Sources of irreproducibility

Odd Erik Gundersen, dr. philos.

Chief AI Officer, Aneo AS

Associate Professor, NTNU

Norwegian Open AI Lab

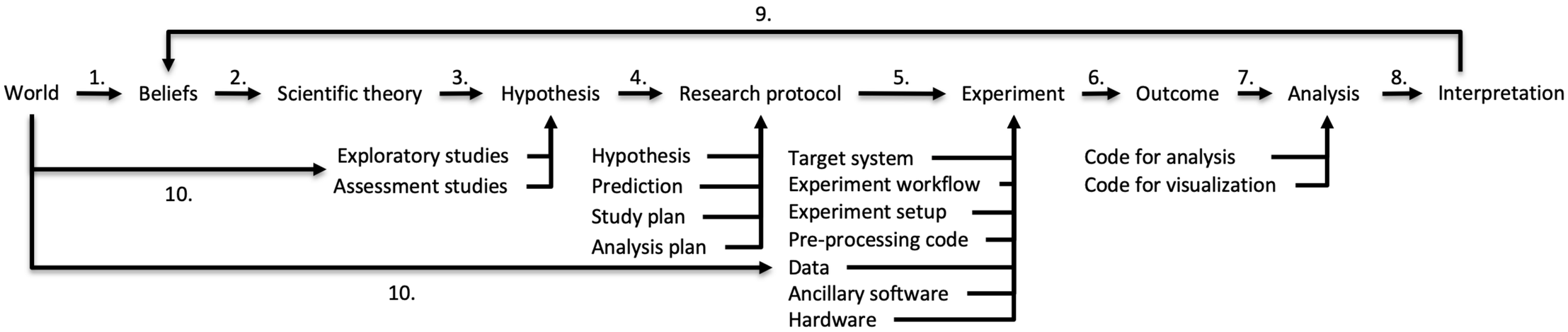
Why do we need a shared definition?

- To properly capture the concept.
- Create a shared understanding.
- Use it for teaching science.
- Use it when we practice science – to guide our research.
- Use it when evaluating research done by others.
- Help point out limitations of conclusions or what went wrong.

What characterizes a good definition?

- It shoud be a mental model that can easily be looked-up
- It should be simple to understand.
- It must have enough depth to be valuable.
- It shoud enable us to operationalize our understanding and help us design and evaluate experiments.
- It needs to be tightly connected to the scientific method, as reproducibility is *“a cornerstone of science”*.

The Scientific Method in Empirical ML



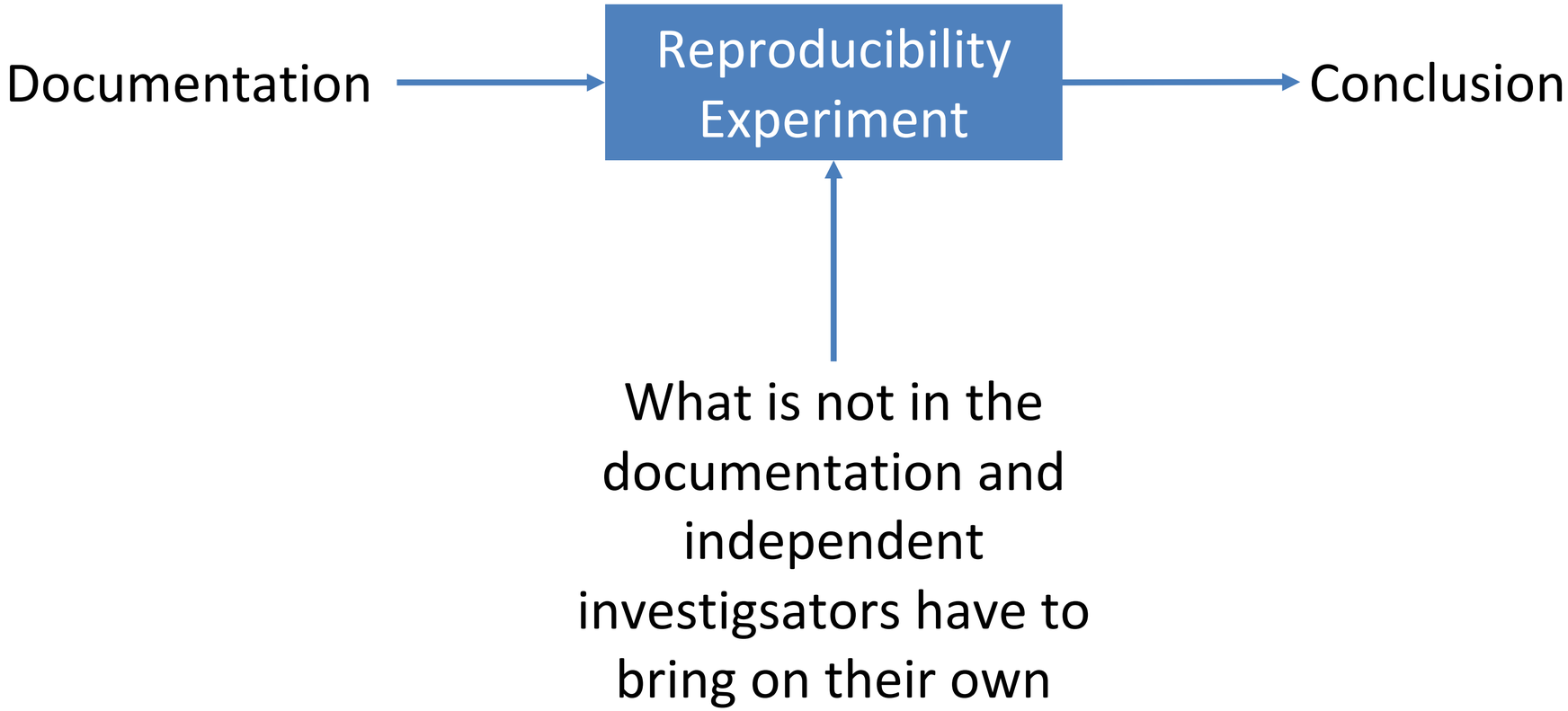
(Gundersen 2021)

Reproducibility

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

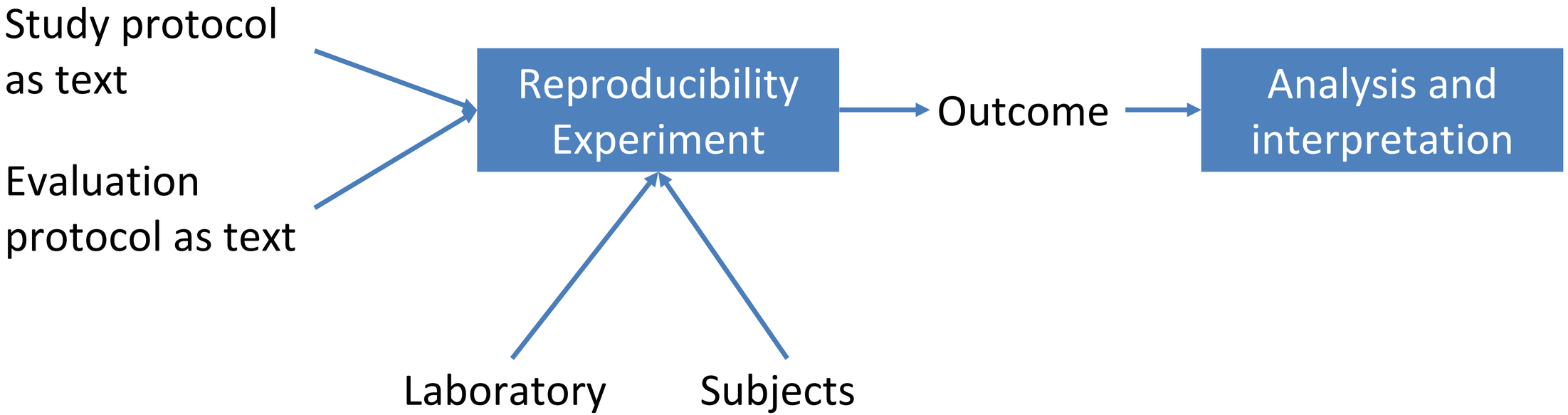
Reproducibility Experiment

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



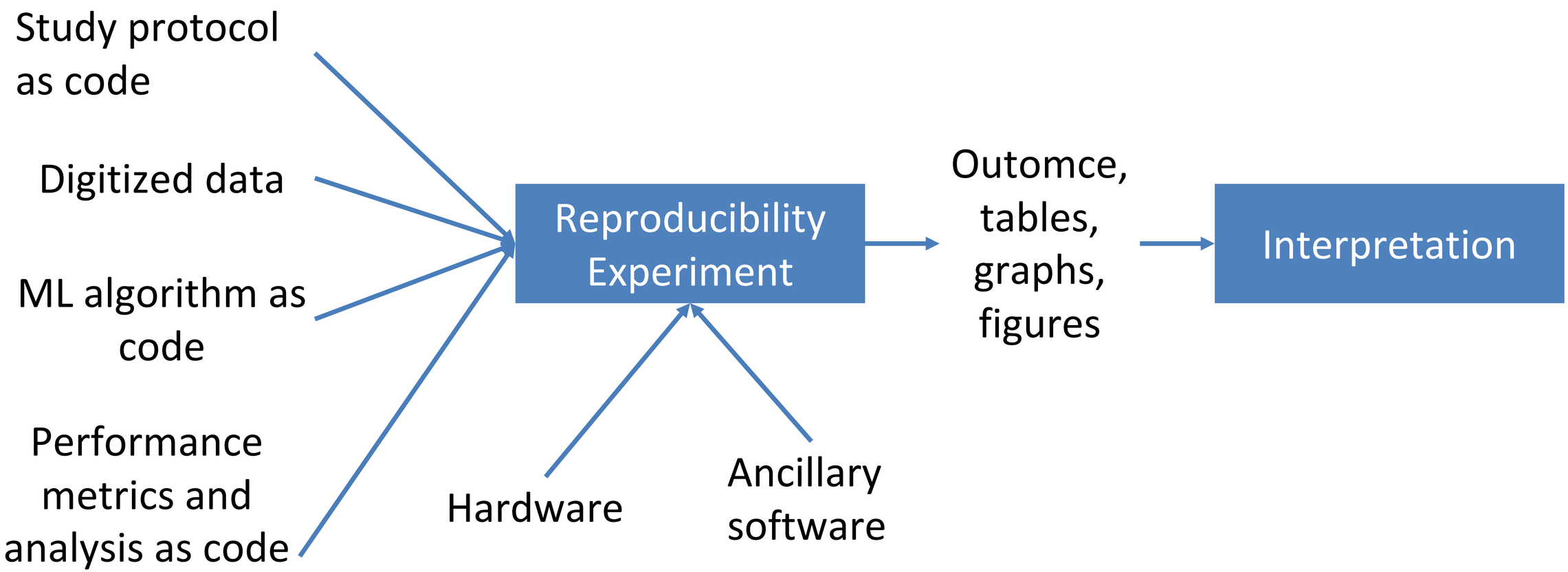
Reproducibility - Psychology

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



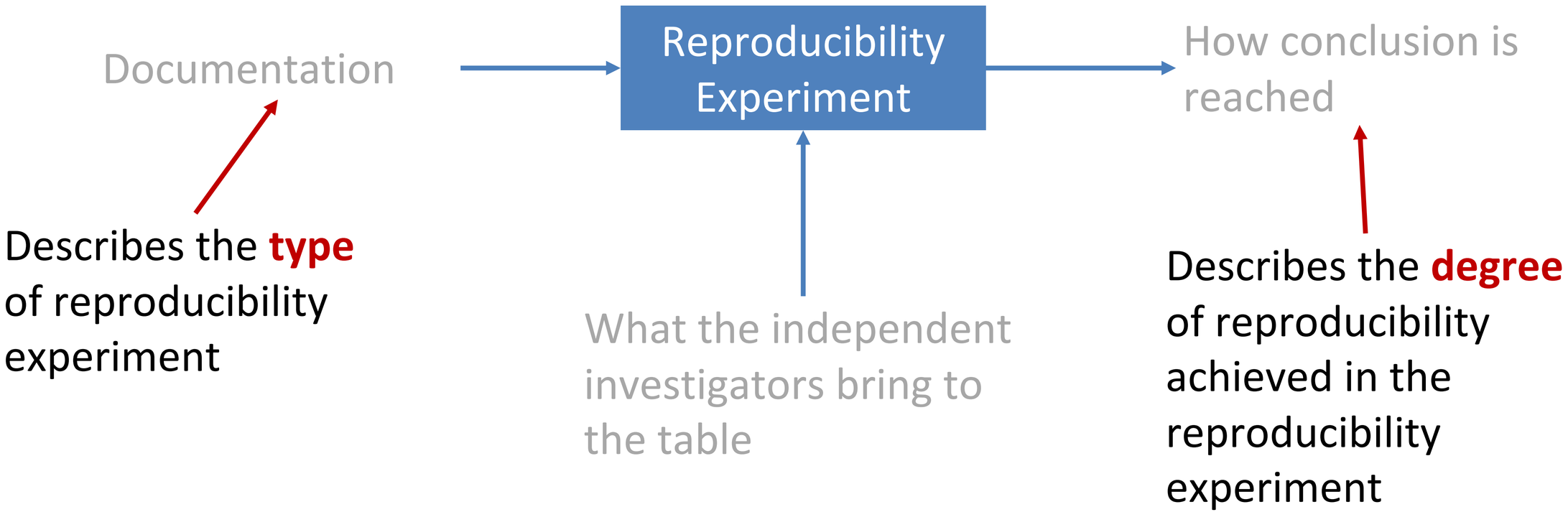
Reproducibility Experiment – ML

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



Reproducibility - Dimensions

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



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)

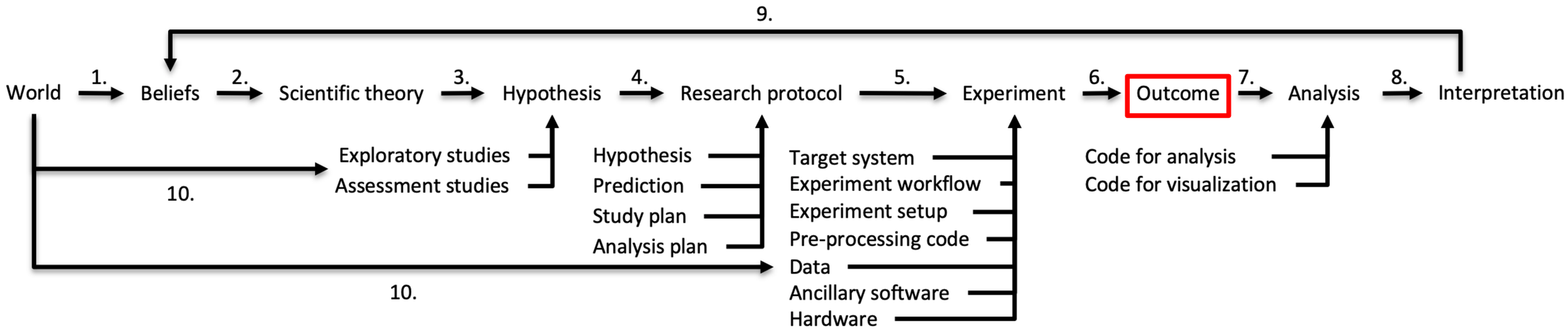
Types of Reproducibility Experiments

	Text	Code	Data
R1 Description	■		
R2 Code	■	■	
R3 Data	■		■
R4 Experiment	■	■	■

(Gundersen 2021)

Outcome Reproducible - Degree

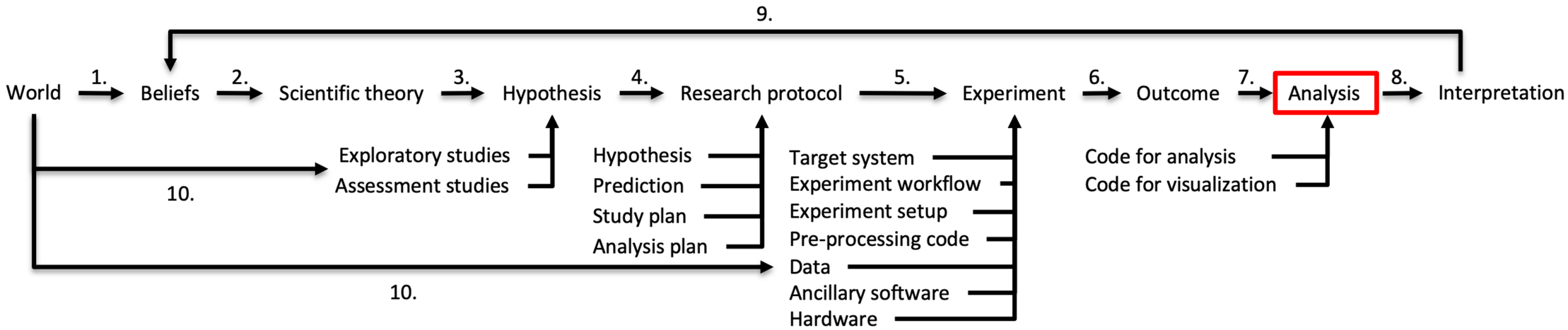
Definition. *Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.*



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

Analysis Reproducible – Degree

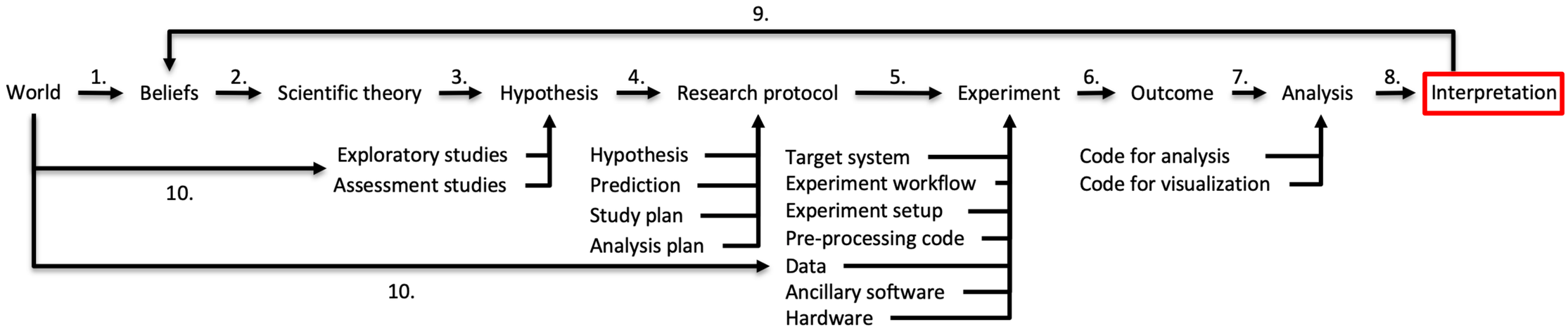
Definition. *Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.*



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

Interpretation Reproducible - Degree

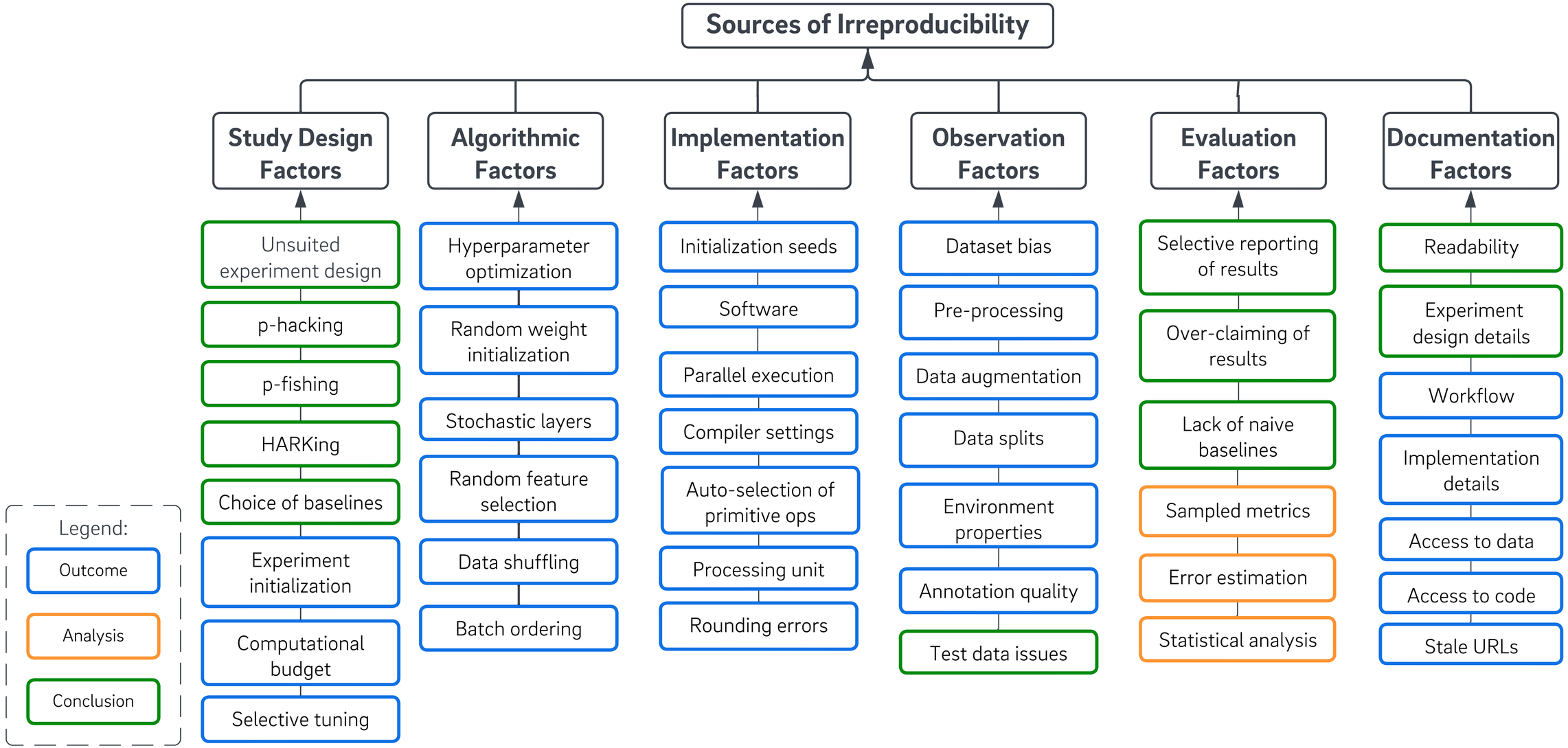
Definition. *Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.*



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

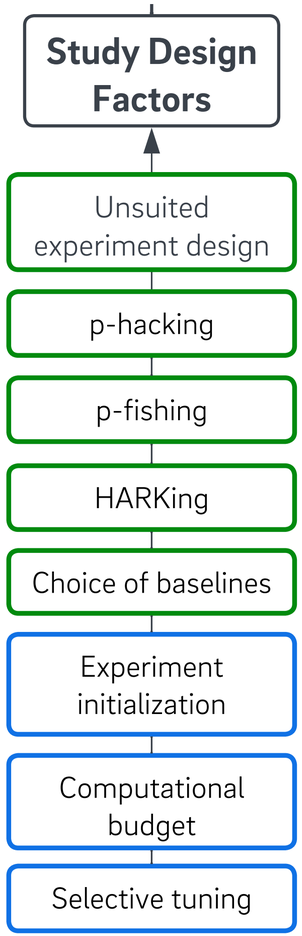
Reproducibility Experiment Classification

	R1 Description (text)	R2 Code (text+code)	R3 Data (text+data)	R4 Experiment (text+code+data)
Outcome Reproducible	OR1	OR2	OR3	OR4
Analysis Reproducible	AR1	AR2	AR3	AR4
Interpretation Reproducible	IR1	IR2	IR3	IR4



(Gundersen, Coakley, Kirkpatrick, Gil, forthcoming)

Study Design Factors



Study design factors capture the decisions that goes into making the high-level plan for how to conduct and analyze an experiment to answer the stated hypothesis and research questions.

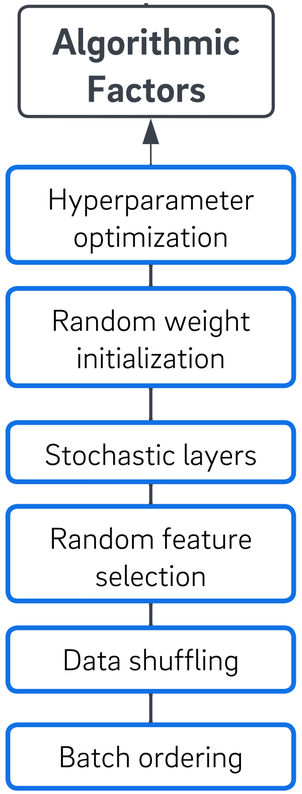
(Gundersen, Coakley, Kirkpatrick, Gil, forthcoming)

HARking

Researchers to explore hypotheses that are different to those that they originally set out to test. This practice is called 'HARKing, which stands for Hypothesizing After the Results are Known, also known as 'outcome switching’

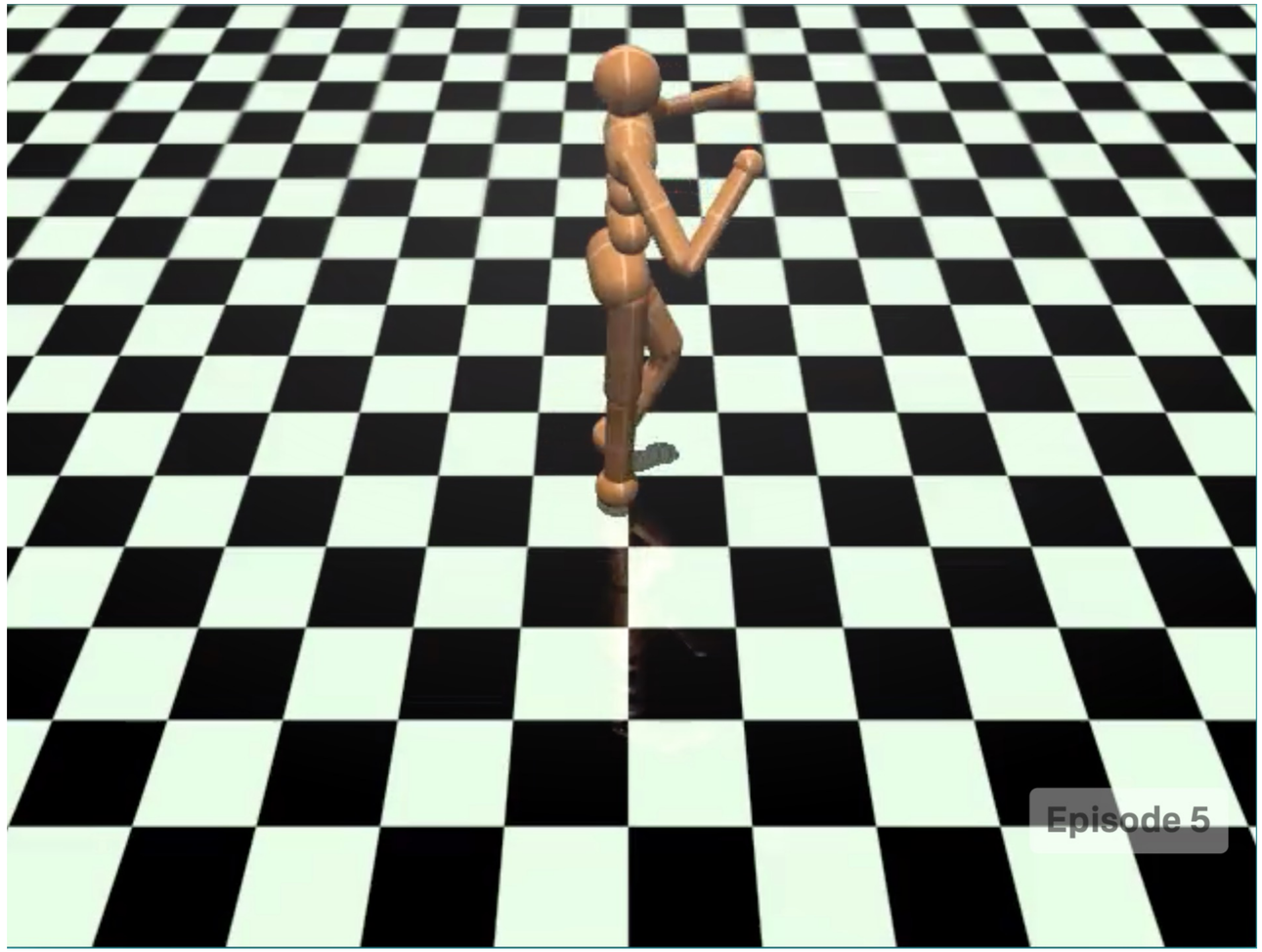
Publication bias, for example, encourages HARKing.

Algorithmic Factors



Algorithmic factors are design decisions to introduce stochasticity in the learning algorithms and training processes, leads to a different outcome for every experiment run.

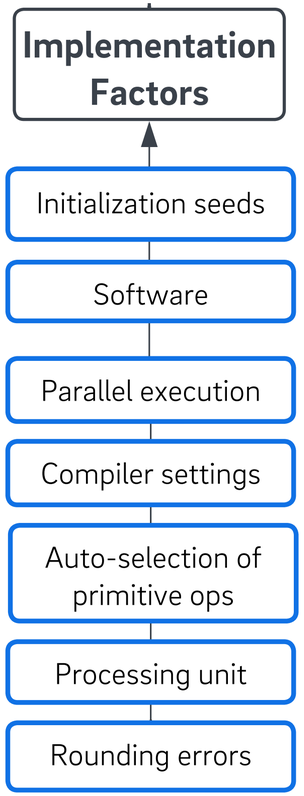
Hyperparameter search



- will have a huge effect on results. Ranges rarely documented properly.

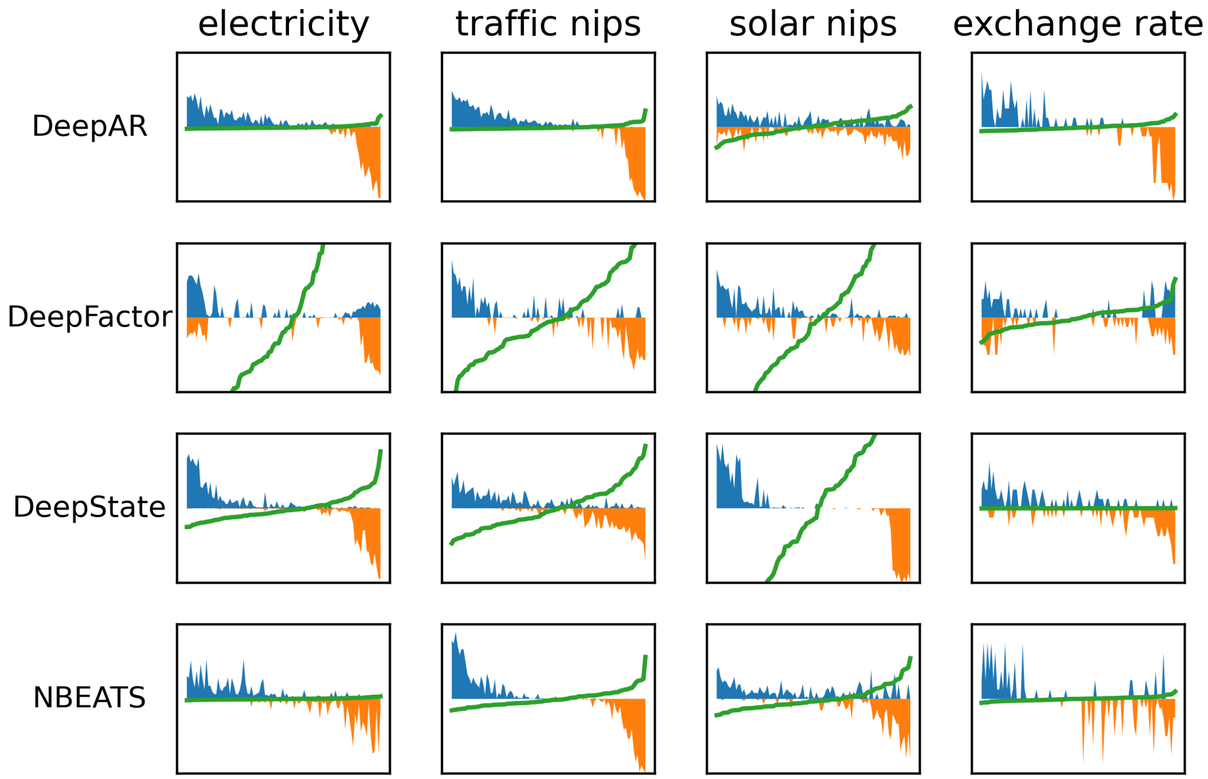
(Henderson et al, AAAI 2018)

Implementation Factors



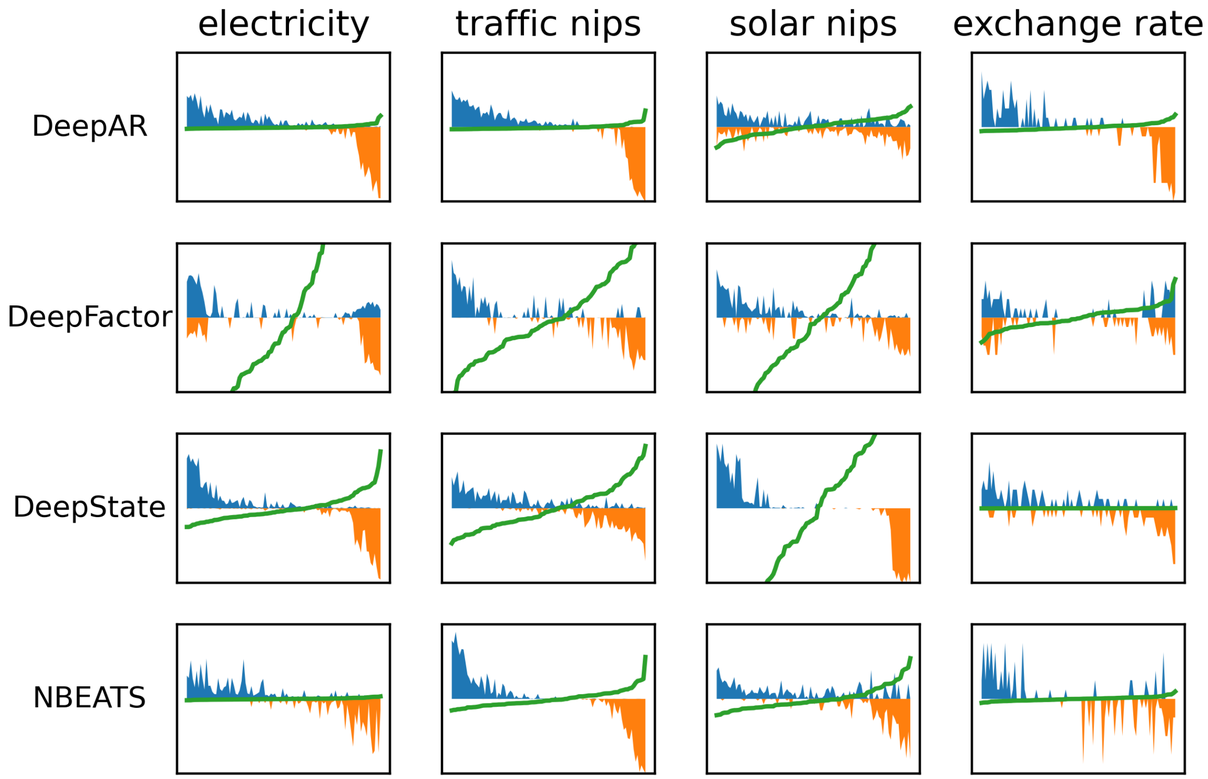
Implementation factors are design choices related to the software and hardware that are used to execute the experiment. These factors mirror the variations in physical sciences experiments that are introduced by conducting the same experiment in different laboratories.

Initialization Seeds I



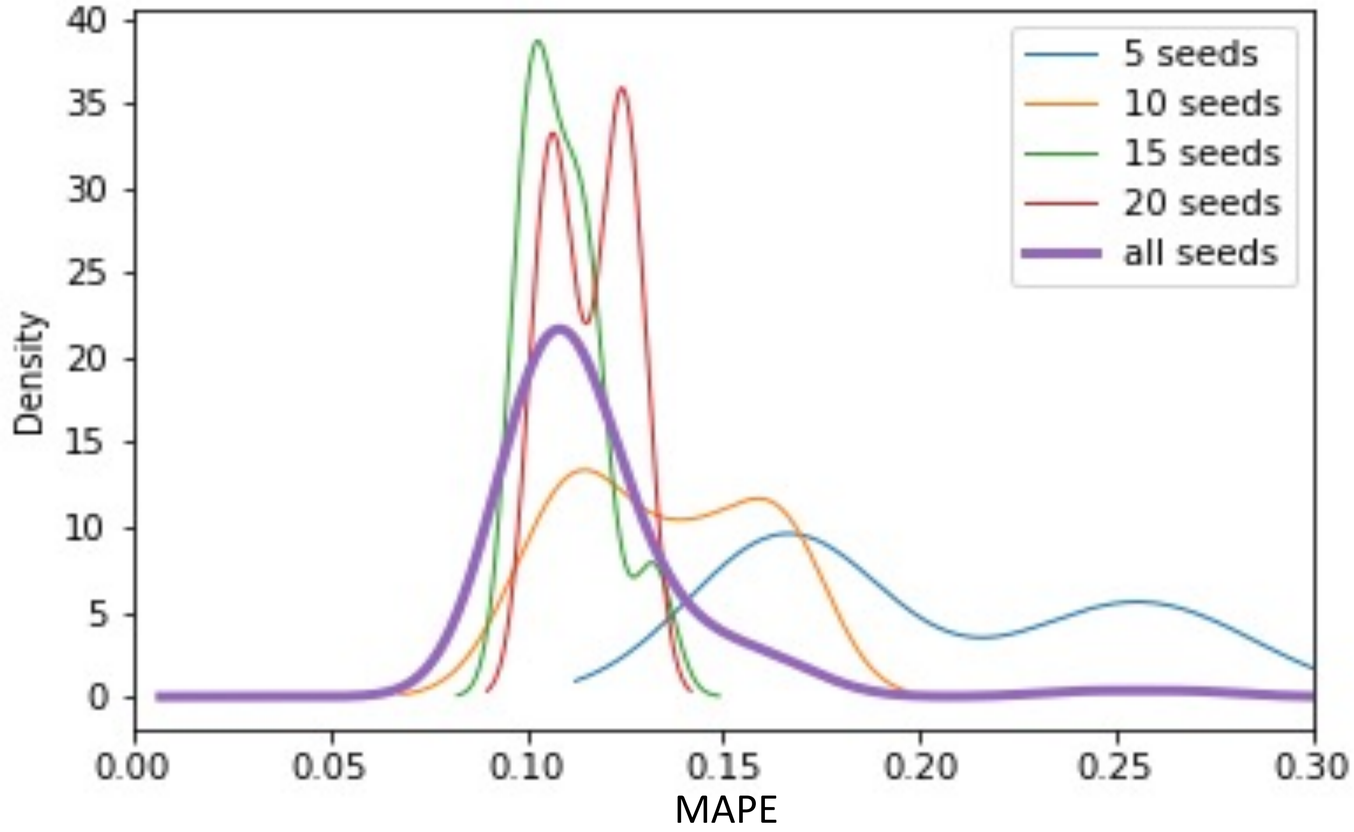
Ran the same experiment 100 times. Only difference was which seeds we used to initialize the pseudorandom number generator

Initialization Seeds I



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

Initialization Seeds II

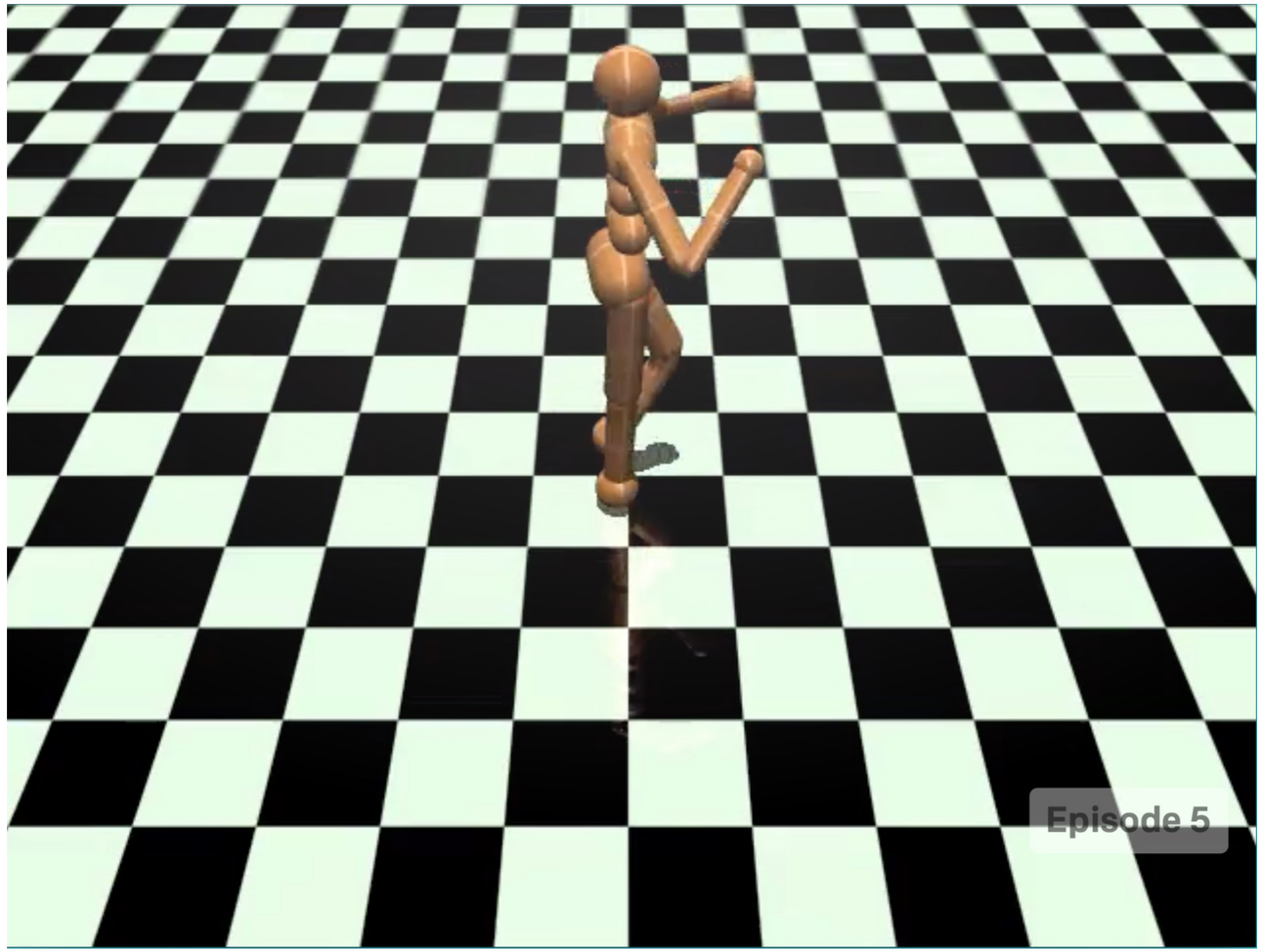


KDE used to smooth out the variance of a selection of seeds.

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.

Deep Learning that Matters



- 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)

CPU and 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

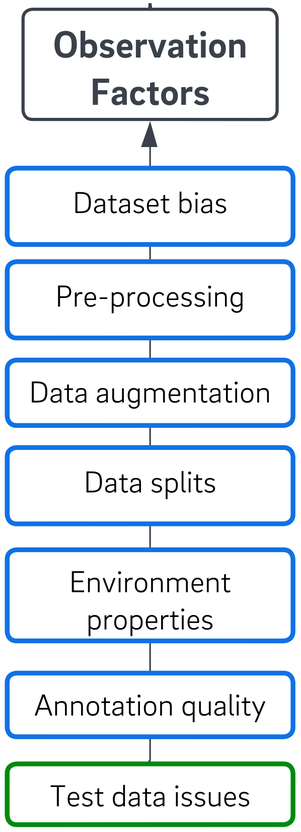
Subject	dv	condition	mediator
11543	70	treatment	5
555	3	control	6
47888	110	placebo	3

After June 24 2016

Subject
11543
555
47888

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

Observation Factors



Observational factors are related to how data is generated, processed and augmented, but also to the properties of environments used for benchmarking, such as agent simulation environments.

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?)*

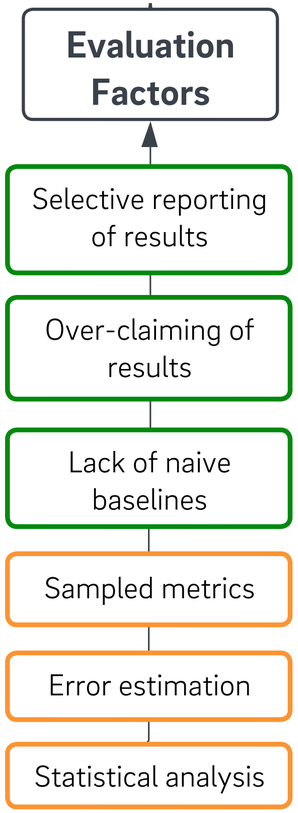
Negative bias *Does the data set contain negative examples as well?*

(Torralba and Efros 2011

Other issues

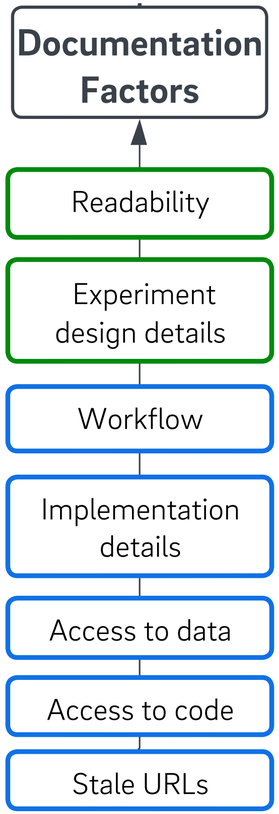
- Data version:
  - Are there different versions of the same dataset?
  - Some software libraries provide standard datasets as well i.e. seaborn and GluonTS.
  - Sometimes these differ from the original ones. Cite the correct version.
  - Sometimes the reported data is not the same as the published data (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.

Evaluation Factors

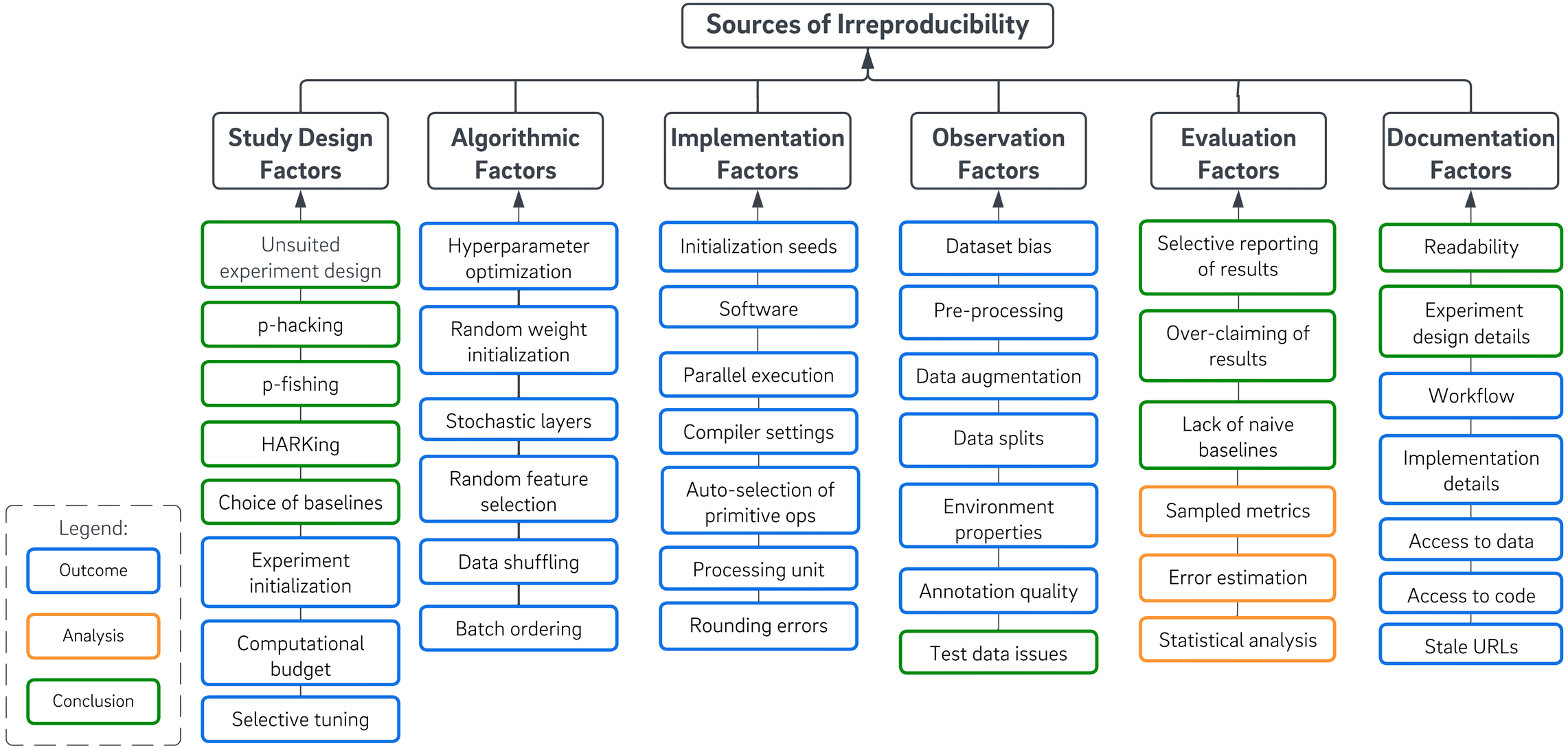


Evaluation factors relate to how the investigators reach the conclusions from doing an experiment.

Documentation Factors

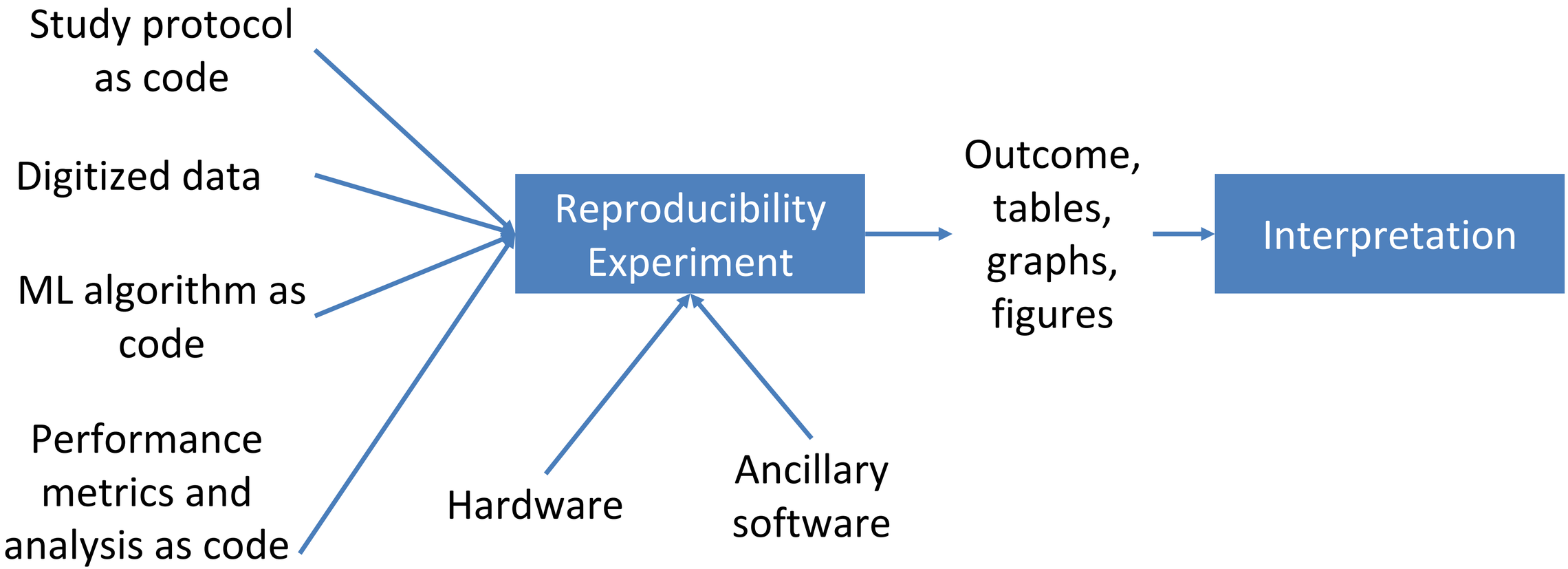


Documentation factors are related to how well an experiment is documented, which means ideally documenting all the choices mentioned above, which can be impractical.



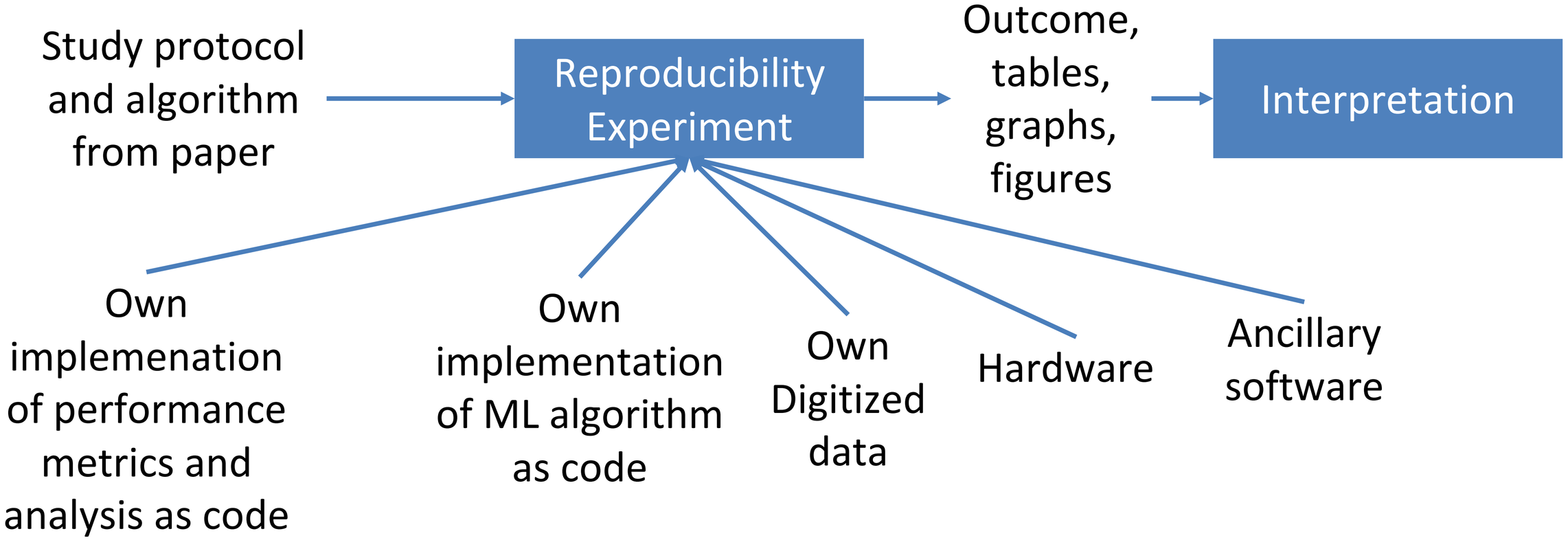
(Gundersen, Coakley, Kirkpatrick, Gil, forthcoming)

Operational use – Example I



OR4: *Outcome reproducible after executing same experiment* – Same outcome after executing code on same data on different hardware and software.

Operational use – Example II



IR1: *Interpretation reproducibible after own implementation* - Different outcome and analysis after implementing experiment fully from paper utilizing different data and different metrics and analysis methods.

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.
- **The Reproducibility Crisis Is Real** Gundersen, O. E., *AI Magazine*, *41*(3), 103-106, 2020.
- **The Case Against Registered Reports**, O. E. Gundersen, AI Magazine, Spring 2021.
- **Do Machine Learning Platforms Provide Out-of-the-box Reproducibility?** O.E. Gundersen, S. Shamaliei and R. Isdahl. Future Generation Computer Systems, Volume147. Elsevier, 2022.
- **Sources of irreproducibility.** O. E. Gundersen, K. Coakley, C. Kirkpatrick, Gil, forthcoming.
- **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

Standing on the Feet of Giants

Odd Erik Gundersen, dr. philos.

Chief AI Officer, Aneo AS

Associate Professor, NTNU

odderik@ntnu.no

Norwegian Open AI Lab