

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...](https://twitter.com/Tim_Dettmers/status/1248811350000000000)

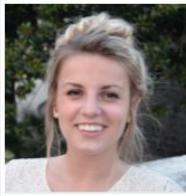
[Show this thread](#)

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](#)



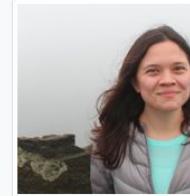
University of Amsterdam



[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

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

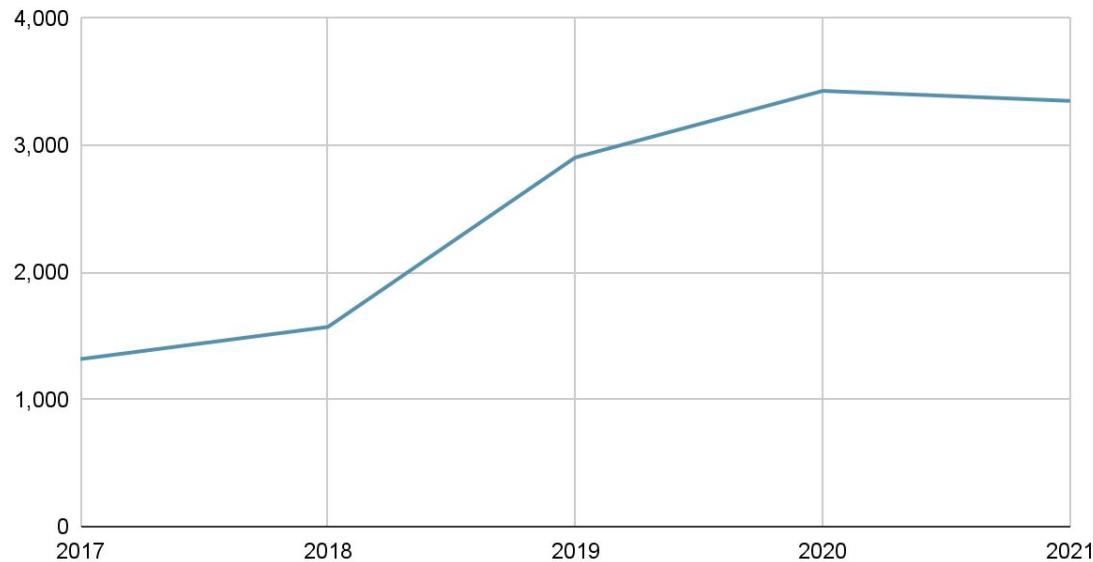
"At the very foundation of scientific inquiry is the process of specifying a hypothesis, running an experiment, analyzing the results and drawing conclusions"

"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



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?

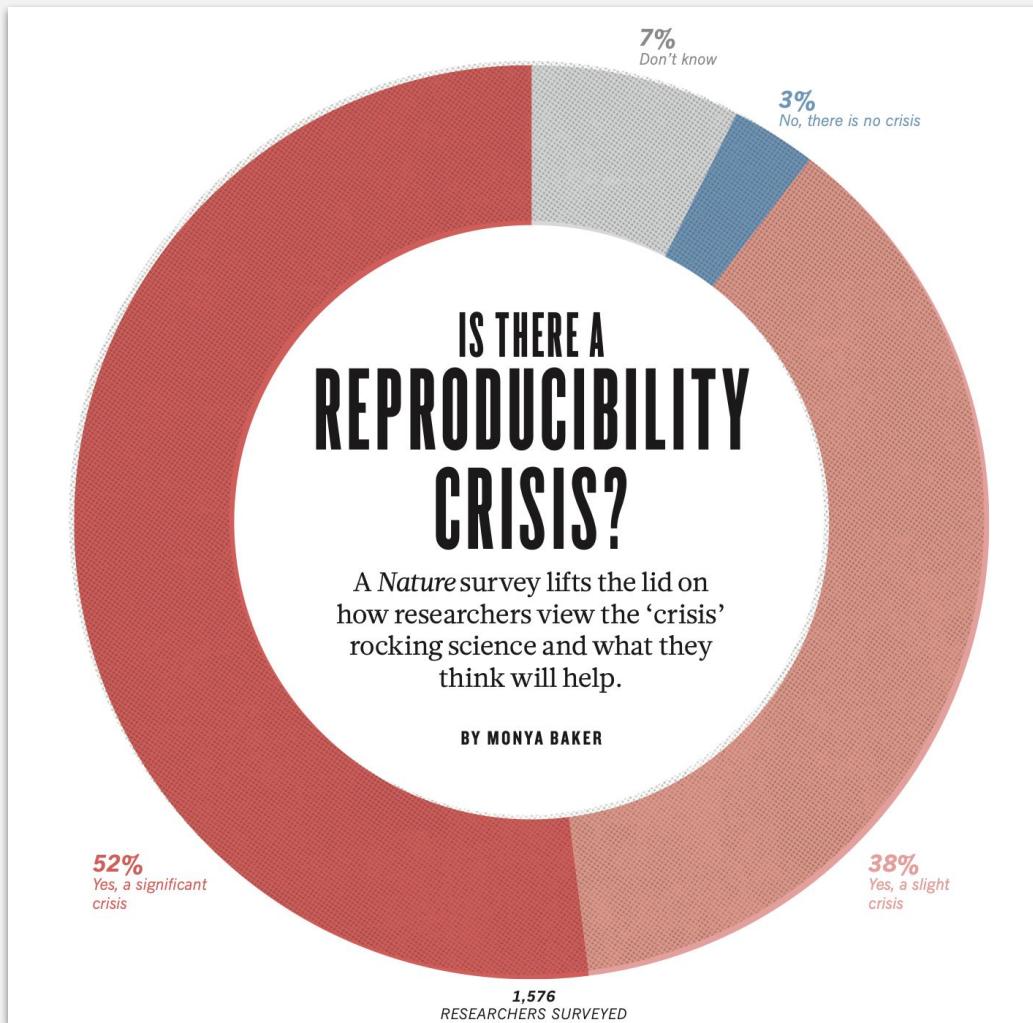
MOTIVATION

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

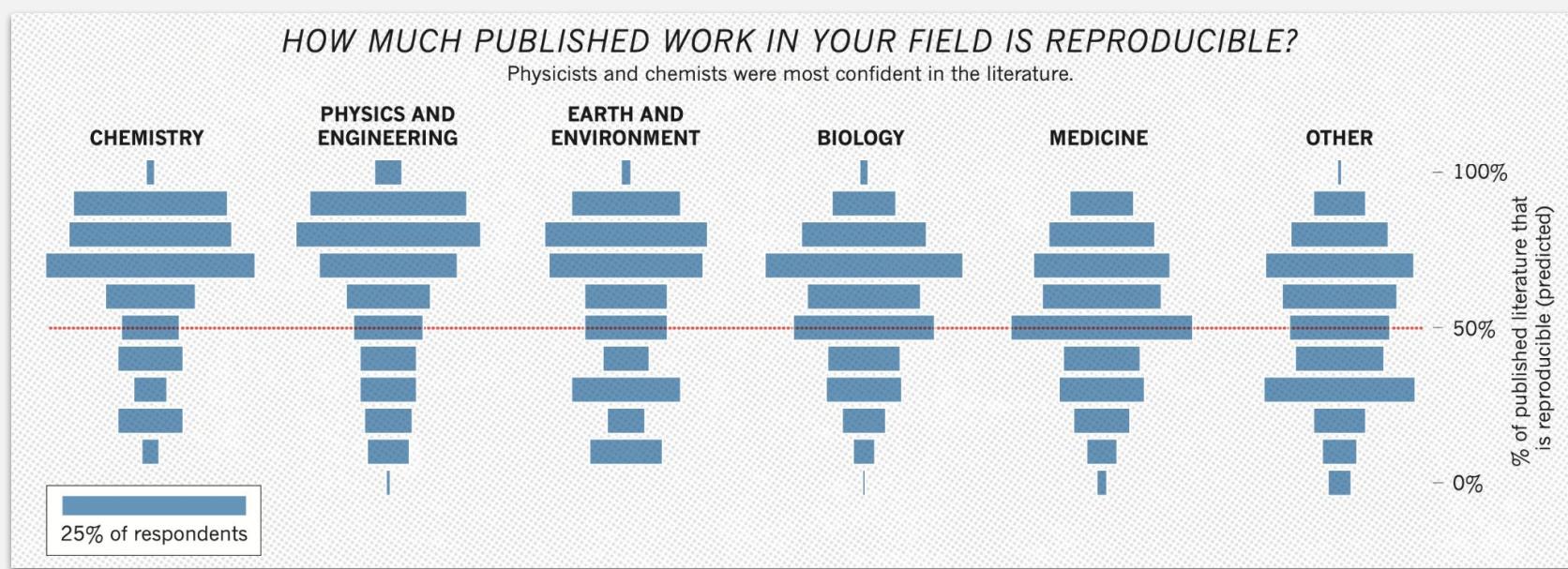
TUTORIAL OVERVIEW

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



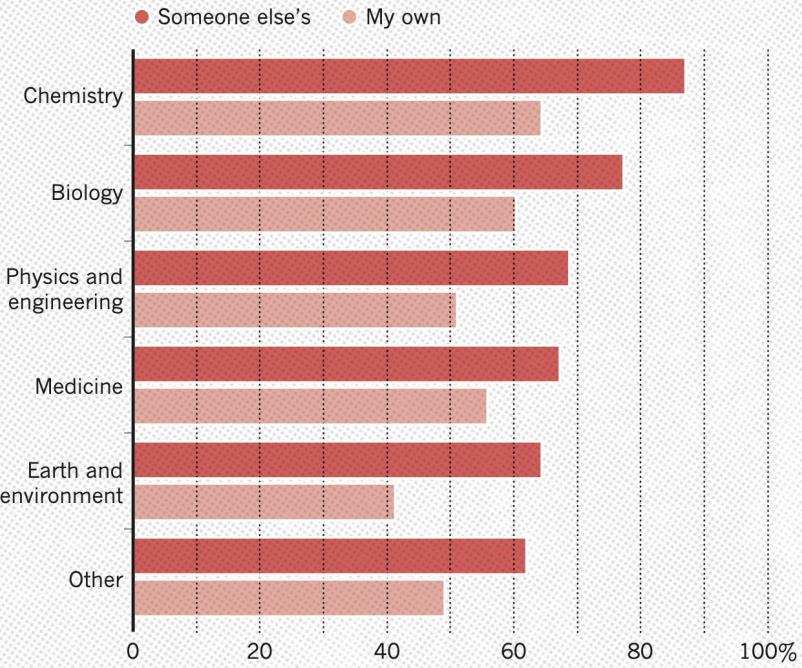
HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.



HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

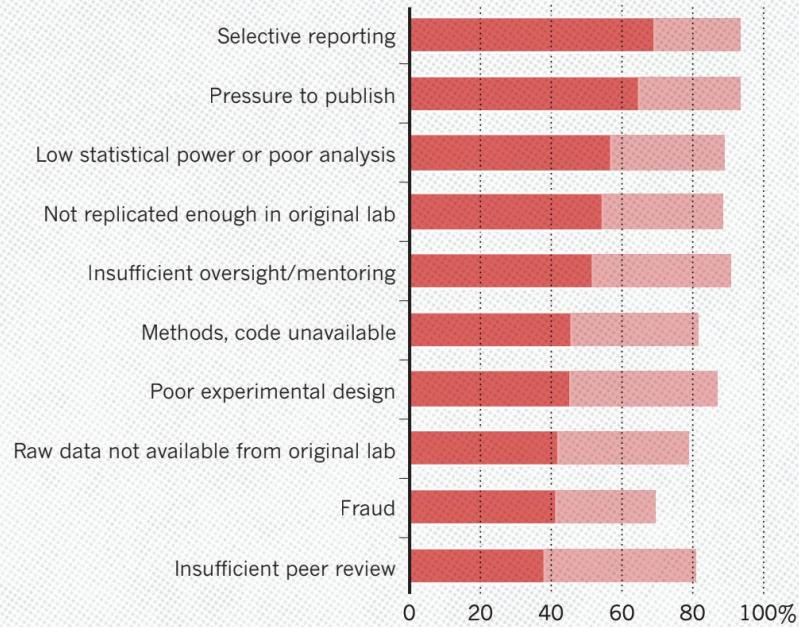
Most scientists have experienced failure to reproduce results.



WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

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

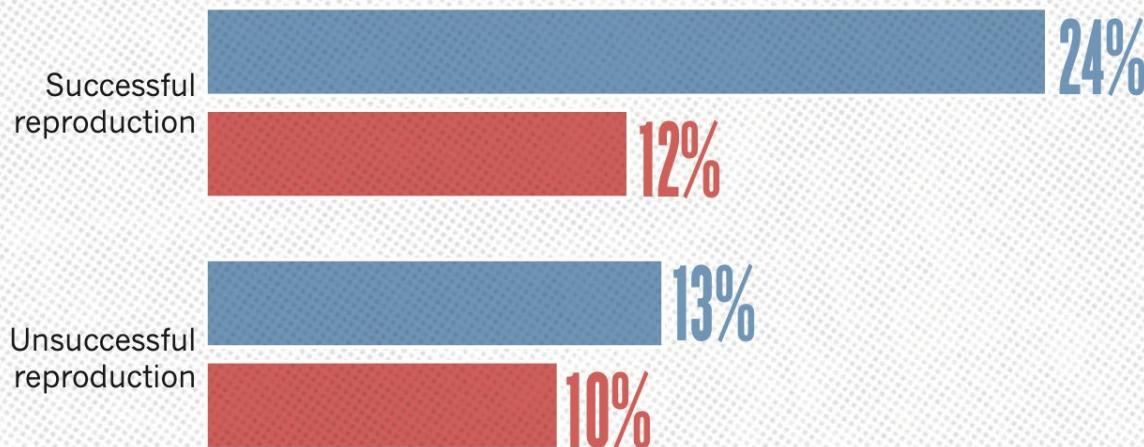
- Always/often contribute
- Sometimes contribute



HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

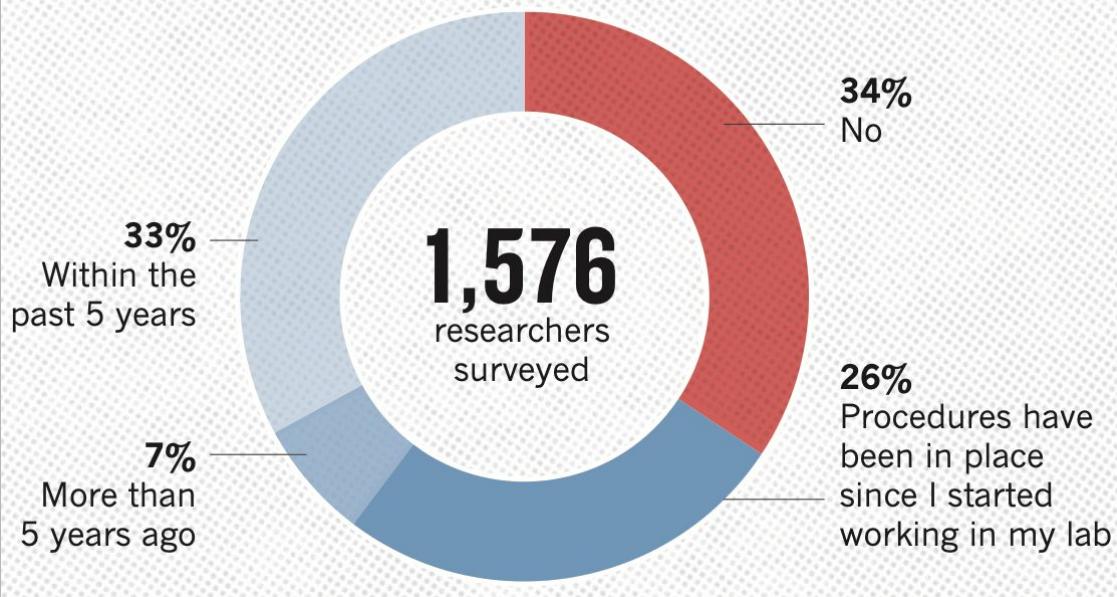
Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

- Published
- Failed to publish



HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



REPRODUCIBILITY IN ML

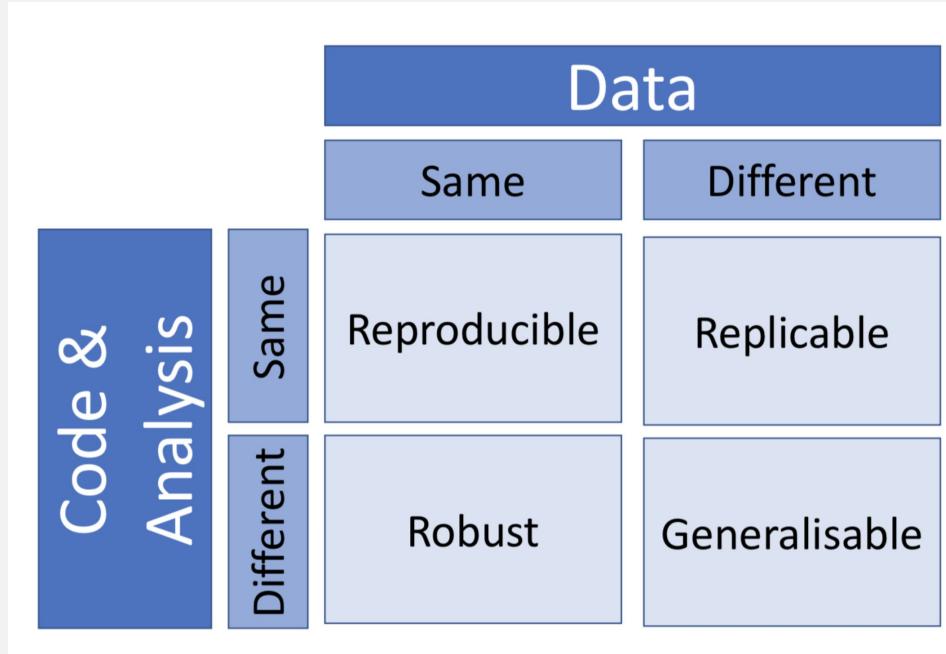
ACM DEFINITIONS

- **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”)

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.

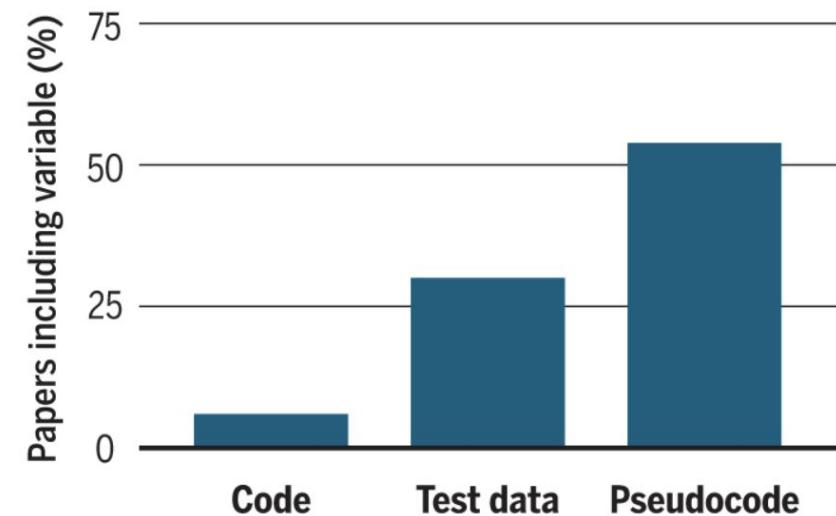
NEURIPS DEFINITIONS



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

REPRODUCIBILITY CRISIS IN ML

Code and Data Associated with this Article



arXiv Links to Code & Data ([What is Links to Code & Data?](#))

Official Code

No official code found; [you can submit it here](#)

Community Code



5 code implementations (in PyTorch and TensorFlow)

Datasets Used



[OpenAI Gym](#)

853 papers also use this dataset



[MuJoCo](#)

831 papers also use this dataset

- 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

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%

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

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.

BIOMEDICAL SCIENCES

Recommendations proposed by Begley and Lee (2012):

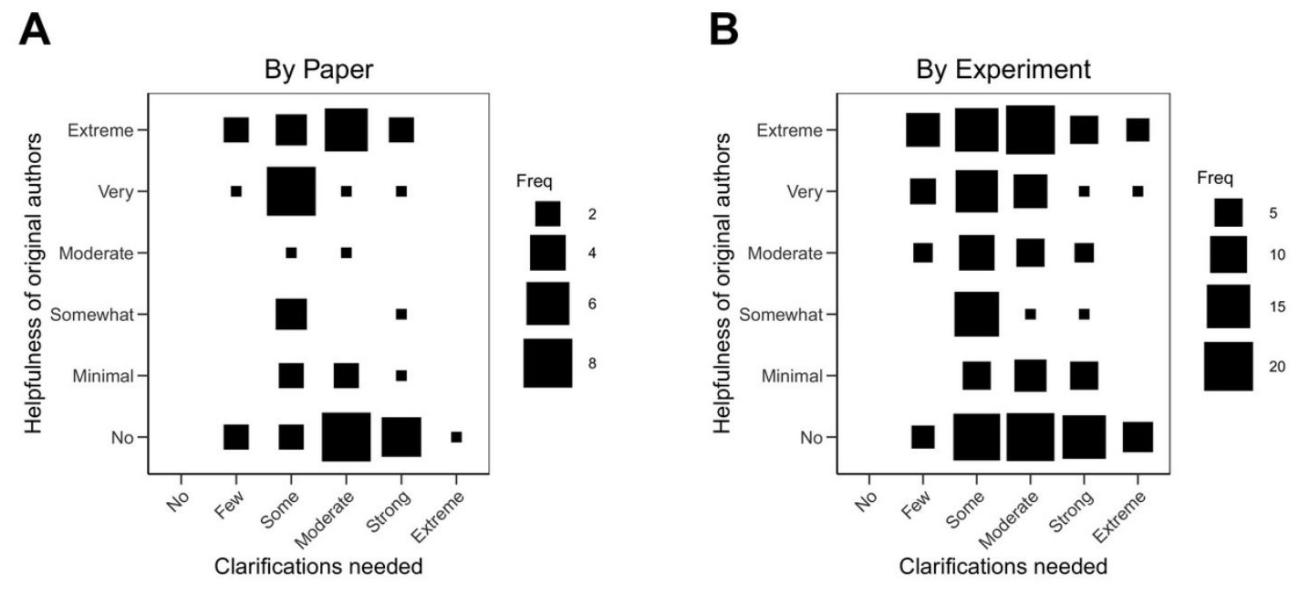
- Require reporting on negative findings
- Encourage reporting on alternative findings that contradict existing work
- Implement transparent mechanisms for reporting unethical practices
- Increase dialogue between physicians, scientists, patient advocates and patients
- Recognize high-quality teaching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

CANCER BIOLOGY

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

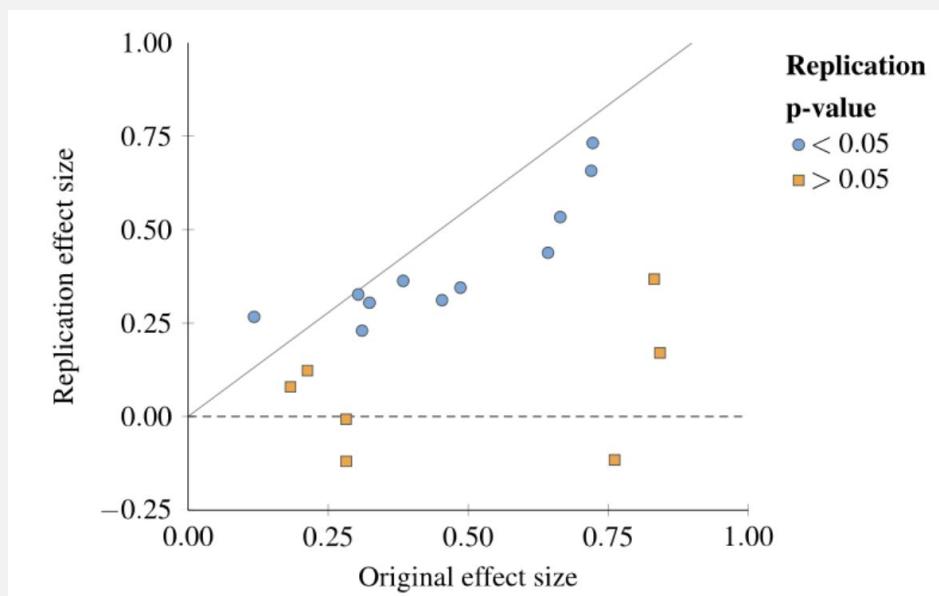
- Only 50/193 experiments from 23 papers were reproduced
- Data was publicly accessible for 4 of 193 papers
- Authors would not share data for 68% of papers
- 32% of authors were rated as "not at all helpful" by researchers reproducing their experiments
- 67% of protocols described in papers needed modifications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

CANCER BIOLOGY



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



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

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 3×3
- The framework by Ulmer et al. 2022 makes use of the Almost Stochastic Order (ASO) test
 - Expresses the amount of violation of stochastic order

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

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 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.

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

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.

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

REPRODUCIBILITY IN NLP

Jesse Dodge, Sasha Luccioni, Jessica Zosa Forde

OVERVIEW

1. The NLP Reproducibility Checklist
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

OVERVIEW

1. **The NLP Reproducibility Checklist**
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

REPRODUCIBLE SCIENCE IS HARD

ML and NLP are driven by experiments.



The most important idea: reporting!



REPRODUCIBLE SCIENCE IS HARD

AI is wrestling with a replication crisis

Artificial intelligence / Machine learning

by Will Douglas Heaven
November 12, 2020

The Importance of Reproducibility in Machine Learning Applications

[Home](#) | The Importance of Reproducibility in Machine Learning Applications

GREGORY BARBER BUSINESS 09.16.2019 07:00 AM

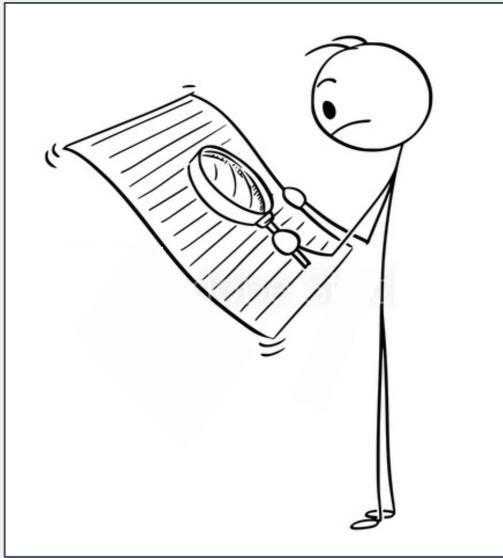
Artificial Intelligence Confronts a 'Reproducibility' Crisis

Machine-learning systems are black boxes even to the researchers that build them. That makes it hard for others to assess the results.

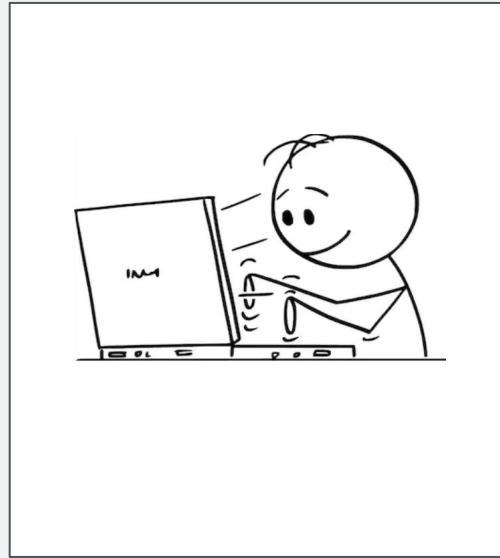




Get an idea



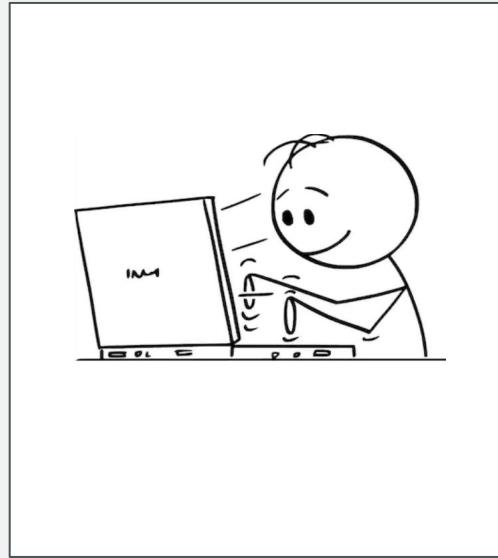
Get an idea



Spend time working



Get an idea



Spend time working

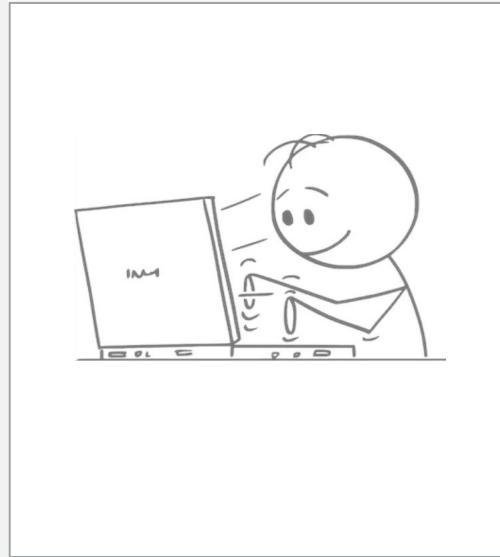


Negative result wasn't reported

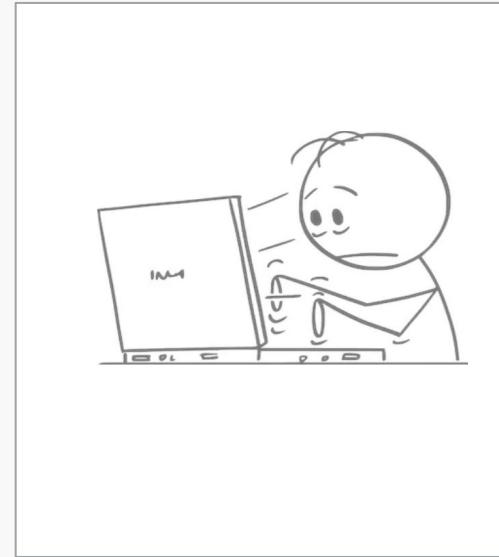
Spurious correlation



Get an idea



Spend time working



Negative result wasn't reported

Spurious correlation

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!

NLP Reproducibility Checklist

EMNLP 2020

NAACL 2021

ACL 2021

EMNLP 2021

Required with submission

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!

NLP Reproducibility Checklist

EMNLP 2020

NAACL 2021

ACL 2021

EMNLP 2021

Required with submission

More than 10,000 submissions
filled it out!

Goal: Remind authors of what
they know they should report

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!

Example items:

For all reported experimental results:

- A description of computing infrastructure used
- The total computational budget used (e.g. GPU hours), average runtime for each model or algorithm, or estimated energy cost

For all results involving multiple experiments, such as hyperparameter search:

- The exact number of training and evaluation runs
- Summary statistics of the results (e.g. expected validation performance, mean, variance, error bars, etc.)

For all datasets used:

- Relevant statistics such as number of examples and label distributions
- Details of train/validation/test splits

NLP REPRODUCIBILITY CHECKLISTS RESULTS - FIRST LOOK

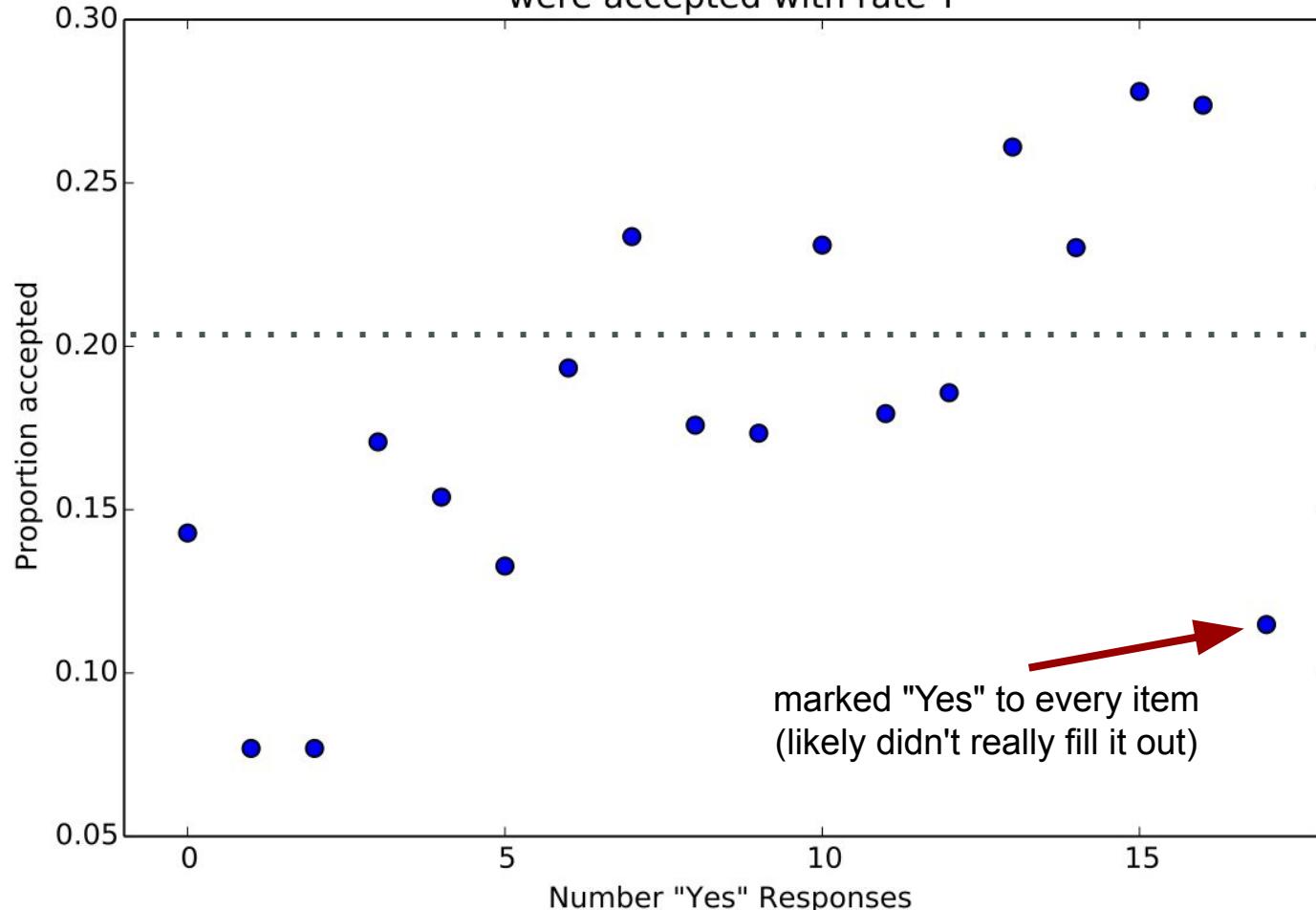
EMNLP 2020

Total submissions: 3,677

Total accepted: 752 (20.4%)

First conference!
Likely different now

Papers that responded "Yes" to X items
were accepted with rate Y



READERS LIKE RELEVANT INFO

Items correlated with acceptance:

- "Average runtime for each approach"
- "Description of computing infrastructure used"

Room for improvement

- "Included all preprocessing steps" -- 11% marked "No"
- "Included a link to download the data" -- only 64% marked "Yes"

Data matters!

OVERVIEW

1. **The NLP Reproducibility Checklist**
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

OVERVIEW

1. The NLP Reproducibility Checklist
2. **The Responsible NLP Checklist**
3. NLP Research on Reproducibility

RESPONSIBLE NLP CHECKLIST

Required with submission to ARR

Combines Reproducibility + Ethics

Collaboration between
NAACL PCs, ARR Editors,
Anna Rogers, Margot Mieskes

Framed in terms of transparency:
“Did you report [information]?”

Goal: Remind authors of what they
know they should report

RESPONSIBLE NLP CHECKLIST

Required with submission to ARR

Combines Reproducibility + Ethics

Collaboration between
NAACL PCs, ARR Editors,
Anna Rogers, Margot Mieskes

Framed in terms of transparency:
“Did you report [information]?”

Goal: Remind authors of what they
know they should report

Marking “No” or “N/A” is not grounds for rejection!

RESPONSIBLE NLP CHECKLIST I

- For every submission:
 - Describe limitations?
 - Describe risks?
 - Abstract and intro summarize main claims?

RESPONSIBLE NLP CHECKLIST 2

- Did you use or create scientific artifacts?
 - Cite creators?
 - Discuss the license or terms?
 - State intended use? Use consistently with creators intended use?
 - Personal information in new data?
 - Documentation of data?
 - Details of train / test / dev?

RESPONSIBLE NLP CHECKLIST 3

- Did you run computational experiments?
 - Number of parameters, total budget (e.g., GPU hours), computing infrastructure?
 - Hyperparameter search?
 - Error bars around results?
 - Details about software packages

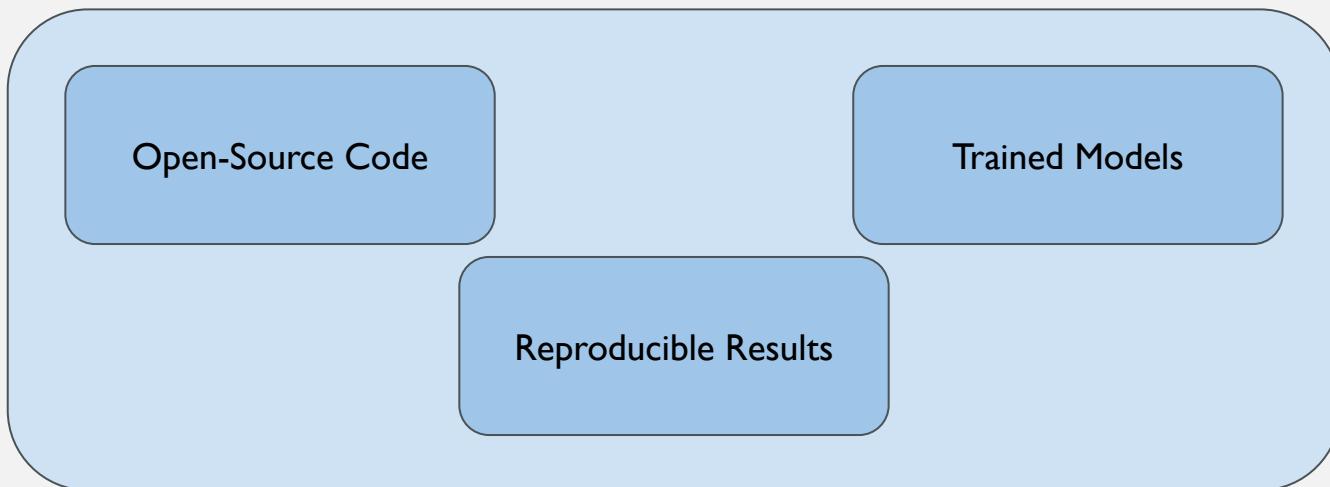
RESPONSIBLE NLP CHECKLIST 4

- Did you use human annotators (e.g., crowdworkers) or research with human participants?
 - Report the full text of instructions?
 - Report information about how you recruited, is payment adequate?
 - Did you get consent for intended use?
 - Approved by IRB?
 - Report the demographic info of annotators?

NLP CHECKLIST CONCLUSIONS

1. Report all the info you can!
2. The checklists are forward looking, cover best practices
3. You can mark “No” or “N/A” (with a good reason)

REPRODUCIBILITY CHAIR AT NAACL 2022



REPRODUCIBILITY CHAIR AT NAACL 2022

Open-Source Code

Trained Models

Reproducible Results

We promote your work!
Authors benefit!

OVERVIEW

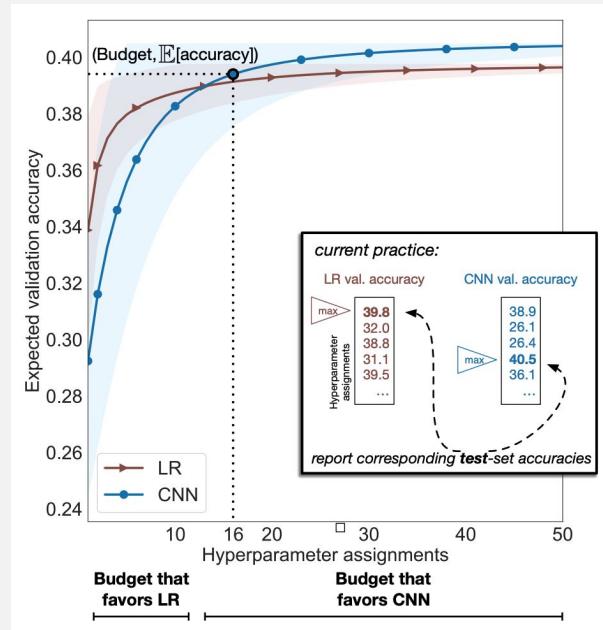
1. The NLP Reproducibility Checklist
2. **The Responsible NLP Checklist**
3. NLP Research on Reproducibility

OVERVIEW

1. The NLP Reproducibility Checklist
2. The Responsible NLP Checklist
3. **NLP Research on Reproducibility**

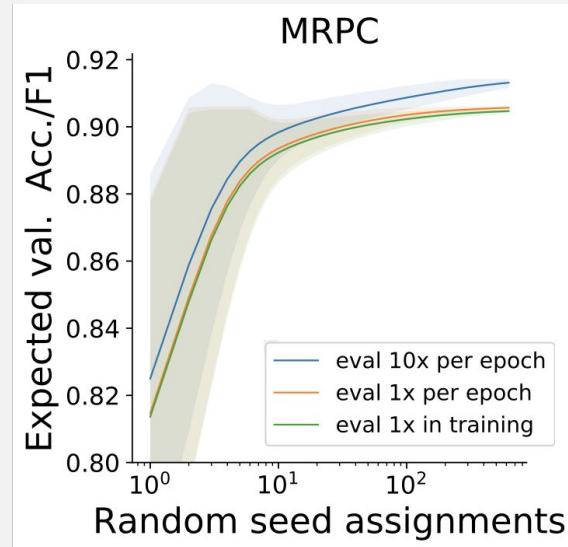
REPORTING OF RESULTS (DODGE ET AL., 2019)

- Different budgets for hparam search lead to different conclusions about which model performs best
- Solution: report expected valid. perf.



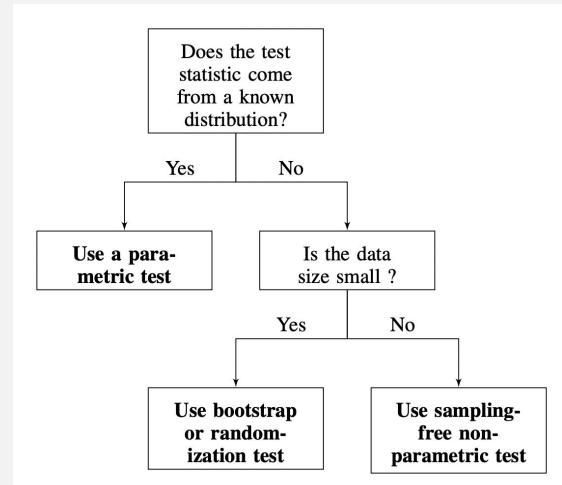
RANDOM SEEDS (DODGE ET AL., 2020)

- Setup: fine-tuning BERT on GLUE tasks (MRPC, SST, CoLA, RTE)
- Conclusion: Surprisingly large variance from random seed!
- Suggestion: Start many runs, stop some early, report error bars



STATISTICAL TESTING (DROR ET AL. 2018)

1. The authors provide a survey of which metrics are used of evaluation, how are metrics reported, and which statistical tests are used in NLP
2. The authors additionally provide a review of statistical tests that are relevant to NLP researchers and a flow chart to help them select a statistical test
3. The authors note that controlling for multiple hypothesis tests (Bonferroni correction) is also an important consideration when conducting statistical tests (Dror et al., 2017)



STATISTICAL POWER (CARD ET AL., 2020)

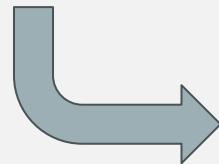
1. Statistical power measures how likely we will correctly reject the null hypothesis of a statistical test.
2. Card et. al analyze the statistical power of experiments in NLP
3. Many experiments lack statistical tests and sufficient statistical power.
4. Power analyses should be included as part of experimental planning.
 - a. Experiments that cannot be conducted with sufficient statistical power may not lead to clear conclusions and should be carefully considered.
5. Code and model release, significance testing, and appropriate sample size can improve the quality of statistical analysis in the field

STANDARD SPLITS (GORMAN & BEDRICK, 2019)

1. Gorman & Bedrick compare utilizing the “standard split” of a provided dataset, versus randomly selecting the train, validation, and testing split for POS tagging task.
2. They utilize statistical testing as recommended in Dror et al. to correct for multiple hypotheses.
3. Many methods for POS are overfitted to the standard split and do not perform as well on a randomly generated split.
4. The authors recommend Bonferroni corrected random split hypothesis testing to confirm that results on the standard split are robust to random split

MACHINE TRANSLATION (MARIE ET AL., 2020)

- A large-scale meta-evaluation of Machine Translation (MT), manually annotating 769 research papers published from 2010-2020.
- It found several issues:
 - the exclusive use of BLEU, a metric with significant limitations
 - the absence of statistical significance testing
 - the comparison of incomparable results from previous work
 - comparing MT systems that do not exploit the same data
- Depending on the metric being used, different systems can be considered as superior.



BLEU	Chinese-to-English (Zh→En)		System
	System	chrF	
36.9	WeChat_AI	0.653	Volctrans
36.8	Tencent_Translation	0.648♦	Tencent_Translation
36.6	DiDi_NLP	0.645♦	DiDi_NLP
36.6	Volctrans	0.644♦	DeepMind
35.9♦	THUNLP	0.643♦	THUNLP

MACHINE TRANSLATION (MARIE ET AL., 2020)

Data differences also impact scores:

- Tokenizer used
- Dataset preprocessing (e.g. max length or language ID used for filtering)

The authors propose guidelines for automatic MT evaluation, including:

1. Metrics other than BLEU
2. Statistical significance testing
3. Reproducing previous scores instead of copying them
4. Ensuring that the data, splits, and preprocessing used are the same

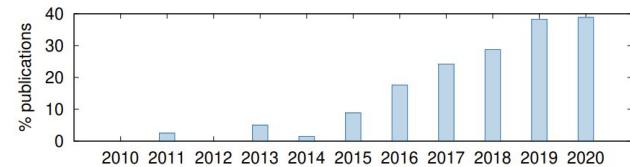


Figure 4: Percentage of papers that compared MT systems using data that are not identical.

TRANSFORMERS (NARANG ET AL., 2021)

- An extensive evaluation of different Transformer modifications in a shared experimental setting in NLP:
 - Activations, normalization, depth, embeddings, parameter sharing, softmax, applied to different Transformer architectures
- They find that many Transformer modifications **do not** result in improved performance, and suggest that changes to Transformers suffer from lack of generalization across different implementations and tasks.
- The authors also found that **hyperparameter tuning** was a major challenge for Transformers given the space of possible combinations

TRANSFORMERS (NARANG ET AL., 2021)

Some proposals made to ensure the robustness of improvements include:

- Trying changes out in multiple codebases
- Applying them to a wide variety of downstream applications, including domains outside of NLP
- Keeping hyperparameters fixed as much as possible, and/or measuring the robustness of the modifications to changes in hyperparameters
- Reporting of results should include mean and standard deviation across multiple trials

REPRODUCIBILITY IN LARGE LANGUAGE MODELS

1. Models such as Transformer XL, Megatron, GPT-Neo, OPT, T0 share code on GitHub
2. Big Science and OPT share model logs
3. Open datasets such as OpenWebText and the Pile aid in pretraining
4. HuggingFace provides a model zoo of pre-trained weights (many shared by the original authors)
5. Checkpoints and replicates such as MultiBert enable researchers to study training dynamics and variability
6. Tools such as evaluationharness, promptsource, codecarbon provide useful evaluation

Ensuring that our research is reproducible remains
an important goal within NLP research

But it is not the only consideration

LIMITATIONS IN REPRODUCIBLE NLP

1. Environmental Impact
2. Depreciation of hardware/software
3. Ethical Challenges

ENVIRONMENTAL IMPACT

1. Reproducing papers from scratch creates additional environmental cost
2. Sharing models and hyperparameters makes it possible to avoid these costs
3. Clearly communicating energy requirements and carbon emissions also makes it possible to take these into account when choosing between different models
4. Tools such as codecarbon, Azure has an upcoming tool¹. Allow for calculations of carbon emissions

DEPRECATION

1. Operating under assumption that we're using the same hardware and software as the original paper, which becomes less likely as time goes on [Mesnard & Barba, 2017]
2. Researchers are not incentivized to maintain their code
3. Dataset deprecation:
 - Versions: e.g. Common Crawl, Wikipedia get updated regularly
 - Datasets removed by creators: TinyImages, Duke MTMC, etc. – but continue being used
 - No centralized identification schema for datasets (e.g. DOI)
 - Current endeavors by conferences like NeurIPS are aiming to create a centralized repository for deprecated datasets

ETHICAL CHALLENGES

1. Reproduction of NLP papers does not happen in a vacuum - considerations when conducting reproduction studies should also take into account the ethical considerations particular to a given methodology
2. The ACL Code of Ethics and ACL Rolling Review Responsible NLP Research checklist provide a useful starting point to help researchers conduct and share their work responsibly
3. Misunderstanding a paper can lead a researcher to make incorrect assumptions when reproducing the paper

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

MECHANISMS FOR REPRODUCIBILITY

Koustuv Sinha, Robert Stojnic, Jessica Zosa Forde

OVERVIEW

1. Papers with Code
2. Reproducibility Challenge
3. Reproducibility Checklists
4. Useful Tools and libraries

PAPERS WITH CODE



- **Goal:** Track all artefacts in ML, create positive incentives for sharing

The screenshot shows the Papers with Code homepage. At the top, there is a navigation bar with a search bar, a 'We are hiring!' message, and social media links for Twitter and GitHub. Below the navigation, there are four filter buttons: 'Top', 'Social', 'New', and 'Greatest'. A 'Subscribe' button is located on the right side of the header. The main content area is titled 'Trending Research' and features a card for a project called 'MVSTER: Epipolar Transformer for Efficient Multi-View Stereo'. The card includes a small diagram of the transformer architecture, the author (jeffwang987), the framework (PyTorch), the date (15 Apr 2022), a star rating of 52, and a rate of 1.29 stars/hour. It also has two buttons: 'Paper' and 'Code'.

PAPERS WITH CODE



- Largest database of papers curated with their code

Code

[Edit](#)

carolineec/EverybodyDanceNow <small>official</small>	★ 508	PyTorch
Lotayou/everybody_dance_now_pytorch	★ 256	PyTorch
VisiumCH/AMLD2020-Dirty-Gancing <small>Quickstart in Colab</small>	★ 17	PyTorch
wjq5446/pytorch-everybody-dance-now	★ 9	PyTorch
Novemser/deep-imitation	★ 9	PyTorch

[See all 14 implementations](#)

PAPERS WITH CODE



- Largest database of datasets, tracking their usage

ImageNet

Edit

Introduced by Jia Deng et al. in [ImageNet: A large-scale hierarchical image database](#)

The **ImageNet** dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection. The publicly released dataset contains a set of manually annotated training images. A set of test images is also released, with the manual annotations withheld. ILSVRC annotations fall into one of two categories: (1) image-level annotation of a binary label for the presence or absence of an object class in the image, e.g., “there are cars in this image” but “there are no tigers,” and (2) object-level annotation of a tight bounding box and class label around an object instance in the image, e.g., “there is a screwdriver centered at position (20,25) with width of 50 pixels and height of 30 pixels”. The ImageNet project does not own the copyright of the images, therefore only thumbnails and URLs of images are provided.

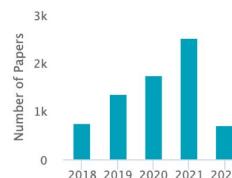
- Total number of non-empty WordNet synsets: 21841
- Total number of images: 14197122
- Number of images with bounding box annotations: 1,034,908
- Number of synsets with SIFT features: 1000
- Number of images with SIFT features: 1.2 million

Source:  [ImageNet Large Scale Visual Recognition Challenge](#)



Source: <https://cs.stanford.edu/people/kar...>

Usage ▲



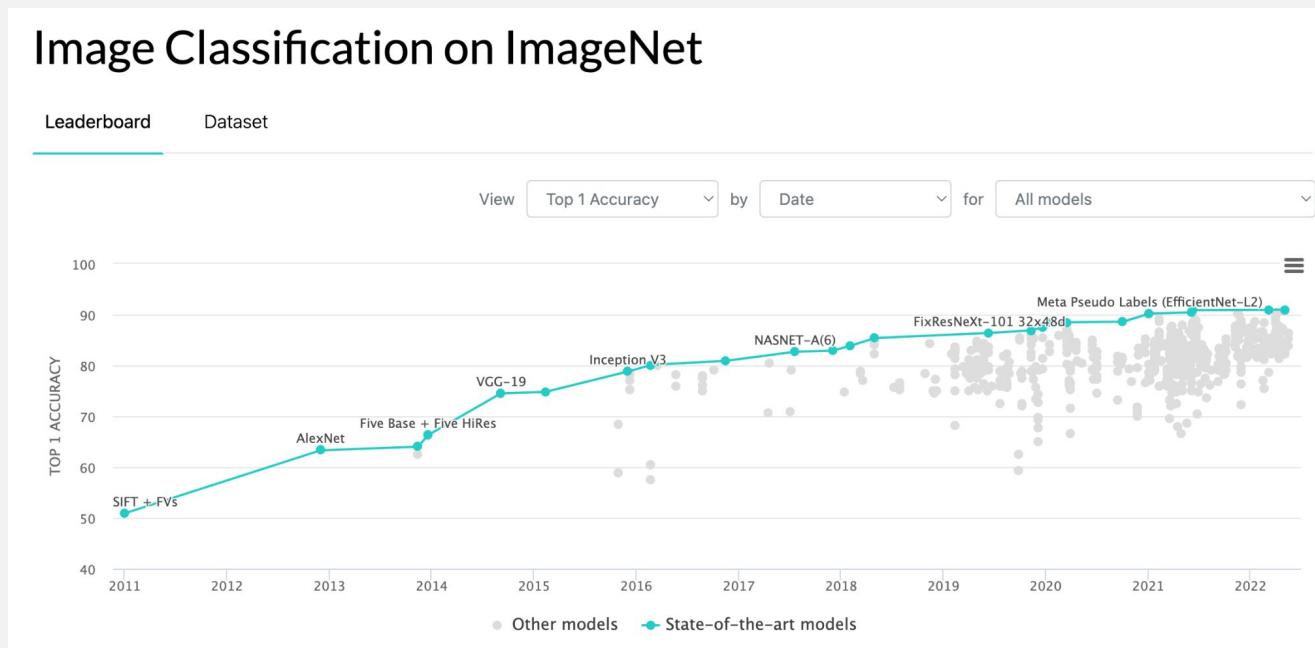
[Homepage](#)

PAPERS WITH CODE



- Largest database of results from published papers

Image Classification on ImageNet



PAPERS WITH CODE



Integrated with:

- arXiv
- ACL anthology
- OpenReview

Bibliographic Tools **Code & Data** Demos Related Papers About arXivLabs

Code and Data Associated with this Article

arXiv Links to Code & Data ([What is Links to Code & Data?](#))

Official Code

 <https://github.com/carolineec/EverybodyDanceNow>

Community Code

 [13 code implementations \(in PyTorch\)](#)

Datasets Used

 [Everybody Dance Now](#)
★ introduced in this paper
7 papers also use this dataset

PAPERS WITH CODE



- Reproducibility reports shown next to original papers

Deep Fair Clustering for Visual Learning

CVPR 2020 · Peizhao Li, Han Zhao, Hongfu Liu · [Edit social preview](#)

Fair clustering aims to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster. Existing work attempts to address this problem by reducing it to a classical balanced clustering with a constraint on the proportion of protected subgroups of the input space...

PDF

Abstract

Reproducibility Reports

Jan 31 2021

[Re] Deep Fair Clustering for Visual Learning

RC 2020 · Pauline Baanders, Chris Al Gerges, Nienke Reints, Tobias Teule

For the MNIST-USPS dataset, we report similar accuracy and NMI values that are within 1.2% and 0.5% of the values reported in the original paper. However, the balance and entropy differed significantly, where our results were within 73.1% and 30.3% of the original values respectively. For the Color Reverse MNIST dataset, we report similar values on accuracy, balance and entropy, which are within 5.3%, 2.6% and 0.2% respectively. Only the value of the NMI differed significantly, name within 12.9% of the original value In general, our results still support the main claim of the original paper, even though on some metrics the results differ significantly.

PAPERS WITH CODE



- Collated resources for publishing research code

Screenshot of the GitHub repository [paperswithcode / releasing-research-code](#) (Public)

Code Issues 2 Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file Code

rstojnic Update README.md a5b2c85 on Mar 19, 2021 120 commits

notebooks Fix graph 2 years ago

templates Update README.md 2 years ago

LICENSE Create LICENSE 2 years ago

README.md Update README.md 14 months ago

README.md

Tips for Publishing Research Code

NEURAL INFORMATION PROCESSING SYSTEMS

Collated best practices from most popular ML research repositories - now official guidelines at NeurIPS 2021!

About

Tips for releasing research code in Machine Learning (with official NeurIPS 2020 recommendations)

machine-learning awesome-list
neurips neurips-2020

Readme MIT License
1.9k stars 53 watching 572 forks

Releases

No releases published Create a new release

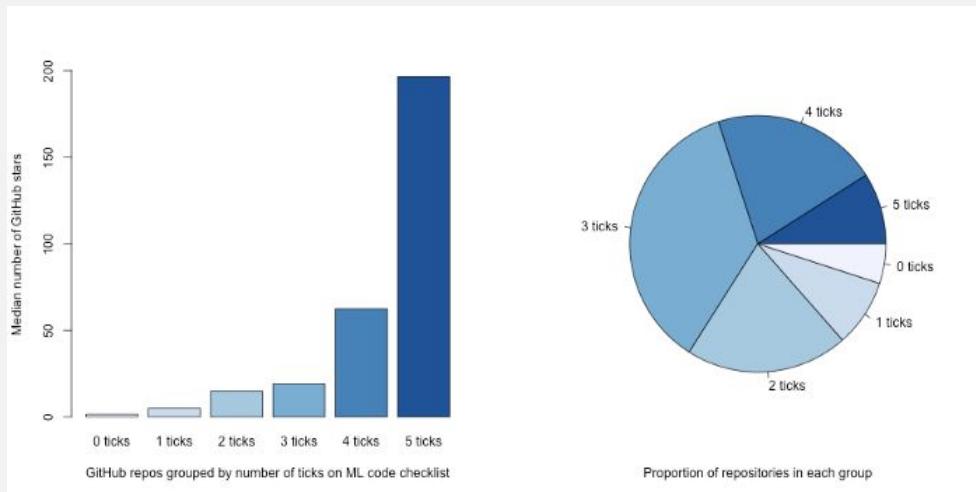
Packages

No packages published

PAPERS WITH CODE



- ML Code Completeness Checklist (Robert Stojnic, 2020)



1. **Dependencies** — does a repository have information on dependencies or instructions on how to set up the environment?
2. **Training scripts** — does a repository contain a way to train/fit the model(s) described in the paper?
3. **Evaluation scripts** — does a repository contain a script to calculate the performance of the trained model(s) or run experiments on models?
4. **Pretrained models** — does a repository provide free access to pretrained model weights?
5. **Results** — does a repository contain a table/plot of main results and a script to reproduce those results?

QUESTIONS?

REPRODUCIBILITY CHECKLISTS

- ML Reproducibility Checklist (Joelle Pineau, 2018)
- Minimal information that should be in a manuscript
- Not necessarily exhaustive
- Part of guidelines for major conferences (NeurIPS, ICLR, ICML)

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 steps.
- 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.

Reproduced from: www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf

REPRODUCIBILITY CHALLENGE

- Started 2018, till date five editions: ICLR 2018, ICLR 2019, NeurIPS 2019, RC 2020, RC 2021
- Task: Choose a submitted paper from a conference, reproduce the central claim of the paper

ML Reproducibility Challenge 2021 Edition

for papers published in:



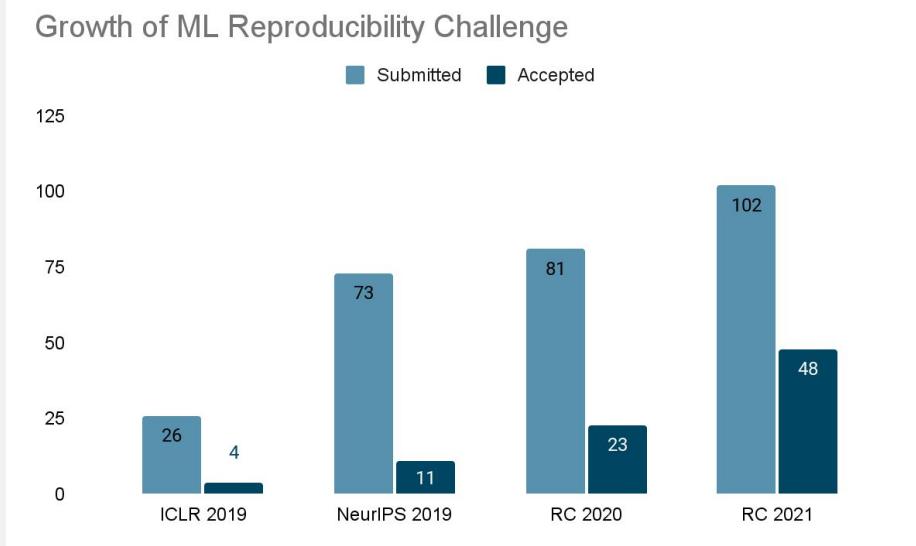
ICML | 2021
Thirty-eighth International Conference on
Machine Learning



EMNLP 2021



REPRODUCIBILITY CHALLENGE



REPRODUCIBILITY CHALLENGE

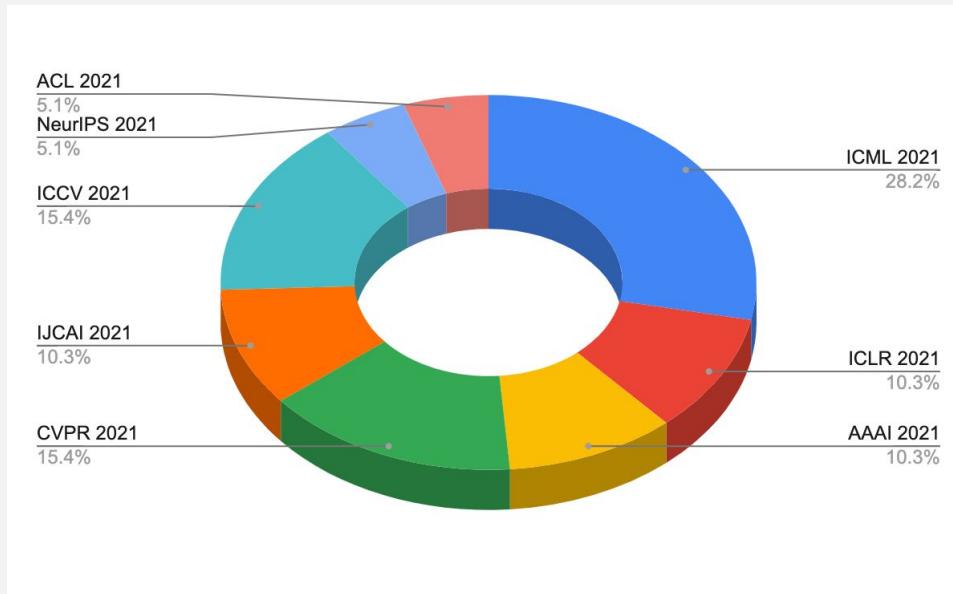
Best Paper Award

- ▶ Reproducibility Study of “Counterfactual Generative Networks”, *Piyush Bagad, Jesse Maas, Paul Hilders, Danilo de Goede*, [Forum](#), [Original Paper \(ICML 2021\)](#)

Outstanding Paper Awards

- ▶ [Re] Learning to count everything, *Matija Teršek, Domen Vreš, Maša Kljun*, [Forum](#), [Original Paper \(CVPR 2021\)](#)
- ▶ [RE] An Implementation of Fair Robust Learning , *Ian Hardy*, [Forum](#), [Original Paper \(ICML 2021\)](#)
- ▶ Strategic classification made practical: reproduction, *Guilly Kolkman, Maks Kulicki, Jan Athmer, Alex Labro*, [Forum](#), [Original Paper \(ICML 2021\)](#)
- ▶ On the reproducibility of "Exacerbating Algorithmic Bias through Fairness Attacks", *Andrea Lombardo, Matteo Tafuro, Tin Hadži Veljković, Lasse Becker-Czarnetzki*, [Forum](#), [Original Paper \(AAAI 2021\)](#)

REPRODUCIBILITY CHALLENGE



Reproducibility Reports accepted to MLRC 2021 by conference

REPRODUCIBILITY CHALLENGE

Volume 7 (2021)

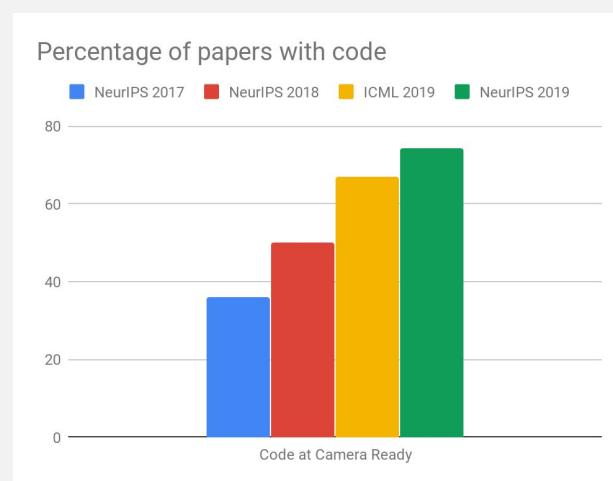
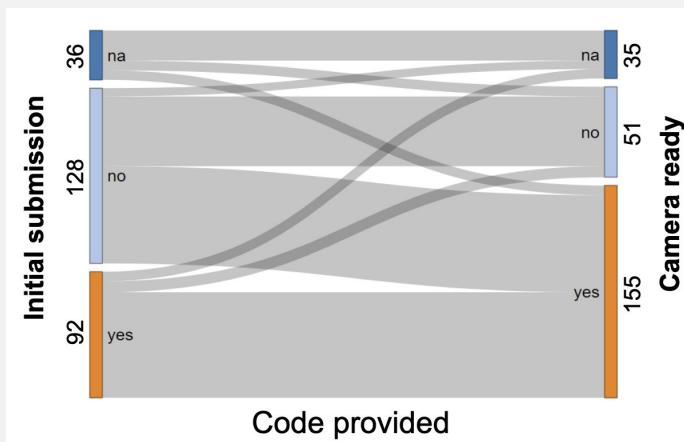
Issue 2 (ML Reproducibility Challenge 2020)

1. **Replication in ML Reproducibility Challenge 2020 (Python)** | 10.5281/zenodo.4835602 | [PDF](#) | [Code](#) | [Review](#) | [BibTeX](#)
VERMA, R., WAGEMANS, J.J.O., DAHAL, P., AND ELFINK, A. 2021. [Re] Explaining Groups of Points in Low-Dimensional Representations. *ReScience C* 7, 2, #24.
2. **Replication in ML Reproducibility Challenge 2020 (Python)** | 10.5281/zenodo.4833219 | [PDF](#) | [Code](#) | [Data](#) | [Review](#) | [BibTeX](#)
ALBANIS, G., ZIOLIS, N., CHATZITOFIS, A., DIMOU, A., ZARPALES, D., AND DARAS, P. 2021. [Re] On end-to-end 6DoF object pose estimation and robustness to object scale. *ReScience C* 7, 2, #2.
3. **Replication in ML Reproducibility Challenge 2020 (python)** | 10.5281/zenodo.4833389 | [PDF](#) | [Code](#) | [Review](#) | [BibTeX](#)
ARVIND, M. AND MAMA, M. 2021. [Re] Neural Networks Fail to Learn Periodic Functions and How to Fix It. *ReScience C* 7, 2, #3.

RESCIENCE C

IMPACT OF CHECKLISTS AND CHALLENGES

- Increase in the amount of code released during submission
- Increased interaction with authors and practitioners after paper publication through OpenReview



USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release

Link to our previous blog post: <https://bit.ly/3LoSuKC>

USEFUL TOOLS AND LIBRARIES

- **Config management**
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



Or even plain
YAML / JSON
files work!

Hydra: <https://hydra.cc>

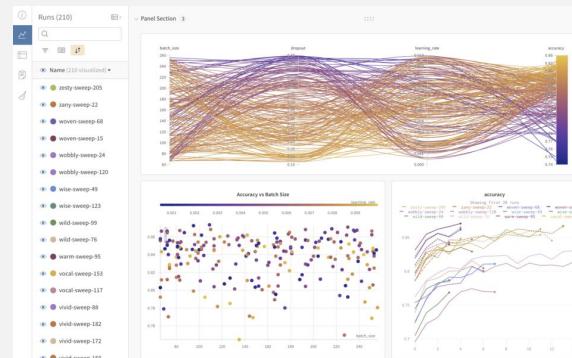
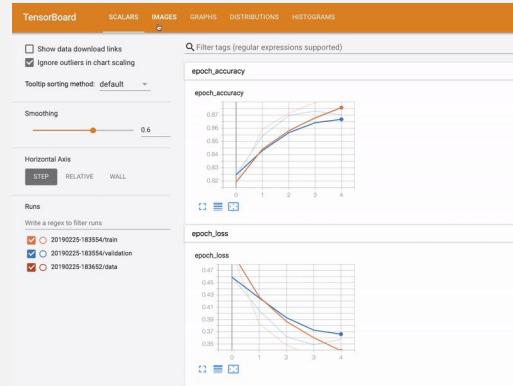
```
general:  
    batch_size: 128  
    data_name: fashionmnist  
    description: This is a sample config  
    device: cuda  
    epochs: 20  
    resume: false  
    logbook:  
        dir: /path/to/log  
        logger_file_path: log.jsonl  
        log_interval: 100  
        project_name: fancy_project  
    model:  
        class_order: 0,1,2,3,4,5,6,7,8,9  
        loss_policy: recon_bce # ce, recon_ce, recon_mse, bce, recon_bce  
        max_class: 10  
        reset_optim: False  
        optim:  
            eps: 1.0e-08  
            learning_rate: 0.001  
            name: Adam  
            scheduler_gamma: 0.999  
            scheduler_patience: 10  
            scheduler_type: plateau  
            weight_decay: 0.0  
        sample_mode: max  
        vae_hidden_dim: 50  
        z_dim: 5  
    resume:  
        in_channels: 1
```

USEFUL TOOLS AND LIBRARIES

- Experimental Config management
- **Logging**
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



Tensorboard



Weights & Biases

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- **Experimental Management**
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release

Sacred

*Every experiment is sacred
Every experiment is great
If an experiment is wasted
God gets quite irate*



Pytorch Lightning



Hugging Face

Trainer

The `Trainer` class provides an API for feature-complete training in PyTorch for most standard use cases. It's used in most of the [example scripts](#).



RAY

mlflow

Tracking

Record and query experiments: code, data, config, results

Projects

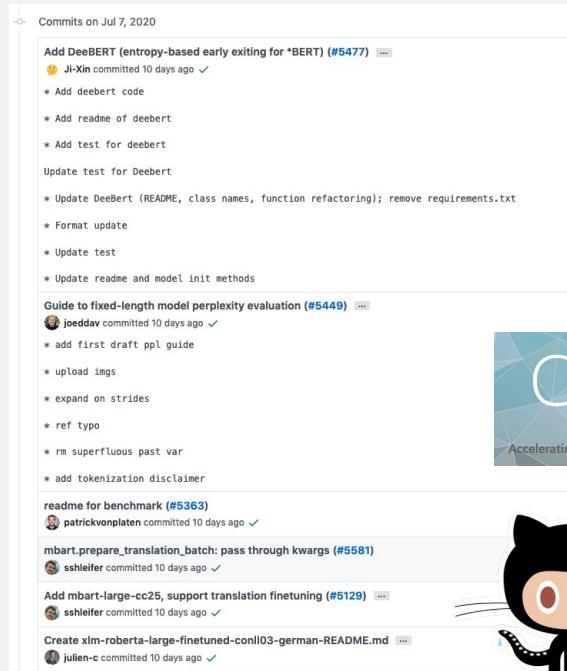
Packaging format for reproducible runs on any platform

Models

General format for sending models to diverse deploy tools

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- **Versioning**
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



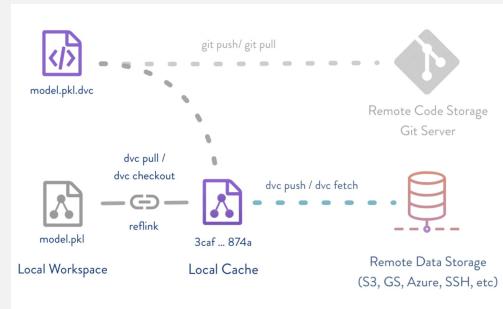
The screenshot shows a GitHub commit history from July 7, 2020. It includes commits for the DeeBERT project, a guide for perplexity evaluation, a README for a benchmark, and a commit for mBART translation batch support. To the right, there is a CodaLab logo and a cartoon cat icon.

Commits on Jul 7, 2020

- Add DeeBERT (entropy-based early exiting for *BERT) (#5477)
 - Ji-Xin committed 10 days ago ✓
 - * Add deebert code
 - * Add readme of deebert
 - * Add test for deebert
 - Update test for DeeBERT
 - * Update DeeBERT (README, class names, function refactoring); remove requirements.txt
 - * Format update
 - * Update test
 - * Update readme and model init methods
- Guide to fixed-length model perplexity evaluation (#5449)
 - joedar committed 10 days ago ✓
 - * add first draft ppl guide
 - * upload imgs
 - * expand on strides
 - * ref typo
 - * rm superfluous past var
 - * add tokenization disclaimer
- readme for benchmark (#5363)
 - patrickvonplaten committed 10 days ago ✓
- mbart.prepare_translation_batch: pass through kwargs (#5581)
 - sshleifer committed 10 days ago ✓
- Add mbart-large-cc25, support translation finetuning (#5129)
 - sshleifer committed 10 days ago ✓
- Create xlm-roberta-large-finetuned-conll03-german-README.md

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- **Data management**
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



DVC, <https://dvc.org/>

Datasheets for Datasets

TIMNIT GEBRU, Google

JAMIE MORGENSTERN, Georgia Institute of Technology

BRIANA VECCHIONE, Cornell University

JENNIFER WORTMAN VAUGHAN, Microsoft Research

HANNA WALLACH, Microsoft Research

HAL DAUMÉ III, Microsoft Research; University of Maryland

KATE CRAWFORD, Microsoft Research; AI Now Institute

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- **Data analysis**
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



Relevant works:

<https://github.com/EleutherAI/lm-evaluation-harness>

master → ParAI / projects / controllable_dialogue / Analysis_n_Graphs.ipynb Go to file ⋮

stephenroller Release remaining Controllable Dialogue Code (#1734) ✓ Latest commit fb4b54d on Jun 2, 2019 History

Ré 1 contributor

2.39 MB Download

Evaluation Analysis (Public Release)

Author: Stephen Roller roller@fb.com. Please direct questions to the ParAI Github issues (<https://github.com/facebookresearch/ParAI/issues>)

This notebook expects to be launched from inside your ParAI installation (typically `~/ParAI`)

You will need to pip install a bunch of things, including pyro and pandas.

General preparation

In [33]: `# bunch of imports and settings
import os
make sure we never see errors on accident in this notebook`

Specificity Control Level (WD)

In [144]: `def plot_resp_wd(metric, figgca, abslim, xaxis='Response-relatedness Control Level (WD)', use_title=False):
 plot_ls = metric['plot_ls'
 modeltype_subset(altered, ["responsive"])),
 modelname_subset(altered, ["repetition++", "baseline_model", "human_eval"])),
 metric,
 fig=figgca,
 xaxis=xaxis,
 xtick_values=[-15, -10, -5, 0, 5, 10, 15],
 xtext_label='-' + '15\nMore\\unrelated', '-10', '-5', '0\\No control', '5', '10', '15\\More\\related'),
 abslim=abslim,
 use_title=use_title,
)

for i, m in enumerate(LIKERT_METRICS):
 fig = plt.figure(figsize=SOLO FIG SIZE)
 plot_resp_wd(m, fig.gca(), None)
 fig.savefig(HOME + "plots/{}_.pdf".format('resp', m), bbox_inches='tight', transparent=True)`

Fluency

Response-relatedness Control Level (WD)

The graph plots Fluency on the y-axis (ranging from 2.4 to 3.6) against Response-relatedness Control Level (WD) on the x-axis (ranging from -15 to 15). Four series are shown: Response-related controlled WD (black line with circles), Beam search baseline (blue dotted line with dots), Human (orange dashed line with triangles), and Repetition-controlled baseline+ (yellow solid line with crosses). The 'Response-related controlled WD' series starts at ~3.2 at -15, peaks at ~3.4 at 0, and drops to ~2.4 at 15. The other three baselines remain relatively flat around 3.0-3.1 across all WD values.

Response-relatedness Control Level (WD)	Response-related controlled WD	Beam search baseline	Human	Repetition-controlled baseline+
-15	~3.2	~3.1	~3.1	~3.1
-10	~3.2	~3.1	~3.1	~3.1
-5	~3.2	~3.1	~3.1	~3.1
0	~3.4	~3.1	~3.1	~3.1
5	~3.2	~3.1	~3.1	~3.1
10	~3.0	~3.1	~3.1	~3.1
15	~2.4	~3.1	~3.1	~3.1

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release

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 steps.
- 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:

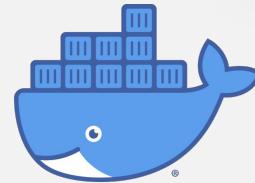
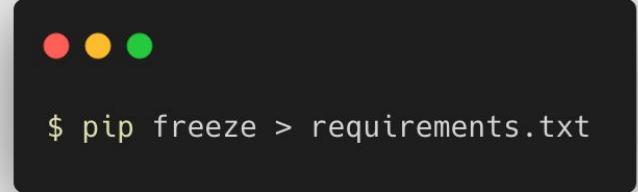
- Specification of dependencies.
- Training code.
- Evaluation code.
- (Pre-)trained model(s).
- README file includes table of results accompanied by precise command to run to produce

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andyrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- **Dependency Management**
- Open Source Release
- Effective Communication
- Test and Release



USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- **Open Source Release**
- Effective Communication
- Test and Release



Language Models are Few-Shot Learners

28 May 2020 • Tom B. Brown • Benjamin Mann • Nick Ryder • Melanie Subbiah • Jared Kaplan • Prafulla Dhariwal • Arvind Neelakantan • Pranav Shyam • Girish Sastry • Amanda Askell • Sandhini Agarwal • Arel Herfort-Voss • Gretchen Kueger • Tom Henighan • Revon Child • Aditya Ramesh • Daniel M. Ziegler • Jeffrey Wu • Clemens Winter • Christopher Hesse • Mark Chen • Eric Sigler • Mateusz Litwin • Scott Gray • Benjamin Chess • Jack Clark • Christopher Berner • Sam McCandlish • Alec Radford • Ilya Sutskever • Dario Amodei

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples... ([read more](#))

[PDF](#) [Abstract](#)

Code

[openai/gpt-3](#)
[sw-yx/gpt3-list](#)
[facebookresearch/anli](#)

[Edit](#)

Tasks

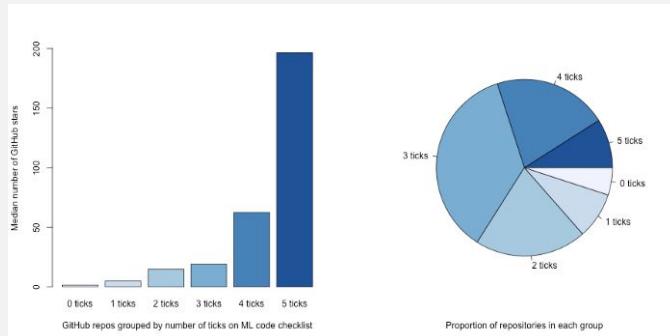
★ 5,107
★ 95
★ 83

[Edit](#)

COMMON SENSE REASONING
COREFERENCE RESOLUTION
DOMAIN ADAPTATION
FEW-SHOT LEARNING

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- **Effective Communication**
- Test and Release



NeurIPS 2019 repositories with 0 ticks had a median of 1.5 GitHub stars. In contrast, repositories with 5 ticks had a median of 196.5 GitHub stars. Only 9% of repositories had 5 ticks, and most repositories (70%) had 3 ticks or less.

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- **Test and Release**



Google Colab



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

REPRODUCIBILITY AS A TEACHING TOOL

Maurits Bleeker, Sam Bhargav

OVERVIEW

How can we mitigate the challenges without reducing the benefits?

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

OVERVIEW

1. Teaching through reproducibility
2. Examples of AI courses utilizing reproducibility as a teaching tool
 - a. Reproducedpapers.org (TU Delft)
 - b. FACT-AI (University of Amsterdam)
3. Guidelines for a successful reproducibility course
4. Lessons learned

TEACHING THROUGH REPRODUCIBILITY

EXAMPLES FROM OTHER ACADEMIC FIELDS

- Learning Networking by Reproducing Research Results (Yan et al. 2017)
 - Stanford CS course on reproducing work on networking systems
- Bringing Replication Into Classroom: Benefits For Education, Science, and Society (Ribotta, Blandine, et al 2022)
 - *"For more than a decade, research in psychology has been struggling to replicate many well-known and highly cited studies"*
- How to Use Replication Assignments for Teaching Integrity in Empirical Archaeology (Marwick, Ben, et al. 2020)
 - *"Here we argue for replications as a core type of class assignment in archaeology courses"*

MOTIVATION

Valuable experience for students:

- Practice implementing and extending existing research
- Recognize the importance (and difficulty) of reproducibility
- Helps students to develop critical thinking skills
 - This also helps with writing research papers
- Can be added to their portfolio, e.g., personal website, blog post, CV
 - Allows students to participate in the community

Contribute to existing research:

- New insights can direct future research
- Results can be published, e.g., in the *ReScience journal*

REPRODUCEPAPERS.ORG
TU Delft

REPRODUCEDPAPERS.ORG

"Is an open online repository for teaching and structuring machine learning reproducibility"

- Primary motivation: there exist several venues for reproducibility but there is a ‘high barrier’ to entry or a focus on ‘short-term’ (alternate years, etc)
- Propose: a low barrier, long term venue focused on reproducibility
- Reproduction aligns with several teaching goals:
 - Reading and critiquing literature
 - Implementing, executing and extending code
 - Comparing, analyzing and presenting results in a clear and concise manner

ONLINE REPOSITORY

The screenshot shows the homepage of the Online Repository. At the top, there is a navigation bar with icons for search, reproductions, papers, help, about, TU Delft logo, and sign in. Below the navigation bar, the page title "Reproductions" is displayed, along with a "Submit Reproduction" button. The main content area features two entries:

Reproduction of "SwinIR: Image Restoration Using Swin Transformer"
by Frans de Boer, Jonathan Borg, Adarsh Denga, Haoran Xia
We explain the technical details of the SwinIR paper in our own words, providing ample detail to understand the authors' contribution and algorithm. Furthermore we explore modifying the architecture used in the paper to allow it to run using reduced resources and thus use less energy.
[Detail](#)

Replicated
Reproduction of "Deep Learning with Differential Privacy"
by Deep Learning CS4240 Group66: Hengkai Zhang, Dong Shen, Yuxin Cheng
Benefits of machine learning techniques based on neuron networks are widely appreciated. While these methods require a large amount of data, sensitive information should be retained. Differential privacy is thus developed. This blog aims to present and describe our efforts to reproduce "Deep... More
[Detail](#)

ONLINE REPOSITORY

- Focus of the project: partial results, minor tweaks, etc.
- Well suited for use in teaching
- Badges (self-labeled):
 - **Replicated:** A full implementation from scratch without using any pre-existing code
 - **Reproduced:** Existing code was evaluated
 - **Hyperparams check:** New evaluation of hyperparameter sensitivity
 - **New data:** Evaluating new datasets to obtain similar results
 - **New algorithm variant:** Evaluating a different variant
 - **New code variant:** Rewrote/ported existing code to be more efficient /readable
 - **Ablation study:** Additional ablation studies

COURSE DETAILS

- Part of MSc CS - Deep Learning course, TU Delft
- Teaching team selects papers with two criteria:
 - Data availability
 - Computational demands
- Projects:
 - Teams should indicate which result to reproduce
 - Groups of 2-4 students, 8 week course
 - $\frac{1}{3}$ of the course time spent on reproduction
- Deliverables:
 - Blog about the repository (private/public)
 - PDF report

24 unique papers, 57 paper reproductions

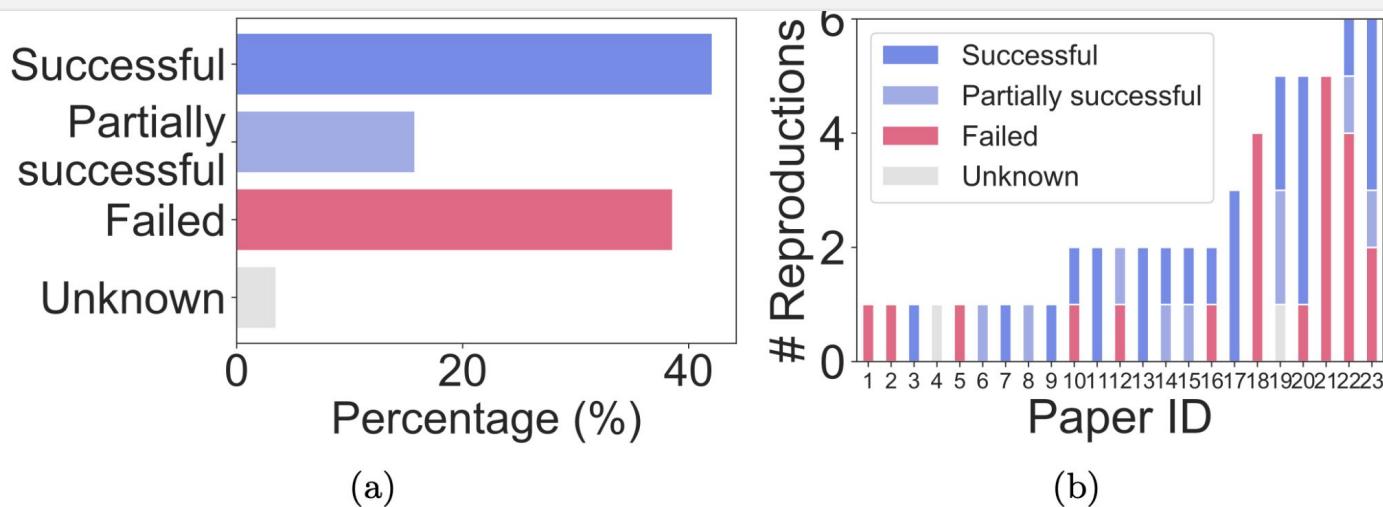
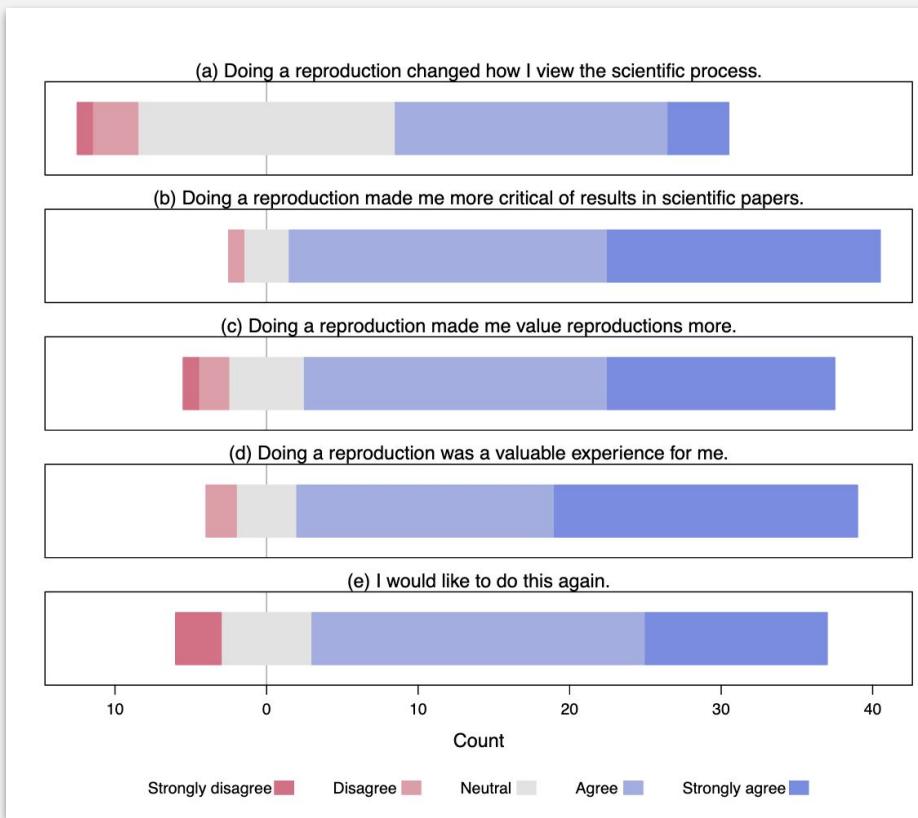


Fig. 2. Current [ReproducedPapers.org](#) statistics. (a) Reproduction success rates; (b) Number of reproductions per paper ID.

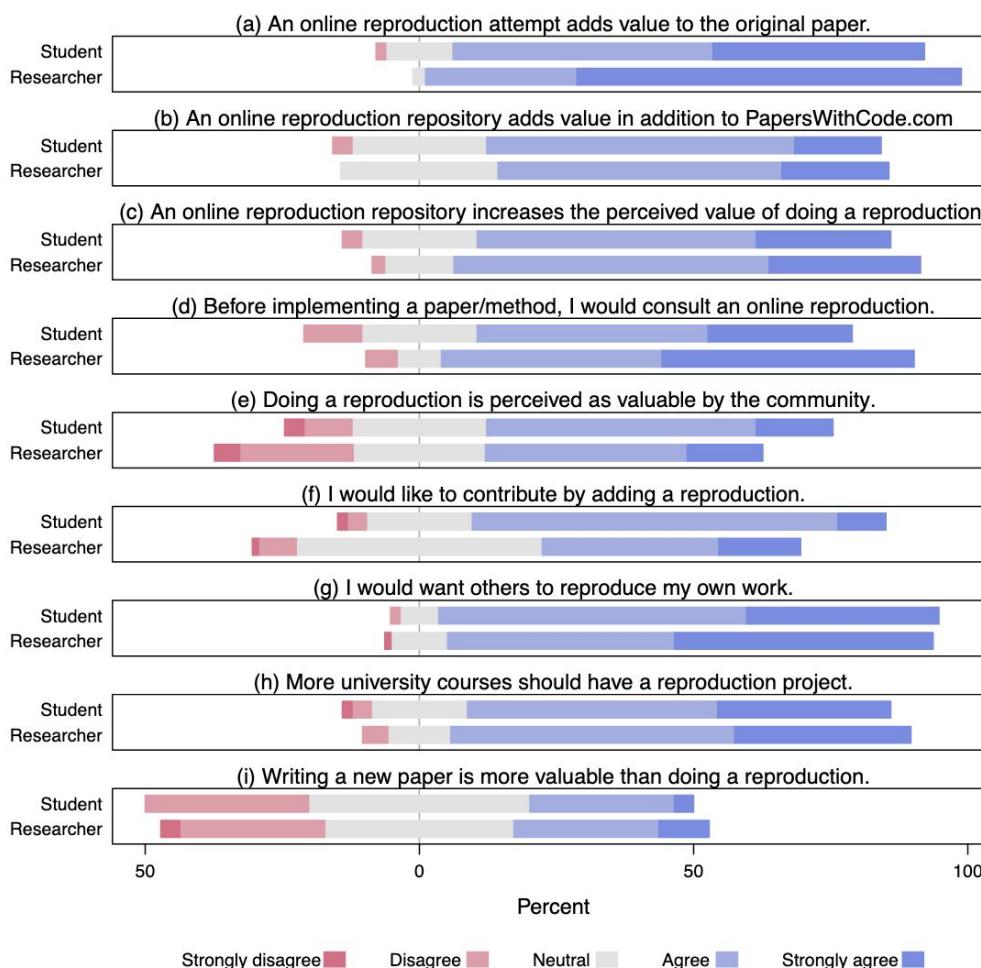
Student survey, N= 43



N = 144

43 course students + 14
other students

87 third-party AI
researchers



CONCLUSION

- Reproduction projects align closely with general course learning goals, and were received positively by most students
- These projects improve perceived value of reproductions, with an added incentive of publishing their work and adding to their portfolios
- *"We finally call on the community to add their reproductions to the website ReproducedPapers.org"*
- *"May the next generation of machine learners be reproducers"*

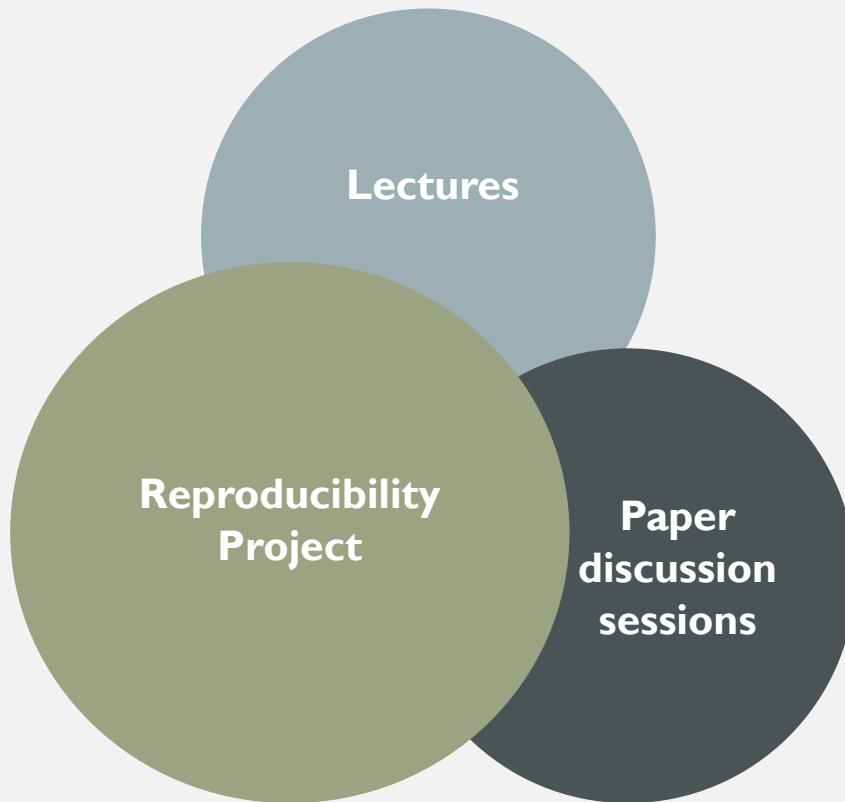
FAIRNESS, ACCOUNTABILITY, CONFIDENTIALITY, AND TRANSPARENCY IN AI COURSE

University of Amsterdam

COURSE MOTIVATION

- In 2019, we designed a new course on Fairness, Accountability, Confidentiality, and Transparency in AI (FACT-AI) at the University of Amsterdam (UvA)
 - Based on the requests of our students in the MSc AI: an increase in interest in ethical issues in AI
- The course aims to make students aware of two types of responsibility:
 - Towards society in terms of potential implications of their research
 - Similar to the NeurIPS Paper Checklist: discuss any potential negative societal impacts of your work
 - Towards the research community in terms of producing reproducible research

COURSE SETUP



LEARNING OBJECTIVES

- **LO #1:** Understanding FACT topics
- **LO #2:** Understanding algorithmic harm
- **LO #3:** Familiarity with FACT methods
- **LO #4:** Reproducing FACT solutions

LEARNING OBJECTIVES

Learning Objective #1: Understanding FACT topics

- Students can explain the major notions of fairness, accountability, confidentiality, and transparency that have been proposed in the literature, along with their strengths and weaknesses

Learning Mechanism:

- General lecture(s) per topic

LEARNING OBJECTIVES

Learning Objective #2: Understanding algorithmic harm

- Students can explain, motivate, and distinguish the main types of algorithmic harm, both in general and in terms of concrete examples where AI is being applied

Learning Mechanism:

- General lectures and guest lectures, where students can ask questions and are encouraged to participate in discussions
- This LO can be used for any AI course

LEARNING OBJECTIVES

Learning Objective #3: Familiarity with FACT methods

- Students are familiar with recent peer-reviewed algorithmic approaches in the FACT-AI literature

Learning Mechanism:

- Paper discussion sessions where students discuss a seminal FACT-AI paper in a small and interactive group, after reading the paper in advance

PAPER DISCUSSION

- Outline of how to dissect a paper ahead of time
 - Examples help!
- For the students, the goal of the paper discussion sessions is to:
 - Learn about prominent methods in the field
 - Reading a technical paper
 - Think critically about the claims made in the papers
 - Understanding a paper's strength and weaknesses
- All these (reading) skills are necessary for a good reproducibility study
 - If students can't understand the paper, how will they reimplement the algorithm?

PAPER DISCUSSION

- Students first read a seminal paper on their own trying to answer the following questions:
 - What are the main claims of the paper?
 - What are the research questions?
 - Does the experimental setup make sense, given the research questions?
 - What are the answers to the research questions? Are these supported by experimental evidence?
- Participate in small discussion sessions (ideally in person) with their peers to discuss their answers
 - Groups of 4 to 5 students

PAPER DISCUSSION

An instructor goes over the same paper, giving an overview of the papers' strengths and weaknesses

- In our case, each session was presented by a different instructor
- This to show:
 - There is no single way of examining a research paper
 - Different researchers will bring different perspectives to their assessment of papers
- We chose papers for their discussion sessions based on their impact on the FACT-AI field

LEARNING OBJECTIVES

Learning Objective #4: Reproducing FACT solutions

- Students can assess the degree to which recent algorithmic solutions are effective, especially with respect to the claims made in the original papers, while understanding their limitations and shortcomings

Learning Mechanism:

- Group project where students work in groups to reproduce FACT-AI papers from top AI conferences

GROUP PROJECT

- The group project is based on **reproducing existing algorithms** from top AI conferences and is the focal point of the course
- In our course, we focused on FACT-AI algorithms
- However, the setup for the course is not specific to FACT-AI and can be tailored to any topic
 - e.g., NLP, computer vision, information retrieval, general ML, etc.

GROUP PROJECT

- The group project is based on **reproducing existing algorithms** from top AI conferences and is the focal point of the course
- Students work in groups to reimplement existing algorithms from papers in top AI conferences (e.g., NeurIPS, ICML, ICLR, AAAI, etc).
 - FACT-AI course: groups of 3-4 students
- Students write up the results and submit reports
 - We encouraged them to submit their reports to the ML Reproducibility Challenge
- In our course, we focused on FACT-AI algorithms. However, the setup for the course is not specific to FACT-AI and can be tailored to any topic
 - e.g., NLP, computer vision, information retrieval, general ML, etc.

GROUP PROJECT

Benefits of participating in the ML Reproducibility Challenge:

- Motivates and incentivizes students
- Reports accepted by the ML Reproducibility Challenge are accepted for publication in the *ReScience* journal
- Exposes students to the paper submission cycle

GROUP PROJECT

Participating in the ML Reproducibility Challenge gives the students the opportunity to experience the whole research pipeline:

1. Reading a technical paper to understand its strength and weaknesses
2. Implementing (and perhaps also extending) the algorithms in the paper
3. Writing up the findings
4. Submitting to a venue with a deadline
5. Obtaining feedback from reviewers
6. Writing a rebuttal
7. Receiving the official acceptance/rejection notification

COURSES PARTICIPATE IN RC2021 FALL EDITION

Courses Participated in RC2021 Fall Edition

- [DD2412 Deep Learning, Advanced](#). KTH (Royal Institute of Technology), Stockholm, Sweden
- [CISC 867 Deep Learning](#), Queen's University, Ontario, Canada
- [Special Topics in CSE: Advanced ML](#), Indian Institute of Technology, Gandhinagar, India
- [FACT: Fairness, Accountability, Confidentiality and Transparency in AI](#), University of Amsterdam, Netherlands
- [CSCI 662 -- Advanced Natural Language Processing](#), University of Southern California, USA
- [Intelligent Systems and Interfaces](#), Indian Institute of Technology, Guwahati, India
- [Intelligent Information Processing Topics](#), Tsinghua University, China
- [Machine learning for data science 2](#), University of Ljubljana, Slovenia
- [EECS 598-005: Randomized Numerical Linear Algebra in Machine Learning](#), University of Michigan, USA
- [SYDE 671 - Advanced Image Processing](#), University of Waterloo, Canada
- [BLG561E Deep Learning](#), Istanbul Technical University, Turkey
- [CS 433 Machine Learning](#), EPFL, Switzerland

RESULTS OF THE ML REPRODUCIBILITY CHALLENGE

- See https://openreview.net/group?id=ML_Reproducibility_Challenge
- ML Reproducibility Challenge 2021
 - $\pm 40\%$ of the accepted papers were from the UvA FACT-AI course
- ML Reproducibility Challenge 2022
 - $\pm 50\%$ of the accepted papers were from the UvA FACT-AI course
 - Best paper award
 - 2 outstanding papers (out of 4)

FEEDBACK

First year MSc AI students

"I appreciate the critical view I have developed on papers as a result of this course. Normally I would easily accept the content of a paper, but I will be more critical from now on, as many papers are not reproducible."

"I really appreciated that this was the first course where students are judging state-of-the-art AI models. In other words, students were able to experience the scientific workfield of AI."

FEEDBACK

First year MSc AI students

"Replicating another study, seeing how (poorly) other research is performed was really eye-opening."

"I think it's really good that we get some practical insights into reproducing results from other papers, not all papers are as good as they seem to be."

QUESTIONS?

GUIDELINES FOR A SUCCESSFUL REPRODUCIBILITY COURSE

GUIDELINES FOR A SUCCESSFUL REPRODUCIBILITY COURSE

- INCLUDE A REPRODUCIBILITY LECTURE
- PAPER REQUIREMENTS
- GRADING
- TEACHING ASSISTANTS
- TIMING OF THE COURSE
- DURATION OF THE COURSE
- ADVANTAGES OF PARTICIPATING IN THE ML REPRODUCIBILITY CHALLENGE

INCLUDE A REPRODUCIBILITY LECTURE

Motivate reproducibility with a general lecture

- Position this lecture (ideally) at the beginning of the course
- Highlight papers examining reproducibility/replicability failures
 - For examples in NLP, see Part 2 of the tutorial
 - Include consequences of failure to reproduce (Part 2)
- Clearly outline scope of the project(s) and potential impact

PAPER REQUIREMENTS

- Choose 10-15 papers from the ML Reproducibility Challenge OpenReview portal that are suitable for your course
- Before the course starts, let the TAs check whether the selected papers are feasible for reproducibility study
 - Hire a team of experienced, graduate-level TAs
- Ideally assign each TA no more than 3-4 papers

PAPER REQUIREMENTS

- Select papers that are computationally feasible to reproduce
 - In our case, we were able to provide one GPU per team
 - Depends the available resources of the course and faculty
- At least one dataset should be publicly available and of a reasonable size
 - If the dataset is too big, it is an option to reproduce the work in a ‘low-resource’ data setting
- Select papers that are relevant to the topics covered in the course
- Emphasize the technical perspective of the sub-field
- It should be reasonable to reimplement the paper within the allotted time

GRADING

- Grading group projects on different papers in a fair manner is challenging
- Try to make the grading criteria as explicit as possible in order to make it clear for the students what is expected
- Organize a grade calibration session with the TAs after grading to align on expectations
- If participating in the ML Reproducibility Challenge, grade reports independently of the reviews

Grade	<= 5 (fail)	6 (sufficient)	7 (satisfactory)	8 (good)	9 (very good)	10 (excellent)
Project (40%)						
	Project Design	Unsystematic and/or no validated use of research and design methodologies. Insufficient explanation. How are the results tested and/or verified?	Adequate use of research and design methodologies. Limited explanation.	Adequate use of research and design methodologies. Explained and justified.	Use of the right research and design methodologies. Well-explained and well justified.	Profound and critical use of research and design methodologies. Very clear and validated design.
	Positioning of project	Project not positioned w.r.t. new literature, the FACT-field and reproducibility papers.	Project is somewhat positioned.	Project is sufficiently positioned in literature.	Project is correctly positioned in literature.	Project is well positioned within literature.
	Creativity	The project does not make an original contribution. E.g. the picked paper is just said to be reproducible or not without any extra insights.	Project does not really make any original contribution. The results are reproducible, with limited effort or not reproducible with limited insights (why is this not working?).	Project team had at least one original contribution to reproduce the work and/or go beyond the original results of the paper.	Project team came up with several original ideas to reproduce the paper and/or go beyond the original results, design options and/or concepts not initiated or thought of by the supervisor.	Project team came up with many original ideas, design options and/or concepts to reproduce the work and/or go beyond the original results. Not initiated or thought of by the supervisor.
Code base (20%)						
	Technical quality	Insufficient	Sufficient	Satisfactory	Good	Very Good
	Reproducability of your results by the TA's.	Not reproducible. The project results should be reproducible by the TAs	N/A	With some effort the results are reproducible by the TAs.	N/A	Without any effort the results are reproducible by the TAs
Paper (30%)						
	Content	Report shows no coherence of content. For example: What questions are you asking? What experiments do you run to answer them? What conclusions can you draw from these experiments?	Report shows sufficient coherence of content.	Report fulfils all requirements in terms of content.	Good report in terms of content.	Very good report in terms of content.
	Form	Structure needs considerable improvement. General presentation of the content (text and figures) not very effective.	Structure needs some improvement. General presentation of the content (text and figures) is sufficient.	Structure is acceptable. General presentation of the content (text and figures) is satisfactory.	Clear structure. Good presentation of the content (text and figures).	Well-structured document. General presentation of the content (text and figures) is effective.
	Quality of writing	Poorly expressed. Document contains serious spelling and grammatical errors.	Reasonably expressed argumentation. Document contains some spelling and grammatical errors.	Sufficiently expressed argumentation. The document contains little spelling and grammatical errors.	Expressed and formulated well. Document has a nice flow. Document contains only minor spelling and grammatical errors.	Expressed and formulated very well. Document has a smooth flow with sufficient transitions. Document is without any spelling and grammatical errors.
Presentation (10%)						
	Content	Presentation lacks detail and does not support conclusions. Irrelevant information presented.	Presentation lacks detail, and is just enough to support conclusions.	Presentation has sufficient detail to support conclusions.	Presentation has a good level of detail to support conclusions.	Presentation has the right level of detail to support the conclusions and to understand the recommendations.
	Form	Presentation is unstructured and not well organized. No (proper) use of visual aids.	Logical structure of presentation is poor. Improvements to the structure should be made. Use of visual aids can be improved.	Logical structure of presentation is reasonable but needs some improvement. Sufficient use of visual aids.	Presentation has good logical structure, the essentials are separated from the ancillary. Good use of visual aids.	Presentation has very good logical structure, the essentials are very well separated from the ancillary. Perfect use of visual aids.
	Performance	Poorly expressed and formulated. Unclearly presented. Audience was ineffectively addressed.	Expression and formulation can be improved. Not always clearly presented.	Expressed and formulated adequately. Most of the time clearly presented. Audience was sufficiently addressed.	Well expressed and formulated. Clearly presented. Audience was well addressed.	Expressed, formulated and presented with great style, clarity and effectiveness. Audience was very well addressed and engaged.

TEACHING ASSISTANTS

- Have the TAs read the papers before the course starts to ensure they have a sufficient, in-depth understanding of their papers
 - Assign papers to TAs based on their interests
- To ease the load for the TAs, have several groups working on the same paper
- Ensure students have regular contact with their TA so no group gets stuck in the process
- Ask students halfway through the course to submit a draft report to their TAs in order to get feedback
 - We found this significantly increased the quality of the final reports

TIMING OF THE COURSE

Students need to have very strong programming skills

Table 1: The first year of the MSc AI program at the University of Amsterdam.

Course	Sem. 1	Sem. 2	EC
Computer Vision 1	■ □ □	□ □ □	6
Machine Learning 1	■ □ □	□ □ □	6
Natural Language Processing 1	□ ■ □	□ □ □	6
Deep Learning 1	□ ■ □	□ □ □	6
Fairness, Accountability, Confidentiality and Transparency in AI	□ □ ■	□ □ □	6
Information Retrieval 1	□ □ □	■ □ □	6
Knowledge Representation and Reasoning	□ □ □	■ □ □	6
Elective 1	□ □ □	□ ■ □	6
Elective 2	□ □ □	□ □ ■	6
Elective 3	□ □ □	□ □ ■	6

DURATION OF THE COURSE

- We strongly recommend to ensure that the students to have enough time to work on the project
- For our course, the students are working one month full-time on the project
 - We found this to be a beneficial setup since students didn't have to worry about any other courses during this time
- If it's not possible to work on the project full-time, then potentially adapt the weight of the course:
 - If students typically have 5 courses in one semester, consider making the reproducibility course worth 2 courses

ADVANTAGES OF PARTICIPATING IN THE ML REPRODUCIBILITY CHALLENGE

- Prioritize the ML Reproducibility Challenge by tying the reproducibility report directly to the grading
 - Students are graded on the same report that they submitted to the challenge therefore, participating is not an extra task
- Submitting to the challenge gives the students the opportunity to experience the whole research pipeline:
 - Submitting to a venue with a strict deadline
 - Obtaining feedback
 - Writing a rebuttal
 - Receiving the official notification

LESSONS LEARNED

SUMMARY OF THE LESSONS LEARNED

In our experiences, we found that the following were important components of a successful course:

- Including extension as part of reproducibility
- Having excellent teaching assistants
- Having students participate in the ML community
- Encouraging communication with the original authors

INCLUDING EXTENSIONS AS PART OF REPRODUCIBILITY

- We argue that the finding "*the original work is (not) reproducible*" is not insightful
- Require students to extend the paper if the source-code is already available
- Either extend the work to:
 - New domains, datasets or a low-resource regime (i.e., less data/compute)
 - New hyper-parameter settings or method different assumptions
 - Different model architecture
- Or explain why the work is not reproducible

INCLUDING EXTENSIONS AS PART OF REPRODUCIBILITY

There are two scenarios possible for the project:

- There already exists an open-source implementation of the selected paper. Students are allowed to use this:
 - The results the students obtain are different as described in the paper
 - The results are reproducible, meaning this method can now be used for further research
- There is no open-source implementation available, meaning the students need to reimplement everything themselves
 - Take this into account when grading

HAVING EXCELLENT TEACHING ASSISTANTS

- It is extremely important for the TAs to have **excellent programming experience** since this is the main aspect students need help with
- Have students meet with the TAs at least twice a week
- We had both second year MSc students and PhD students
 - PhD students are preferred, if possible
- Have the TAs help students with writing the rebuttal, since this is a new experience for them

HAVING EXCELLENT TEACHING ASSISTANTS

Since this is probably the first time the students are submitting a research paper, try to prevent the following common mistakes:

- Submitting single blind
- Referring to the course project in the introduction
- Motivation: "We had to do this for a course project"
- Submitting a non-anonymized code-base

HAVING STUDENTS PARTICIPATE IN THE ML COMMUNITY

- It is a motivating factor for students to create concrete output that is beneficial to the broader ML research community
- FACT-AI course 2019--2020
 - Creating a public repository with the best algorithm implementations
- FACT-AI course 2020--2021 and 2021--2022:
 - Participating in the ML Reproducibility Challenge

ENCOURAGING COMMUNICATION WITH THE ORIGINAL AUTHORS

- We strongly encourage students to contact the original authors
- It is beneficial for students to interact with scientists in the field
- It improves the papers' credibility, readability, and reproducibility
- Give the students some instructions how to do this:
 - Be aware that the authors are busy
 - Prevent that multiple teams are emailing at the same time
 - Have the TAs coordinate this

SUMMARY OF THE LESSONS LEARNED

In our experiences, we found that the following were important components of a successful course:

- Including extension as part of reproducibility
- Having excellent teaching assistants
- Having students participate in the ML community
- Encouraging communication with the original authors

QUESTIONS?

CONCLUSION

CONCLUSION

- We have shown two successful examples of graduate-level AI courses that focus on reproducibility with their course project
- We provided guidelines to successfully run a reproducibility project for any graduate-level AI course
- Implementing a course centred on a reproducibility project is fairly straightforward for the instructor and has many benefits for students
 - The course naturally "refreshes" itself every year when a new batch of papers is chosen

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

KEY TAKEAWAYS

SUMMARY OF TUTORIAL

In this tutorial, we've aimed to address the issue of **ensuring research results are reproducible**

- Part 1: We gave an introduction to reproducibility and presented some examples of (ir)reproducible results, both from within CS and from other disciplines
- Part 2: We went over some checklists in NLP as well as some examples of reproducibility research in NLP
- Part 3: We investigated existing mechanisms for reproducibility in ML/NLP such as Papers with Code and the ML Reproducibility Challenge
- Part 4: We discuss how to teach reproducibility to the next generation of AI researchers

BEST PRACTICES TO KEEP IN MIND

1. **Report** as much as much information as you can
 - Different types of papers have different requirements -- when creating a new dataset, consider the annotators! When running experiments, do a hyperparameter search!
2. **Share** dependency config files
3. **Release** code
 - If an experiment didn't work or provides evidence that doesn't support your main hypothesis (e.g., that your model is better than previous models), you should still report it!
4. **Run** multiple experiments (with different random seeds, or different data orders, etc.) and report error bars.
5. **Record** your carbon emissions
 - You can use tools like [CodeCarbon](#) or the [ML CO₂ Calculator](#)
6. **Fill out** reproducibility checklists correctly, try to do any items that are appropriate (though we recognize the checklist isn't perfect)