

# Machine Learning Reproducibility: An update from the NeurIPS 2019 Reproducibility Co-Chairs

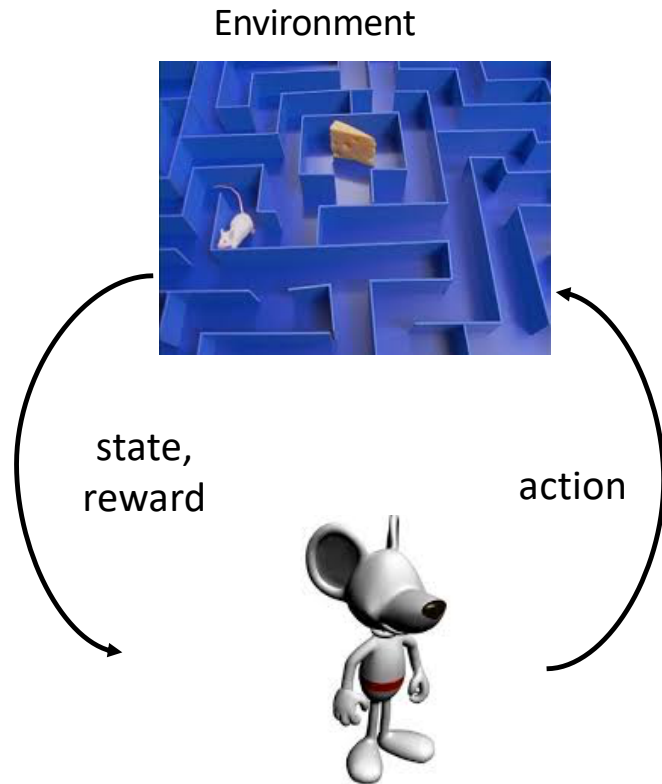
Reproducibility chairs: Joelle Pineau (McGill / FAIR), Koustuv Sinha (McGill)

Collaborators: Jessica Forde, Hugo Larochelle, Vincent Larivière, Philippe Vincent-Lamarre

Workshop on Systems for ML @ NeurIPS 2019

December 13 2019, Vancouver CANADA

# Reinforcement learning (RL)



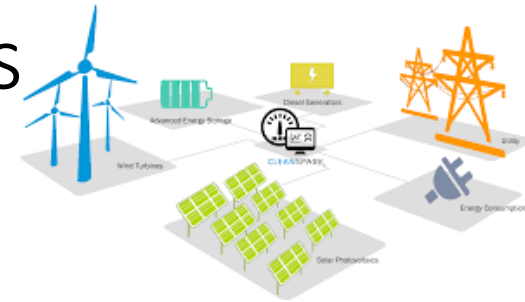
Learn  $\pi = \text{strategy to find this cheese!}$

- Very general framework for sequential decision-making!
- Learning by trial-and-error, from sparse feedback.
- Improves with experience, in real-time.

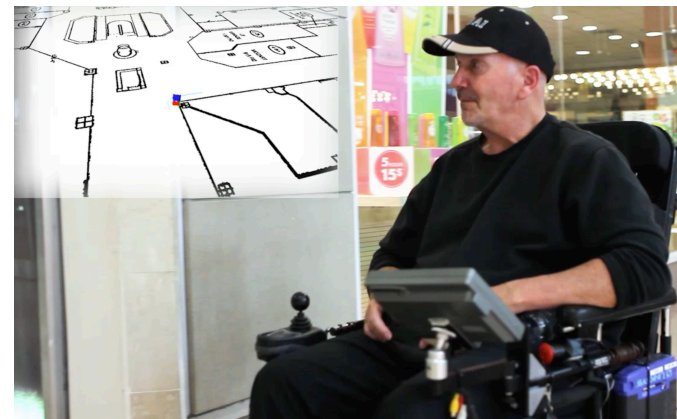
# Impressive successes in games!



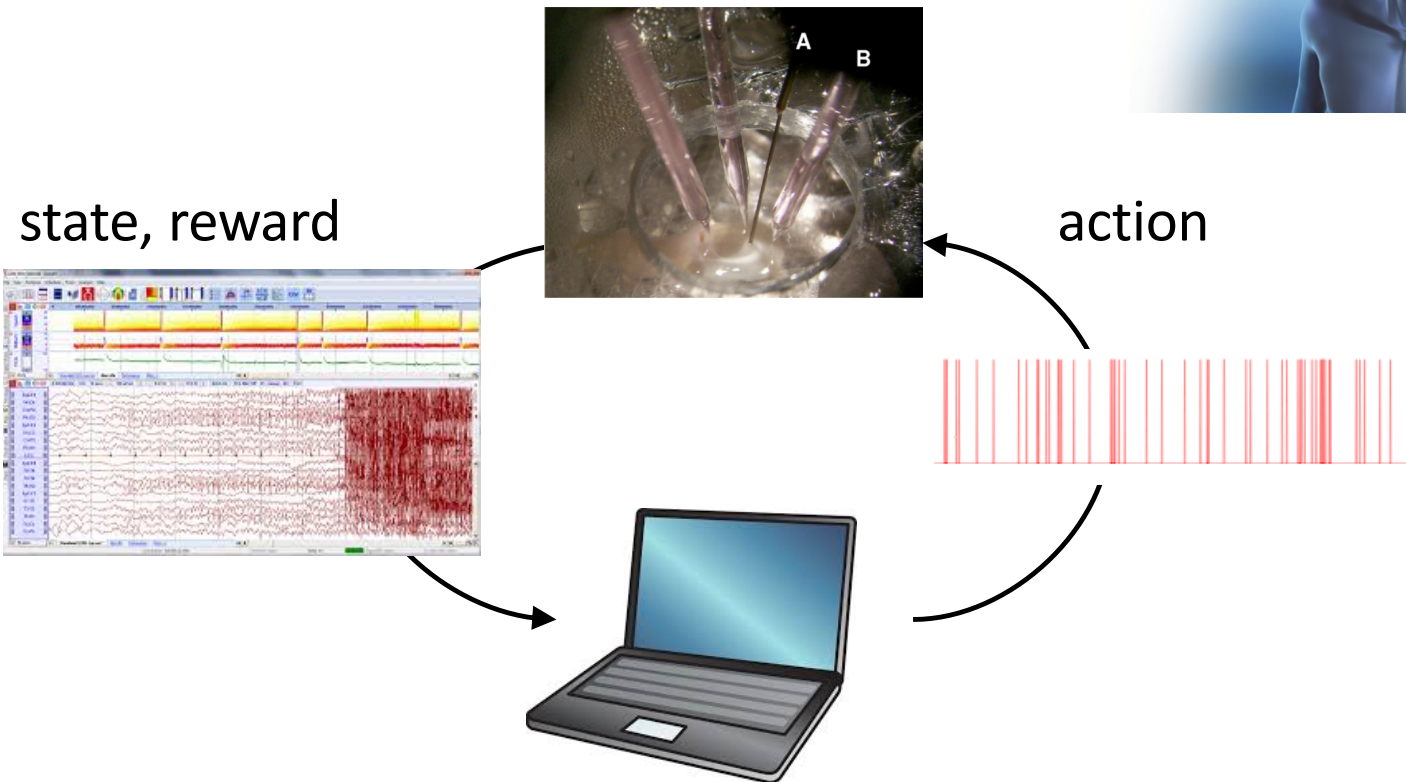
# RL applications beyond games

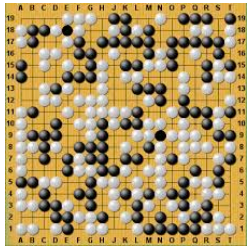
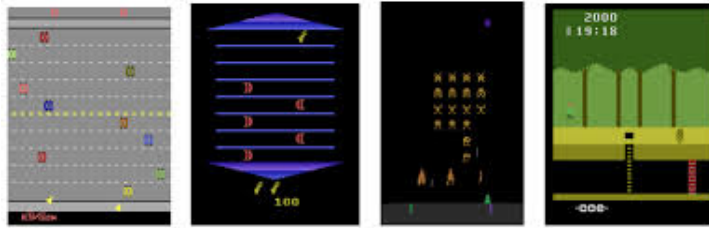


- Robotics
- Video games
- Conversational systems
- Medical intervention
- Algorithm improvement
- Crop management
- Personalized tutoring
- Energy trading
- Autonomous driving
- Prosthetic arm control
- Forest fire management
- Financial trading
- Many more!



# Adaptive neurostimulation

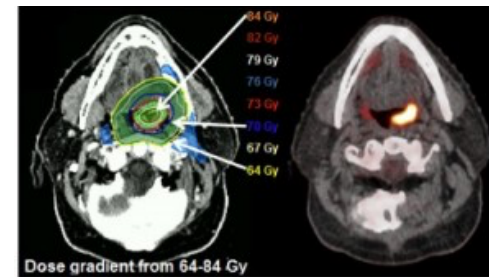
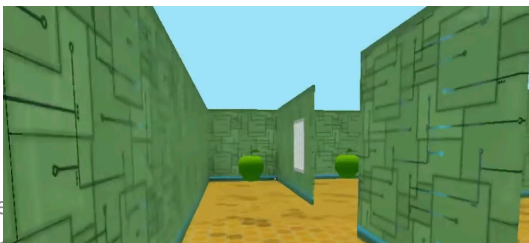
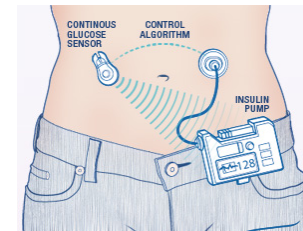
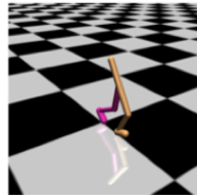




RL in simulation

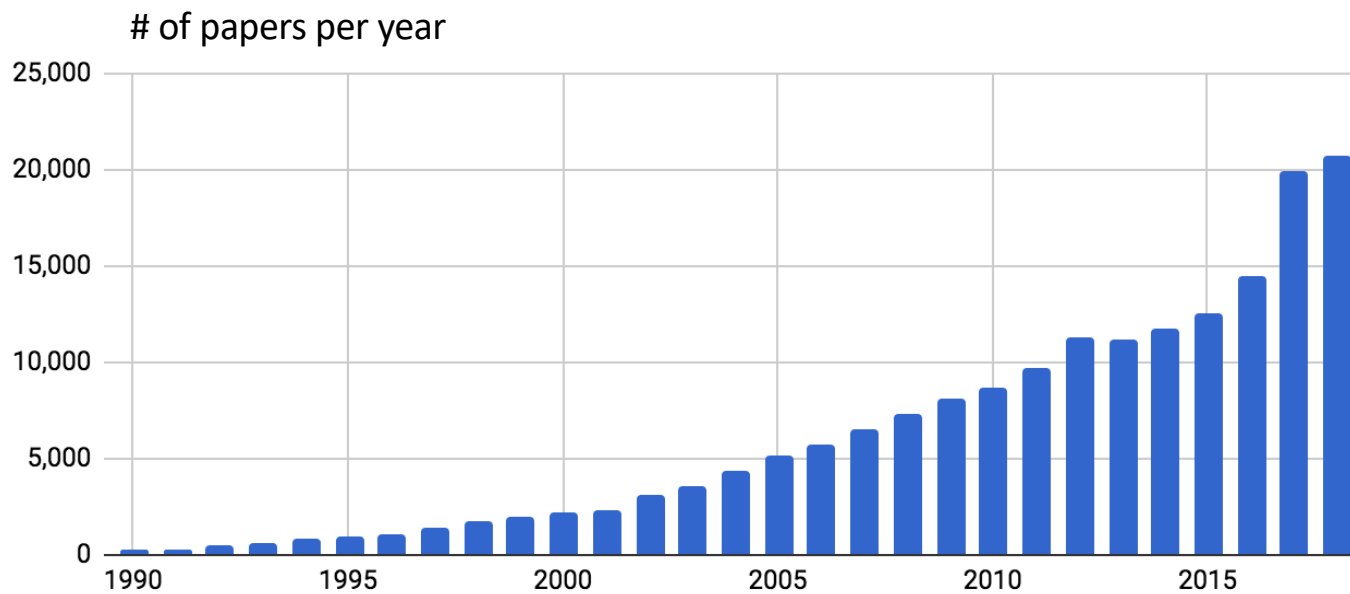


RL in real-world  
from  $\sim 10^1 - 10^2$  trials



Improving he  
Joelle Pineau

# 25+ years of RL papers



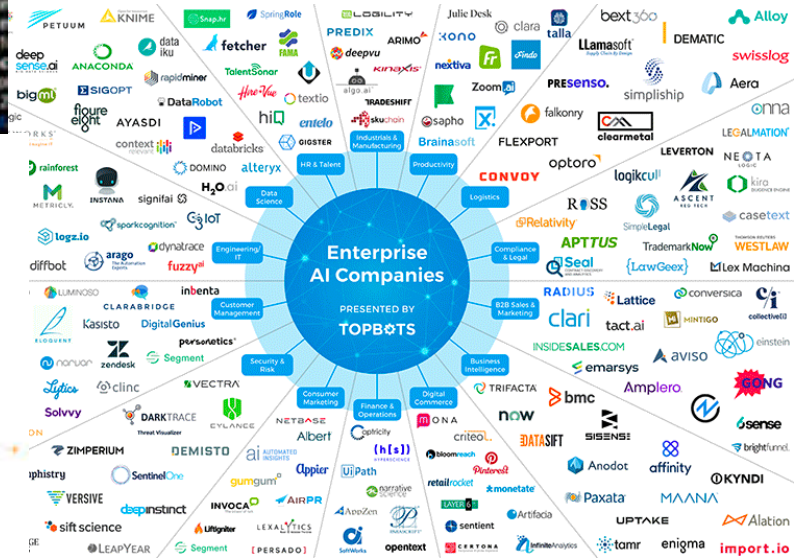
P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger.  
*Deep Reinforcement Learning that Matters. AAI 2017 (+updates).*

# Machine learning systems



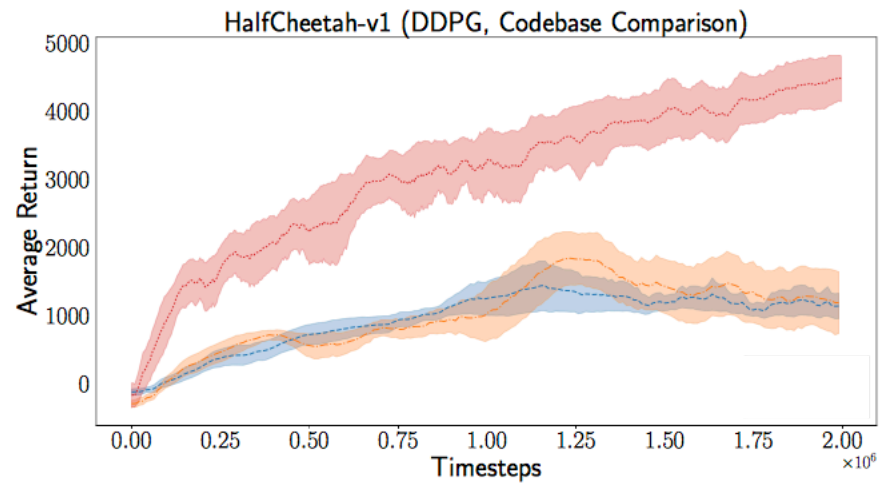
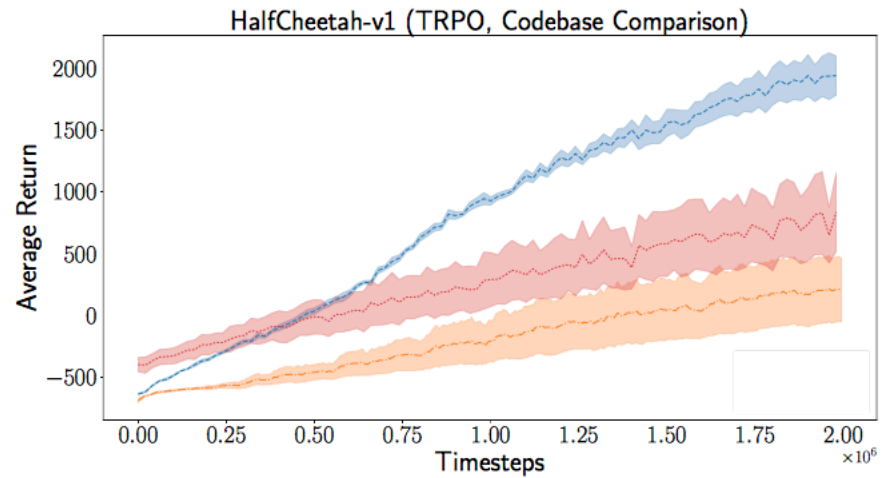
```

    style.visibility="hidden"; } res1 = arg2.toString() args = arg; var while(args>1) sms = do
    999) ElementFrc arg1 = parseInt(args/2); res1 = arg2.toString(); ("dumdiv"); if (Lans ==
    args.byte; } } res1 != 999) window.onload=chk; a_fase = (b_fase - dayBreak)*24;
    rands() args = arg1; </script> (var str=span.firstChild.data;+res1.toString(); var if(a
    removeChild if (data.substring(i, i+1)=="") (span.+res1.toString(); firstChild; for var
    res1 == fun(sp) ) (var theSpan=document.createElement("Blind");false if(res1 == 99
    (res1 = args.toString() document.createTextNode(str.charAt(i)); span.appendChild
    percent!++;window.status=" % complete"; fid1=window.setTimeout (if(percent < 1
    ofForm = Math.floor(secTimeCode); sec.ctrl.innerHTML<=;break; Math.abs deg;
    on Seconds(data) { ;var ll = return(data.substring(i+1,data.length); res1.length; M
    r.while(ll%4 != 0) var sd = name.value; bhspres1 = 0; =hsp return(data.substring
    360); else color.length=span.firstChild.data.length; light span=span;function chang
    (cube) { string.speed=(spd==fun(bar) if (isNum(sdl) Math.abs(spd); x=Math.floor re
    = decimalToBin(sdl); sqr.binc= fork.deg/this.length; charm.br=br; if (percent1 < 1
    nit:function(x{value = result; sort.ctrl.setAttribute("Source", ct) 121;Math.abs(br)
    turn res1; } sort times=null;toSpans(spon); merge.moveColor = function()
    (data.substring(i, i+1)=="") function changer(IfmoveColor = function()
  
```





- Same algorithm
  - Top graph: TRPO
  - Bottom graph: DDPG
- Same domain
- Simulation environment
- Different implementations



# December 2018....

## *Reproducible, Reusable, and Robust Reinforcement Learning*

**Joelle Pineau**

Facebook AI Research, Montreal  
School of Computer Science, McGill University



Neural Information Processing Systems (NeurIPS)  
December 5, 2018



# We surveyed 50 RL papers from 2018

(published at NeurIPS, ICML, ICLR)

	<u>Yes:</u>
• Paper has experiments	100%
• Paper uses neural networks	90%
• All hyperparams for proposed algorithm are provided.	90%
• All hyperparams for baselines are provided.	60%
• Code is linked.	55%
• Method for choosing hyperparams is specified	20%
• Evaluations on some variation of a hold-out test set	10%
• Significance testing applied	5%



# Behind the Program for Reproducibility at NeurIPS 2019



Neural Information Processing Systems Conference [Follow](#)

Sep 27 · 3 min read



The Neural Information Processing Systems (NeurIPS) conference has long been the leading venue for new and exciting machine learning research. This year, the spirit of innovation and scientific leadership goes beyond the conference content, with the creation of the new role of Reproducibility Chair within the program committee. This guest blog post is written by the Reproducibility Chairs for NeurIPS 2019, [Joelle Pineau](#) and [Koustuv Sinha](#), to explain the reproducibility program being rolled out to support high-quality scientific contributions at the conference and beyond.

One of the challenges in machine learning research is to ensure that presented and published results are sound and reliable. Reproducibility,

# The NIPS Experiment

[edit]

16 December 2014

Just back from NIPS where it was really great to see the results of all the work everyone put in. I really enjoyed the program and thought the quality of all presented work was really strong. Both Corinna and I were particularly impressed by the work that put in by oral presenters to make their work accessible to such a large and diverse audience.

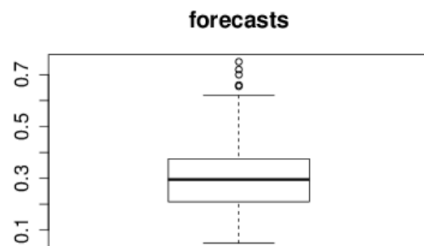
We also released some of the figures from the NIPS experiment, and there was a lot of discussion at the conference about what the result meant.

As we announced at the conference the consistency figure was 25.9%. I just wanted to confirm that in the spirit of openness that we've pursued across the entire conference process Corinna and I will provide a full write up of our analysis and conclusions in due course!

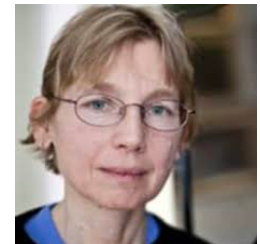
Some of the comment in the existing debate is missing out some of the background information we've tried to generate, so I just wanted to write a post that summarises that information to highlight its availability.

## Scicast Question

With the help of [Nicolo Fusi](#), [Charles Twardy](#) and the entire Scicast team we launched [a Scicast question](#) a week before the results were revealed. The comment thread for that question already had [an amount of interesting comment](#) before the conference. Just for informational purposes before we began reviewing Corinna forecast this figure would be 25% and I forecast it would be 20%. The box plot summary of predictions from Scicast is below.



Neil Lawrence



Corinna Cortes

# NeurIPS 2019 Reproducibility program

- Code submission policy
- NeurIPS 2019 Reproducibility challenge
- Machine Learning Reproducibility Checklist



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

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

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

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

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

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

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

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

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

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

# Any objections?

- Dataset confidentiality
- Proprietary software
- Computation infrastructure
- Replication of mistakes



## Papers with link to code

NeurIPS 2018 < 50%

ICML 2019 67%

NeurIPS 2019 75%

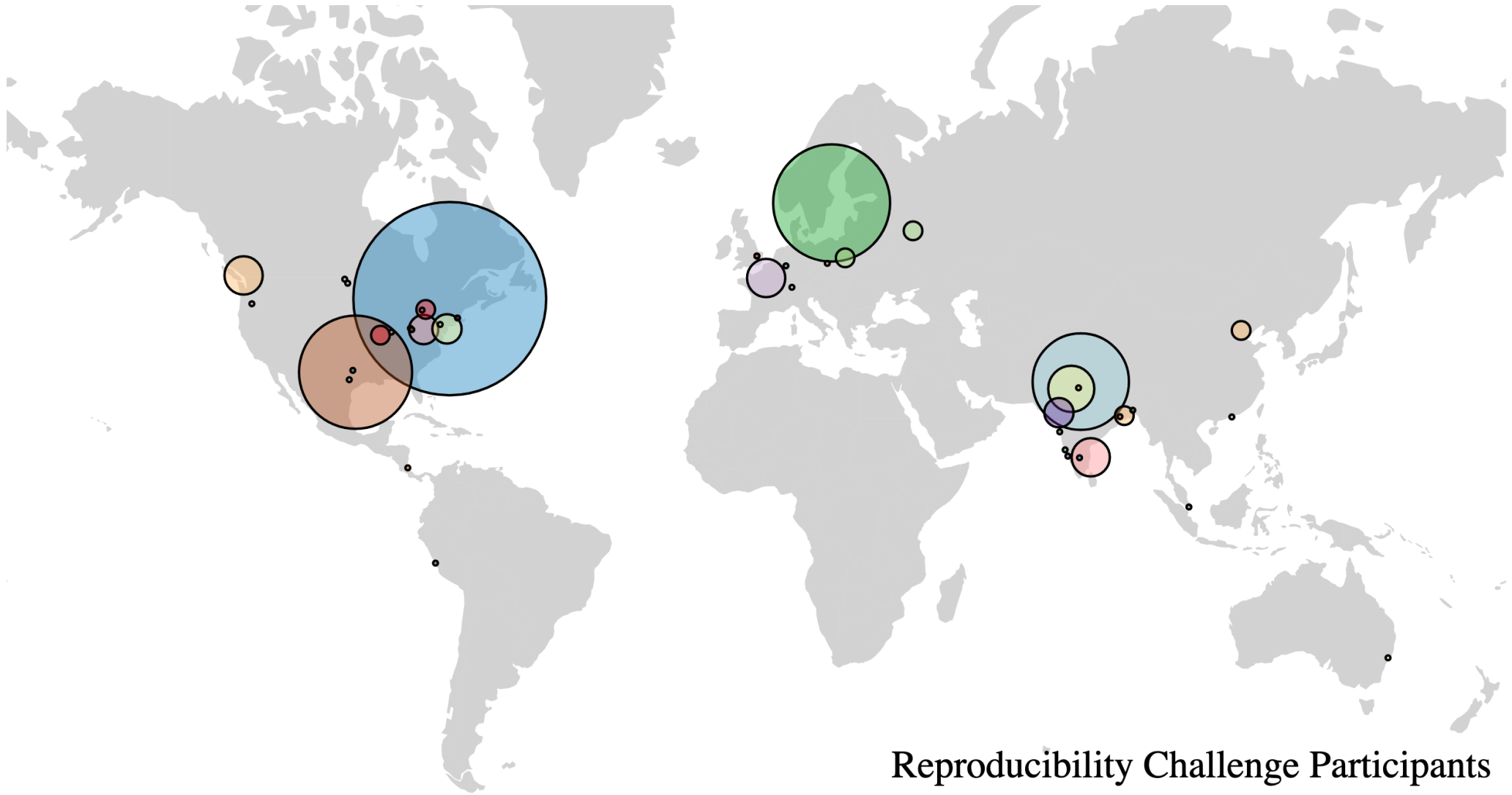
# Reproducibility Challenge @ NeurIPS 2019

NeurIPS 2019 papers claimed for reproducibility challenge 173

vs ICLR 2019 papers claimed 90

Max claims for a single paper = 5

*T Ginart, MY Guan, G Valiant, J Zou "Making AI Forget You: Data Deletion in Machine Learning"*



**Reproducibility Challenge Participants**  
63 from universities + 10 from industry



# NeurIPS 2019 Reproducibility Challenge

Vancouver, Canada December 13-14, 2019 <https://reproducibility-challenge.github.io/neurips2019/date...>

## Here are some instructions

Submission Claims accepted from 2019 Aug 7 to 2019 Nov 1 (GMT)

[Unclaimed](#)[Claimed](#)[All Reports](#) 

## Re: Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

Manjot Singh, Yiyu Wang

02 Dec 2019 (modified: 12 Dec 2019) NeurIPS 2019 Reproducibility Challenge Blind Report Readers: Everyone 0 Replies

[Show details](#)

## [Re] Learning to Learn By Self-Critique

Isac Arnekvist, Dmytro Kalpakchi

02 Dec 2019 (modified: 12 Dec 2019) NeurIPS 2019 Reproducibility Challenge Blind Report Readers: Everyone 1 Reply

[Show details](#)

## [Replication] A Unified Bellman Optimality Principle Combining Reward Maximization and Empowerment

Akhil Bagaria, Seungchan Kim, Alessio Mazzetto, Rafael Rodriguez-Sanchez

02 Dec 2019 (modified: 10 Dec 2019) NeurIPS 2019 Reproducibility Challenge Blind Report Readers: Everyone 0 Replies

[Show details](#)

## [Re] Unsupervised Object Segmentation by Redrawing [Rev2], NeurIPS 2019 Reproducibility Challenge



← Go to [NeurIPS 2019 Reproducibility Challenge homepage](#)

# [Re] No Press Diplomacy: Modeling Multi-Agent Gameplay

*Daniel Ritter, Dylan Sam, Kevin Du, Shamay G Samuel, Cody West, Aaron Zhang*

02 Dec 2019 (modified: 09 Dec 2019) NeurIPS 2019 Reproducibility Challenge Blind Report Readers:  Everyone

**Abstract:** Diplomacy is a strategic board game where different powers battle over control of supply centers in Europe. The original authors [1] developed supervised learning and reinforcement learning models to learn to play the No Press version of Diplomacy, beating the existing state of the art rule-based bots. The original paper utilizes various different machine and reinforcement learning techniques, including attention, encoder and decoder blocks, graph convolutional networks (GCN), LSTM, and FiLM [2]. Their implementation and code built off of extensive existing software frameworks like DAIDE [3], developed by the Diplomacy research community for interfacing with other bots. Furthermore, the authors have also developed a game engine that provides a simple interface for playing Diplomacy games. Because the authors of the paper released all their code for their models, the paper is not entirely comprehensive with their implementation details. Without being able to refer to their code, these ambiguities proved to make replication fairly difficult. We relied on communication with the paper authors in order to resolve a variety of ambiguities. Ultimately, this report details our attempts to reproduce the paper. We failed to reproduce the results for many reasons, including architecture ambiguities, expensive training times/compute resources required that were unmentioned in the original paper, and the complexity of this project given a 2-month time frame.

**Track:** Replicability

**NeurIPS Paper Id:** [https://openreview.net/forum?id=B1ETuVrgUr&noteId=ByxN8Pfx\\_H](https://openreview.net/forum?id=B1ETuVrgUr&noteId=ByxN8Pfx_H)

**0 Replies**


---

Editorial

# ICLR Reproducibility Challenge 2019

Joelle Pineau<sup>1,2,3</sup>, Koustuv Sinha<sup>1,2,3</sup>, , Genevieve Fried<sup>1</sup>, Rosemary Nan Ke<sup>2,3</sup>, and Hugo Larochelle<sup>4</sup>

<sup>1</sup>School of Computer Science, McGill University, Montreal, Canada – <sup>2</sup>Montreal Institute of Learning Algorithms (Mila), Montreal, Canada – <sup>3</sup>Facebook AI Research (FAIR), Montreal, Canada – <sup>4</sup>Google Brain, Montreal, Canada – <sup>5</sup>Polytechnique Montréal, Montreal, Canada

**Edited by**  
Nicolas Rougier 

**Received**  
04 May 2019

**Published**  
22 May 2019

**DOI**  
10.5281/zenodo.3158244

Welcome to this special issue of the ReScience C journal, which presents results of the 2019 ICLR Reproducibility Challenge (2nd edition). One of the challenges in machine learning research is to ensure that published results are sound and reliable. *Reproducibility*, that is obtaining similar results as presented in a paper, using the same code and data (when available), is a necessary step to verify research findings. Reproducibility is also an important step to promote open and accessible research, thereby allowing the scientific community to quickly integrate new findings and convert ideas to practice. Reproducibility also promotes use of robust experimentation workflows, which can potentially reduce unintentional errors.

**The Challenge** – In support of this, the goal of this challenge was to investigate reproducibility of empirical results submitted to the 2019 International Conference on Learning Representations (ICLR). Primarily, the aim was to assess if the experiments reported

- ✓ Code submission policy
- ✓ NeurIPS 2019 Reproducibility challenge
- Machine Learning Reproducibility Checklist

# ML Reproducibility Checklist

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

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

- A clear description of the mathematical setting, algorithm, and/or model.
- An analysis of the complexity (time, space, sample size) of any algorithm.
- A link to a downloadable source code, with specification of all dependencies, including external libraries.

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

- A statement of the result.
- A clear explanation of any assumptions.
- A complete proof of the claim.

For all **figures** and **tables** that present empirical results, check if you include:

- A complete description of the data collection process, including sample size.
- A link to a downloadable version of the dataset or simulation environment.
- An explanation of any data that were excluded, description of any pre-processing step.
- An explanation of how samples were allocated for training / validation / testing.
- The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- The exact number of evaluation runs.
- A description of how experiments were run.
- A clear definition of the specific measure or statistics used to report results.
- Clearly defined error bars.
- A description of results with **central tendency** (e.g. mean) & **variation** (e.g. stddev).
- A description of the computing infrastructure used.



MA\_link: *For all models and algorithms presented, indicate if you include:  
A clear description of the mathematical setting, algorithm, and/or model.*

**Yes: 97%**

*FT\_exp: For all figures and tables that present empirical results, indicate if you include: A description of how experiments were run.*

**Yes: 89%**

FT\_exp: *For all figures and tables that present empirical results, indicate if you include: A description of how experiments were run.*

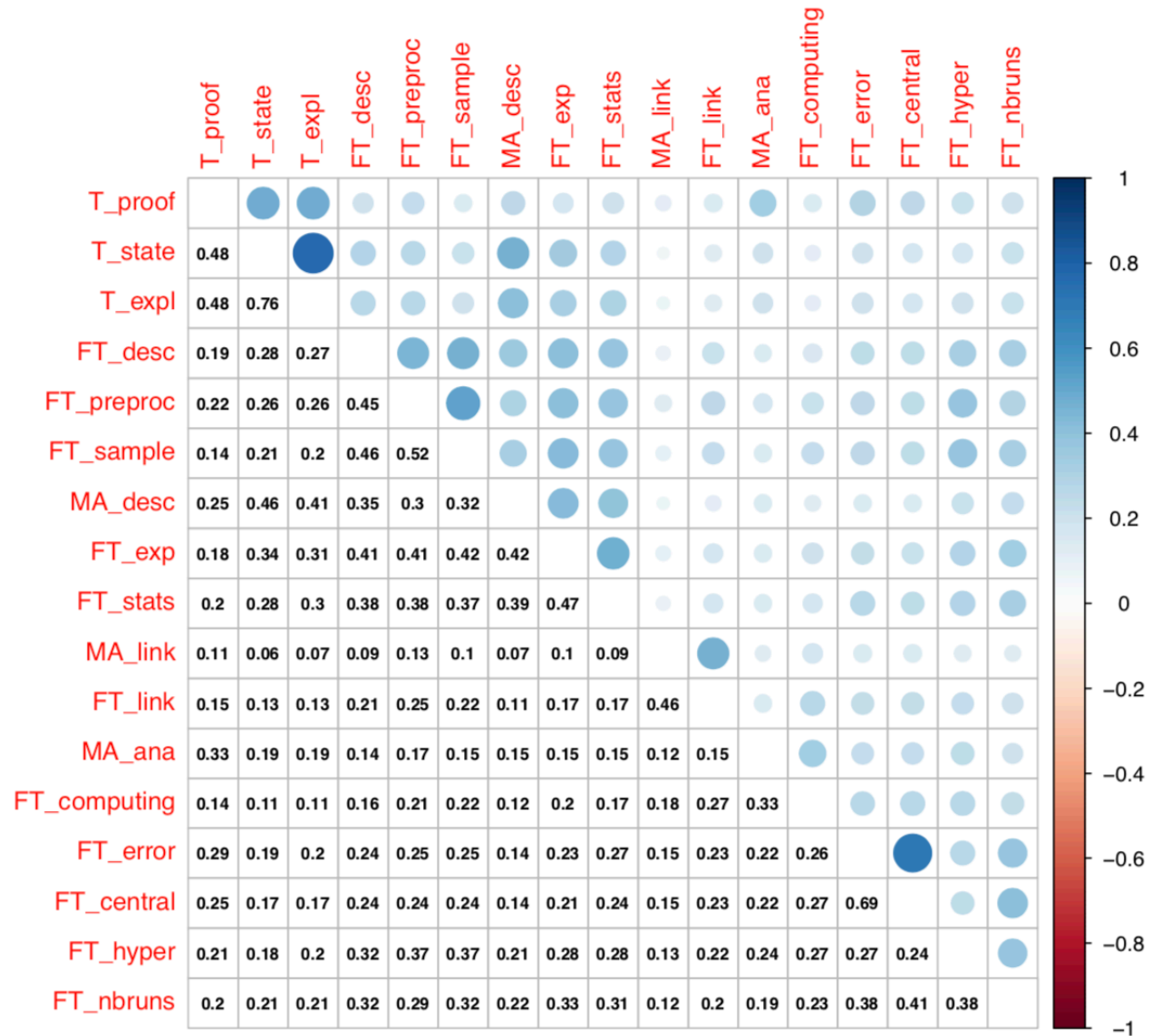
**Yes: 89%**

About 9% of papers indicate **Theory** as the primary subject area.

Association (phi coefficient)  
between checklist questions.

Weak to moderate  
associations between some  
variables.

But overall it seems each  
question captures non-  
redundant information  
about the submissions.



For all figures and tables that present empirical results, indicate if you include:

FT\_stats: *A clear definition of the specific measure or statistics used to report results.*

Yes: 87%  
No: 2%  
N/A: 11%

For all figures and tables that present empirical results, indicate if you include:

FT\_stats: *A clear definition of the specific measure or statistics used to report results.*

Yes: 87%  
No: 2%  
N/A: 11%

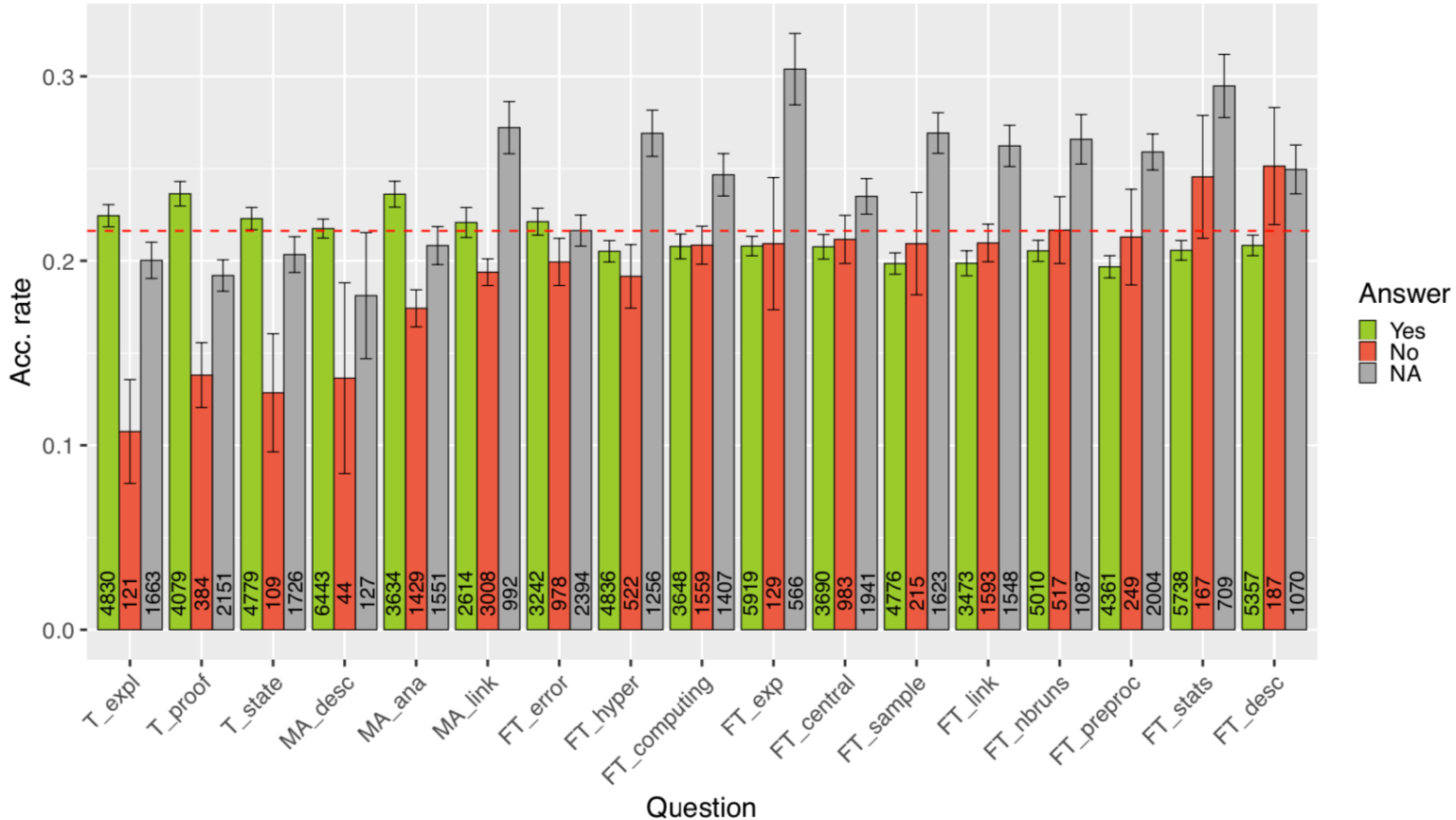
FT\_error: *Clearly defined error bars.*

Yes: 49%  
No: 15%  
N/A: 36%

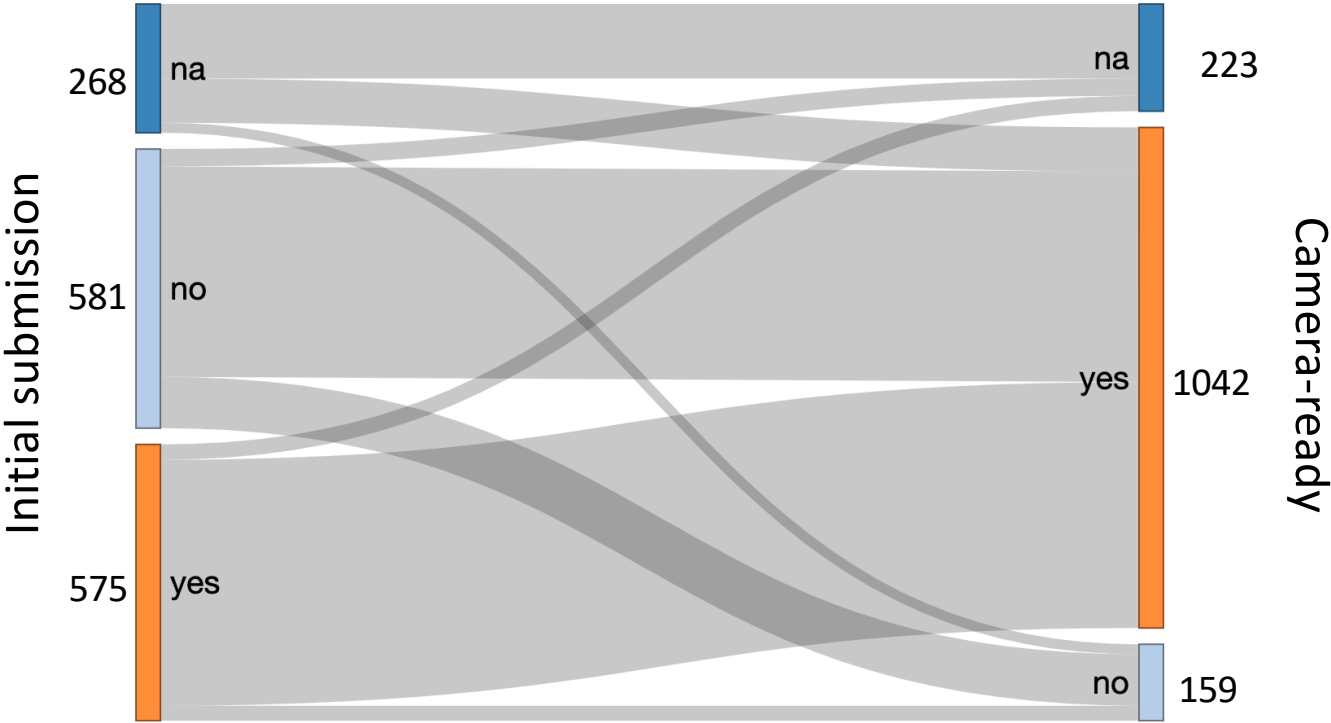
FT\_central: *A description of results with central tendency (e.g. mean) & variation (e.g. stddev).*

Yes: 56%  
No: 15%  
N/A: 29%

# Association between answer & acceptance rate

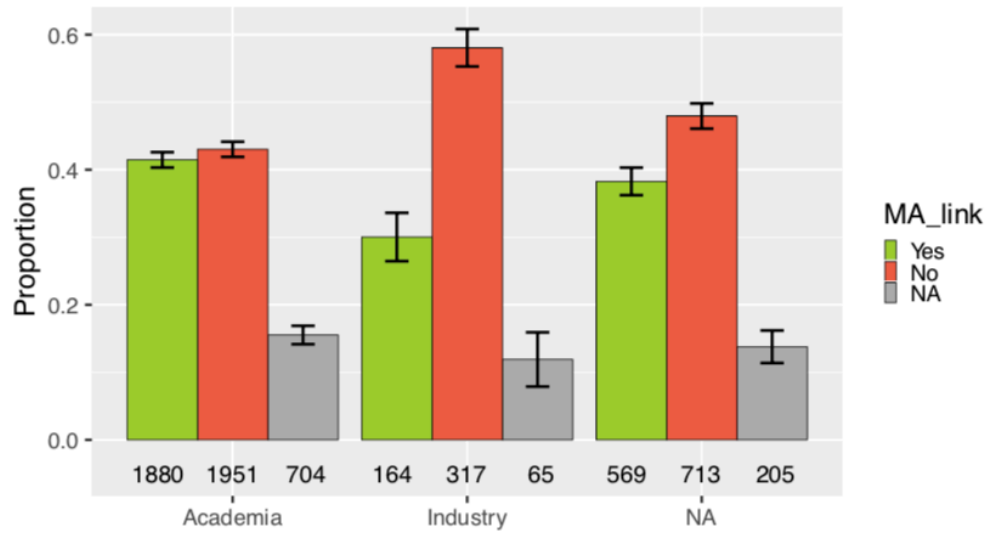


MA\_link: For all models and algorithms presented, indicate if you include: A link to a downloadable source code, with specification of all dependencies, including external libraries.

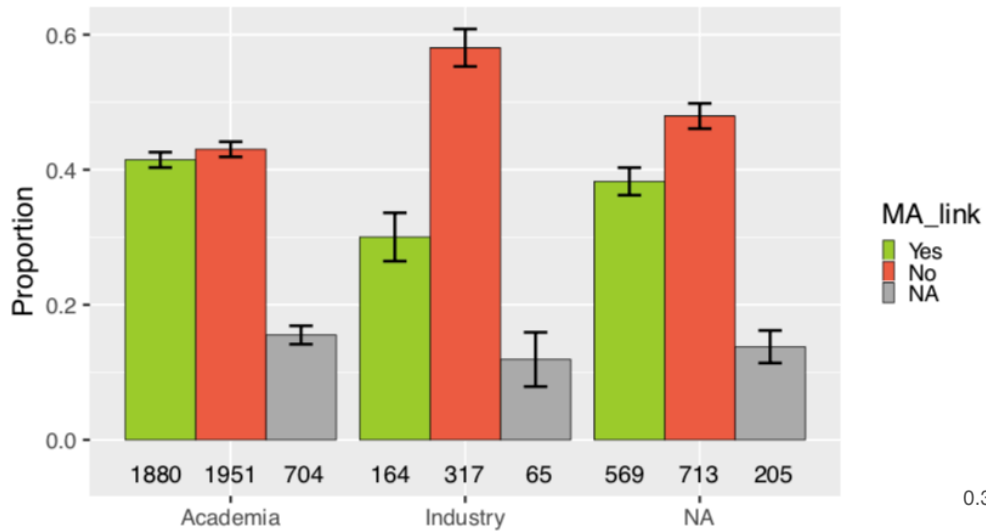




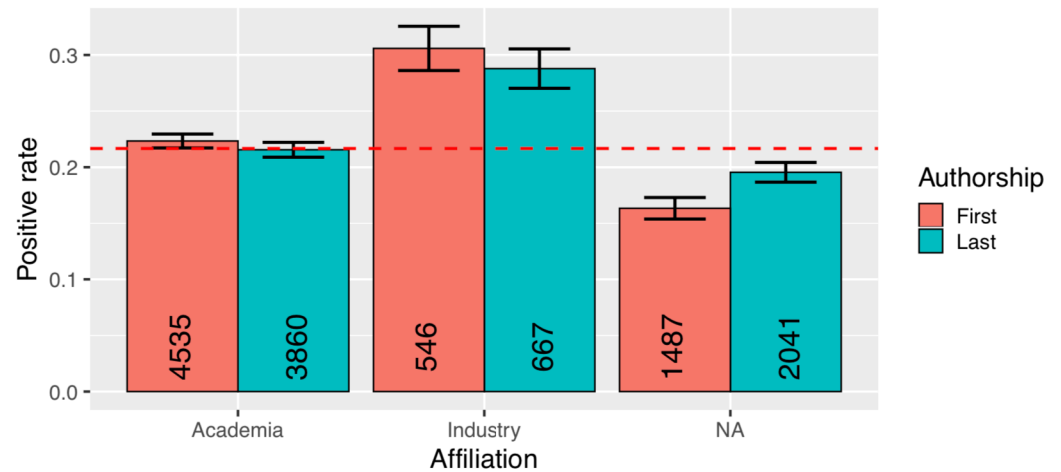
Lower code availability from industry authors (at submission).



Lower code availability from industry authors (at submission).



But... acceptance rate remains higher for (first / last) authors from industry.



What did the reviewers think?

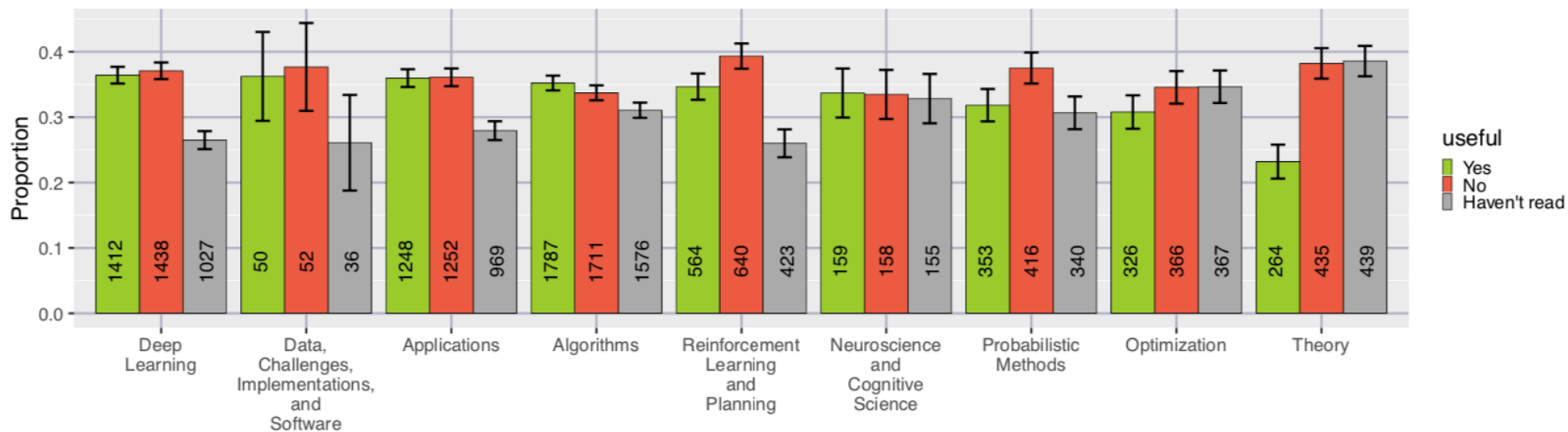
Review form question: *Were the Reproducibility Checklist answers  
useful for evaluating the submission?*

Yes: 34%

Review form question: *Were the Reproducibility Checklist answers*

*useful for evaluating the submission?*

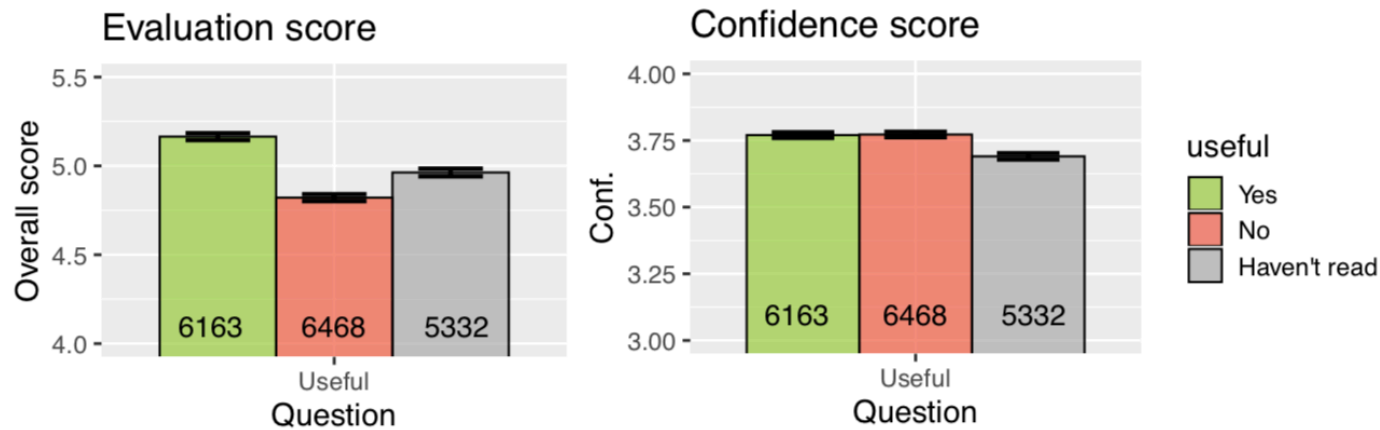
Yes: 34%



Review form question: *Were the Reproducibility Checklist answers*

*useful for evaluating the submission?*

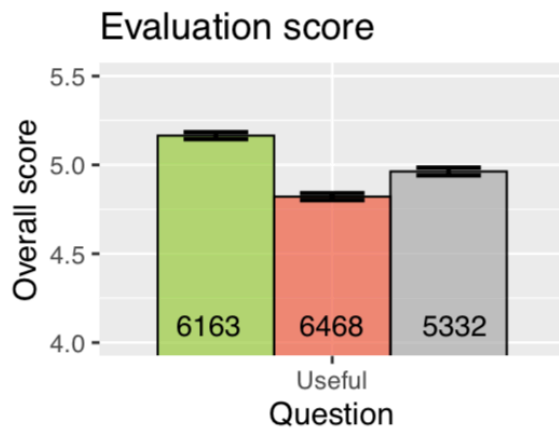
Yes: 34%



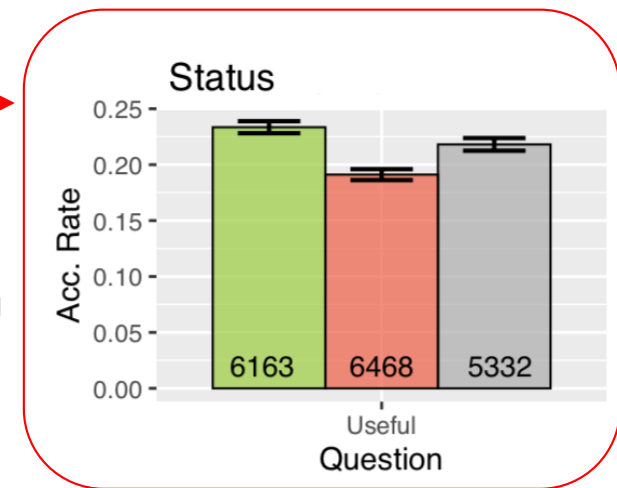
Review form question: *Were the Reproducibility Checklist answers*

*useful for evaluating the submission?*

Yes: 34%



useful  
Yes  
No  
Haven't read



From review form:

*Was code provided (e.g. in the supplementary material)?*

*Yes: 5298*



From review form:

*Was code provided (e.g. in the supplementary material)?* *Yes: 5298*

*If provided, did you look at the code?* *Yes: 2255*

*If provided, was the code useful in guiding your review?* *Yes: 1315*

From review form:

*Was code provided (e.g. in the supplementary material)?* *Yes: 5298*

*If provided, did you look at the code?* *Yes: 2255*

*If provided, was the code useful in guiding your review?* *Yes: 1315*

*If not provided, did you wish code had been available?* *Yes: 3881*

**1e-08**

p-value of code availability on reviewer score

*Indicate if you include a description of computing infrastructure used.*

**Yes: 88%**

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

From: <https://github.com/WhitakerLab/ReproducibleResearch>

# Next steps?

- Reproduce the reproducibility program at other major conferences.
  - ICML 2020, NeurIPS 2020.
  - Variants are being explored at other conferences.
- Open discussion with the community on best practices for conducting research, reporting findings, reviewing & evaluation.
- New models for research verification & certification.
- Zone in on a few specific questions to run controlled experiments.

Thank you!