

Reprodukowalność

“Non-reproducible single occurrences are of no significance to science.”

Karl Popper (1959) “The logic of scientific discovery”, p. 66

“We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results.”

Ronald Fisher (1935) “The Design of Experiments”, p. 14

Reproducibility is a major principle of the scientific method. It means that a result obtained by an experiment or observational study should be achieved again with a high degree of agreement when the study is replicated with the same methodology by different researchers.

Only after one or several such successful replications should a result be recognized as scientific knowledge.



Reproducibility

- ensures that the results are correct
- ensures transparency
- gives us confidence in understanding exactly what was done



reproducibility/replicability/independent reproducibility

A: The terms are used with no distinction between them.

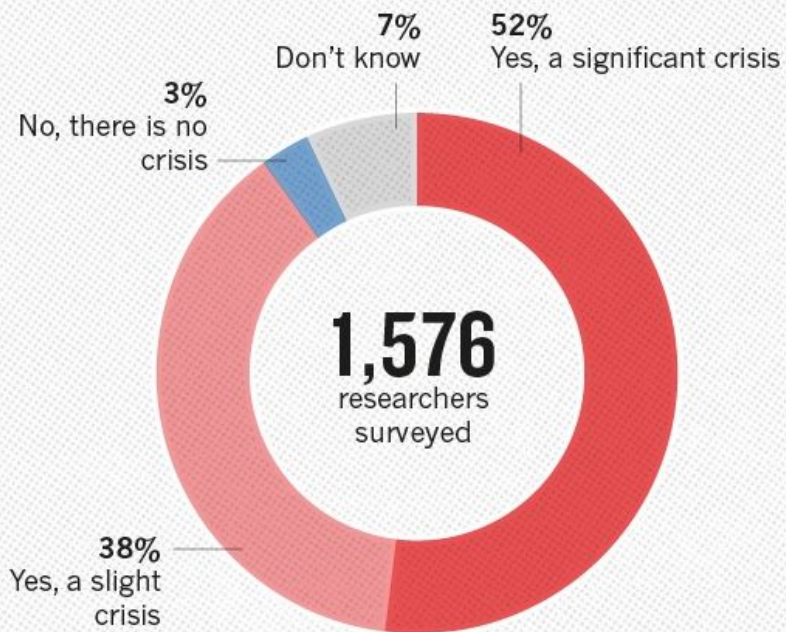
B1: “Reproducibility” refers to instances in which the original researcher's data and computer codes are used to regenerate the results, while “replicability” refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.

B2: “Reproducibility” refers to independent researchers arriving at the same results using their own data and methods, while “replicability” refers to a different team arriving at the same results using the original author's artifacts.

A	B1	B2
Political Science	Signal Processing	Microbiology, Immunology (FASEB) Computer Science (ACM)
Economics	Scientific Computing	
	Econometry	
	Epidemiology	
	Clinical Studies	
	Internal Medicine	
	Physiology (neurophysiology)	
	Computational Biology	
	Biomedical Research	
	Statistics	

		Data	
		Same	Different
Code & Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

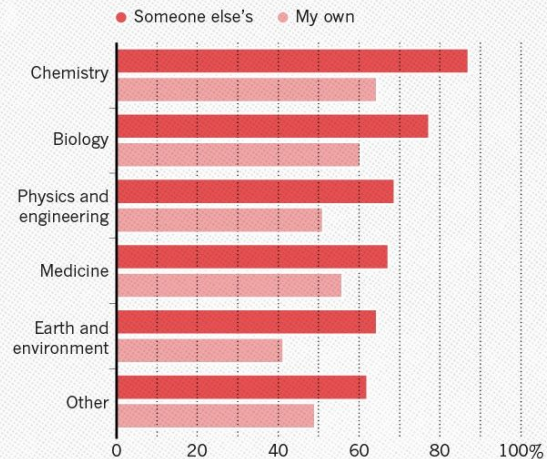
IS THERE A REPRODUCIBILITY CRISIS?



©nature

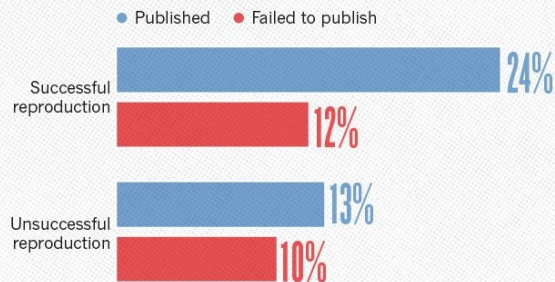
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

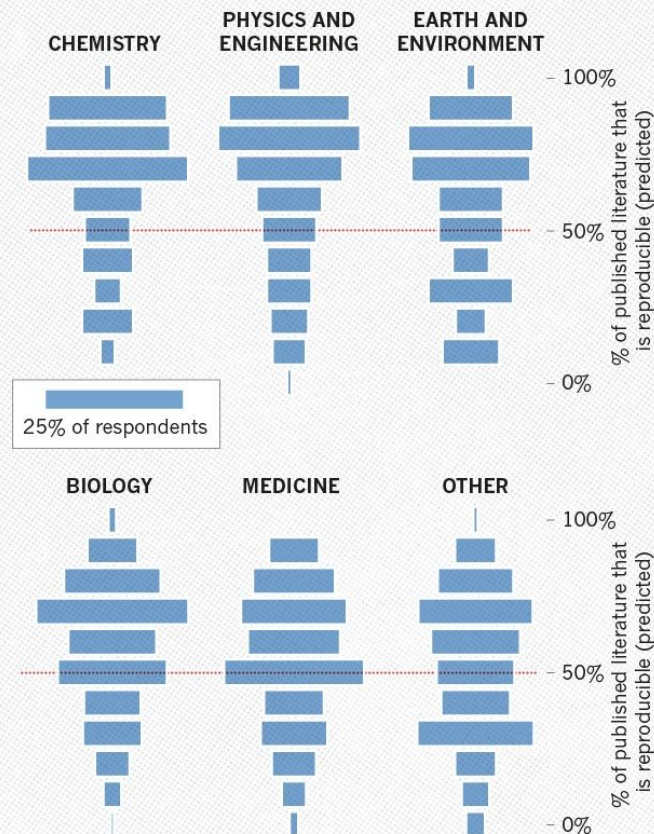


Number of respondents from each discipline:
Biology 703, Chemistry 106, Earth and environmental 95,
Medicine 203, Physics and engineering 236, Other 233

enature

HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.



Number of respondents from each discipline:
Biology 703, Chemistry 106, Earth and environmental 95,
Medicine 203, Physics and engineering 236, Other 233

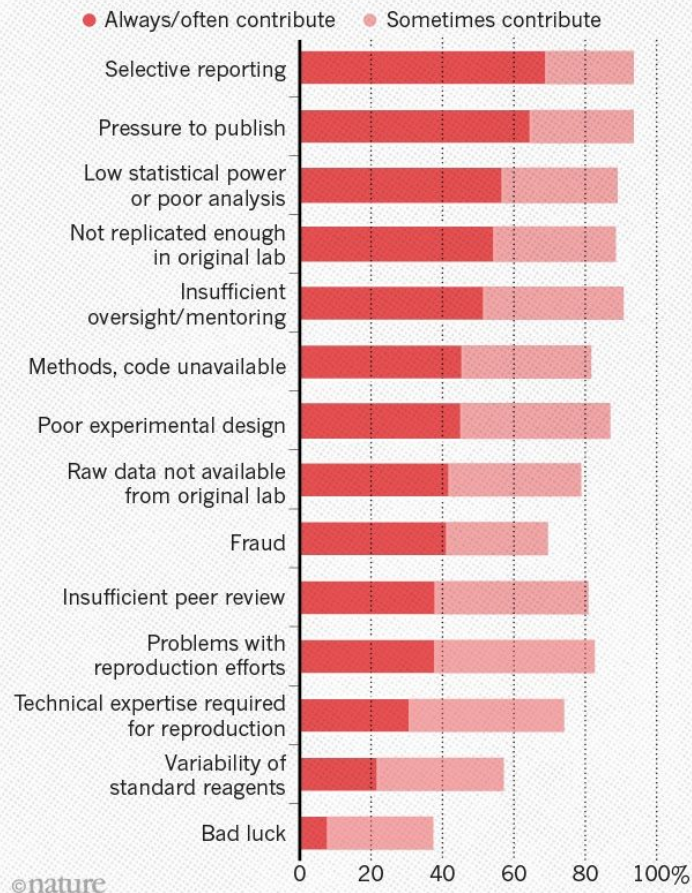
enature

Why is this crisis so *big*?

- positive publication bias
- publish or perish

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



Why is this crisis so *big*?

Methods Reproducibility is defined as the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results as in an original work

Causes of non-determinism:

1. Random initialization of layer weights
2. Shuffling of datasets
3. Changes in machine learning frameworks
4. Non-deterministic GPU floating-point calculations



Results Reproducibility is defined as the ability to produce corroborating results in a new (independent) study having followed the same experimental procedures

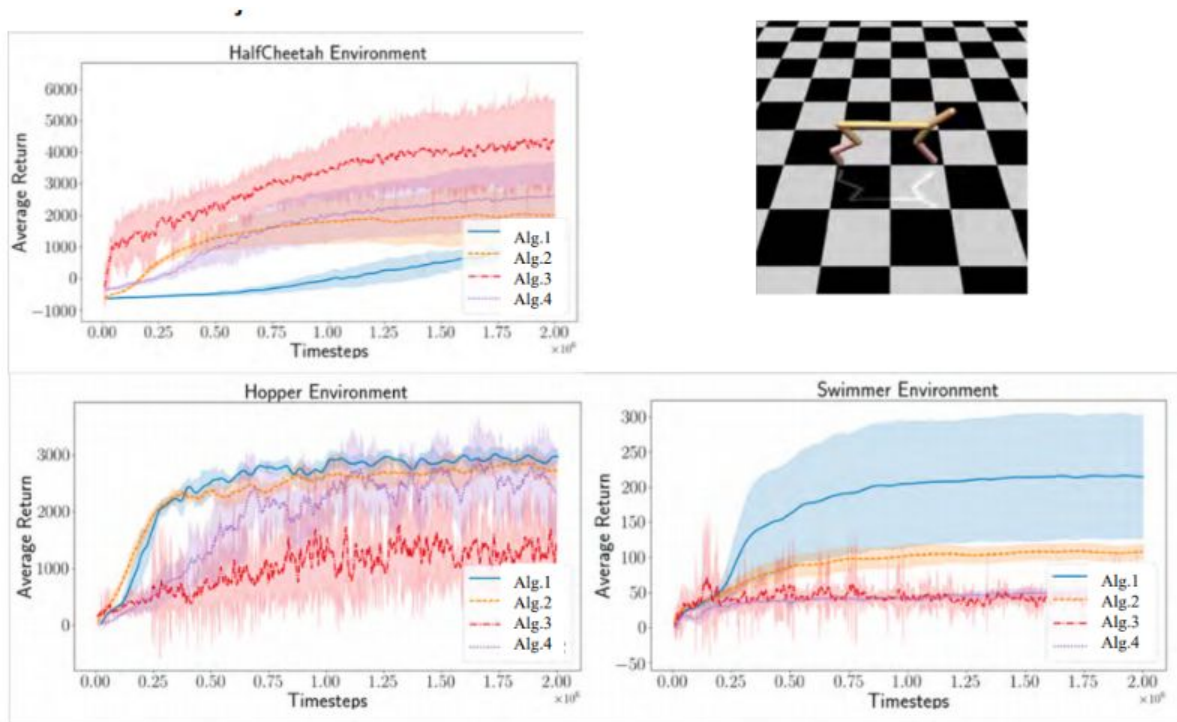


Figure 3: Comparison of performances of the four baseline algorithms on three different environments: HalfCheetah, Hopper, and Swimmer. For the purpose of reproducibility, it is not necessary to know which baseline algorithm each curve or color corresponds to. On HalfCheetah, the red algorithm performs the best, while on Hopper red seems to

Inferential Reproducibility- A study is deemed to have inferential reproducibility when an independent replication of the study or a reanalysis of it arrive at qualitatively similar conclusions to that of the original study

- bias to show that hypothesis is correct may try to reject the null hypothesis by manipulating the data and justify it in the name of outlier removal
- researcher tries out many different hypotheses and finally finds and reports a small subset of these hypotheses which are statistically significant.

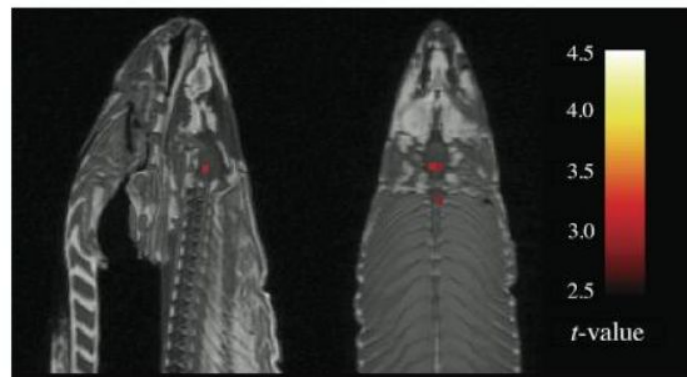


Figure 7: Significant voxels identified by Bennett et al. [13] are shown in red. Uncorrected $p < 0.001$ for each significant voxel.

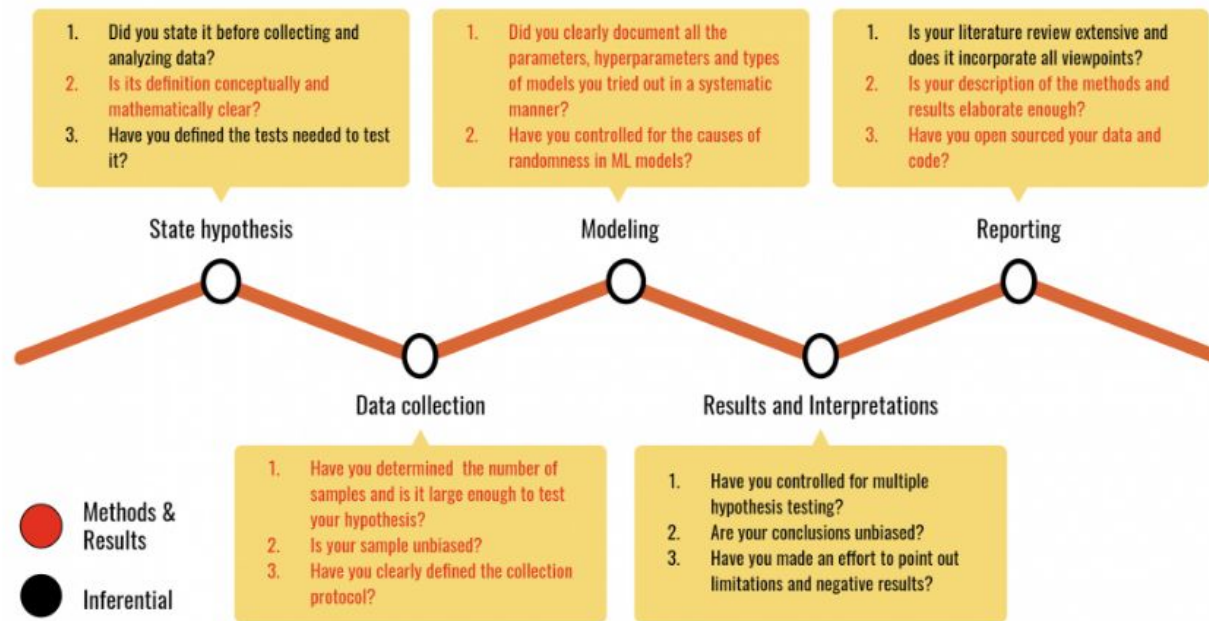


Figure 1: Summary diagram highlighting the different steps in a data analysis project and the questions one should ask at each step in order to overcome the barriers to various types of reproducibility.

Gaps in ML research:

- Lack of access to the same training data / differences in data distribution
- Misspecification or under-specification of the model or training procedure
- Lack of availability of the code necessary to run the experiments, or errors in the code
- Under-specification of the metrics used to report results



Gaps in ML research:

- Improper use of statistics to analyze results, such as claiming significance without proper statistical testing or using the wrong statistic test
- Selective reporting of results and ignoring the danger of adaptive overfitting
- Over-claiming of the results, by drawing conclusions that go beyond the evidence presented (e.g. insufficient number of experiments, mismatch between hypothesis & claim).

ML challenges (NeurIPS 2019)

- an insufficient exploration of the variables e.g. experimental conditions, hyperparameters that might affect the conclusions of a study.
- the proper documentation and reporting of the information necessary to reproduce the reported results

A recent report indicated that 63.5% of the results in 255 manuscripts were successfully replicated ([Raff, 2019](#)).

Rozwiązania



Reproducible code

- Tips for Publishing Research Code:
- Specification of dependencies
- Training code
- Evaluation code
- Pre-trained models
- README file including table of results accompanied by precise commands to run/produce those results

https://www.cs.mcgill.ca/~ksinha4/practices_for_reproducibility/

- <https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Reproducible-Code-2019.pdf>



As an experiment, NeurIPS-2019 will use the following Code Submission Policy.

1. The policy only applies to papers that **contribute and present experiments with a new algorithm (or a modification to an existing algorithm)**. That is, a paper is **not** covered by this policy if:

- a. The paper is not claiming the contribution of any novel algorithm.
- b. The paper presents a new algorithm but only analyzes it theoretically (i.e., no experimental results are presented).

2. Code submission for papers covered by this policy is **expected but not enforced**.

3. The policy **accepts a reimplementation** by the authors that isn't the code originally run to produce the results reported in the paper (what is instead requested is the equivalent of an official implementation of the paper's contribution).


4. The policy **accepts code that isn't "executable" as is** as it has dependencies going beyond the algorithm itself and that cannot be released. Such dependencies would include

- a. Dataset that cannot be released (e.g., for privacy reasons).
- b. Specialized hardware that might not be commonly accessible (e.g., specialized accelerators or robotic platforms).
- c. Non-open sourced or non-free libraries, which do not include the algorithm that is claimed as the scientific contribution of the paper (e.g., paid-for mathematical programming solvers, commercial simulators, MATLAB).

The authors will be asked to explain what dependencies are not released and why.

5. The policy expects code **only for accepted papers**, and only **by the camera-ready deadline (October 27, 2019)**.

After the camera-ready deadline, NeurIPS intends to measure the percentage of accepted papers for which code was not released, despite being covered by the policy.



The Machine Learning Reproducibility Checklist (Version 1.2, Mar.27 2019)

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.

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.

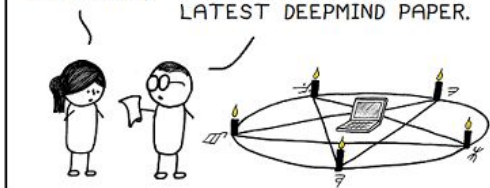
Conference	# papers submitted	% papers accepted	% papers w/code at submission	% papers w/code at camera-ready	Code submission policy
NeurIPS 2018	4856	20.8%		<50%	"Authors may submit up to 100MB of supplementary material, such as proofs, derivations, data, or source code."
ICML 2019	3424	22.6%	36%	67%	"To foster reproducibility, we highly encourage authors to submit code. Reproducibility of results and easy availability of code will be taken into account in the decision-making process."
NeurIPS 2019	6743	21.1%	40%	74.4%	"We expect (but not require) accompanying code to be submitted with accepted papers that contribute and present experiments with a new algorithm." See Appendix, Fig. 7

Table 1: Code submission frequency for recent ML conferences. Source for number of papers accepted and acceptance rates: <https://github.com/lixin4ever/Conference-Acceptance-Rate>. ICML 2019 numbers reproduced from the ICML 2019 Code-at-Submit-Time Experiment.

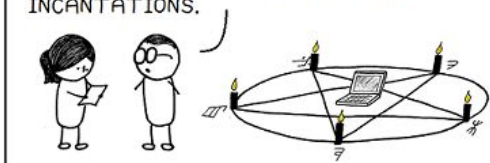
HYPERPARAMETERS REVELIO!
STOCHASTIC GRADIENT DESCENDO!



WHAT ARE YOU DOING?
I'M TRYING TO REPRODUCE
THE RESULTS FROM THE
LATEST DEEPMIND PAPER.



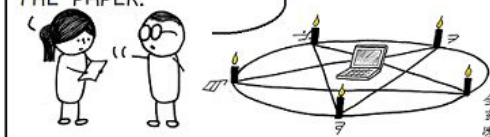
SO FAR NONE OF MY SPELLS SEEM TO
BE WORKING. I THINK I MAY NEED TO
INVOKE SOME EVEN MORE POWERFUL
INCANTATIONS.



OR, Y'KNOW,
MAYBE JUST
FOLLOW THEIR
DESCRIPTION IN
THE PAPER.

I'M IN NO MOOD
FOR YOUR JOKES.

SILENCIO!



DON'T WORRY,
YOU DON'T HAVE
TO START YOUR
CODE FROM
SCRATCH.



YOU CAN RE-USE THE
SOFTWARE THAT THE
PREVIOUS PERSON
ON THE PROJECT
WROTE SEVERAL
YEARS AGO.



ARE THERE
INSTRUCTIONS FOR
HOW TO USE IT?

I DOUBT IT.

IS THE CODE
COMMENTED?

NOT LIKELY.

WHERE ARE
THE FILES?

WHO KNOWS.



THIS IS GOING
TO BE PAINFUL,
ISN'T IT?

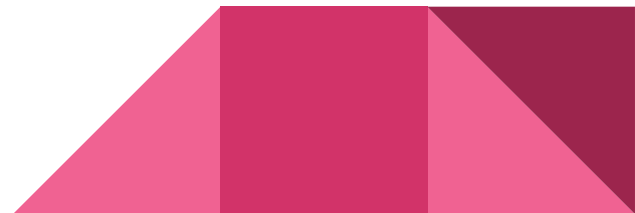
JUST A
SCRATCH.



- [Paper With Code](#)
- ICLR and NeurIPS Reproducibility Challenge

[ML Reproducibility Challenge 2020 and Spring 2021](#)

- Nature: reusability report
- Nature: kontenery ze środowiskiem - [Code Ocean](#) (dodatkowy edytor)



1. We need to have a clear description of the algorithm along with the complexity analysis (space, time, sample size). Sample size becomes important in case of an independent replication study.
2. We need to include links to downloadable source code and dataset along with the dependencies.
3. We need to provide a clear description about the data collection process, and how samples were allocated for training, testing, and validation.
4. We need to specify the range of the hyperparameters considered and the method employed to select the best hyperparameters. Finally, we need to have a specification of the hyperparameters.
5. We need to include a clear definition of statistics used to report the results, description of results including central tendency, variance, error bars as well as number of evaluation runs.
6. We also need to include the computing infrastructure used.

