Booz | Allen | Hamilton

A STEP TOWARD QUANTIFYING INDEPENDENTLY REPRODUCIBLE MACHINE LEARNING RESEARCH

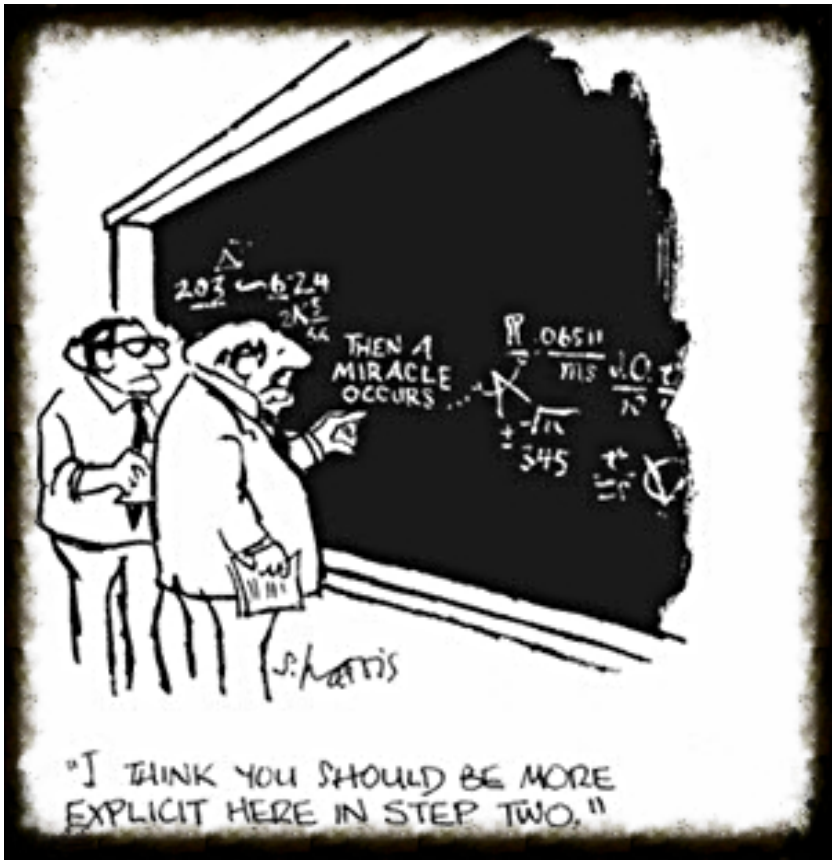
Edward Raff

12/2019, NEURAL INFORMATION PROCESSING SYSTEMS

REPRODUCIBLE MACHINE LEARNING

The machine learning community is rightfully putting a greater emphasis on reproducible research.

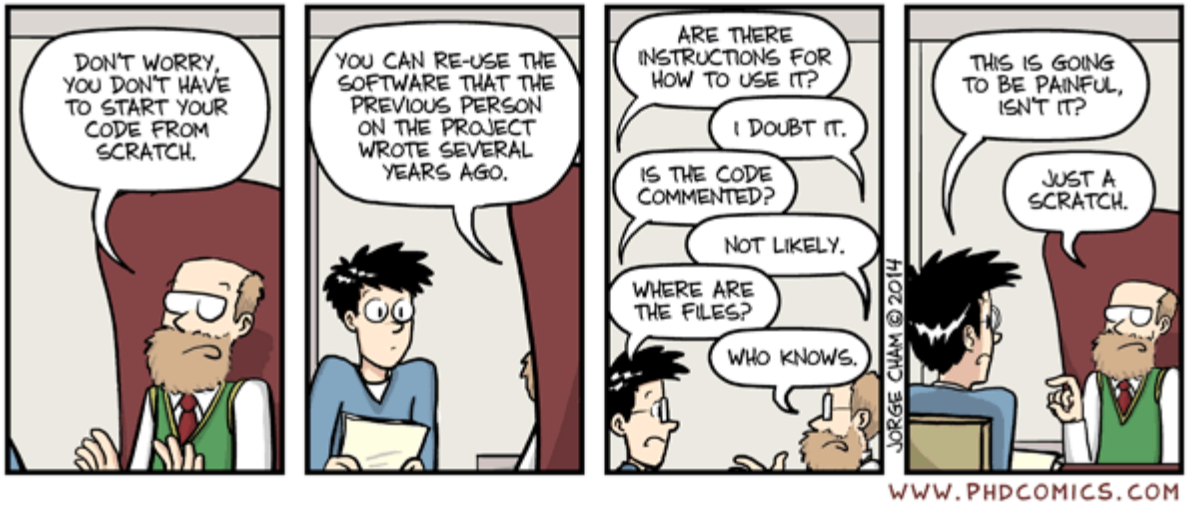
- "*The booming field of artificial intelligence (AI) is grappling with a replication crisis*" - Hutson, Matthew (2018)
  doi:10.1126/science.359.6377.725
- Our results require code and data, which can be shared electronically. It seems like this should be easier for us.
- Many works are being conducted around this belief. Better tools for hyper-parameter tuning in a reproducible way, sharing code, dockerizing artifacts, etc.
- Unfortunately, most of this work is going off intuition. All the current effort is valuable and should be lauded, but how do we quantify these questions?



Cartoon created by Sidney Harris (The New Yorker).

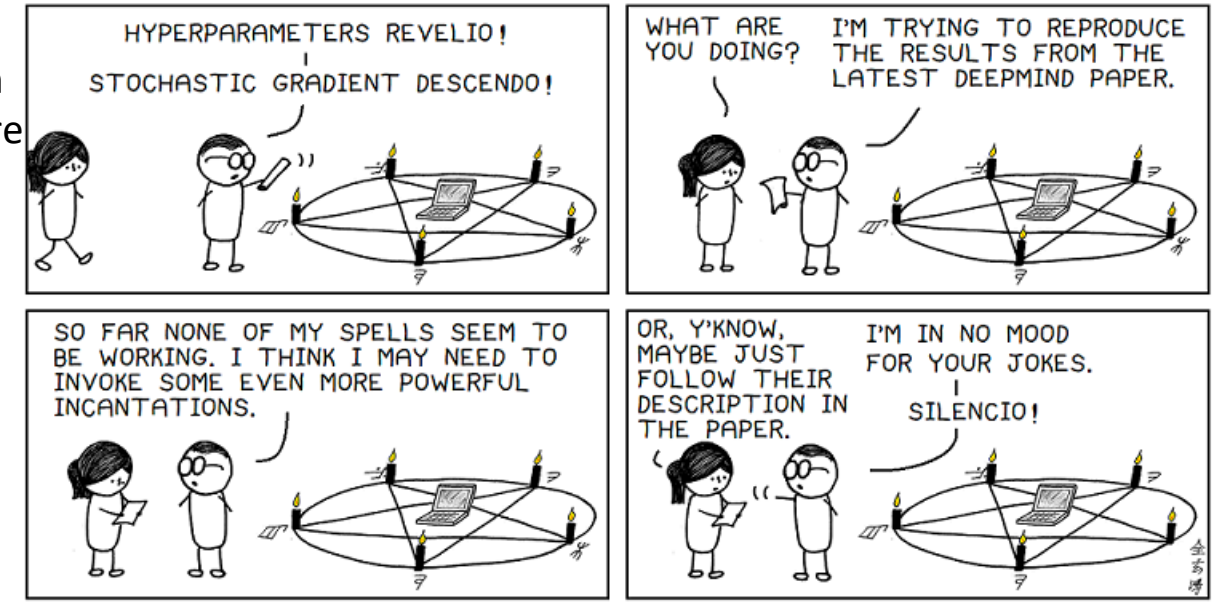
INDEPENDENTLY REPRODUCIBLE

- If authors release code and data, *replicating* their results we enter a software engineering problem. This is valuable and good. But is it *sufficient*?
- We argue no, it is not. If a paper is scientifically sound it should be possible to reproduce the results without use of the author's code.
  - See *Replicability is not Reproducibility: Nor is it Good Science* (2009)
- We want to quantify what we will call *independent reproducibility*, where we seek to reproduce the results of a paper without using that paper's code.
- To do this, we need to attempt reproductions of several papers, while simultaneously quantifying information about each paper. We did this for 255 papers.



OUR STUDY DESIGN

- Attempt to independently reproduce results of 255 paper, succeeded 63.5% of the time.
- Papers published from 1984-2017, reproduction attempts performed from 2012-2017
- If we ever looked at another implementation before reproduction, the attempt was disqualified
- Developed 26 quantifications, grouped by Objective, Mild Subjective, & Subjective
  - Developed a protocol for every feature to minimize subjectivity
- Study made possible by paper organization & note taking software that was used early on.
- Results analyzed with non-parametric statistical hypothesis testing



https://abstrusegoose.com/588

SOME RESULTS, AT A HIGH LEVEL

There is no apparent correlation with the year we attempted to reproduce a paper. This makes our analysis easier. Some results with too little discussion:

- *No relation between reproduction and year attempted, suggesting issues are perhaps not new or fears overblown – depending on perspective*
- *Papers that have significant empirical emphasis, are more reproducible than ones that emphasize proofs and theorems in their work.*
- *The emphasis on hyper-parameter specification is well placed by the community.*
- *Having no pseudo code is just as reproducible as having code-like descriptions. Describing your method as high-level steps is worse.*
- *Authors replies result in 85% reproduction rate. No reply goes down to 4%.*

Table 1: Significance test of which paper properties impact reproducibility. Results significant at $\alpha \leq 0.05$ marked with"*".

Feature	p-value
Year Published	0.964
Year First Attempted	0.674
Venue Type	0.631
Rigor vs Empirical*	1.55×10^{-9}
Has Appendix	0.330
Looks Intimidating	0.829
Readability*	9.68×10^{-25}
Algorithm Difficulty*	2.94×10^{-5}
Pseudo Code*	2.31×10^{-4}
Primary Topic*	7.039×10^{-4}
Exemplar Problem	0.720
Compute Specified	0.257
Hyperparameters Specified*	8.45×10^{-6}
Compute Needed*	8.75×10^{-5}
Authors Reply*	6.01×10^{-8}
Code Available	0.213
Pages	0.364
Publication Venue	0.342
Number of References	0.740
Number Equations*	0.004
Number Proofs	0.130
Number Tables*	0.010
Number Graphs/Plots	0.139
Number Other Figures	0.217
Conceptualization Figures	0.365
Number of Authors	0.497

STUDY DEFICIENCIES

There are more results here than we have time to discuss, and our paper has likely not yet elucidated all insights that could be obtained from the data. But, we must also take all results with some salt due to study biases.

- All reproductions attempts where done by one author, who is not an expert in all the topic areas attempted, and does not have unlimited time.
- Papers studied are not randomly sampled, but biased toward personal interests, as well as what has become popular over time.
- We have not yet factored into our analysts anything about the authors of the papers under analysis, which would likely have a significant impact on the results.

In particular, after performing this work, we note a fundamental problem with the question framing: that a reproducibility is a binary property that paper has or does not have. One particular paper under analysis took 4.5 years to successfully reproduce.

In this light, perhaps we should look at reproducibility as a kind of survival analysis? Reproduction is the "death" of a paper, and a paper that fails reproduction "survives" indefinitely. The survival rate becomes the effort and time needed to reproduce, conditioned on properties of both the paper (e.g., what we have quantified) as well as the author and their resources.

QUESTIONS?

We've performed the first quantification of what makes a machine learning paper reproducible by an independent party.

We expect this to lead to debate, and do not claim to authoritatively answer these questions.

This is the start point, and we need more people to start quantifying and tracking this information from their own efforts. So that we can form a less biased study and further our field.

Raff_Edward@bah.com
@EdwardRaffML
EdwardRaff.com