HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

On Integrity of Insight in AI

Reproducibility and Replicatability in Machine Learning Research

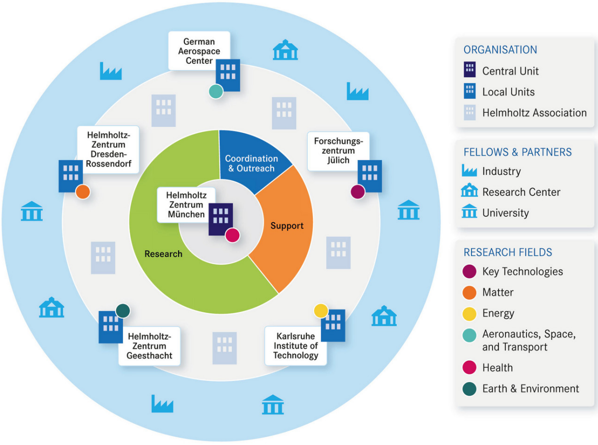
HELMHOLTZ AI

Peter Steinbach

Helmholtz-Zentrum Dresden-Rossendorf */ Helmholtz-CAS Workshop @ WIKOOP-INFRA Project, June 5, 2023*

Where am I coming from?

Helmholtz.AI: a Helmholtz Network across Germany



from www.helmholtz.ai

- running over 7 years 2019 - 2026
- each local unit:
  - **young investigator group**
  - **consultant team**
- planned staff:
  - 37 FTEs science
  - 35 FTEs consulting
  - 6 FTEs coordination, outreach, management

Challenges in Machine Learning

Reproducibility, Replicability, Re ... What? [Plesser, 2018] [Barba, 2018]

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Fig. 5 How the Turing Way defines reproducible research

Let's use the definitions that we teach! [Community, 2021]

Recap: Image Classification using Deep Learning

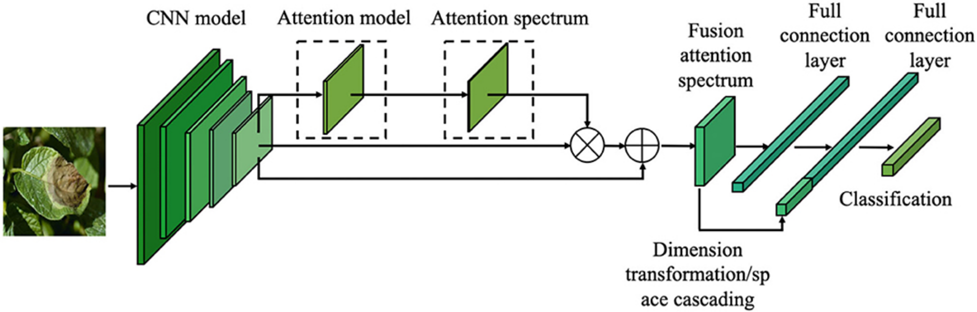
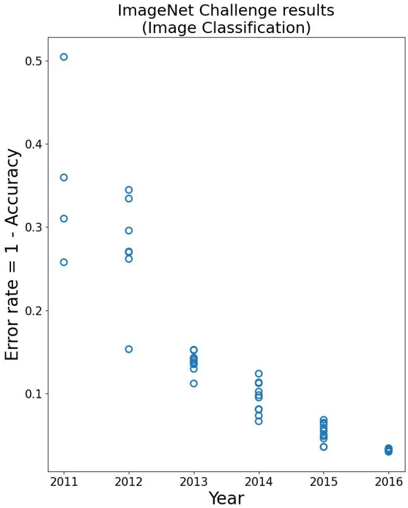
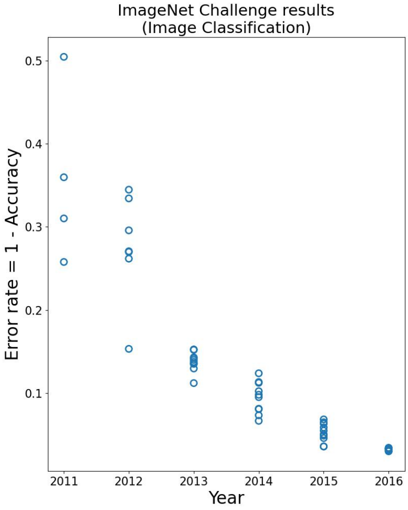


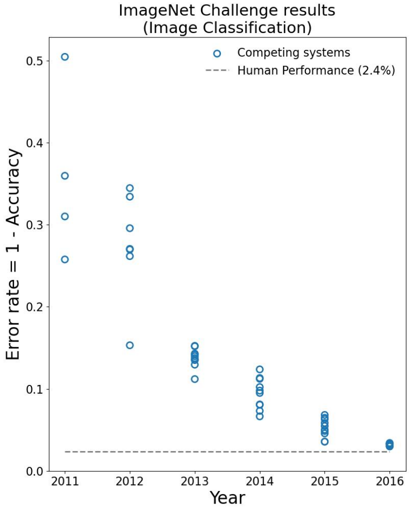
figure 4 from [Yang et al., 2020]

Human Expectations: The Imagenet Moment 2012



Human Expectations: The Imagenet Moment 2012



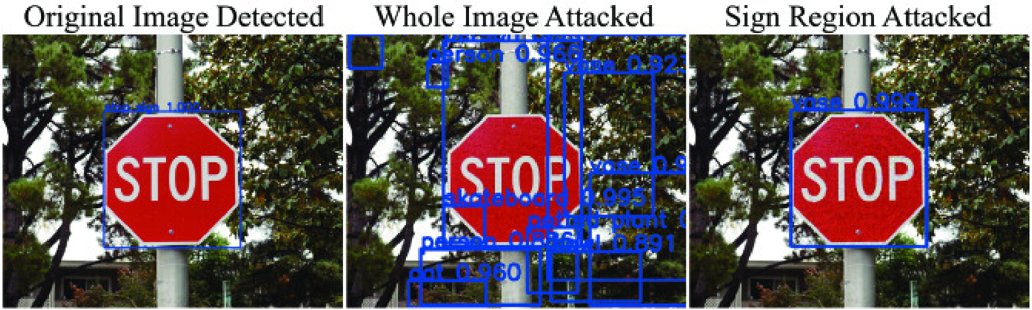


Inspired by Wikipedia: ImageNet Error Rate History

ML can be brittle: without human intervention [Geirhos et al., 2020]

Is this a cow? **“AI” says “No! It’s a horse.”** Replicability?

ML can be brittle: with deliberate human intervention [Lu et al., 2017]



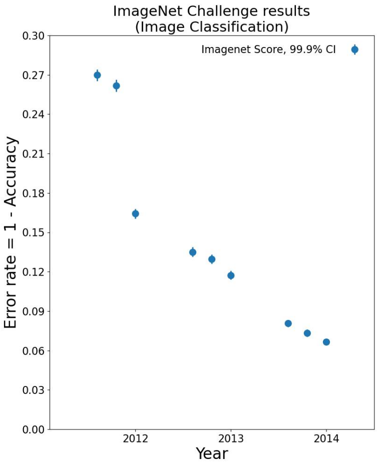
Stop sign identified as **vase.**

Consequences for autonomous driving?

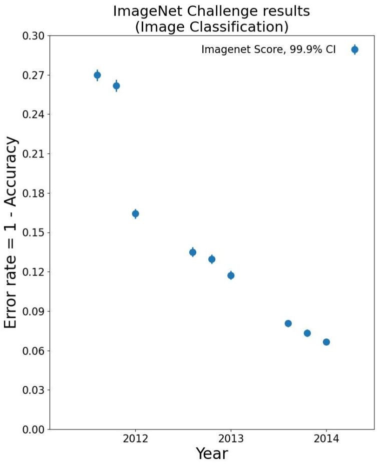
(replicability in real-life applications)

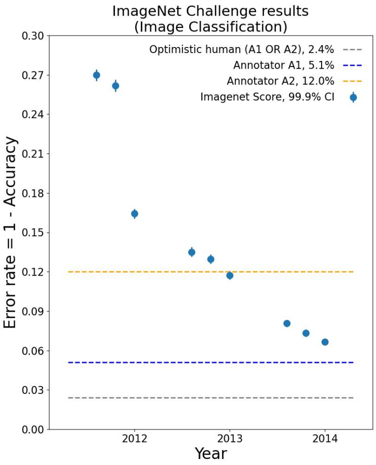
Quality of Evidence in the Digital Age

Uncertainties, expectations and myths: ImageNet[Russakovsky et al., 2015] again

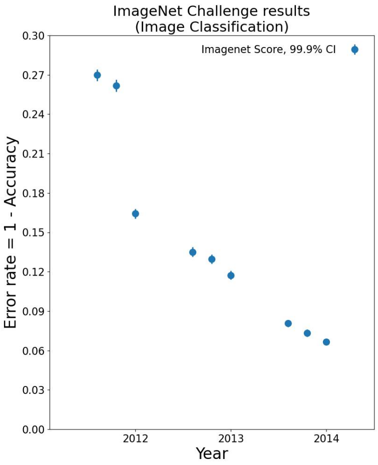


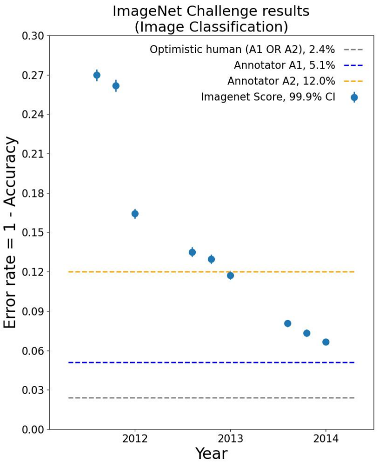
Uncertainties, expectations and myths: ImageNet[Russakovsky et al., 2015] again





Uncertainties, expectations and myths: ImageNet[Russakovsky et al., 2015] again

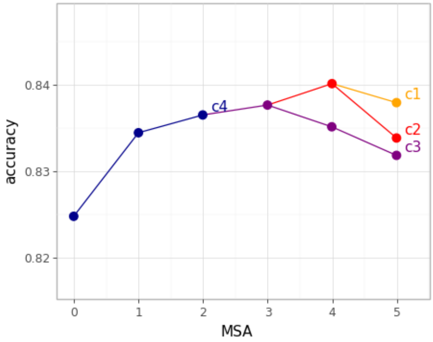




Quality of evidence: did we really reach super-human performance?

(reliability of scientific interpretation)

Uncertainties = guarantees for reproducibility and replicability



Uncertainties = guarantees for reproducibility and replicability

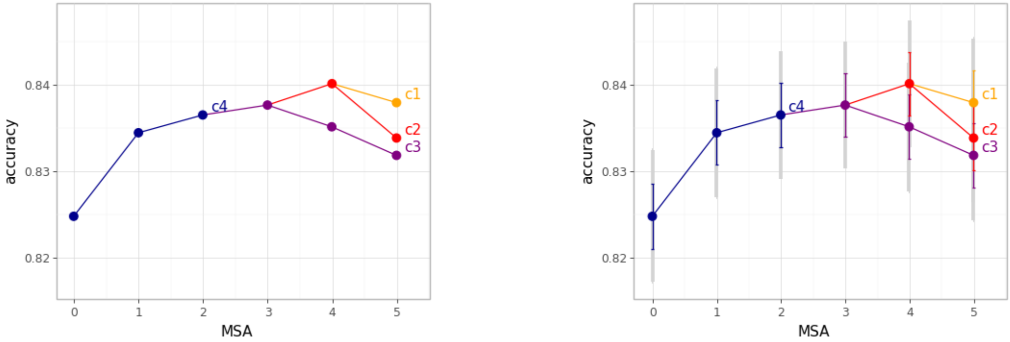


Figure 2: Reproduction of figure 12a from [Park and Kim, 2022] (left). Augmentation of the same figure with estimated accuracy calculated using eq. (1) from [Steinbach et al., 2022] using a one-sigma 68.2% (colored) and two-sigma 95% (grey) confidence interval (right). Data to reproduce these figures was obtained by using [Rohatgi, 2021] on the figures from the preprint PDF.

Uncertainties = guarantees for reproducibility and replicability

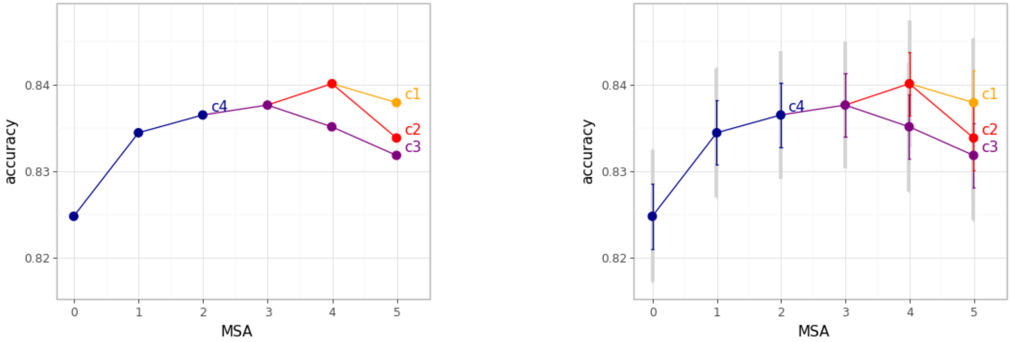


Figure 2: Reproduction of figure 12a from [Park and Kim, 2022] (left). Augmentation of the same figure with estimated accuracy calculated using eq. (1) from [Steinbach et al., 2022] using a one-sigma 68.2% (colored) and two-sigma 95% (grey) confidence interval (right). Data to reproduce these figures was obtained by using [Rohatgi, 2021] on the figures from the preprint PDF.

Quality of evidence: Are the interpretations replicable?

Next steps? Connect!

Exchange, Educate, Communicate (journals, conferences), ...

Summaries

Conclusions

- if you use a computer for doing ML:
  Check for reproducibility of your work!

Conclusions

- if you use a computer for doing ML:
  Check for reproducibility of your work!
- trustworthiness[Spiegelhalter, 2020] of ML (as a product or in science):
  let's get reproducibility/replicability right from the start!

Conclusions

- if you use a computer for doing ML:
  Check for reproducibility of your work!
- trustworthiness[Spiegelhalter, 2020] of ML (as a product or in science):
  let's get reproducibility/replicability right from the start!
- scientific integrity:
  progress speed versus scientific rigor

Conclusions

- if you use a computer for doing ML:
  Check for reproducibility of your work!
- trustworthiness[Spiegelhalter, 2020] of ML (as a product or in science):
  let's get reproducibility/replicability right from the start!
- scientific integrity:
  progress speed versus scientific rigor

Happy to hear your feedback, questions or comments!

@helmholtz_ai helmholtz.ai linkedin.com

References

References I

Lorena A. Barba. Terminologies for reproducible research, 2018. URL https://arxiv.org/abs/1802.03311.

Siddharth Bhat. Everything you know about word2vec is wrong, 2019. URL https://bollu.github.io/everything-you-know-about-word2vec-is-wrong.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *CoRR*, abs/2103.03098, 2021. URL https://arxiv.org/abs/2103.03098.

The Turing Way Community. The Turing Way: A handbook for reproducible, ethical and collaborative research, November 2021. URL https://doi.org/10.5281/zenodo.6533831.

References II

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

Damien Irving, Kate Hertweck, Luke Johnston, Joel Ostblom, Charlotte Wickham, and Greg Wilson. *Research Software Engineering with Python: Building software that makes research possible*. Chapman and Hall/CRC, 2021. URL https://merely-useful.tech/py-rse/.

Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494*, 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Namuk Park and Songkuk Kim. How do vision transformers work?, 2022. URL https://arxiv.org/abs/2202.06709.

References III

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, ASE '20, page 771–783, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367684. doi: 10.1145/3324884.3416545. URL https://doi.org/10.1145/3324884.3416545.

Hans E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2017.00076. URL https://www.frontiersin.org/article/10.3389/fninf.2017.00076.

Edward Raff and Andrew L. Farris. A siren song of open source reproducibility, 2022. URL https://arxiv.org/abs/2204.04372.

References IV

Ankit Rohatgi. Webplotdigitizer: Version 4.5, 2021. URL https://automeris.io/WebPlotDigitizer.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

David Spiegelhalter. Should We Trust Algorithms? *Harvard Data Science Review*, 2(1), jan 31 2020. https://hdsr.mitpress.mit.edu/pub/56lnenzj.

Peter Steinbach, Felicita Gernhardt, Mahnoor Tanveer, Steve Schmerler, and Sebastian Starke. Machine learning state-of-the-art with uncertainties, 2022. URL https://arxiv.org/abs/2204.05173.

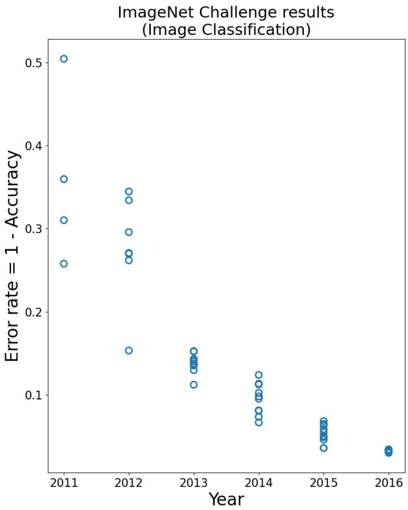
References V

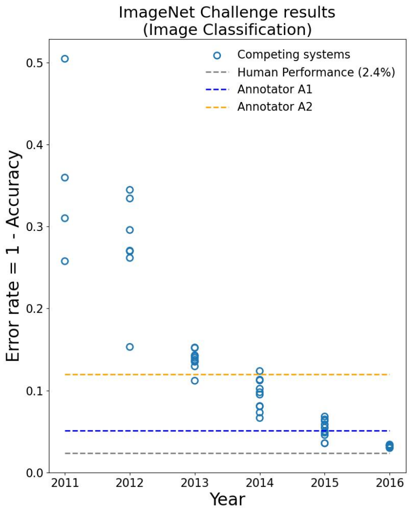
Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):1–20, 06 2017. doi: 10.1371/journal.pcbi.1005510. URL https://doi.org/10.1371/journal.pcbi.1005510.

Guofeng Yang, Yong He, Yong Yang, and Beibei Xu. Fine-grained image classification for crop disease based on attention mechanism. *Frontiers in Plant Science*, 11:600854, 2020. doi: 10.3389/fpls.2020.600854.

Backup

Expectations: The Imagenet Moment 2012





Inspired by Wikipedia: ImageNet Error Rate History

Humans in the loop: the Word2Vec story

2013

- Natural Language Processing algorithm (paper: [Mikolov et al., 2013], code)
- uses a neural network model to learn word associations from a large corpus of text
- purpose:
  - detect synonymous words
  - suggest additional words for a partial sentence

Humans in the loop: the Word2Vec story

2013

- Natural Language Processing algorithm (paper: [Mikolov et al., 2013], code)
- uses a neural network model to learn word associations from a large corpus of text
- purpose:
  - detect synonymous words
  - suggest additional words for a partial sentence

2019

- open-source code does not match algorithmic recipe [Bhat., 2019]
- community used C implementation (unquestioned)
- algorithmic understanding disparate

Humans in the loop: the Word2Vec story

2013

- Natural Language Processing algorithm (paper: [Mikolov et al., 2013], code)
- uses a neural network model to learn word associations from a large corpus of text
- purpose:
  - detect synonymous words
  - suggest additional words for a partial sentence

2019

- open-source code does not match algorithmic recipe [Bhat., 2019]
- community used C implementation (unquestioned)
- algorithmic understanding disparate

  *“**Is this academic dishonesty?** I don’t know the answer, and that’s a heavy question. But I’m frankly incredibly pissed, and this is probably the last time I take a machine learning paper’s explanation of the algorithm **seriously** again ...”*

Humans in the loop: the Word2Vec story

2013

- Natural Language Processing algorithm (paper: [Mikolov et al., 2013], code)
- uses a neural network model to learn word associations from a large corpus of text
- purpose:
  - detect synonymous words
  - suggest additional words for a partial sentence

2019

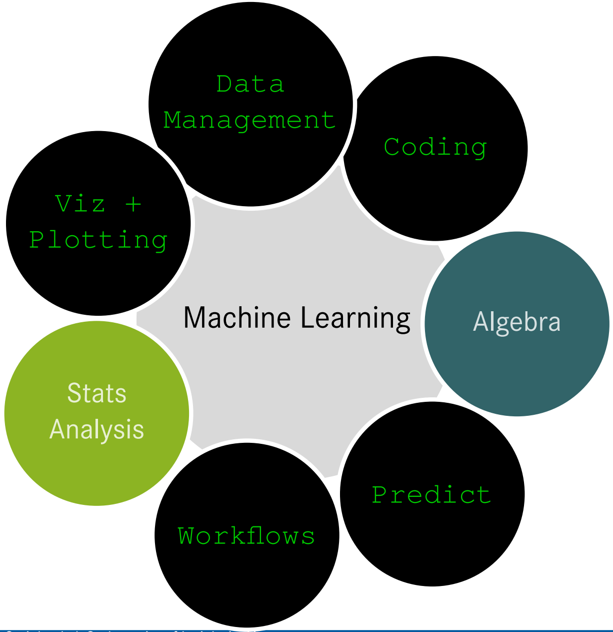
- open-source code does not match algorithmic recipe [Bhat., 2019]
- community used C implementation (unquestioned)
- algorithmic understanding disparate

  *“**Is this academic dishonesty?** I don’t know the answer, and that’s a heavy question. But I’m frankly incredibly pissed, and this is probably the last time I take a machine learning paper’s explanation of the algorithm **seriously** again ...”*

2022: **Do we need only open-source for reproduction?**[Raff and Farris, 2022]

A way forward?

ML requires good (enough) software engineering [Irving et al., 2021]



- your domain decides what good-enough is [Wilson et al., 2017]
- code needs to be reproducible
- lack of software engineering $\rightarrow$ more brittleness

ML = hardware + code + data + data + data

Hardware

Software + Code

Data for training today	Data at inference today	Data at inference tomorrow

ML = hardware + code + data + data + data

Hardware

Software + Code

Data for training today	Data at inference today	Data at inference tomorrow

- constant **retraining and introspection** required for ML products
- data today can be **corrupted** (by human or device)
- data tomorrow can be subject to **drifts** (in feature space, in concept space)
- **crucial:** flexible MLops and data science monitoring

Helmholtz-Zentrum Dresden-Rossendorf

1200 staff, infrastructure+research: life science, energy and matter

Figure: HZDR/Oliver Killig

ML = a minefield for reproducibility?

Our ML software stack yields variance [Pham et al., 2020]:

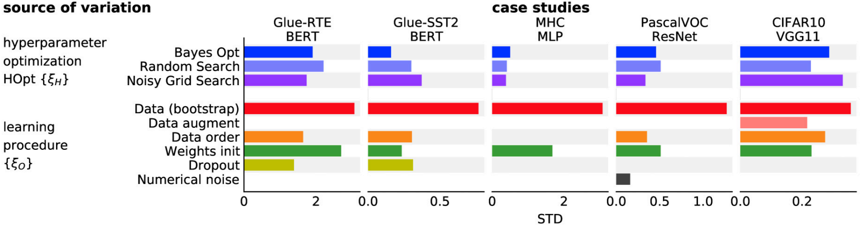
- 10.8% accuracy variation across DL library stack
- up to 52.4% per-class accuracy variation due to DL library stack
- $755/901$ authors unaware/unclear about code-level variance

ML = a minefield for reproducibility?

Our ML software stack yields variance [Pham et al., 2020]:

- 10.8% accuracy variation across DL library stack
- up to 52.4% per-class accuracy variation due to DL library stack
- $755/901$ authors unaware/unclear about code-level variance

ML has a lot of intrinsic variance! [Bouthillier et al., 2021]:



Our Helmholtz AI Network across Germany

by James Kahn, Helmholtz AI @ KIT

- 6 centers host Helmholtz AI units across Germany
- innovation: combine science teams and consulting teams
- total: 78 FTEs running
- consulting client base: 28.000 scientists

Helmholtz AI Consultant Team at HZDR

Mahnoor Tanveer

Helene Hoffmann

Steve Schmerler

Sebastian Starke