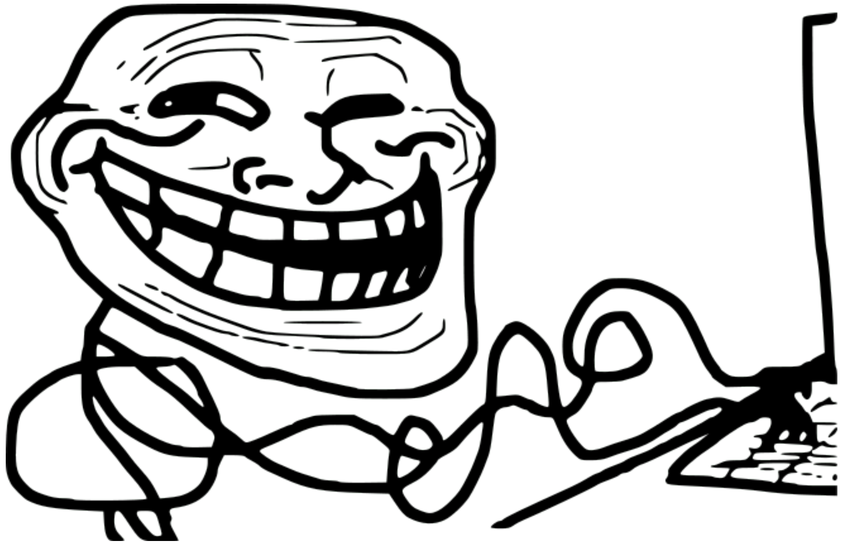
PASC23
Davos Switzerland | 26-28 June 2023

Anti Patterns of Scientific Machine Learning to Fool the Masses

A Call for Open Science

LorenaABarba. @labarba@fosstodon.org



About me

http://lorenabarba.com

- SC19 Reproducibility Chair; JupyterCon 2020 General Chair
- NASEM committee “Reproducibility and Replicability in Science” and NASEM committee “Open Source Software Policy Options for NASA”
- NumFOCUS Board of Directors, 2014-2021
- Founding editor and past AEiC of The Journal of Open Source Software
- Editor-in-Chief of IEEE Computing in Science and Engineering
- Author “Reproducibility PI Manifesto”

Scientific Machine Learning for Complex Systems: Beyond Forward Simulation to Inference and Optimization

Noyce Conference Room
Workshop
All day
October 10, 2022 – October 12, 2022

Organizers

Karen Willcox

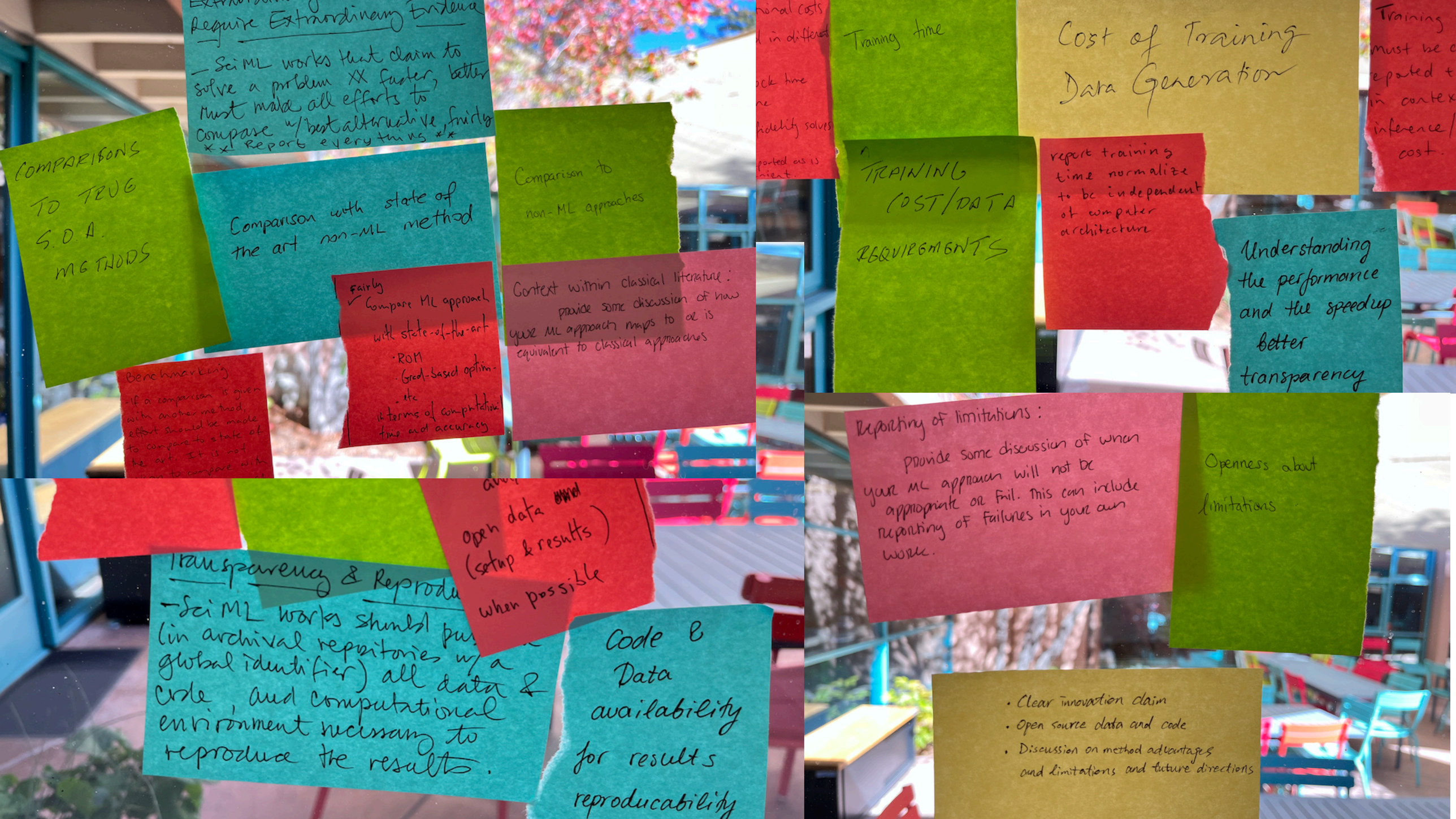
Science Board, External Professor

Omar Ghattas

Professor of Mechanical Engineering and Geological Sciences at the University of Texas at Austin

Joaquim Martins

Bart van Bloemen Waanders

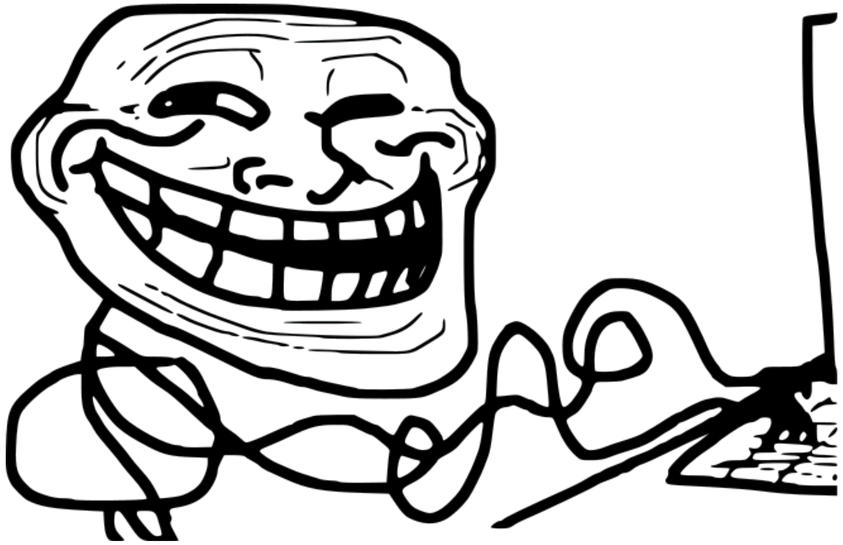

Require Extraordinary Evidence
- SciML works that claim to solve a problem XX faster, better must make all efforts to compare w/ best alternative, fairly
** Report everything **
COMPARISONS TO TRUE S.O.A. METHODS
Comparison with state of the art non-ML method
Comparison to non-ML approaches
Benchmarking
If a comparison is given with another method, effort should be made to compare to state of the art. It is not
Fairly
Compare ML approach with state-of-the-art
- ROM
- Grad-based optim-
- etc
in terms of computational time and accuracy
Context within classical literature:
provide some discussion of how your ML approach maps to or is equivalent to classical approaches
Training time
Cost of Training Data Generation
TRAINING COST/DATA REQUIREMENTS
report training time normalize to be independent of computer architecture
Understanding the performance and the speedup better transparency
Reporting of limitations:
provide some discussion of when your ML approach will not be appropriate or fail. This can include reporting of failures in your own work.
Openness about limitations
Transparency & Reproducibility
- SciML works should publish (in archival repositories w/ a global identifier) all data & code, and computational environment necessary to reproduce the results.
open data (setup & results) when possible
Code & Data availability for results reproducability
• Clear innovation claim
• Open source data and code
• Discussion on method advantages and limitations and future directions

PASC23
Davos Switzerland | 26-28 June 2023

Anti Patterns of Scientific Machine Learning to Fool the Masses

A Call for Open Science

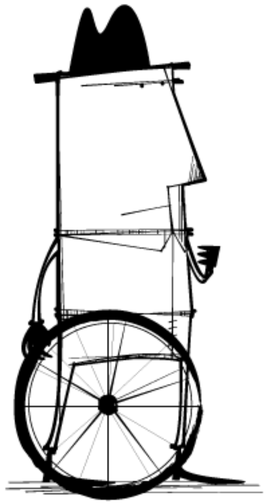
LorenaABarba. @labarba@fosstodon.org

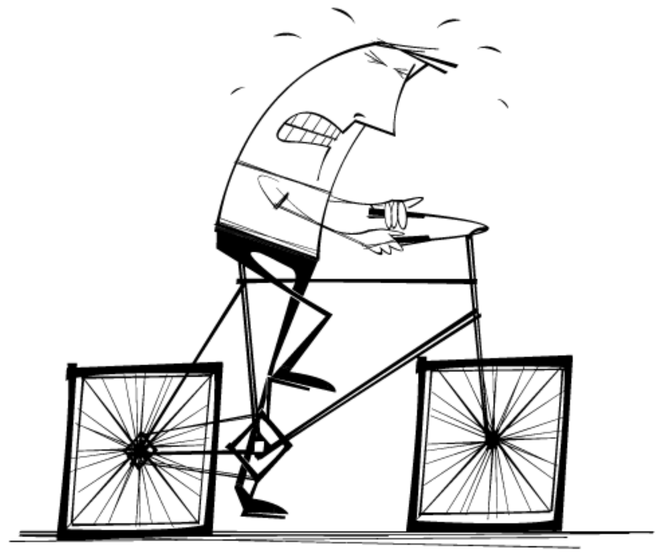


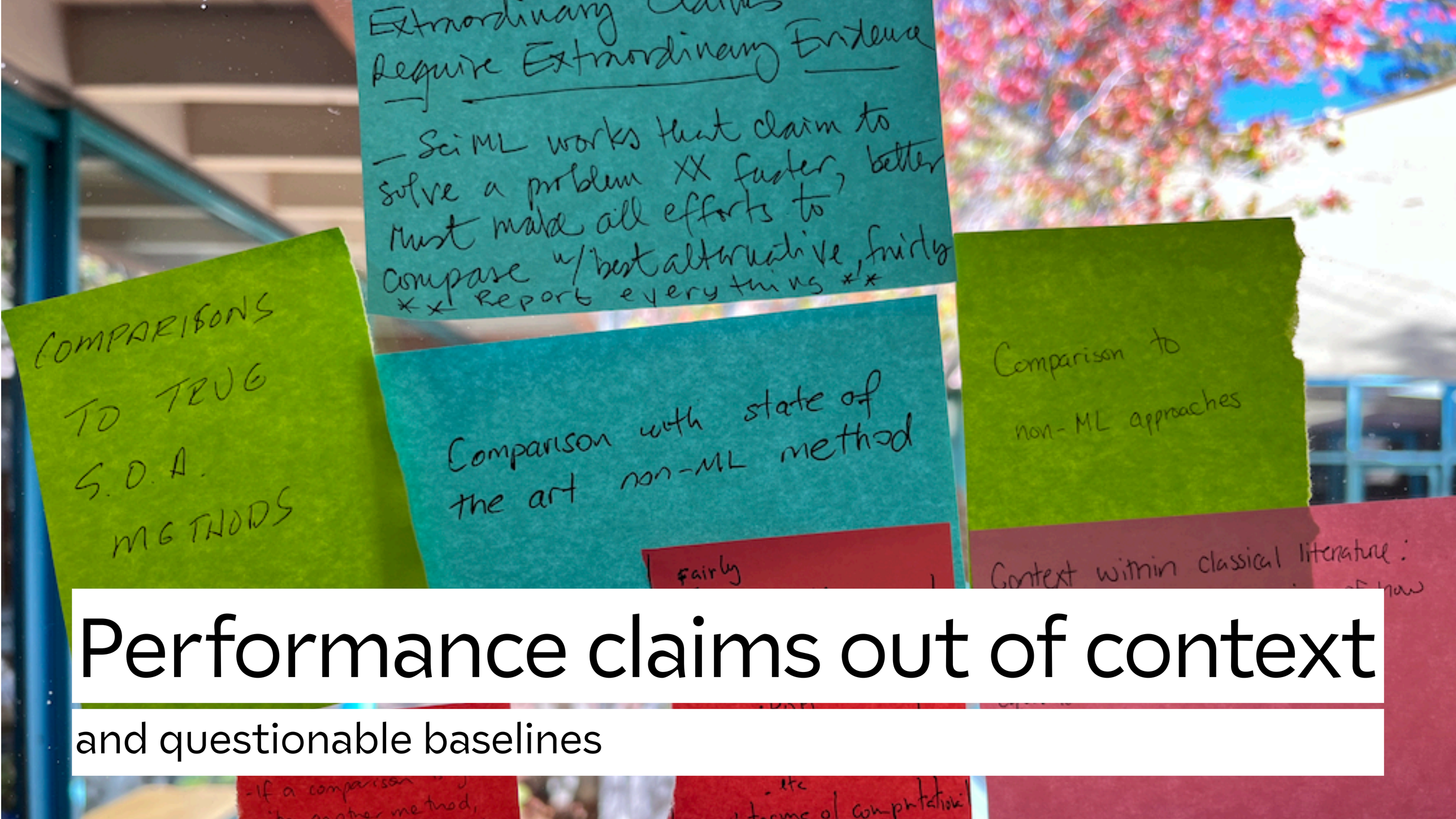
Patterns and anti-patterns

Terms from software engineering

- It is recurrent (“rule of three”)
- It has bad consequences
- A better solution exists





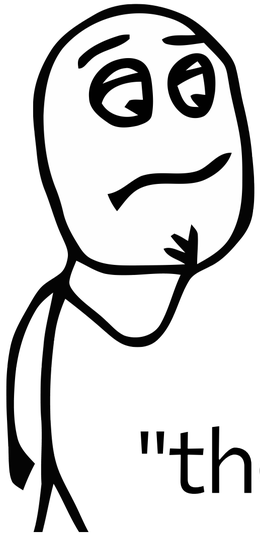


Performance claims out of context

and questionable baselines

“data-driven 🔥 gives accurate solutions with a dramatic drop in required resolution … 4x to 8x coarser than is possible with standard methods”

- coarser resolution … but at what cost?
- what are those “standard methods” you speak of?



"the learned model is clearly far superior to the polynomial approximation, demonstrating that the spatial resolution required ... can be greatly reduced..."

- far superior to a method that is known to be poor: bad “baseline"
- better methods shown, but all claims are compared with the worst method as “baseline”

Ryan Abernathey · Mar 30, 2021
@rabernat · **Follow**
Replying to @LorenaABarba @shoyer and @NatComputSci

This paper (on which Stephan is a co-author) is a great example:

pnas.org
Learning data-driven discretizations for partial differential equations | ...

Timo Betcke @TimoBetcke@fosstodon.org
@BetckeTimo · **Follow**

Just skimmed through this. It seems that ML in this case was compared as approximation tool to a bad way of doing the approximation, namely polynomial interpolation in what looks like from the figure equispaced points (sorry if I am wrong, haven't read the details).

12:49 PM · Apr 1, 2021

Reply **Copy link**

Read 1 reply

Lorena Barba @labarba@fosstodo...
@LorenaABarba · **Follow**

As a CFD expert, I am disappointed when a paper claims their method is “far superior” to others, or it can “greatly reduce” grid resolutions... without giving readers all the numbers, straight up, to go along with those claims.

12:33 PM · Apr 1, 2021

4 **Reply** **Copy link**

“the overall agreement between [NN-based method] and [commercial solver] is very good”

- shows line plot for a quantity of interest with each method: "eyeball metric”
- no mention of runtimes at all

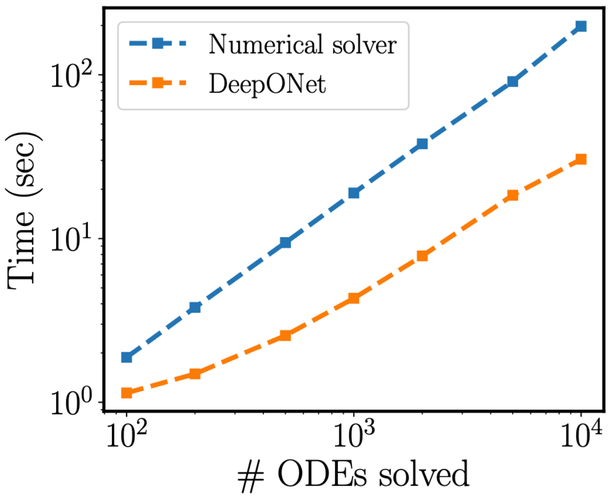
“...a novel and fast approach (1000x) to learning the solution operator of a PDE...”

- what does that 1000x mean? what was it compared against?
- is the comparison point a competitive implementation within its own class of methods?

"...a new approach... effective in performing accurate long-time simulations for a wide range of parametric ODE and PDE systems..."

- what is the numerical solver being compared against?

In fact, Julia's numerical solver is 7,000x faster, just running on CPU
(Source: Chris Rackauckas, MIT)



Incomplete reporting

E.g., full computational cost, data generation

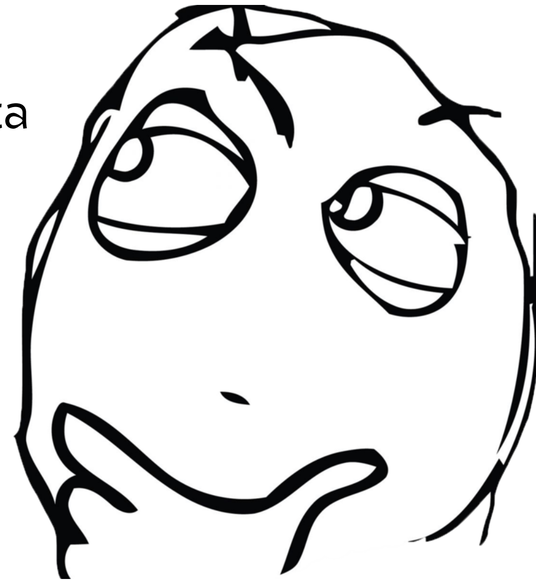
“we first generate a training set of high-resolution data and then learn…”

- how is the data generated, and at what cost?
- what is the cost of training?

Step 4 of the “12 steps to Navier-Stokes”

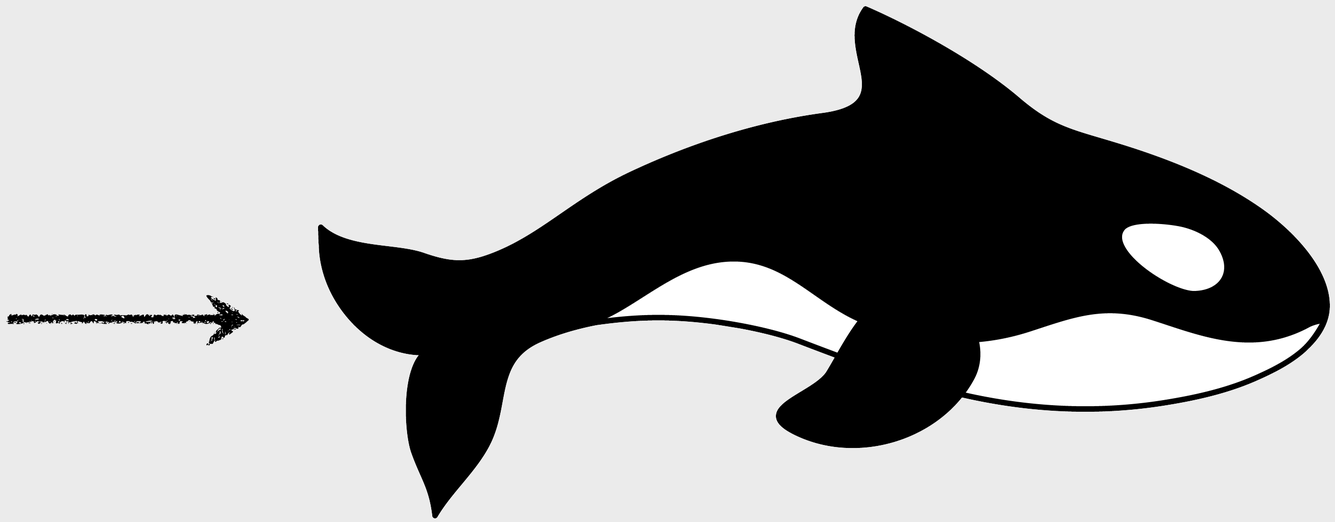
“the data for the N-S equation is obtained by the direct numerical simulation…”

- zero discussion of anything about the DNS solver used to generate training data
- no mention of computational cost of generating data



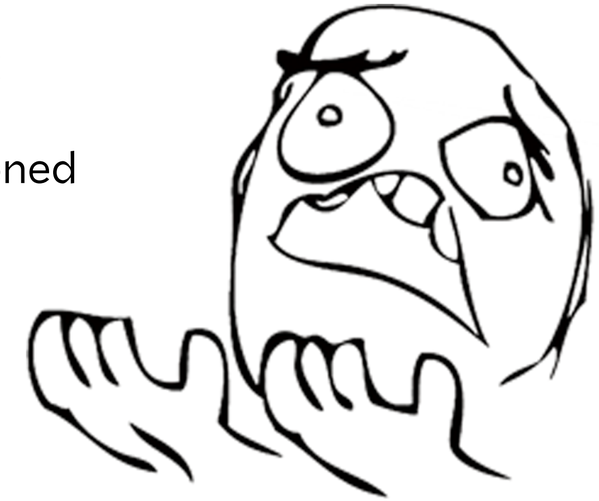
Renaming old things

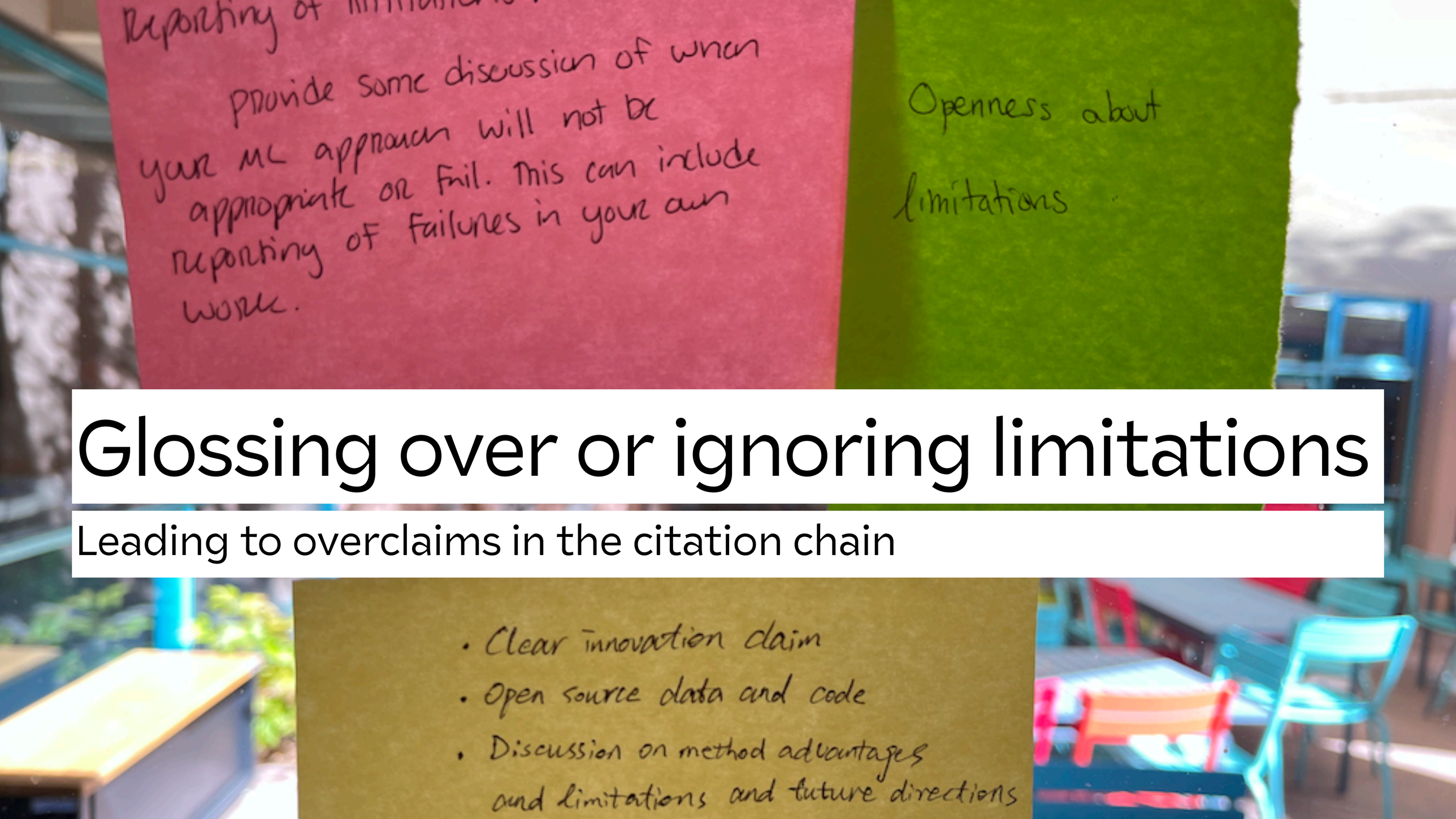
just add a NN somewhere and call it “deep”

Panda fish 

“Deep random vortex method… a novel physics-informed machine learning framework…”

- it is the classic random vortex method (Chorin 1973): vorticity equation + random walk
- use a NN to represent the velocity field (obtained from vorticity via integral equation)
- state-of-the-art is to compute velocity with fast multipole method at O(N): not mentioned





Glossing over or ignoring limitations

Leading to overclaims in the citation chain

We use 🔥 method “to directly simulate incompressible flows… including…two-dimensional cylinder wake”

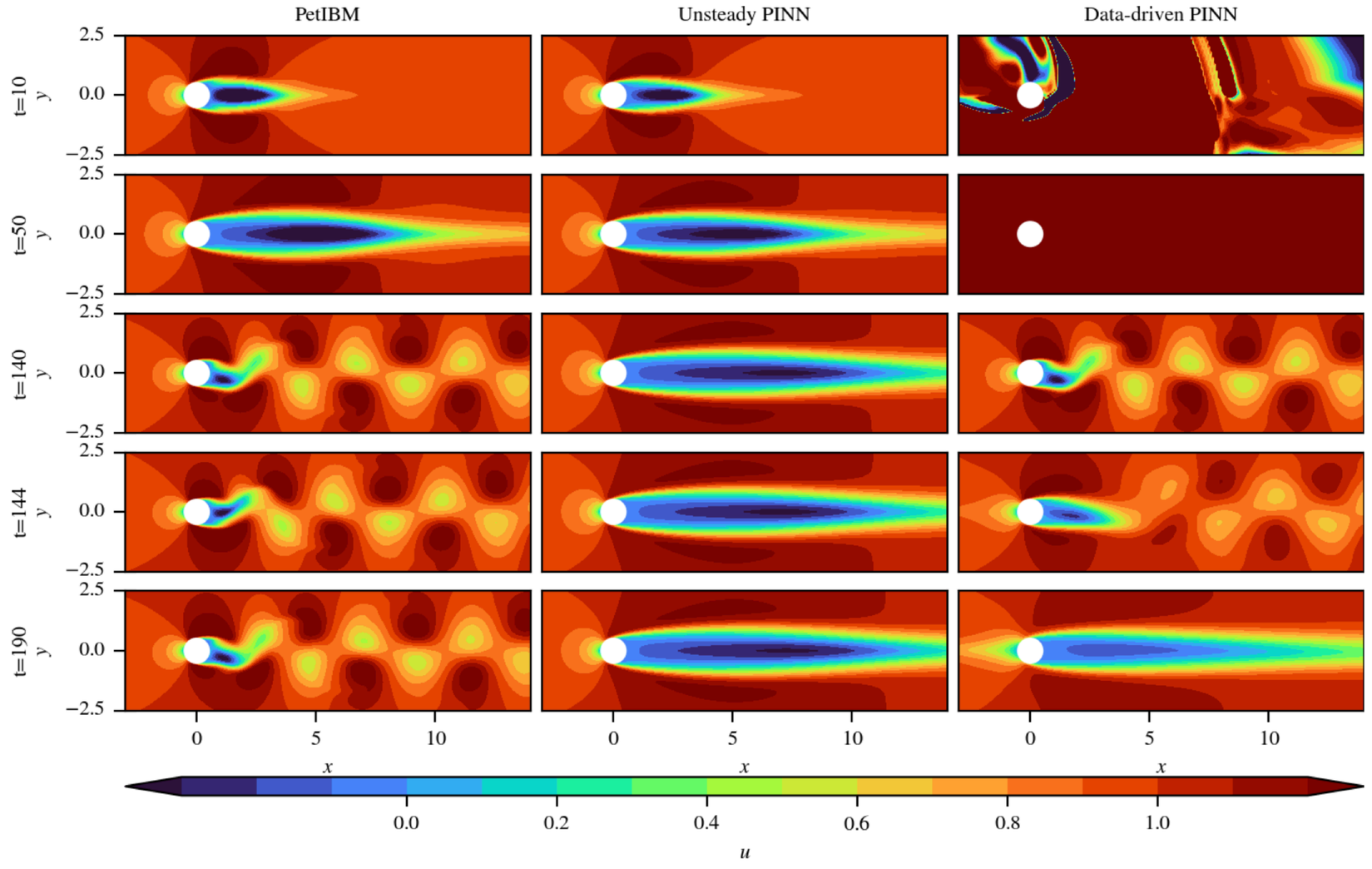
- DNS data provided boundary conditions for the training
- the cylinder was not even present in the domain
- no discussion of this limitation

Computer Science > Computational Engineering, Finance, and Science

[Submitted on 31 May 2023]

Predictive Limitations of Physics-Informed Neural Networks in Vortex Shedding

Pi-Yueh Chuang, Lorena A. Barba



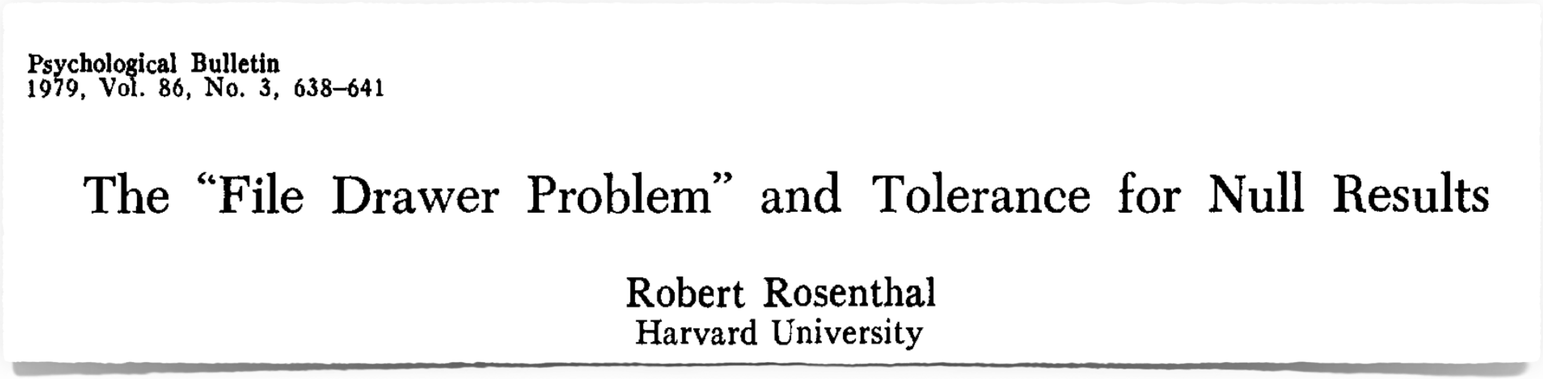
Closet failures

A.k.a., the file-drawer problem (publication bias)

Publication bias

"the file-drawer problem"

- Publish positive results
- File away negative results



Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

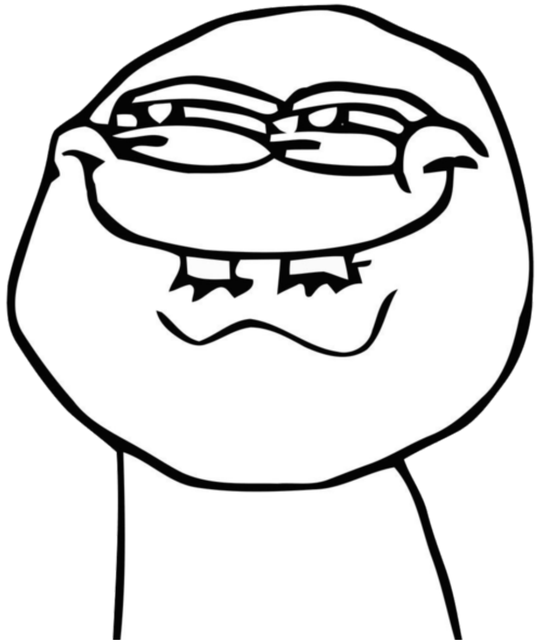
Robert Rosenthal
Harvard University

Lack of transparency

and irreproducible results

“Data available upon reasonable request”

Leaving the data or code preparation for a later time of "request" is too late!



PNAS

An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden[a,1], Jennifer Seiler[b], and Zhaokun Ma[b]

...only 44% of requests led to receiving data and/or code from the original authors

https://doi.org/gc8gkw

PLOS ONE

RESEARCH ARTICLE

A funder-imposed data publication requirement seldom inspired data sharing

Jessica L. Couture[1,2]*, Rachael E. Blake[2,3], Gavin McDonald[1,4], Colette L. Ward[2,5]

...could recover data in just 26% (N=315) of cases

https://doi.org/gdts9v

To be FAIR, a GitHub code repo is not enough

FAIR = findable, accessible, interoperable, reusable

- Findable means an archival deposit with a DOI (e.g., Zenodo)
- Accessible means retrievable by the identifier using open protocols
- Interoperable means well structured metadata that is machine-actionable
- Reusable implies a proper license

None of these is achieved by Supplementary Materials!
(where data goes to die)

Lorena Barba @labarba@fosstodon.org
@LorenaABarba · **Follow**

The SIAM Journal on Scientific Computing (SISC) now offers a "Reproducibility Badge: code and data available" for eligible articles. But what they consider "available" is not up to #reproducibility standards—an archival deposit with DOI should be required! epubs.siam.org/journal/sisc/i...

Guidelines for SISC Reproducibility Badges

- Authors can request the **"SISC Reproducibility Badge: code and data available"** at the time of manuscript submission
- **Criteria for obtaining the "code and data available" badge:**
  - Authors make all computer code and data publicly available that implement the computational methods proposed
  - Authors should aim to include all parameter settings, either in the code or in separate data files, that allow readers to reproduce all numerical results presented in the paper (including all tables, figures, ...)
  - In a README file, authors include a brief description of the material provided and how to use it
- **Acceptable mechanisms for making the code and data available**:
  - Publicly available permanent repository such as github, bitbucket, or similar
  - Supplementary materials that appear with the published SISC paper
    Note: authors' academic websites and similar are not eligible locations.
- **Guidelines for public repositories** such as github, bitbucket:
  - Provide the URL to your github or bitbucket repository when requesting the badge during manuscript submission

1:34 PM · Feb 14, 2023

Open code and open data are not enough

How to achieve transparency of the research workflow?

- Data provenance, stewardship, documentation, version control
- Computational environment, including all library versions (better: standard env file)
- Tools for reproducing results via virtualization, cloud computing, packaging, containers (e.g., Docker, Singularity/Apptainer)
- Automatic capture of computational details; workflow management systems

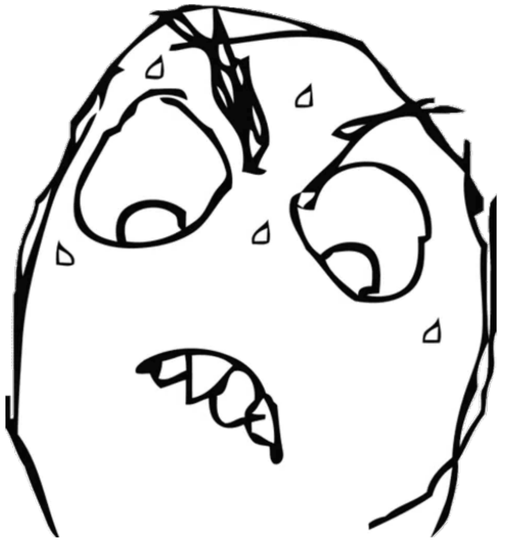
Gatekeeping

(Not just a SciML thing.)

Hypothetical scenario

You are new to this, but your talented PhD student is working on 🔥

- Months of painstaking work. Results disappointing.
- Why does it not work? Let’s write it up anyway.
- Prepare to present at a conference. Post preprint on arXiv.
- 24h later: you get an angry email from big shot about your “erroneous paper” — and it is copied to 15 people (including your department chair!)



Brian Nosek (@briannosek@nerdcu... · May 25, 2023

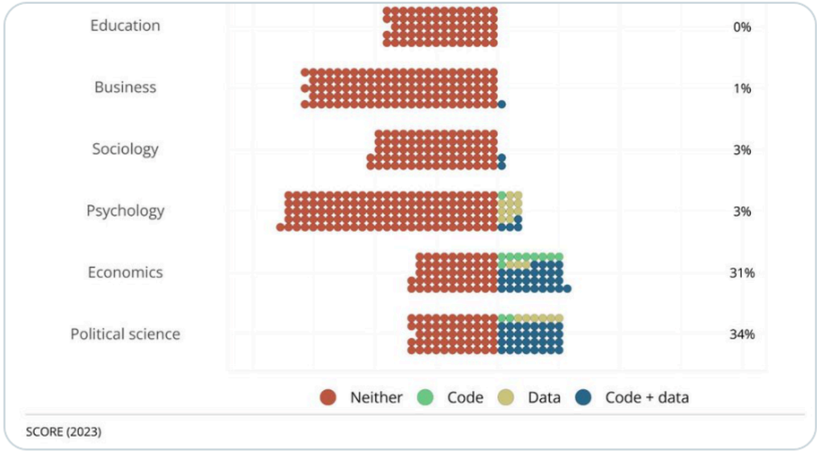
@BrianNosek · **Follow**

Replying to @BrianNosek

Chapter 1: Tim Errington summarized the challenge -- in every field that has looked, reproducibility, robustness, and replicability are weaker than expected or desired.

It included a sneak preview of SCORE reproduction and replication results.

15-min: youtube.com/watch?v=oHpzm8...

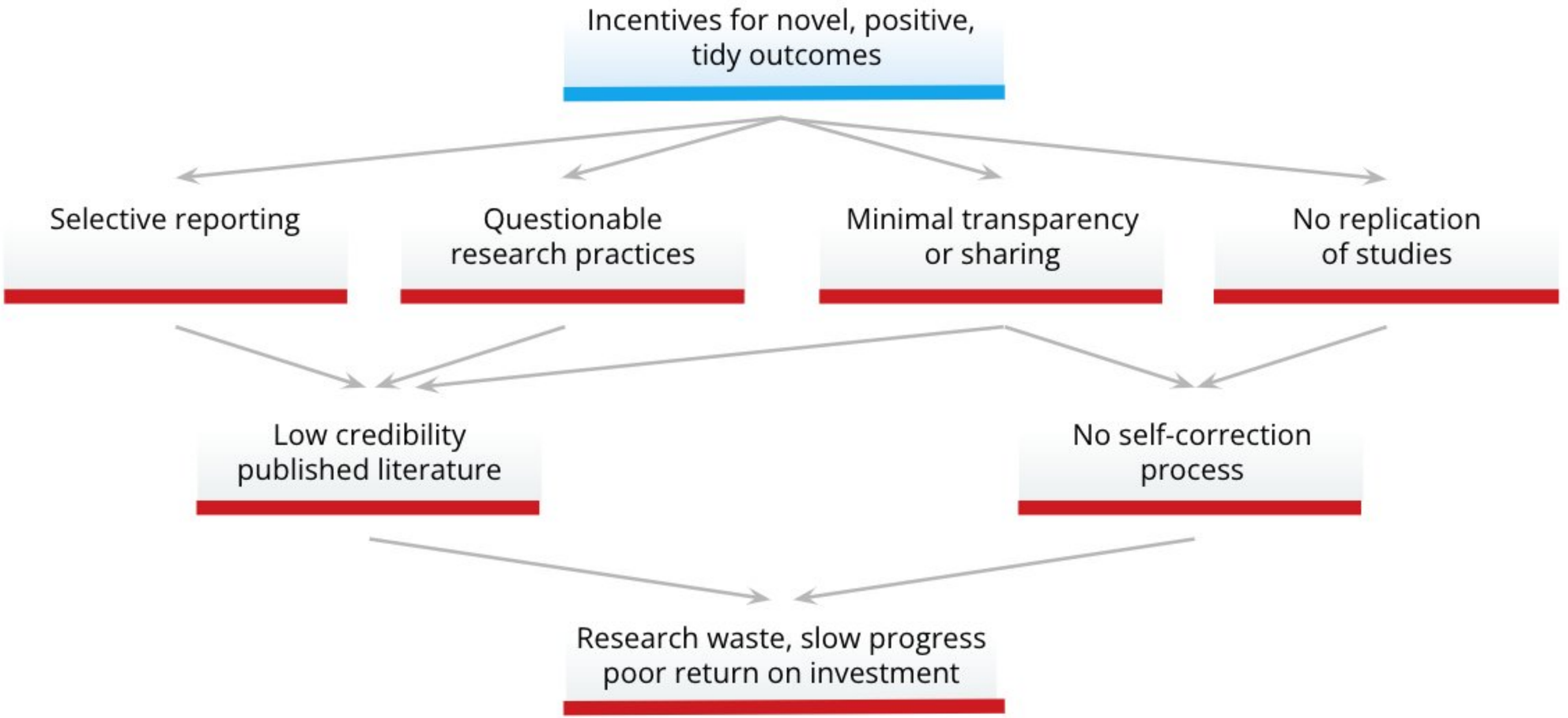


Brian Nosek (@briannosek@nerdculture.de)

@BrianNosek · **Follow**

Chapter 2: I discussed why the known solutions to these challenges have not been adopted, and laid responsibility for the intransigence on the reward system.

A dysfunctional reward system



Credit: Brian Nosek, Center for Open Science

THE WHITE HOUSE
WASHINGTON

YEAR OF
OPEN SCIENCE

✦ NASA ✦ NSF ✦ NOAA ✦
✦ DOE ✦ GSA ✦ NEH ✦ NIH ✦
✦ NIST ✦ USDA ✦ USGS ✦

A call for Open Science

We are in the Year of Open Science!

What is Open Science?

Open science “aims to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms that were used to generate those data”

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. Open Science by Design: Realizing a Vision for 21st Century Research. https://doi.org/gfxzc4

Vision for EU 2016

“Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools.”

https://doi.org/gk7tw3

Openness is about the possibilities of communicating with other people. It’s not about *stuff*, what you do with stuff. It’s about what you do with each other

— Stephen Downes, 2017

https://youtu.be/FPHYAFcUziA

“Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks “

Vicente-Saez, R. and Martinez-Fuentes, C., 2018. Open Science now: A systematic literature review for an integrated definition. Journal of Business Research, 88, pp.428-436. https://doi.org/gc5sjb

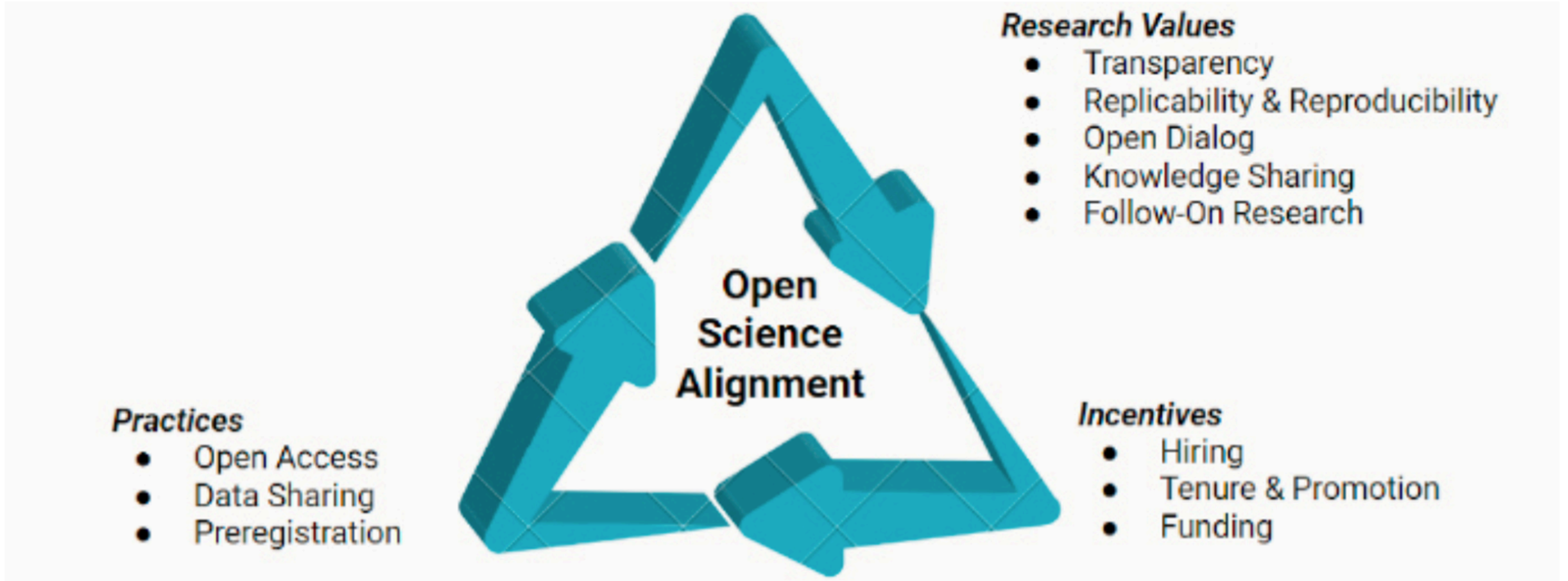


FIGURE 1-1 Open science alignment.

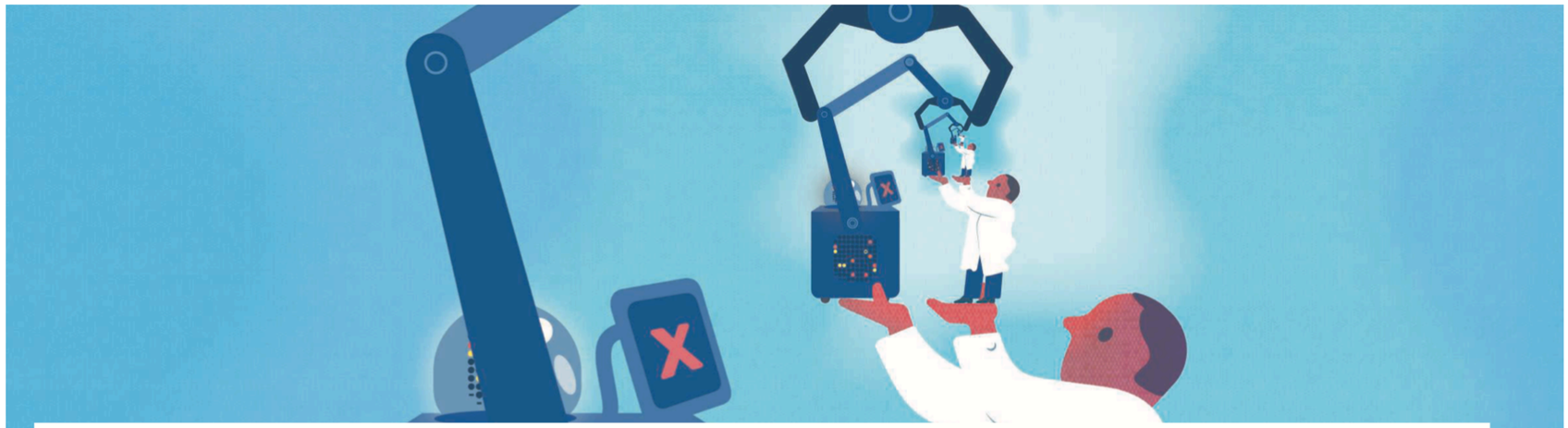
NASEM (National Academies of Sciences, Engineering, and Medicine). 2021. Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop. https://doi.org/10.17226/26308

“Making scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society...”

definition in the UNESCO Recommendation on Open Science (2021)
https://www.unesco.org/en/open-science

“principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity.”

NASA definition of Open Science, 2023



Defining the Role of Open Source Software in Research Reproducibility

Lorena A. Barba, The George Washington University

Barba, L.A., 2022. Defining the role of open source software in research reproducibility. Computer, 55(8), pp.40-48. DOI: 10/kggw

PASC23
Davos Switzerland | 26-28 June 2023

Anti Patterns of Scientific Machine Learning to Fool the Masses

A Call for Open Science

LorenaABarba. @labarba@fosstodon.org