How can we measure reproducibility of IR experiments?

Maria Maistro, mm@di.ku.dk

The 20th Dutch-Belgian Information Retrieval Workshop (DIR 2021 ½), February 4, 2022

Today's Agenda

About Reproducibility

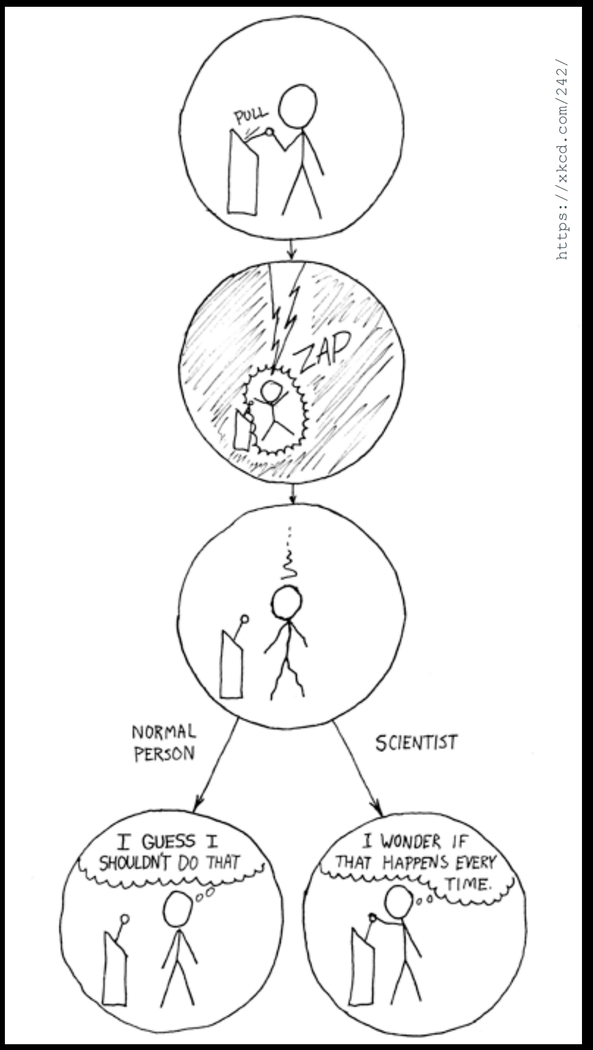
- Examples of reproducibility;
- Motivations;
- Terminology;
- Challenges;
- Measure reproducibility;
- Some Initiatives.

We all agree that reproducibility is important…

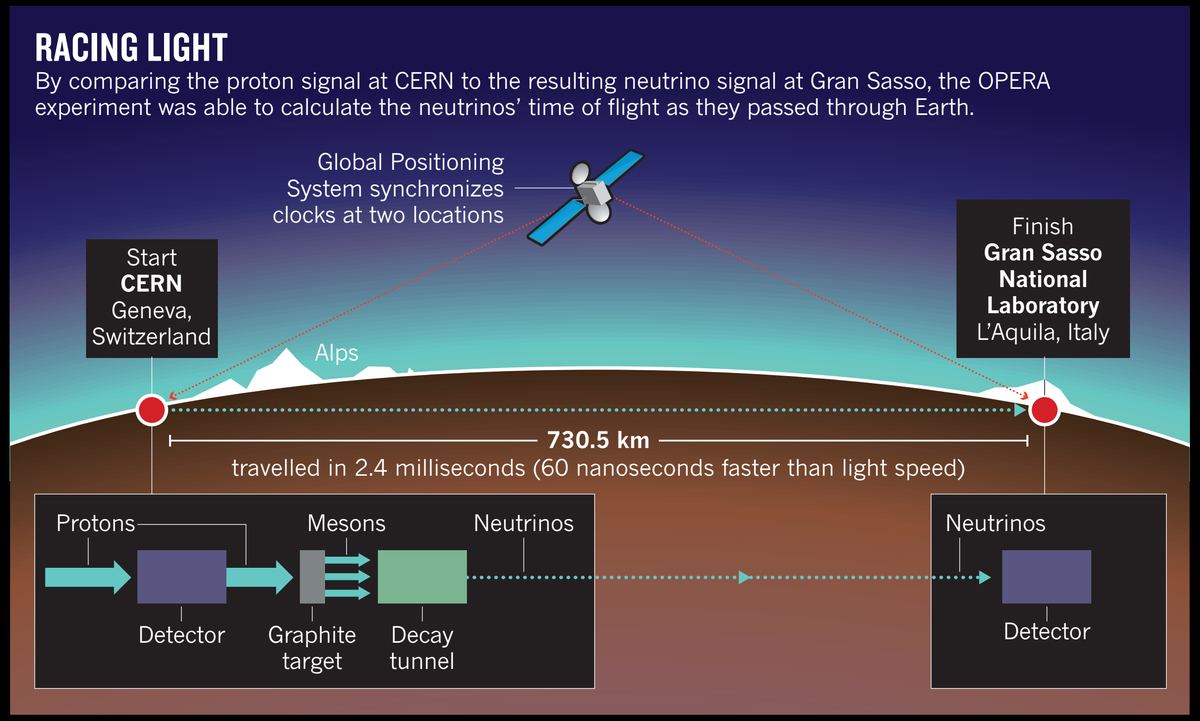
Science is not the static knowledge written in textbooks.

“An experimental result is not fully established unless it can be independently reproduced.”

ACM Artifact Review and Badging



Speedy Neutrinos Challenge Physicists



Shocking discovery: neutrinos travelled 60 nanoseconds faster than light speed.

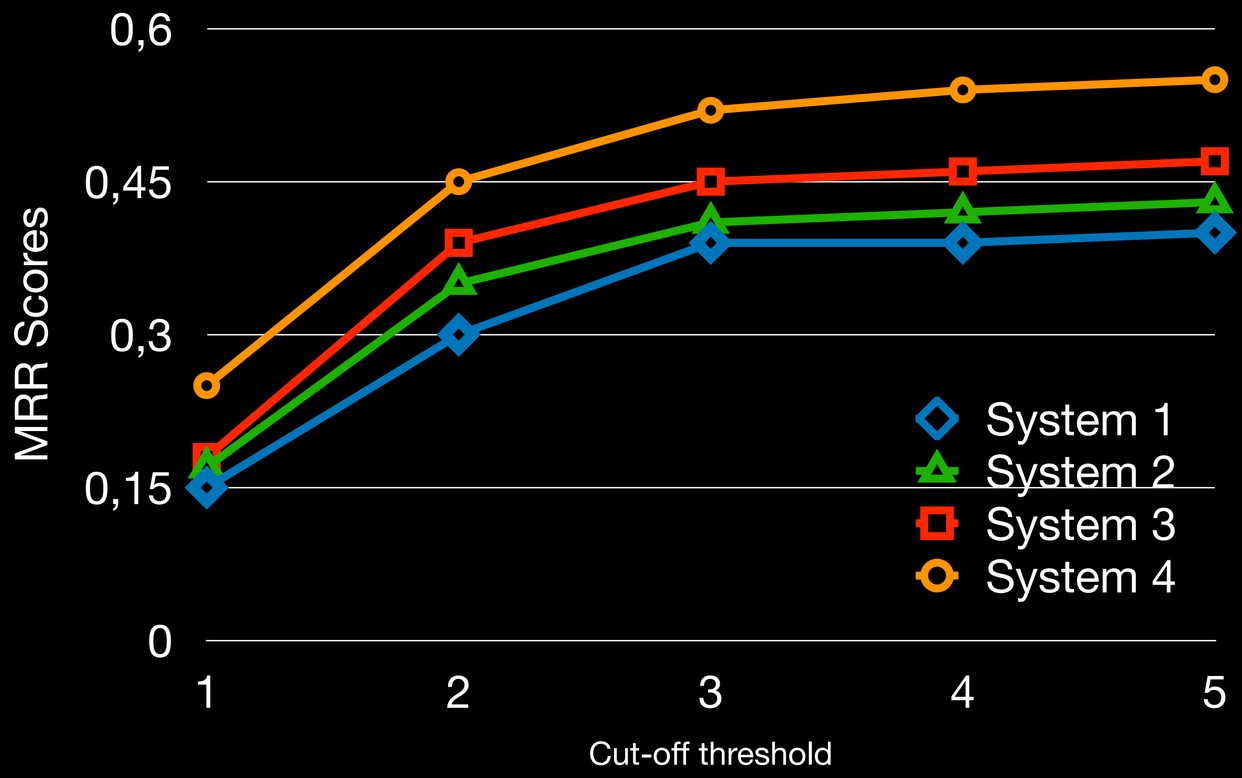
Reich, E. S. (2011). Speedy Neutrinos Challenge Physicists: Experiment Under Scrutiny as Teams Prepare to Test Claim that Particles can Beat Light Speed. Nature, 477(7366), 520-521.

Computing Reciprocal Rank

Reproducibility of an Evaluation Measure

$$\mathrm{MRR} = \frac{1}{|U|} \sum_{u=1}^{U} \frac{1}{\mathrm{k}_u}$$

`trec_eval`



What is Reproducibility?

Research Misconduct


The Washington Post
Democracy Dies in Darkness

Health

Researcher who spiked rabbit blood to fake HIV vaccine results slapped with rare prison sentence

The researcher spiked rabbit blood samples with human HIV antibodies so that the vaccine appeared to have caused the animals to develop immunity to the virus.

https://www.washingtonpost.com/news/to-your-health/wp/2015/07/01/researcher-who-spiked-rabbit-blood-to-fake-hiv-vaccine-results-slapped-with-rare-prison-sentence/

What is Reproducibility?

The “R Words”

- Scientific method – reproducible, repeatable, replicable, reusable
- Access – referenceable, retrievable, reviewable
- Understanding – replayable, reinterpretable, reprocessable
- New use – recomposable, reconstructable, repurposable
- Social – reliable, respectful, reputable, revealable
- Curation – recoverable, restorable, reparable, refreshable

D. De Roure. 2014. The future of scholarly communications. Insights 27, 3 (November 2014), 233–238.

ACM Terminology

- **Repeatability** (Same team, same experimental setup): the measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- **Reproducibility** (Different team, same experimental setup): the measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- **Replicability** (Different team, different experimental setup): the measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

ACM Artifact Review and Badging: https://www.acm.org/publications/policies/artifact-review-badging

What is the Status of Reproducibility?

Raise your hand…

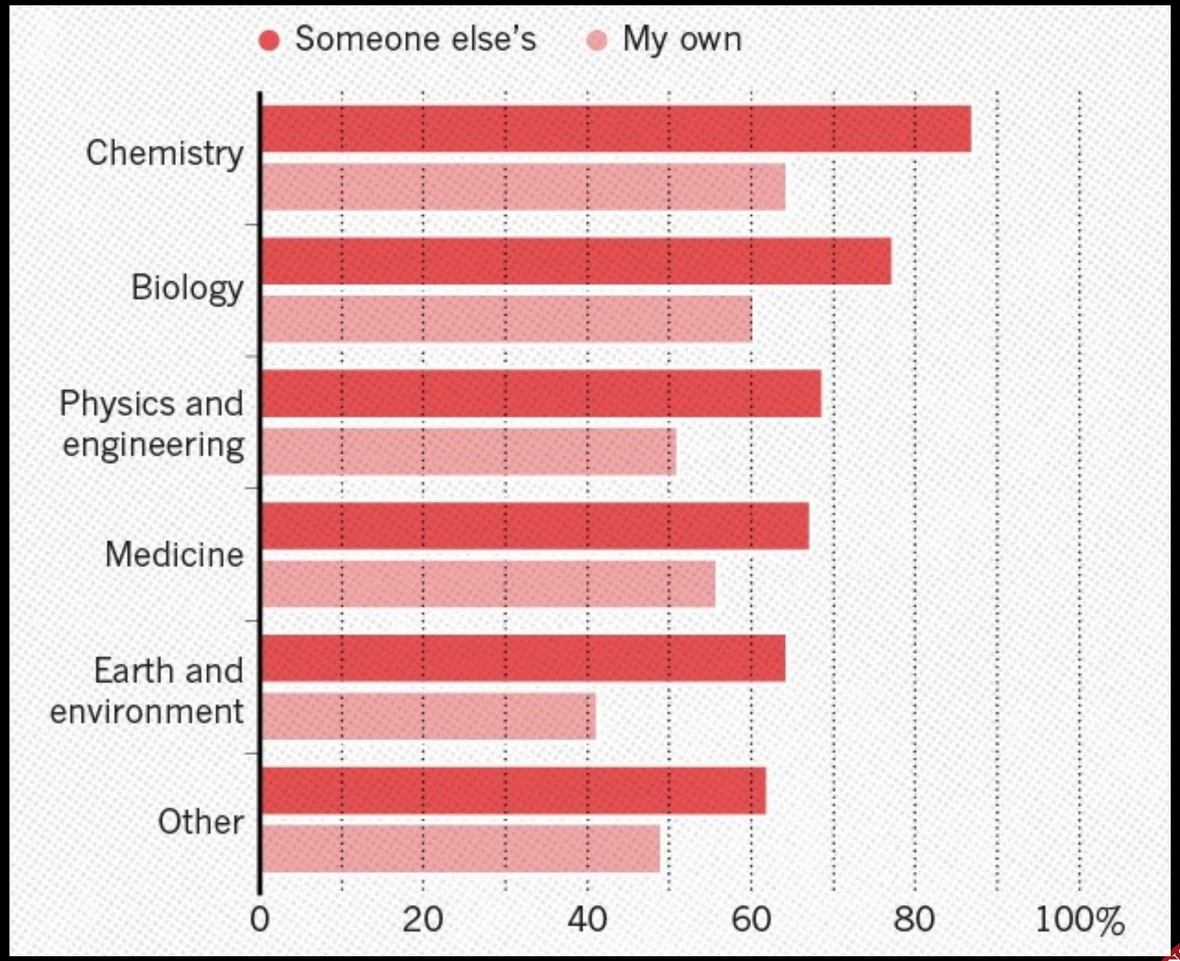
"Failing to reproduce results is a rite of passage"

Marcus Munafo, biological psychologist at te university of Bristol, UK

- Have you ever tried and failed to reproduce another scientist experiment?
- Have you ever failed to repeat your own experiment?

Have you Failed to Reproduce an Experiment?

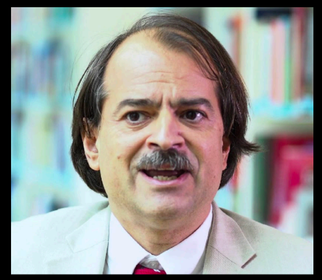
More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to repeat their own experiments



Baker, M. (2016). Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help. Nature, 533(7604), 452-455.

“Most published research findings are false.”

John Ioannidis, Stanford University, PLoS Med 2005;2(8): e124.



Reproducibility is a core issue to (almost) any scientific discipline:

- 39% (39/100) in psychological studies[1]
- 21% (14/67) in pharmacological studies[2]
- 11% (6/53) in cancer studies[3]
- 46% (12/26) in deep learning for recommendation[4]

A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research

MAURIZIO FERRARI DACREMA, SIMONE BOGLIO, and PAOLO CREMONESI,
Politecnico di Milano, Italy
DIETMAR JANNACH, University of Klagenfurt, Austria

20

The design of algorithms that generate personalized ranked item lists is a central topic of research in the field of recommender systems. In the past few years, in particular, approaches based on deep learning (neural) techniques have become dominant in the literature. For all of them, substantial progress over the state-of-the-art

[1] Baker, M. (2015). First Results from Psychology’s Largest Reproducibility Test. *Nature News*.
[2] Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how Much can we Rely on Published Data on Potential Drug Targets?. Nature reviews Drug discovery, 10(9), 712.
[3] Begley, C. G., and Ellis, L. M. (2012). Raise Standards for Preclinical Cancer Research. Nature, 483(7391), 531-533.
[4] Ferrari Dacrema, M., Boglio, S., Cremonesi, P., and Jannach, D. (2021). A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. ACM TOIS.

Reproducibility in IR

Survey on the SIGIR implementation of ACM Artifact Review & Badging:

- What about introducing badges?

  ➡ 75% supportive or very supportive, only 10% negative answers

- Would you submit your paper to be revised for a badge?

  ➡ 70% consider to submit their paper, only 10% would not submit

- Would badges change your way to do research?

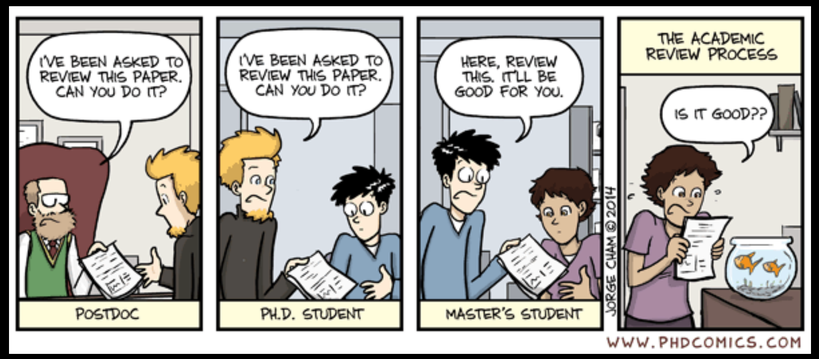
  ➡ 40% yes and 40% no.

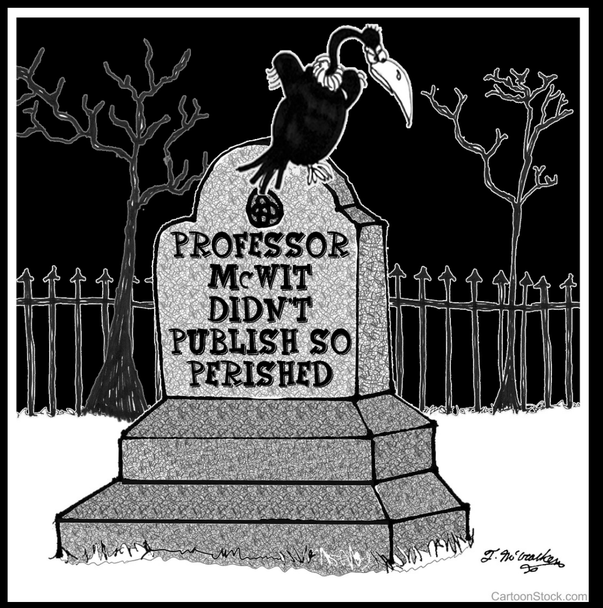
Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. SIGIR Forum.

Why is it so Difficult to Achieve Reproducibility?

Science Overload

- >50 Million: total number of science papers published from 1665 to 2009[1];
- Publishing ~3 millions articles per year (estimated in 2018)[2];
- Google Scholar was estimated to index between 100[3] and 160[4] million documents in 2014.
- Peer review process?





"I don't mind your thinking slowly; I mind your publishing faster than you think."

Wolfgang Ernst Pauli

Nobel Prize in Physics

[1] Jinha, A. E. (2010). Article 50 Million: an Estimate of the Number of Scholarly Articles in Existence. Learned Publishing, 23(3), 258-263.
[2] Johnson, R., Watkinson, A., and Mabe, M. (2018). The STM report. An Overview of Scientific and Scholarly Publishing. 5th Edition, October.
[3] Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., and López-Cózar, E. D. (2014). About the Size of Google Scholar: Playing the Numbers. arXiv preprint arXiv:1407.6239.
[4] Khabsa, M., & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. PloS one, 9(5), e93949.
Image Credit: http://phdcomics.com/comics.php?f=1760 and CartoonStock.com

Why Most Published Research Findings are False

A research finding is less likely to be true when:

- The studies conducted in a field are small;
- There is bias: manipulation in the analysis and selective or distorted reporting;
- There is a great flexibility in designs, definitions, outcomes and analytical modes;
- There is great financial interest and prejudice.

Ioannidis, J. P. (2005). Why Most Published Research Findings are False. PLoS medicine, 2(8), e124.

Luckily we have Evaluation Campaigns

- No “small” studies and the same track can run multiple times;
- No bias, experimental evaluation is performed by organizers;
- No flexibility, same experimental set-up for all participants;
- No financial interest or prejudice;
- Publicly available dataset and sometimes source code.

What happens when we reproduce a system?

The Score is Close Enough

1. Pick a model you would like to reproduce;
2. If possible, use the same dataset(s) as in the original paper;
3. Reimplement the model or re-use the source code;
4. Compare the scores obtained with the ones in the original paper;
5. Adjust your implementation until the performance score is close enough.

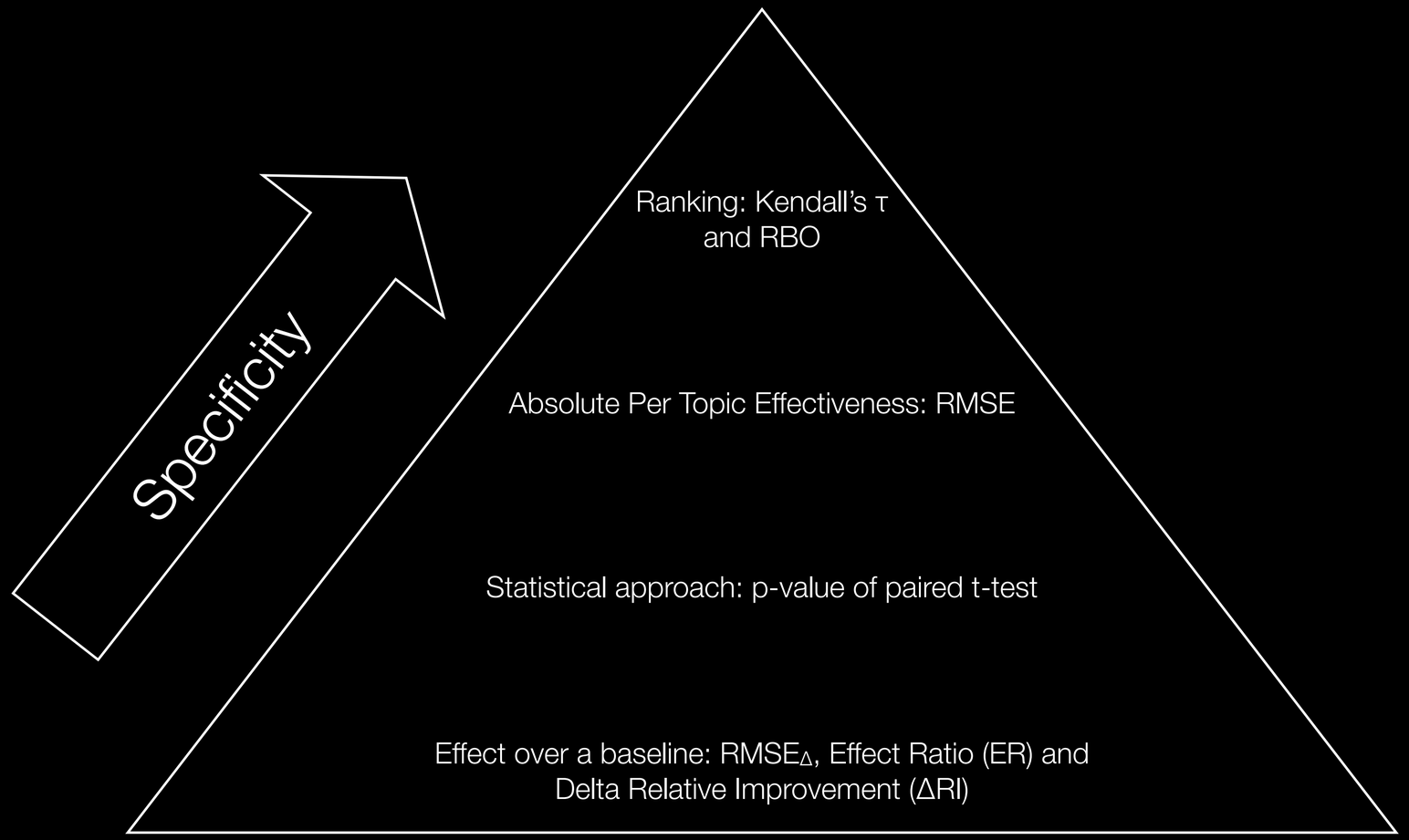
Can we Measure Reproducibility?

The Goal

- Input: an original run r and a reproduced run r’;
- Goal: measure the similarity between r and r’.
- Close enough approach: Delta Average Retrieval Performance (ARP).

$$\Delta\mathrm{ARP} = \overline{M(r)} - \overline{M(r')} = \frac{1}{T}\sum_{t=1}^{T} M(r_t) - \frac{1}{T}\sum_{t=1}^{T} M(r'_t)$$

Reproducibility Measures



Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., and Soboroff, I. (2020). How to Measure the Reproducibility of System-oriented IR Experiments. In SIGIR 2020.

Ranking Level

- Kendall's Tau Union (KTU):

$$\mathrm{KTU}_t(l_t, l'_t) = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}}$$

- Rank Biased Overlap (RBO):

$$\mathrm{RBO}_t(r_t, r'_t) = (1 - \phi) \sum_{k=1}^{\infty} \phi^{k-1} \cdot A_k$$

Per Topic Effectiveness

- Root Mean Square Error (RMSE):

$$\mathrm{RMSE}(r, r', M) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(M(r_t) - M(r'_t))^2}$$

- M is any IR effectiveness measure (e.g., Average Precision)

Statistical Approach

- Two-tailed paired t-test between the scores of the original and reproduced runs;
- The p-value as an indicator of reproducibility;
- The smaller the p-value, the stronger the evidence that the reproduced run is different from the original run.

Effect over a Baseline

- Effect Ratio (ER): comparison between baseline and advanced run

$$\mathrm{ER}(a, a', b, b', M) = \frac{\frac{1}{T}\sum_{t=1}^{T} M(a'_t) - M(b'_t)}{\frac{1}{T}\sum_{t=1}^{T} M(a_t) - M(b_t)}$$

- Delta Relative Improvement (DeltaRI):

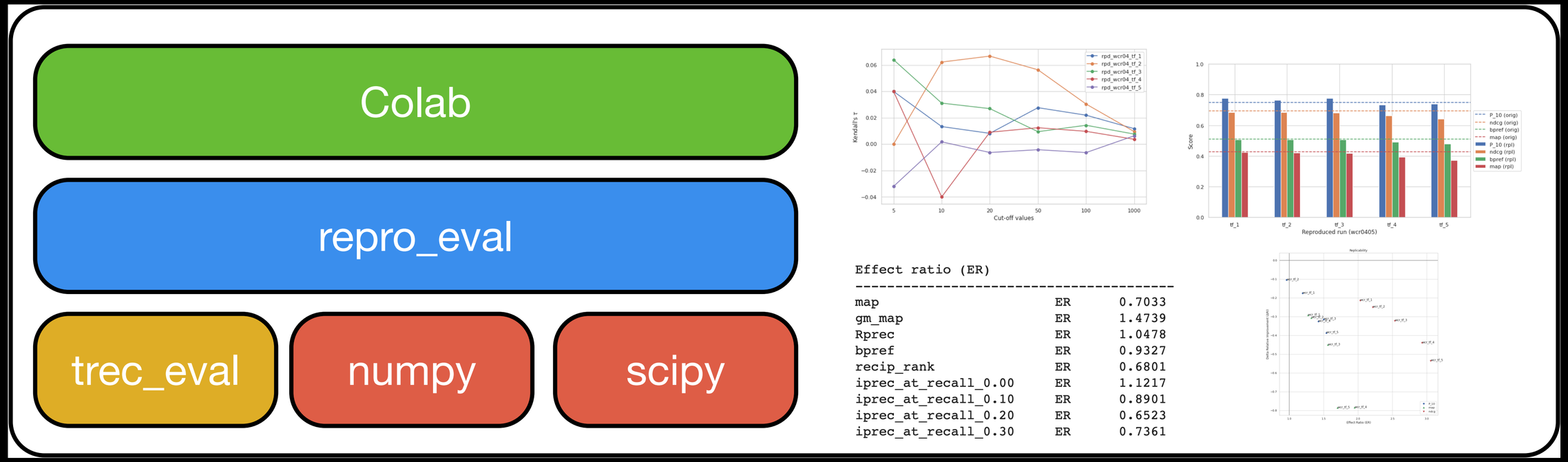
$$\begin{aligned}\Delta\mathrm{RI}(a, a', b, b', M) &= \mathrm{RI}(a, b, M) - \mathrm{RI}(a', b', M) \\ &= \frac{\overline{M(a)} - \overline{M(b)}}{\overline{M(b)}} - \frac{\overline{M(a')} - \overline{M(b')}}{\overline{M(b')}}\end{aligned}$$

Experimental Set-up

- Reproducibility dataset;
- `WCrobust04` and `WCrobust0405`, submitted by Grossman and Cormack[1] to the TREC 2017 Common Core track;
- Systematically change parameters: excluding pre-processing steps, varying the generation of the vocabulary, applying different tf-idf formulations, etc.
- A total of 100 runs;
- `rpl_wcr04_tf`: incrementally reduce the vocabulary size (5 runs).

Grossman, M. R. and Cormack, G. V. (2017). MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track. In TREC 2017.

Repro_eval: Library for Reproducibility

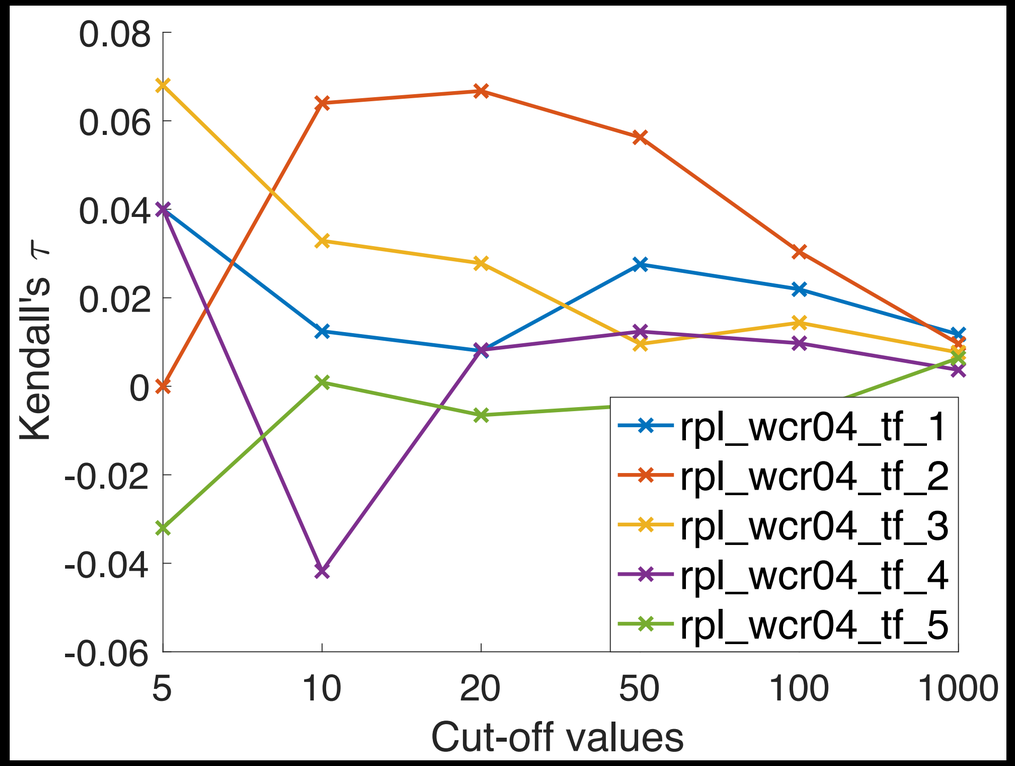



```
Effect ratio (ER)
-----------------------------------------
map                       ER        0.7033
gm_map                    ER        1.4739
Rprec                     ER        1.0478
bpref                     ER        0.9327
recip_rank                ER        0.6801
iprec_at_recall_0.00      ER        1.1217
iprec_at_recall_0.10      ER        0.8901
iprec_at_recall_0.20      ER        0.6523
iprec_at_recall_0.30      ER        0.7361
```


Breuer, T., Ferro, N., Maistro, M., and Schaer, P. repro_eval: A Python Interface to Reproducibility Measures of System-oriented IR Experiments. In ECIR 2021

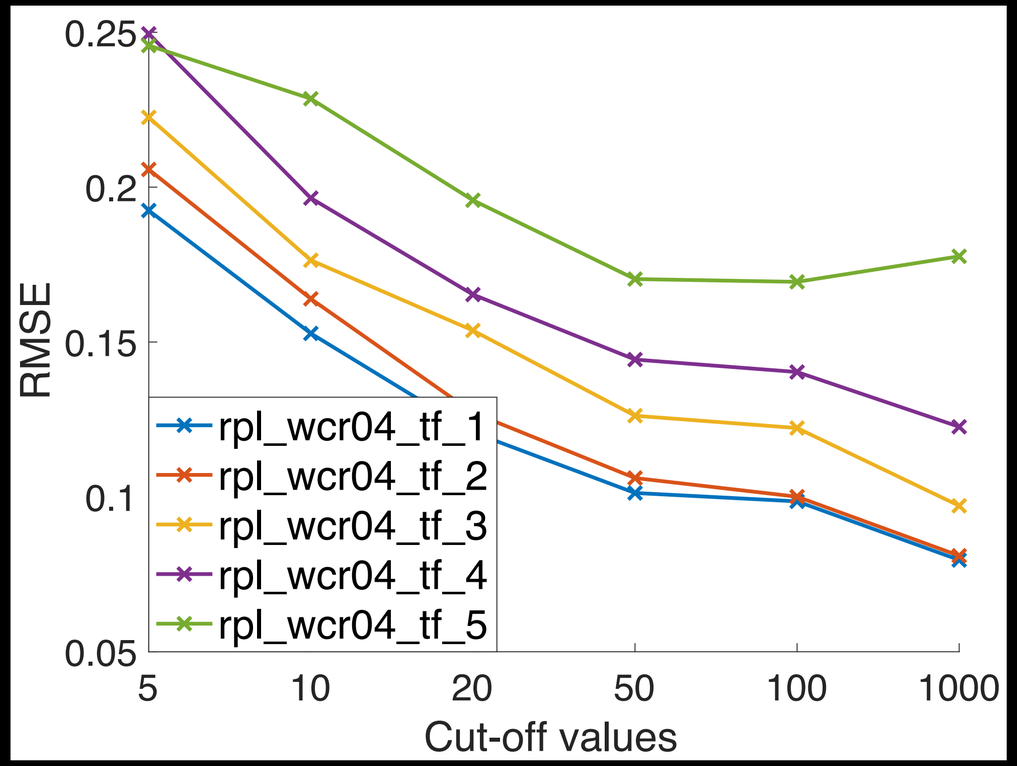
Ranking Level

run	ARP P@10	ARP AP	ARP nDCG	Correlation τ	Correlation RBO
WCrobust04	0.6460	0.3711	0.6371	1	1
rpl_wcr04_tf_1	0.6920	0.3646	0.6172	0.0117	0.5448
rpl_wcr04_tf_2	0.6900	0.3624	0.6177	0.0096	0.5090
rpl_wcr04_tf_3	0.6820	0.3420	0.6011	0.0076	0.4372
rpl_wcr04_tf_4	0.6680	0.3106	0.5711	0.0037	0.3626
rpl_wcr04_tf_5	0.6220	0.2806	0.5365	0.0064	0.2878



Per-topic Effectiveness & p-values

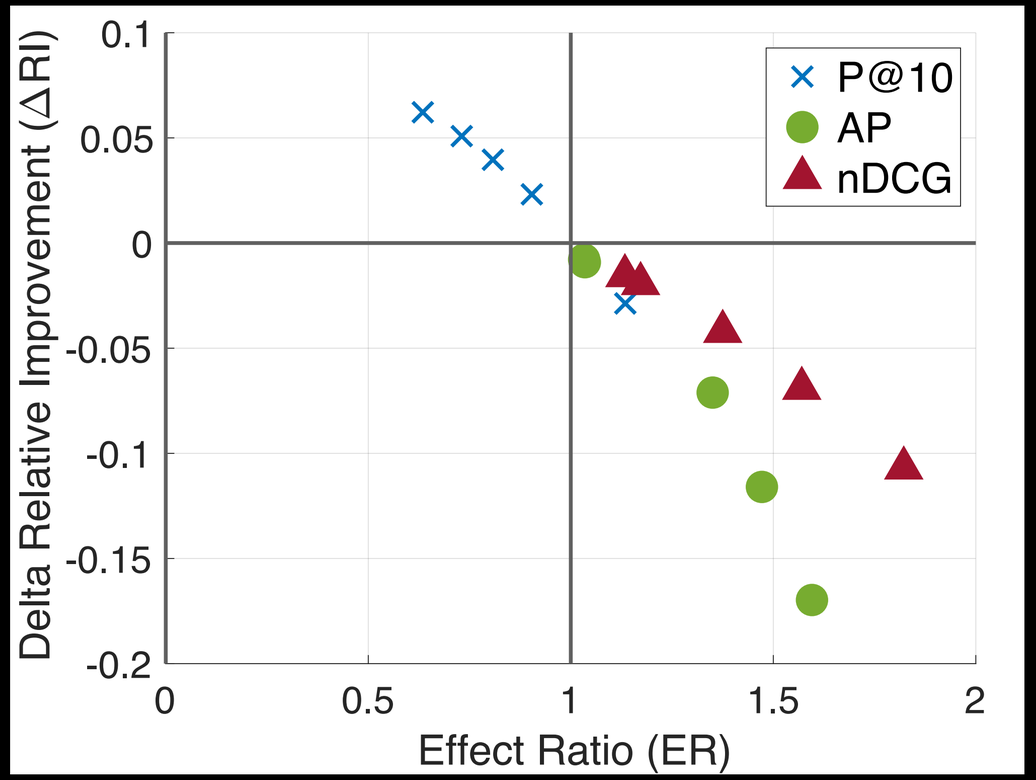
run	ARP P@10	ARP AP	ARP nDCG	RMSE P@10	RMSE AP	RMSE nDCG	p-value P@10	p-value AP	p-value nDCG
WCrobust04	0.6460	0.3711	0.6371	0	0	0	1	1	1
rpl_wcr04_tf_1	0.6920	0.3646	0.6172	0.2035	0.0755	0.0796	0.110	0.551	0.077
rpl_wcr04_tf_2	0.6900	0.3624	0.6177	0.2088	0.0799	0.0810	0.137	0.445	0.090
rpl_wcr04_tf_3	0.6820	0.3420	0.6011	0.2375	0.1083	0.0971	0.288	0.056	0.007
rpl_wcr04_tf_4	0.6680	0.3106	0.5711	0.2534	0.1341	0.1226	0.544	$9E{-}04$	$4E{-}05$
rpl_wcr04_tf_5	0.6220	0.2806	0.5365	0.2993	0.1604	0.1777	0.575	$1E{-}05$	$1E{-}05$



Effect Over a Baseline

- ER = 1 → perfect reproducibility;
- DeltaRI = 0 → perfect reproducibility.

run	replicability		
	P@10	AP	nDCG
rpl_tf_1	0.8077	1.0330	1.1724
rpl_tf_2	0.7308	1.0347	1.1336
rpl_tf_3	0.9038	1.3503	1.3751
rpl_tf_4	0.6346	1.4719	1.5703
rpl_tf_5	1.1346	1.5955	1.8221



Correlation Analysis of Reproducibility Measures

- High correlation (> 0.8):
  - ARP, RMSE, p-value with AP and nDCG;
  - ARP and p-value with all measures (P@10, AP and nDCG);
  - ARP and RMSE with AP and nDCG;
- Low correlation (< 0.3):
  - KTU will all other measures;
  - ER with ARP and p-values.

What about Replicability?

- Replicability: different team, different experimental setup;
- Statistical approach: two-tailed unpaired t-test;
- Effect ratio and delta relative improvement;
- Even harder than reproducibility;
- None of our runs could achieve good reproducibility scores on TREC Common Core 2018;
- Even when we new that the runs were generated by the same system.

Conclusions on How to Measure Reproducibility

- Comparing average scores might not be enough;
- Differences in the actual ranking of documents → impact on the user?
- Different effectiveness measures might lead to different results → which measure to use for reproducibility?
- Top heaviness affects the results → what are important features for reproducibility?

Reproducibility is challenging and replicability even more!

How can we Ease Reproducibility?

Reproducibility Initiatives in IR

- ACM Artifacts Badging Policy
- Qualitatively assessed in review forms (SIGIR, ECIR, TOIS, …);
- Since 2015 ECIR track devoted to it and now also SIGIR;
- SIGIR 2015 RIGOR Workshop;
- CENTRE evaluation across CLEF/NTCIR/TREC (2018 - present);
- Weak open-source baselines;
- The Open-Source IR Replicability Challenge (OSIRRC 2019) at SIGIR 2019.

ACM Badging Artifacts

- Artifacts have successfully completed an independent audit:
  - Functional
  - Reusable
- Artifacts have been made permanently available for retrieval:
  - Available
- The main results of the paper have been successfully obtained by a person or team other than the author:
  - Results Reproduced
  - Results Replicated

Write Reproducible Papers

- Datasets, experimental procedures, and code publicly available (FAIR Principles);
- Dockers or other “containers” for source code;
- Open-runs[1];
- Describe implementation choices and experimental set-up, even tiny details;
- Follow some simple rules to ease reproducibility of the experimental results.

The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

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

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

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

- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:

- ❑ The relevant statistics, such as number of examples.
- ❑ The details of train / validation / test splits.
- ❑ An explanation of any data that were excluded, and all pre-processing step.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:

- ❑ Specification of dependencies.
- ❑ Training code.
- ❑ Evaluation code.
- ❑ (Pre-)trained model(s).
- ❑ README file includes table of results accompanied by precise command to run to produce those results.

For all reported **experimental results**, check if you include:

- ❑ 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 training and evaluation runs.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. error bars).
- ❑ The average runtime for each result, or estimated energy cost.
- ❑ A description of the computing infrastructure used.

https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

[1] Voorhees, E.M., Rajput, S., Soboroff, I. (2016). Promoting Repeatability through Open Runs. EVIA 2016.

Reproducibility: Some Needs

- Shift in culture:
  - More work needed to put reproducibility in action;
  - Acknowledgment in careers;
  - Training future scientists: “Reproducible and Collaborative Data Science”;
- Systematic but focused approach:
  - How to choose what to reproduce?
- Quantitative assessment:
  - When do we consider something as “reproduced”?
- Infrastructures (evaluation campaigns?):
  - Lightweight tools and protocols… but they need adoption!

https://berkeley-stat159-f17.github.io/stat159-f17/

Special Thank

Timo Breuer

Nicola Ferro

Norbert Fuhr

Tetsuya Sakai

Philipp Schaer

Ian Soboroff

Thank you!
Any Questions or Comments?