How can we measure reproducibility of IR experiments?

Maria Maistro, mm@di.ku.dk
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
Shocking discovery: neutrinos travelled 60 nanoseconds faster than light speed.

Computing Reciprocal Rank
Reproducibility of an Evaluation Measure

\[ \text{MRR} = \frac{1}{|U|} \sum_{u=1}^{U} \frac{1}{k_u} \]

MRR Scores

<table>
<thead>
<tr>
<th>Cut-off threshold</th>
<th>System 1</th>
<th>System 2</th>
<th>System 3</th>
<th>System 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

`trec_eval`
What is Reproducibility?
Research Misconduct

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.

What is Reproducibility?

![Diagram of a scientific experiment with a cross sign, indicating incorrect reproducibility.]

![Diagram of a person reading a paper at a desk, with a green check mark, indicating correct reproducibility.]

---

**Yummy Beer Batter Love and Chocolate Ice Cream**

**Ingredients:**
- 6 oz. flour
- 2 tbsp. sugar
- 1 tbsp. baking powder
- 1/2 tsp. salt
- 1/2 cup of milk
- 1 egg
- 1/2 cup of brown sugar
- 1/2 cup of heavy cream

1. Preheat oven to 350°F.
2. Mix flour, sugar, baking powder, and salt in a bowl.
3. Combine milk, brown sugar, and heavy cream in a bowl.
4. Add the wet ingredients to the dry ingredients and mix well.
5. Chill the batter for 1 hour in the refrigerator.
7. Serve immediately.
The “R Words”

• **Scientific method** – reproducible, repeatable, replicable, reusable

• **Access** – referenceable, retrievable, reviewable

• **Understanding** – replayable, reinterpretable, reprocessable

• **New use** – recomposable, reconstructable, repurposeable

• **Social** – reliable, respectful, reputable, revealable

• **Curation** – recoverable, restorable, reparable, refreshable

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 the 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.

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\(^4\)
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.

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?

---


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.

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).

\[
\Delta \text{ARP} = \overline{M(r)} - \overline{M(r')} = \frac{1}{T} \sum_{t=1}^{T} M(r_t) - \frac{1}{T} \sum_{t=1}^{T} M(r'_t)
\]
Reproducibility Measures

- Ranking: Kendall’s τ and RBO
- Absolute Per Topic Effectiveness: RMSE
- Statistical approach: p-value of paired t-test
- Effect over a baseline: $\text{RMSE}_\Delta$, Effect Ratio (ER) and Delta Relative Improvement (ΔRI)

Ranking Level

- Kendall’s Tau Union (KTU):

\[ KTU_t(l_t, l'_t) = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}} \]

- Rank Biased Overlap (RBO):

\[ RBO_t(r_t, r'_t) = (1 - \phi) \sum_{k=1}^{\infty} \phi^{k-1} \cdot A_k \]
Per Topic Effectiveness

- Root Mean Square Error (RMSE):

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

- \(M\) 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

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

- **Delta Relative Improvement (DeltaRI):**

\[
\Delta RI(a, a', b, b', M) = RI(a, b, M) - RI(a', b', M)
\]

\[
= \frac{M(a) - M(b)}{M(b)} - \frac{M(a') - M(b')}{M(b')}
\]
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\_wc\_04\_tf: incrementally reduce the vocabulary size (5 runs).

Repro_eval: Library for Reproducibility

Ranking Level

<table>
<thead>
<tr>
<th>run</th>
<th>P@10</th>
<th>ARP</th>
<th>AP</th>
<th>nDCG</th>
<th>Correlation</th>
<th>RBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCrobust04</td>
<td>0.6460</td>
<td>0.3711</td>
<td>0.6371</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>rpl_wcr04_tf_1</td>
<td>0.6920</td>
<td>0.3646</td>
<td>0.6172</td>
<td>0.0117</td>
<td>0.5448</td>
<td></td>
</tr>
<tr>
<td>rpl_wcr04_tf_2</td>
<td>0.6900</td>
<td>0.3624</td>
<td>0.6177</td>
<td>0.0096</td>
<td>0.5090</td>
<td></td>
</tr>
<tr>
<td>rpl_wcr04_tf_3</td>
<td>0.6820</td>
<td>0.3420</td>
<td>0.6011</td>
<td>0.0076</td>
<td>0.4372</td>
<td></td>
</tr>
<tr>
<td>rpl_wcr04_tf_4</td>
<td>0.6680</td>
<td>0.3106</td>
<td>0.5711</td>
<td>0.0037</td>
<td>0.3626</td>
<td></td>
</tr>
<tr>
<td>rpl_wcr04_tf_5</td>
<td>0.6220</td>
<td>0.2806</td>
<td>0.5365</td>
<td>0.0064</td>
<td>0.2878</td>
<td></td>
</tr>
</tbody>
</table>
Per-topic Effectiveness & p-values

<table>
<thead>
<tr>
<th>run</th>
<th>ARP P@10</th>
<th>AP</th>
<th>nDCG</th>
<th>RMSE P@10</th>
<th>AP</th>
<th>nDCG</th>
<th>p-value P@10</th>
<th>AP</th>
<th>nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCrobust04</td>
<td>0.6460</td>
<td>0.3711</td>
<td>0.6371</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>rpl_wcr04_tf_1</td>
<td>0.6920</td>
<td>0.3646</td>
<td>0.6172</td>
<td>0.2035</td>
<td>0.0755</td>
<td>0.0796</td>
<td>0.110</td>
<td>0.551</td>
<td>0.077</td>
</tr>
<tr>
<td>rpl_wcr04_tf_2</td>
<td>0.6900</td>
<td>0.3624</td>
<td>0.6177</td>
<td>0.2088</td>
<td>0.0799</td>
<td>0.0810</td>
<td>0.137</td>
<td>0.445</td>
<td>0.090</td>
</tr>
<tr>
<td>rpl_wcr04_tf_3</td>
<td>0.6820</td>
<td>0.3420</td>
<td>0.6011</td>
<td>0.2375</td>
<td>0.1083</td>
<td>0.0971</td>
<td>0.288</td>
<td>0.056</td>
<td>0.007</td>
</tr>
<tr>
<td>rpl_wcr04_tf_4</td>
<td>0.6680</td>
<td>0.3106</td>
<td>0.5711</td>
<td>0.2534</td>
<td>0.1341</td>
<td>0.1226</td>
<td>0.544</td>
<td>9E-04</td>
<td>4E-05</td>
</tr>
<tr>
<td>rpl_wcr04_tf_5</td>
<td>0.6220</td>
<td>0.2806</td>
<td>0.5365</td>
<td>0.2993</td>
<td>0.1604</td>
<td>0.1777</td>
<td>0.575</td>
<td>1E-05</td>
<td>1E-05</td>
</tr>
</tbody>
</table>
Effect Over a Baseline

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

<table>
<thead>
<tr>
<th>run</th>
<th>replicability</th>
<th>P@10</th>
<th>AP</th>
<th>nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpl_tf_1</td>
<td></td>
<td>0.8077</td>
<td>1.0330</td>
<td>1.1724</td>
</tr>
<tr>
<td>rpl_tf_2</td>
<td></td>
<td>0.7308</td>
<td>1.0347</td>
<td>1.1336</td>
</tr>
<tr>
<td>rpl_tf_3</td>
<td></td>
<td>0.9038</td>
<td>1.3503</td>
<td>1.3751</td>
</tr>
<tr>
<td>rpl_tf_4</td>
<td></td>
<td>0.6346</td>
<td>1.4719</td>
<td>1.5703</td>
</tr>
<tr>
<td>rpl_tf_5</td>
<td></td>
<td>1.1346</td>
<td>1.5955</td>
<td>1.8221</td>
</tr>
</tbody>
</table>

![Graph showing effect ratio (ER) vs. delta relative improvement (ΔRI)]
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?

- **Re repliciability**: 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 knew 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

5 Papers in 2021
Write Reproducible Papers

- Datasets, experimental procedures, and code publicly available (FAIR Principles);
- Dockers or other “containers” for source code;
- Open-runs¹;
- 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.


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?