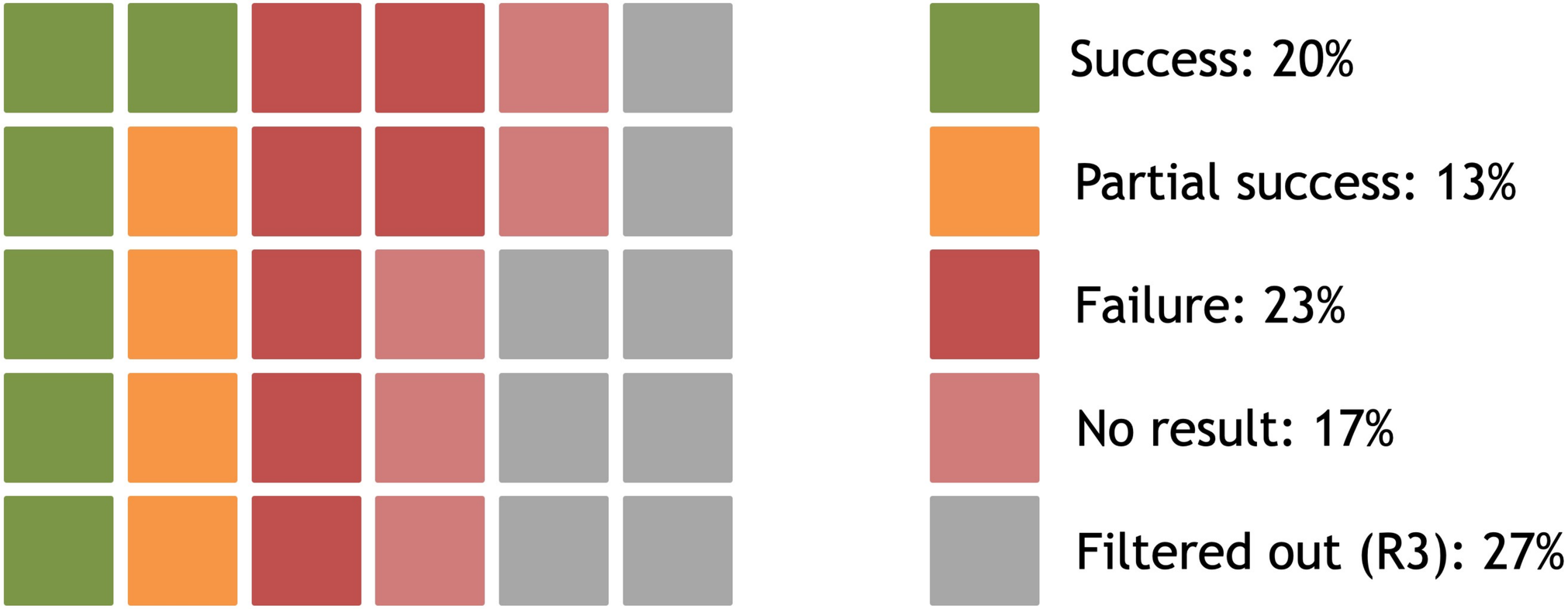
PoV of Original Researchers

(Gundersen, Gil and Aha, 2018)

PoV of Independent Researchers

(Gundersen, Gil and Aha, 2018)

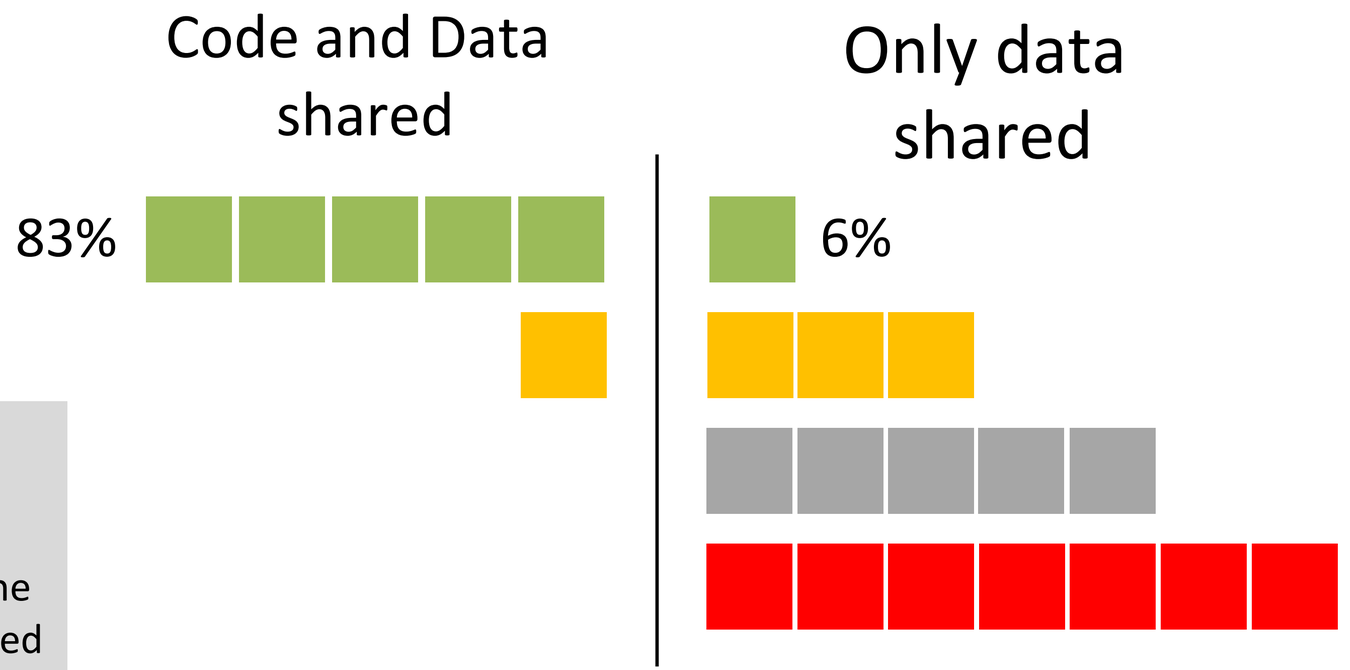
Reproducibility Experiment



(Gundersen et al, forthcoming)

The value of sharing both code and data

We tried to reproduce 30 of the top-cited papers from 2012, 2014 and 2016. These are the results: *Sharing both code and data is really effectfull.*



Success (green), Partial success (orange), Failure (red) and no result (grey) when reproducing experiments with and without code. Each box represents an aggregate of the experiments reported in one paper (most cited AI papers from Scopus).

(Gundersen et al, forthcoming)

CONCLUSION: WHAT IF YOU CANNOT DO EVERYTHING?

PART V

Important to Remember

State of the Art: Reproducibilty in Artificial Intelligence

Odd Erik Gundersen and Sigbjørn Kjensmo
Department of Computer Science
Norwegian University of Science and Technology

Comment from Hacker News

Hacker News new | threads | past | comments | ask | show | jobs | submit

▲ State of the Art: Reproducibility in Artificial Intelligence [pdf] (aaai.org)

43 points by capablemonkey on Oct 6, 2018 | hide | past | favorite | 6 comments

▲ sgt101 on Oct 6, 2018 [–]

I think that the result is overcooked. Their hypothesis 1 is somewhat falsifiable in that I don't think that there is a widespread reproducibility crisis. I have been unable to reproduce results a couple of times in my career, but I think that each time that was due to naughtiness (deliberate) on the authors part or incompetence by me. Almost always you can reproduce and when I have run into trouble I've found that the authors almost always help out (most people are just delighted that you are interested!) On the other hand this paper is very useful in that I think it will be used to establish better criteria for papers in the future. I often reject papers because they make no claim and have no results, contribution or conclusions (this makes reviewing them quick so I really like papers like this !) I think that it would be harsh to outright reject a paper because the hardware set up is poorly documented, but it would be reasonable to ask for that change before publication (for example). I agree with the authors that their criteria are useful.

One issue though, open sourcing software is a good aspiration, but it's not always possible due to IP and licensing - also export controls in some cases (not always US -> other places too). If the community insists on opensource pre-publication some important stuff is not going to get published.

producibility scores decrease with increased documentation requirements. Improvement over time is found. **Conclusion:** Both hypotheses are supported.

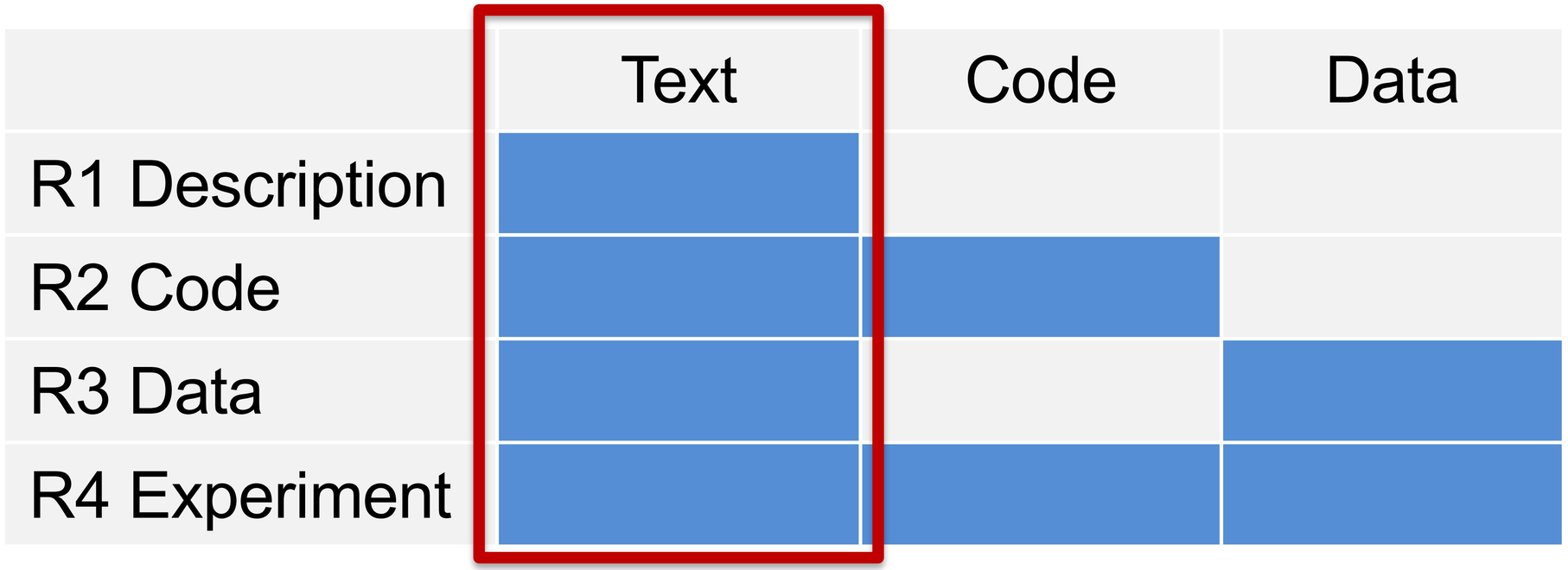
…andve *et al.* 2013; …d focus on repro- …option of data and … *et al.* 2013). Still, …ucibility see little … time required to …lutions (Gent and …gues that automa- …or machine learn- … a computer. De- …at is reproducible …ficial intelligence …kkens *et al.* 2013;

…producibility; *"if …ment and get the same result as the original researchers, then they refute the hypothesis" (Oates 2006, p. 285)*. Hence, the inability to reproduce results affects the trustworthiness of science. To ensure high trustworthiness of AI and machine learning re-

Many people believes that the reason that they are not able to reproduce results is their own incompetency.

This leads to false claims not being refuted!

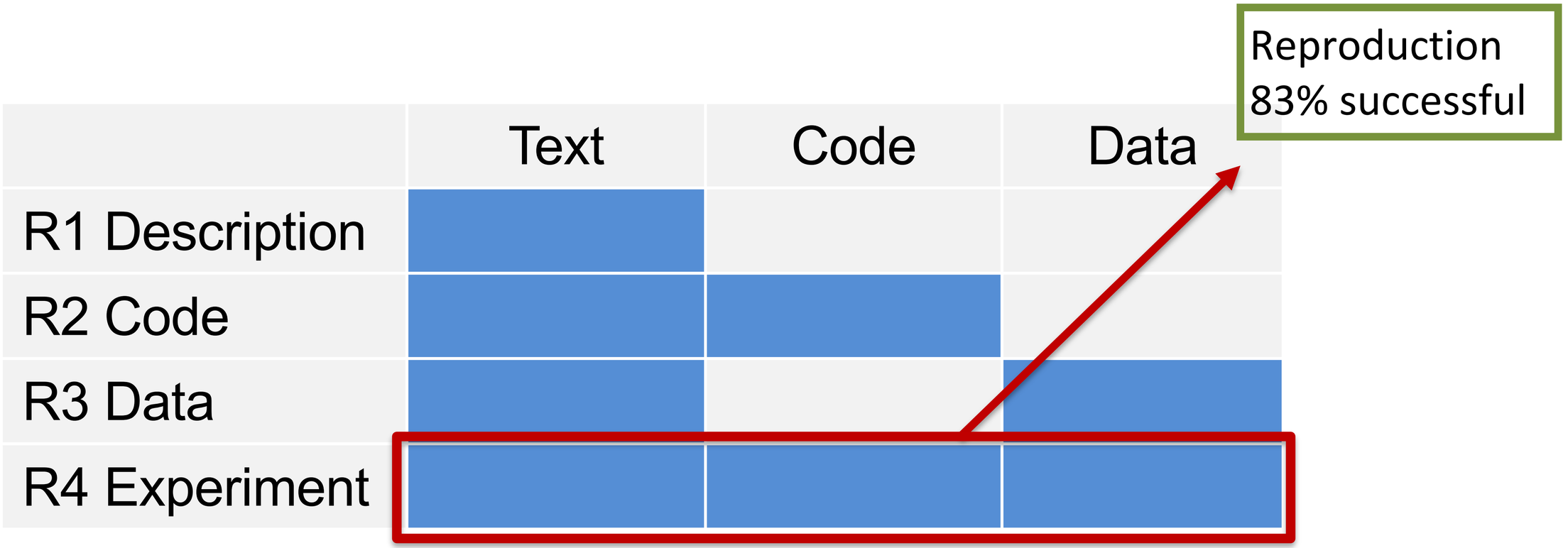
Conclusion I: If one has to choose



Newton did not share code and data

Writing a good paper that describes the experiment well and is fully transparent is **most important**!

Conclusion II: Sharing is Caring



- If you do not have time to document and tidy up the code and data, it is **still better to share** the code and data than not to.
- Sharing is more important than good documentation.

Standing on the Feet of Giants

Odd Erik Gundersen, dr. philos.

Chief AI Officer, TrønderEnergi AS

Adjunct Associate Professor, NTNU

odderik@ntnu.no

TrønderEnergi®

Norwegian Open AI Lab

References

- Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., & Zitnick, C. L. (2019). Elf opengo: An analysis and open reimplementation of alphazero. ICML. *arXiv preprint arXiv:1902.04522*.

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.
- **A Method for Assessment and a Survey of the Reproducibility Support of ML Platforms**, R. Isdahl and O. E. Gundersen, eScience 2019.
- **The Reproducibility Crisis Is Real,** O. E. Gundersen, AI Magazine, 41(3), 103-106, 2020.
- **The Case Against Registered Reports**, O. E. Gundersen, AI Magazine, Spring 2021.
- **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

References

- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.
- Tian et al (2019), ELF OpenGo: An Analysis and Open Reimplementation of AlphaZero, ICML 2019, URL: https://arxiv.org/abs/1902.04522
- P. Henderson et al. (2018). Deep Reinforceement Learning that Matters, AAAI 2018.
- M. Baker (2016), Is There a Reproducibility Crisis?, Nature, 2016.
- Song-You Hong, Myung-Seo Koo, Jihyeon Jang, Jung- Eun Esther Kim, Hoon Park, Min-Su Joh, Ji-Hoon Kang, and Tae-Jin Oh (2013), An evaluation of the software system dependency of a global atmospheric model. *Monthly Weather Review*, 141(11):4165–4172, 2013.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv preprint arXiv:2003.12206*.
- Raff, E. (2020). Research Reproducibility as a Survival Analysis. *arXiv preprint arXiv:2012.09932*.
- Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011* (pp. 1521-1528). IEEE.
- D. Goldberg (1991). What Every Computer Scientist Should Know About Floating-Point Arithmetic, Computing Surveys, 1991.

NTNU Odd Erik Gundersen odderik@ntnu.no

Resources

- **Ten Year Reproducibility Challenge**, Rescience C,URL: https://rescience.github.io/ten-years/
- **Challenge to scientists: does your ten-year-old code still run?,** J. M. Perkel, Nature (2020), URL: https://www.nature.com/articles/d41586-020-02462-7
- **General Reproducibility Guidelines for AI**, Gundersen, Gil, Mausam, 2020, URL: https://folk.idi.ntnu.no/odderik/reproducibility_guidelines.pdf
- **The Machine Learning Reproducibility Checklist**, J. Pineau, 2020, URL: https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist. Pdf
- **AlphaGo – the Movie**, Youtube, URL: https://youtu.be/WXuK6gekU1Y