TOWARDS REPRODUCIBLE ML RESEARCH IN NLP

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Robert Stojnic

ACL 2022

Yann LeCun @ylecun · Apr 3, 2020

The Transformer-XL results from Google Brain on language modeling could not be reproduced by some top NLP researchers (and the authors are not helping).
@srush_nlp offers a bounty for whoever can reproduce the results.
(I assume the authors are excluded from the challenge!).

> **Sasha Rush** @srush_nlp · Apr 2, 2020
>
> Open-Science NLP Bounty: ($100 + $100 to charity)
>
> Task: A notebook demonstrating experiments within 30(!) PPL (<84) of this widely cited LM baseline on PTB / WikiText-2 using any non-pretrained, word-only Transformer variant.
>
> Context: twitter.com/Tim_Dettmers/s...
>
> Show this thread

Model	#Param	PPL
Inan et al. (2016) - Tied Variational LSTM	24M	73.2
Zilly et al. (2016) - Variational RHN	23M	65.4
Zoph and Le (2016) - NAS Cell	25M	64.0
Merity et al. (2017) - AWD-LSTM	24M	58.8
Pham et al. (2018) - Efficient NAS	24M	58.6
Liu et al. (2018) - Differentiable NAS	23M	56.1
Yang et al. (2017) - AWD-LSTM-MoS	22M	55.97
Melis et al. (2018) - Dropout tuning	24M	55.3
Ours - Transformer-XL	24M	**54.52**

TUTORIAL OVERVIEW

- **Part 1: Introduction to Reproducibility**
  - ML reproducibility crisis, examples from non-CS fields, how to conduct reproducible research
- **Part 2: Reproducibility in NLP**
  - NLP paper checklists, reproducibility research in NLP
- **Part 3: Mechanisms for Reproducibility**
  - Papers with Code, ML Reproducibility Challenge, useful tools and libraries
- **Part 4: Reproducibility as a Teaching Tool**
  - How to incorporate an ML reproducibility project into a course

TEACHING TEAM

Ana Lucic

University of Amsterdam

Maurits Bleeker

University of Amsterdam

Samarth Bhargav

University of Amsterdam

Jessica Zosa Forde

Brown University

Koustuv Sinha

McGill University

Jesse Dodge

Allen Institute for AI

Sasha Luccioni

HuggingFace

Robert Stojnic

Facebook AI Research

Tutorial website: https://acl-reproducibility-tutorial.github.io

INTRODUCTION TO REPRODUCIBILITY

Ana Lucic

OVERVIEW

1. Motivation
2. Reproducibility Crisis in ML
3. Reproducibility in Non-CS Fields
4. Conducting Reproducible Research

MOTIVATION

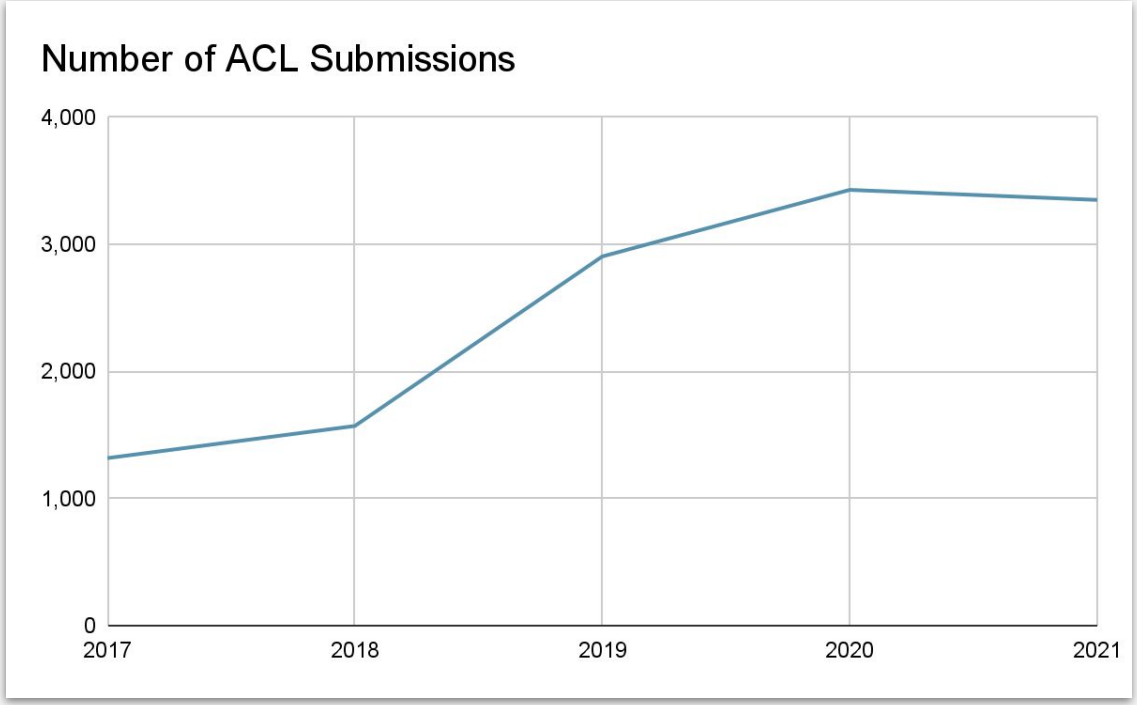
MOTIVATION

"At the very foundation of scientific inquiry is the process of specifying a hypothesis, running an experiment, analyzing the results and drawing conclusions"

"Scientists have used this process to build our collective understanding of the natural world and the laws that govern it. However, for the findings to be valid and reliable, it is important that the experimental process be repeatable, and yield consistent results and conclusions"

-- Pineau et al, 2020.

MOTIVATION


Number of ACL Submissions
4,000
3,000
2,000
1,000
0
2017
2018
2019
2020
2021

MOTIVATION

- As a field, we've made considerable progress by increasing the amount of computation used in our experiments:
  - Better performance
  - Easier to explore ideas
- This has also come with some challenges:
  - Running baselines can be very expensive
  - Results are not always reproducible

MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

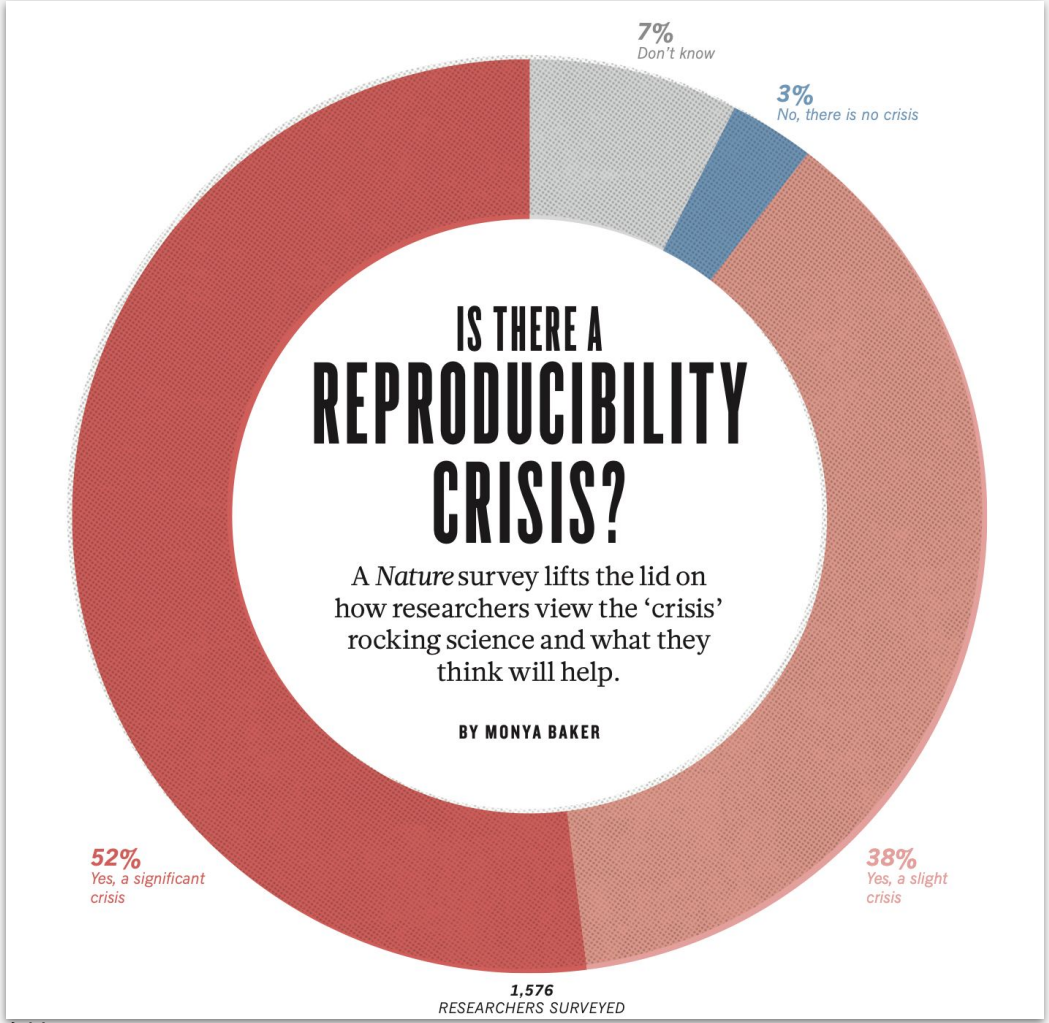
MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

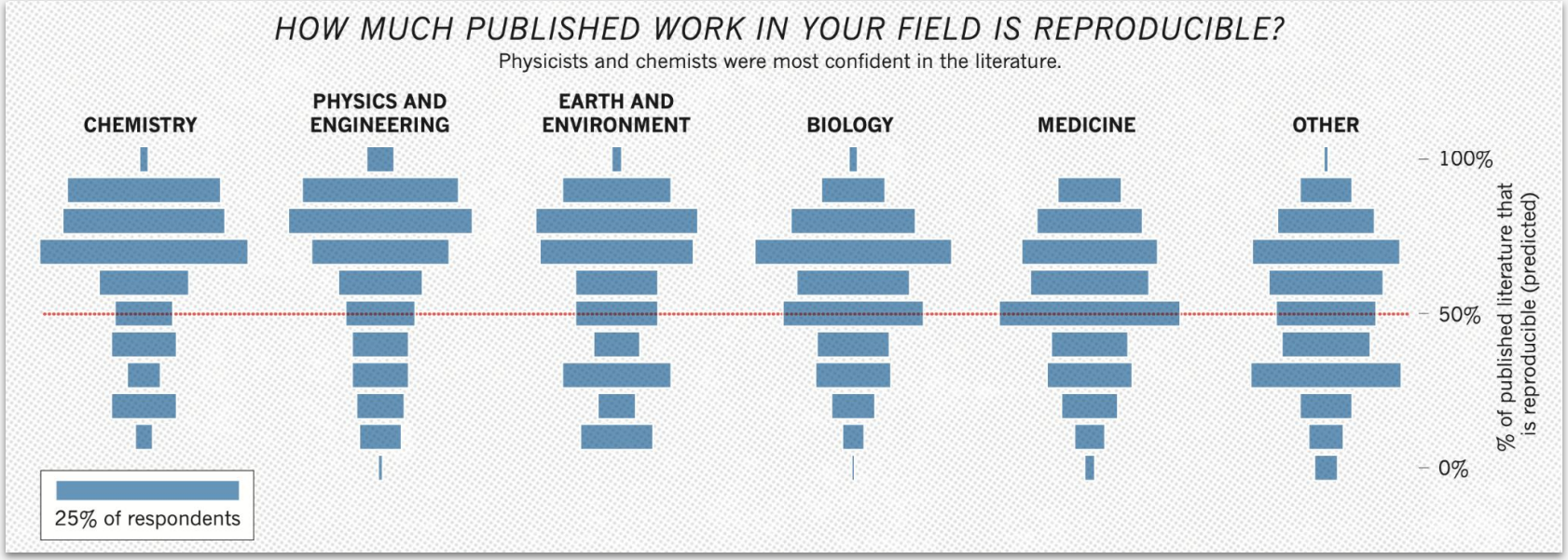
In this tutorial, we focus on the challenge of ensuring research results are reproducible

TUTORIAL OVERVIEW

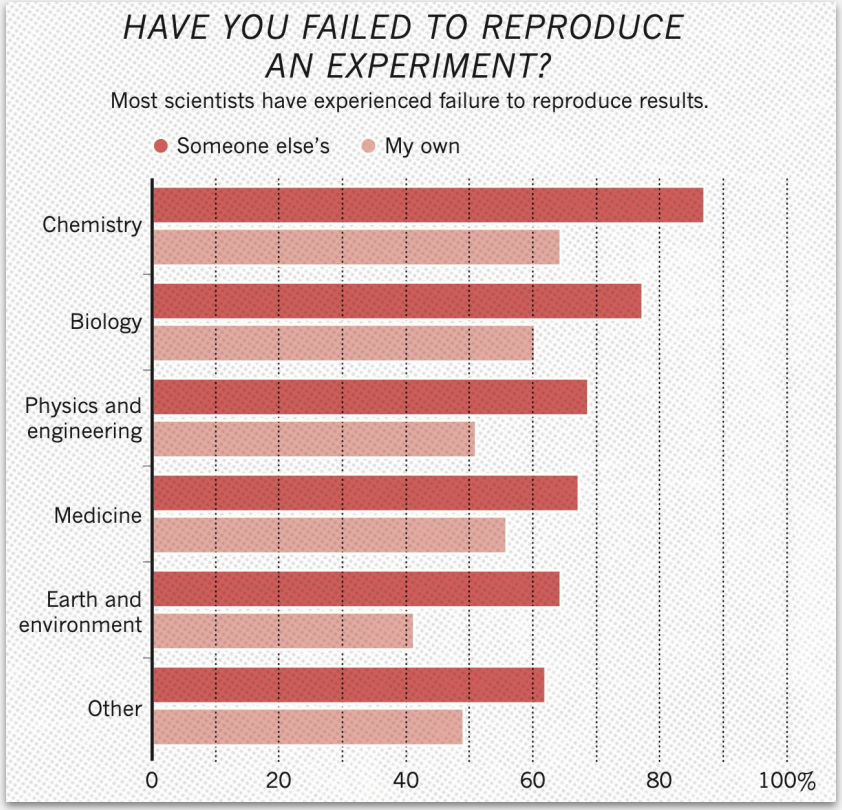
1. Introduction to Reproducibility
2. Reproducibility in NLP
3. Mechanisms for Reproducibility
4. Reproducibility as a Teaching Tool



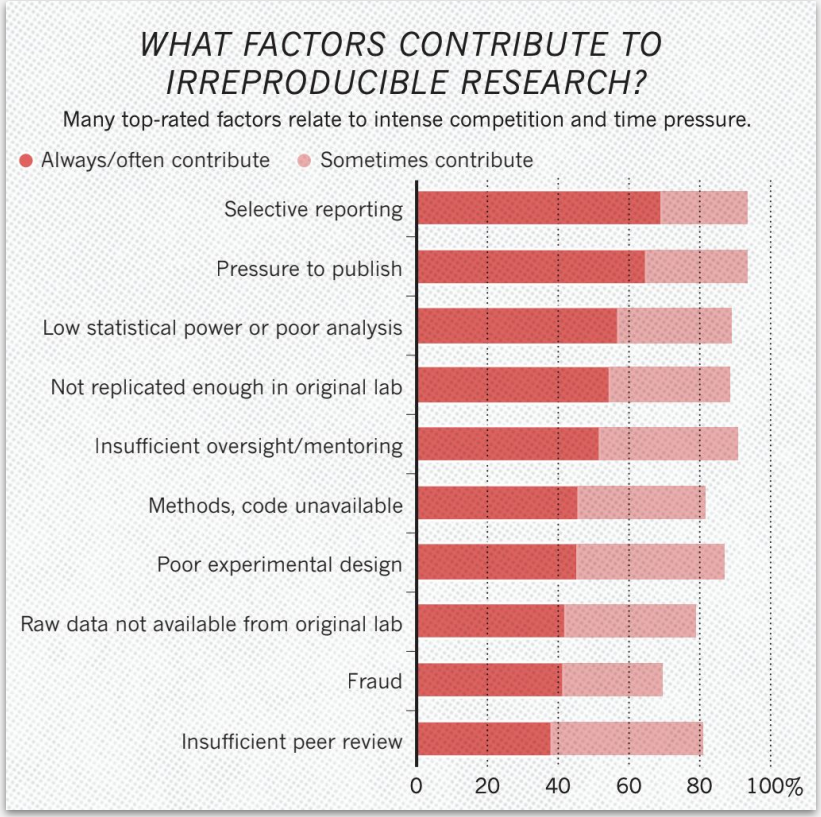
Baker. 2016. Is there a reproducibility crisis? *Nature*.



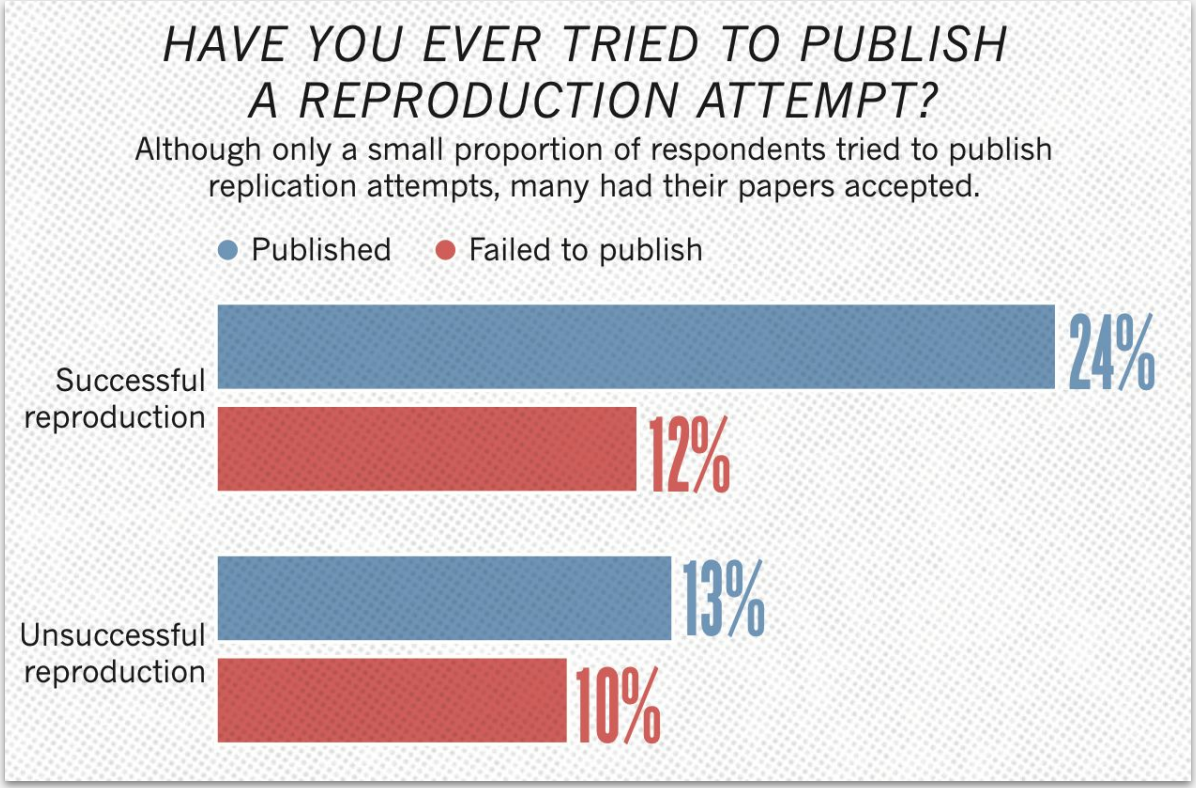
Baker. 2016. Is there a reproducibility crisis? *Nature.*



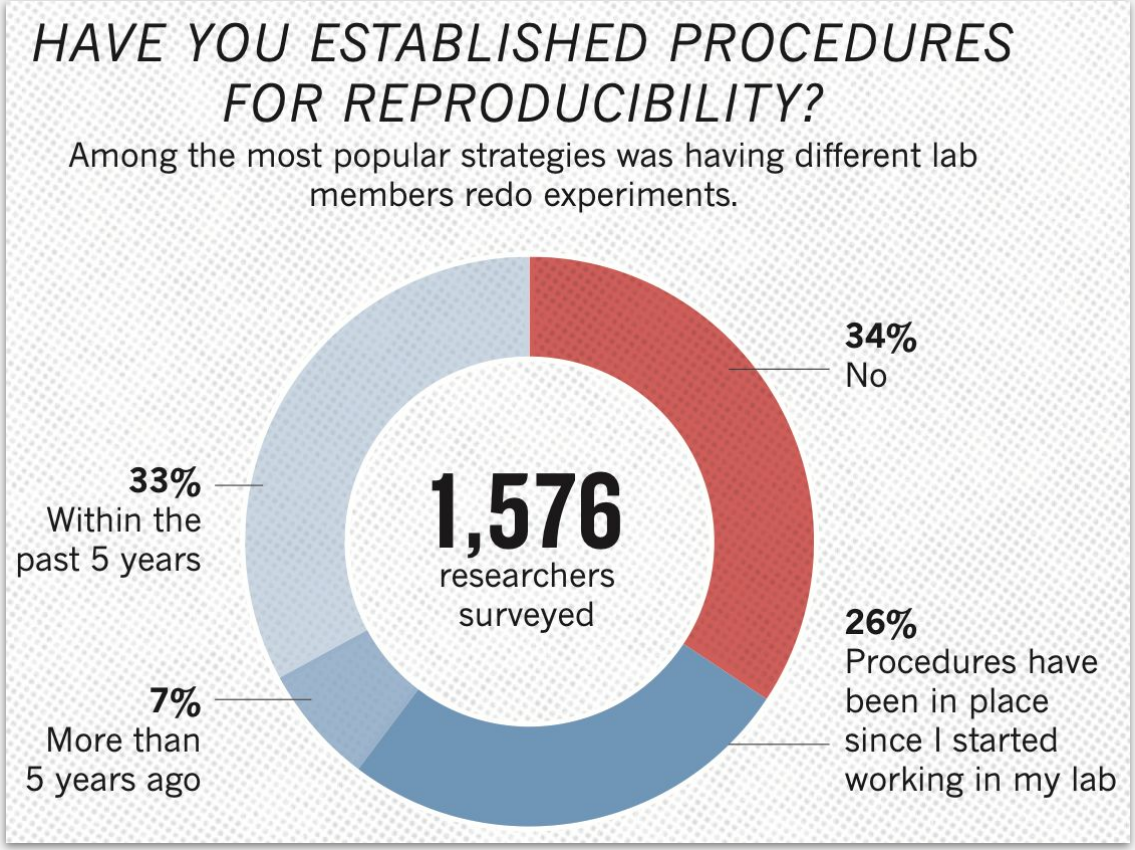
Baker. 2016. Is there a reproducibility crisis? *Nature.*



Baker. 2016. Is there a reproducibility crisis? *Nature*.



Baker. 2016. Is there a reproducibility crisis? *Nature.*



Baker. 2016. Is there a reproducibility crisis? *Nature.*

REPRODUCIBILITY IN ML

ACM DEFINITIONS

- **Repeatable:** a researcher can obtain the same results for their own experiment under exactly the same conditions, i.e., they can reliably repeat their own experiment ("Same team, same experimental setup")

- **Replicability:** a different researcher can obtain the same results for an experiment under exactly the same conditions and using exactly the same artifacts, i.e., another independent researcher can reliably repeat an experiment of someone other than herself ("Different team, same experimental setup")

- **Reproducibility:** a different researcher can obtain the same results for an experiment under different conditions and using their self-developed artifacts ("Different team, different experimental setup")

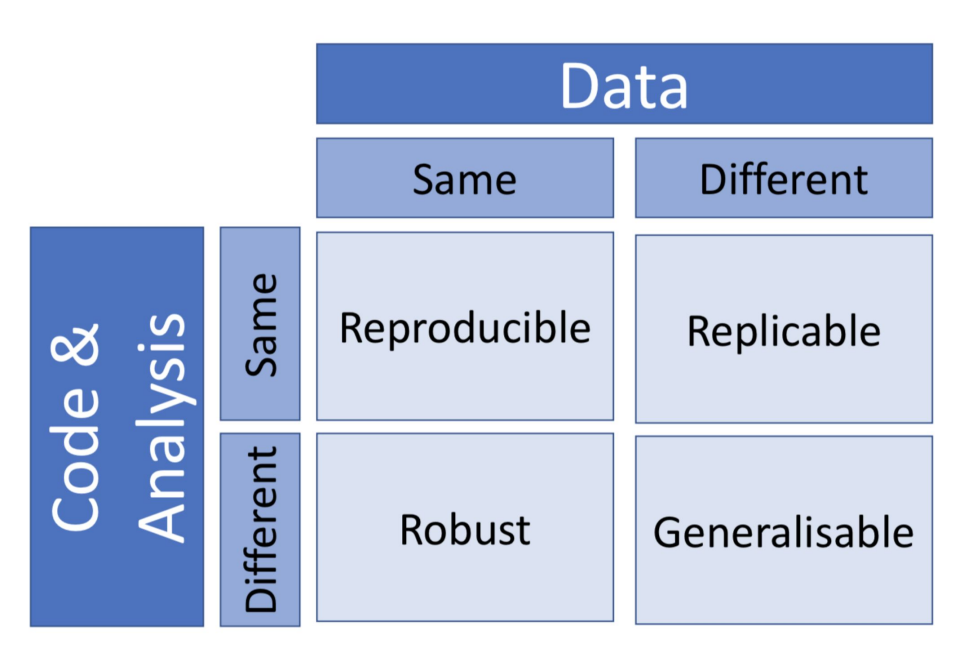
ACM. 2018. Artifact Review and Badging.

NEURIPS DEFINITIONS

- **Reproducible:** same conclusions are drawn when re-doing an experiment with the same data and same analytical tools
- **Replicable:** same conclusions are drawn when re-doing an experiment with a different dataset, but the same tools
- **Robust:** same conclusions are drawn when re-doing an experiment with the same data but different tools (i.e., different code implementations)
- **Generalizable:** same conclusions are drawn when re-doing an experiment with different data and different tools.

Pineau et al. 2020. Improving Reproducibility in ML Research

NEURIPS DEFINITIONS

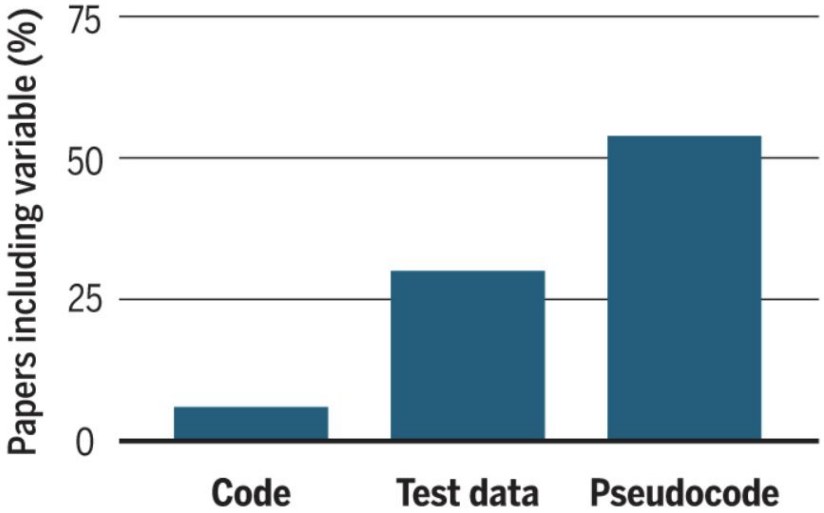


Pineau et al. 2020. Improving Reproducibility in ML Research

REPRODUCIBILITY CRISIS IN ML

Code break

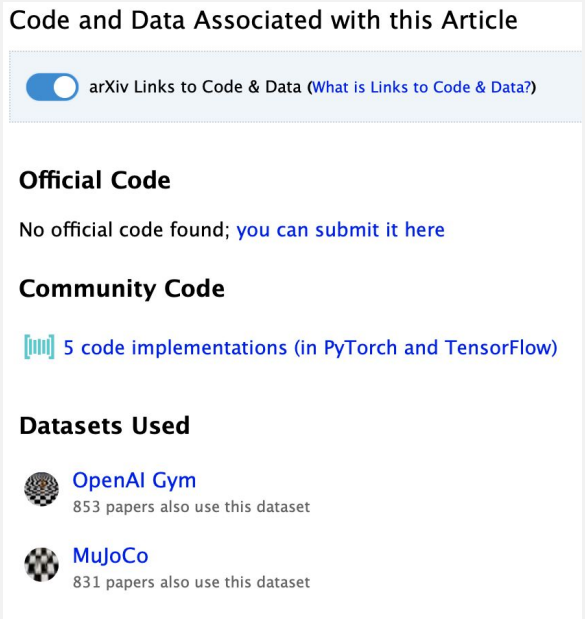
In a survey of 400 artificial intelligence papers presented at major conferences, just 6% included code for the papers' algorithms. Some 30% included test data, whereas 54% included pseudocode, a limited summary of an algorithm.



2018

Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science*

REPRODUCIBILITY CRISIS IN ML



- Since 2018, we've made some progress
- Many conferences strongly encourage or even require code submissions
- Can get links to code repositories and datasets through arXiv thanks to Papers with Code
- Reproducibility checklists at conferences

COMMON REPRODUCIBILITY ISSUES IN ML

- 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
- Improper use of statistics to analyze results
- Selective reporting or over-claiming of results

Pineau et al. 2020. Improving Reproducibility in ML Research

QUESTIONS?

REPRODUCIBILITY IN NON-CS FIELDS

PSYCHOLOGY

- The Open Science Collaboration conducted 100 replications of studies from 3 psychology journals
    - In total, there are 270 authors on the paper published in Science
- Found a significant proportion of replications produced weaker evidence despite using materials provided by authors
- Mean effect size of replication was found to be half of the original
    - Original: 97% significant ($p < 0.05$) vs Study: 36%

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*

BIOMEDICAL SCIENCES

- Clinical trials in oncology have some of the highest failure rates in comparison to other therapeutic areas
- Begley and Lee (2012) claim this is due to the lack of robustness in preclinical trials i.e., drug development
- Out of 53 "landmark" studies, only 6 could be reproduced
- Non-reproducible papers are still heavily cited since they are considered to be "part of the literature", contributing to failing clinical trials

Begley and Lee. 2012. Raise standards for preclinical cancer research. *Nature*

BIOMEDICAL SCIENCES

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Begley and Lee. 2012. Raise standards for preclinical cancer research. *Nature*

BIOMEDICAL SCIENCES

Recommendations proposed by Begley and Lee (2012):

- Require reporting on negative findings
- Encourage reporting on alternative findings that contradict existing work
- Implement transparent mechanisms for reporting unethical practices
- Increase dialogue between physicians, scientists, patient advocates and patients
- Recognize high-quality teaching and mentoring as valuable
- Funding organizations should facilitate development and access to new tools

Begley and Lee. 2012. Raise standards for preclinical cancer research. *Nature*

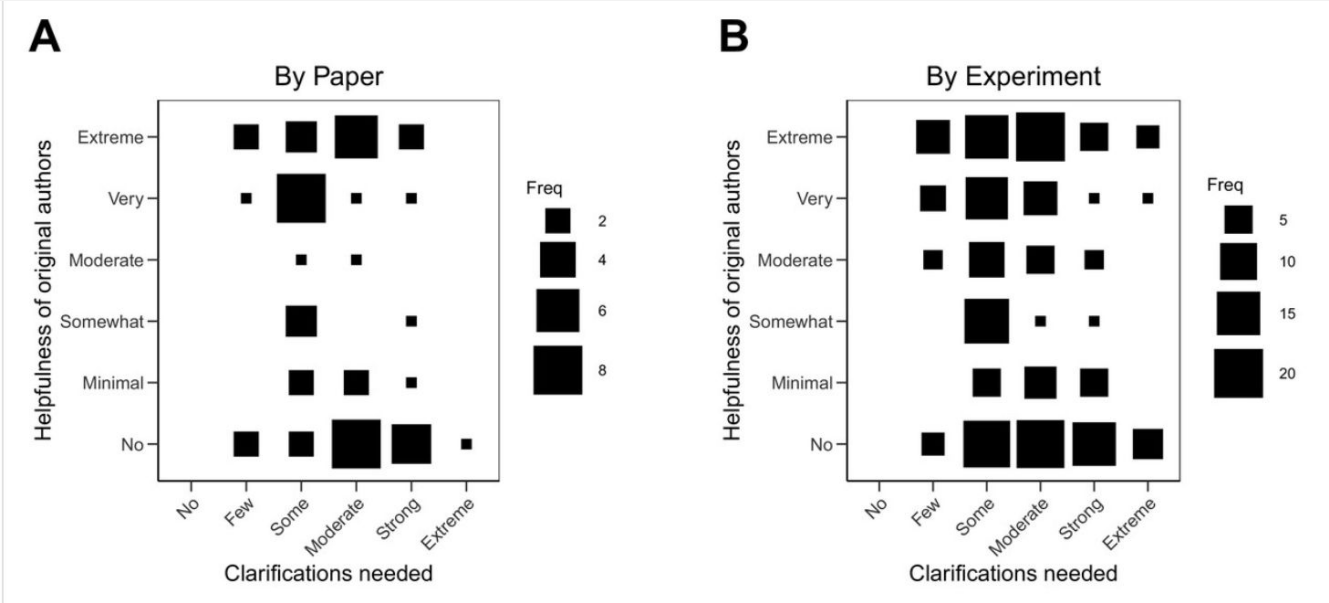
CANCER BIOLOGY

Errington et al (2020) conduct a reproduction of 193 experiments from 53 high impact papers in preclinical cancer biology:

- Only 50/193 experiments from 23 papers were reproduced
- Data was publicly accessible for 4 of 193 papers
- Authors would not share data for 68% of papers
- 32% of authors were rated as "not at all helpful" by researchers reproducing their experiments
- 67% of protocols described in papers needed modifications
    - Only 41% of those modifications could be implemented

Errington et al. 2021. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. *eLife*

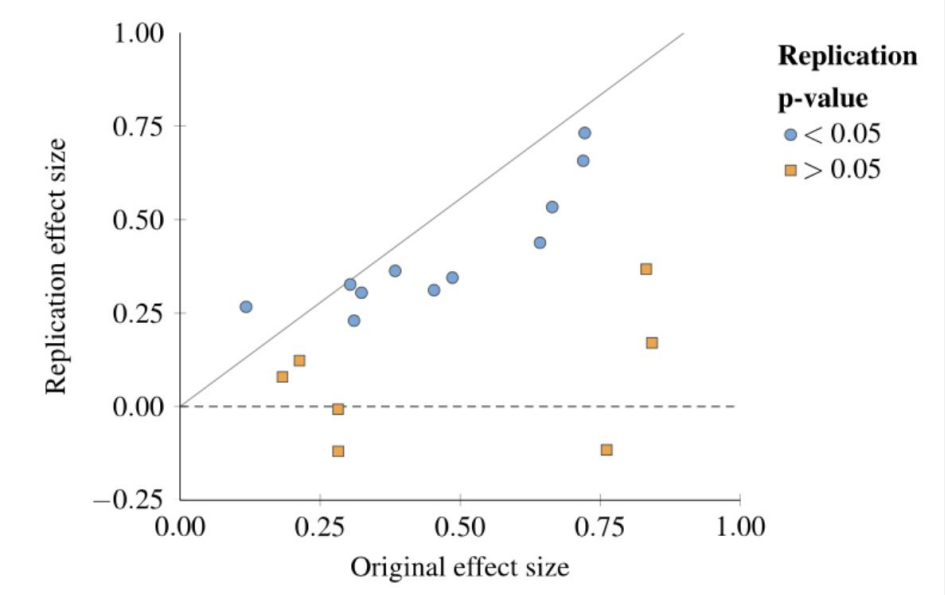
CANCER BIOLOGY



Errington et al. 2021. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. *eLife*

ECONOMICS

- Camerer et al (2016) analyze 18 studies in economics:
- They find that 61% of studies detect the original effect size in the same direction at alpha = 0.05
- However, the replicated effect size is 66% of the original, on average



Camerer et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science*

CONDUCTING REPRODUCIBLE RESEARCH

CONDUCTING REPRODUCIBLE RESEARCH

1. Hypothesis testing
2. Randomness
3. Statistical testing
4. Open-source code
5. Model cards
6. Datasheets

HYPOTHESIS TESTING

- In ML/NLP, we often get started with running experiments right away due to the low barrier to entry, which can result in:
  - Unclear research questions
  - Unclear conclusions
  - Wasted time, effort and computation power
- Formulating (some version of) the RQs before starting with experimentation can help alleviate some of these issues

RANDOMNESS

Deep Neural Networks display highly non-convex loss surfaces and therefore the performance of a model depends on several factors:

- Specific hyperparameters
- Dropout applied during training
- Weight initialization
- Order of the training data
- Randomly sampled data augmentations

It is important identify all sources of potential randomness in order to try to compensate for them in your experiments

STATISTICAL TESTING

- Comparing the means of two models is not enough to conclude model A is better than B
- It is important to choose the appropriate statistical test to determine whether or not your results are significant. Some resources:
  - Ulmer et al. 2022. Deep-Significance: Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks.
  - Dror et al. 2019. Deep Dominance: How to Properly Compare Deep Neural Models.

STATISTICAL TESTING

- **Scenario 1:** Comparing multiple runs of two models
  - Scores from a model **A** and a baseline **B** on a dataset, stemming from N model runs with different random seeds
  - Comparing multiple runs will *always* be preferable
- **Scenario 2:** Comparing multiple runs across datasets
  - When comparing models across datasets, formulate one null hypothesis per dataset
  - N model runs with different random seeds

Ulmer et al. 2022. Deep-Significance: Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks.

STATISTICAL TESTING

- **Scenario 3:** Comparing sample-level scores
  - If only one run is available, comparing sample-wise score distribution can be an option
- **Scenario 4**: Comparing more than two models
  - For instance, for three models, we can create a matrix 3x3
- The framework by Ulmer et al. 2022 makes use of the Almost Stochastic Order (ASO) test
  - Expresses the amount of violation of stochastic order

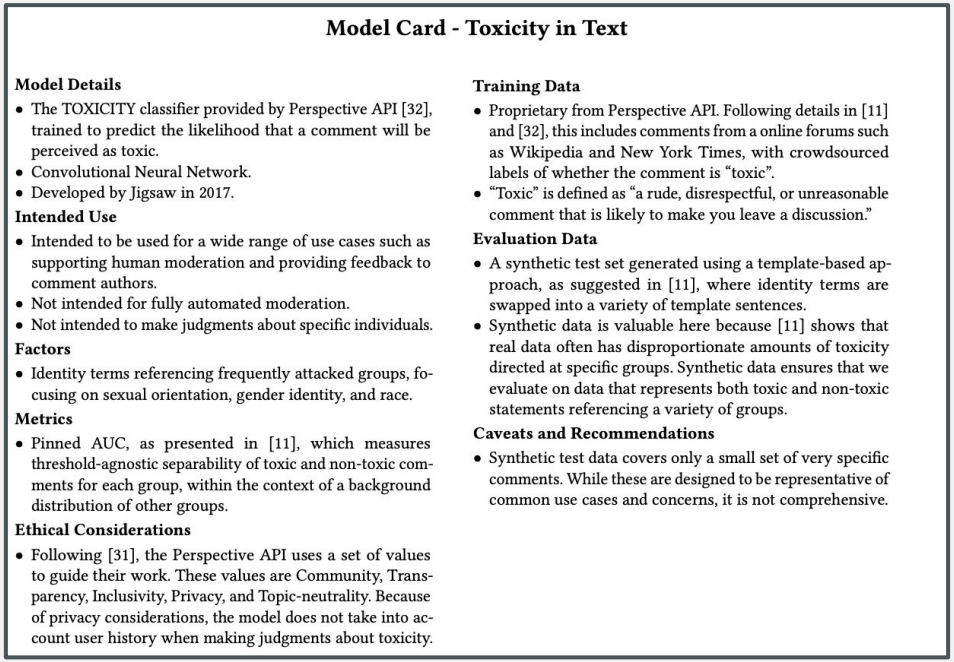
Ulmer et al. 2022. Deep-Significance: Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks.

OPEN-SOURCE CODE

- When possible, it is beneficial to open source your code and data in order to promote open and reproducible science
- Templates such as the ML Code Completeness Checklist can help you arrange your repository before publishing it publicly
  - More details in Part 3 of the tutorial
- Open-source code provides insights into:
  - The underlying implementation of a formal idea
  - Many hyperparameters and minor details that are not discussed in the paper

https://github.com/paperswithcode/releasing-research-code

MODEL CARDS



Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Mitchell et al. 2019. Model Cards for Model Reporting. *ACM FAccT*.

DATASHEETS

Datasheets were proposed as a mechanism to standardize documentation practices for ML datasets. They include ~50 questions on the following topics:

- Motivation
- Composition
- Collection Process
- Preprocessing/cleaning/labelling
- Uses
- Distribution
- Maintenance

Gebru et al. 2020. Datasheets for Datasets. *ACM FaccT*

DATASHEETS

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on Bo Pang's webpage at Cornell: http://www.cs.cornell.edu/people/pabo/movie-review-data. The dataset does not have a DOI and there is no redundant archive.

When will the dataset be distributed?

The dataset was first released in 2002.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: *Thumbs up? Sentiment classification using machine learning techniques.* Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, 2002.

Gebru et al. 2020. Datasheets for Datasets. *ACM FaccT*

RECOMMENDATIONS FOR CONDUCTING REPRODUCIBLE RESEARCH

1. Formulate hypothesis prior to starting experiments
2. Identify appropriate statistical tests
3. Identify stochastic components of experiments and account for randomness
4. Open-source your code with clear instructions on how to run it
5. Clearly document your contribution with a model card and/or a datasheet

QUESTIONS?

MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

In this tutorial, we focus on the challenge of ensuring research results are reproducible

TUTORIAL OVERVIEW

1. Introduction to Reproducibility
2. Reproducibility in NLP
3. Mechanisms for Reproducibility
4. Reproducibility as a Teaching Tool