The Benchmarking Epistemology

What inferences can scientists draw from competitive comparisons of prediction models?

Timo Freiesleben[1] Sebastian Zezulka[1]

[1]University of Tübingen, Cluster of Excellence 'Machine Learning for Science'

February 21, 2025

Benchmarks in ML: An Epistemology?

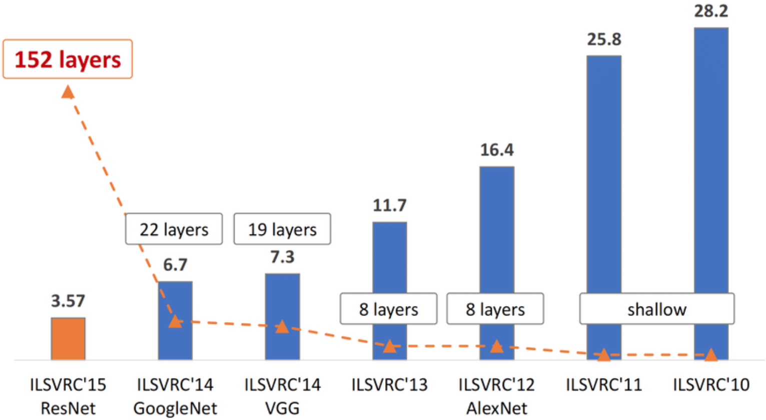

Figure 1: ImageNet Classification top-5 error (%) in [Nguyen et al., 2017]

“The iron rule of machine learning” [Hardt, 2024]

Scientific progress in ML: Whatever works, judged by benchmark results.

“The iron rule of machine learning” [Hardt, 2024]

Scientific progress in ML: Whatever works, judged by benchmark results.

Definition (Benchmark)

1. Predictive tasks $T = \{T_1, \ldots, T_r\}$, specified by input and output features.
2. Standardised datasets $D = (D_{train}, D_{leaderboard})$.
3. Evaluation metrics $L = \{L_1, \ldots, L_q\}$.
4. Public leaderboard with model ranking and/or scores.

There is a gap between benchmark island and real world inferences.

How to bridge the gap?

That is, how can we use benchmark results for scientific inferences?

How to bridge the gap?

That is, how can we use benchmark results for scientific inferences?
Machine learning benchmarks are very similar to tests in educational or psychological research:

1. We operationalize a *latent* skill as a concrete prediction task.
2. The test items are represented by data.
3. We assign skill scores based on empirical risk.

Construct Validity

There is a whole research field that is concerned with the validity of inferences based on test scores called *construct validity*. See, for example, [Cronbach and Meehl, 1955], [Messick, 1995], [Strauss and Smith, 2009], [Tal, 2020].

Inference I & II: Model and algorithm comparison

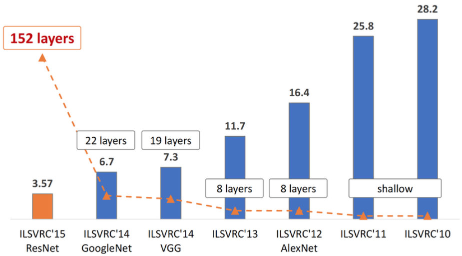

Figure 2: ImageNet Classification top-5 error (%) in [Nguyen et al., 2017]

Do improvements on the ImageNet leaderboard imply progress in image classification?

Typical inferences

- Ranking models.
- Inferring model skill scores.
- Ranking learning algorithms.

Inference I & II: Model and algorithm comparison

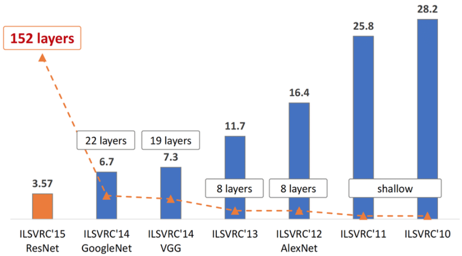

Figure 2: ImageNet Classification top-5 error (%) in [Nguyen et al., 2017]

Do improvements on the ImageNet leaderboard imply progress in image classification?

Typical inferences

- Ranking models.
- Inferring model skill scores.
- Ranking learning algorithms.

Empirical work by [Recht et al., 2019] and [Salaudeen and Hardt, 2024] indicates that model and algorithm *rankings* on ImageNet are robust at the task level, but not the skill scores.

Inference III: Deployment decisions

Should we deploy the weather forecasting model GraphCast for energy planing?

Typical Inferences

- Deployment rankings.
- Deployment utility.

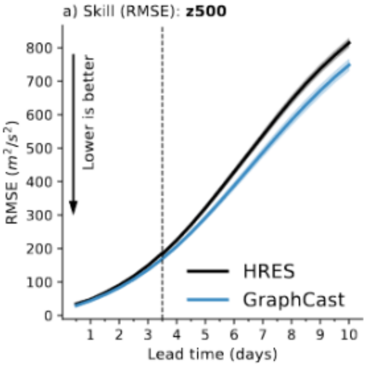

Figure 3: WeatherBench RMSE on z500 in [Lam et al,. 2023].

Inference IV: Predictability

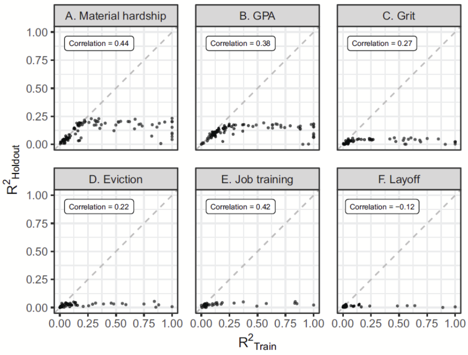

Figure 4: Results of the Fragile Families Challenge in [Salganik et al., 2019].

How predictable are life outcomes at the age of 15 from survey data?

Typical Inferences

- Bayes risk of a prediction task.
- Predictability of an outcome.
- Model selection: Theory development based on predictive performance.
- Finding relevant features.

Inference IV: Predictability

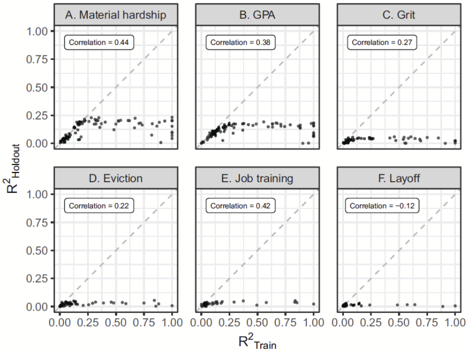

Figure 4: Results of the Fragile Families Challenge in [Salganik et al., 2019].

How predictable are life outcomes at the age of 15 from survey data?

Typical Inferences

- Bayes risk of a prediction task.
- Predictability of an outcome.
- Model selection: Theory development based on predictive performance.
- Finding relevant features.

The results of the Fragile Families Challenge indicate that life outcomes (at the age of 15) are poorly predictable, especially for a subset of families.

Summary

- Benchmarks are the central evaluation and model comparison method in ML.
- From measurement theory to ML: The theory of *construct validity* allows us to explicate required assumptions to support valid inferences from benchmarks.
- From ML to the empirical sciences: We can utilize the benchmark methodology in empirical research.
- Benchmark results form the basis for various scientific inferences:
  - Model and algorithm comparison.
  - Deployment decisions.
  - Predictability.
  - ...

How do you use benchmark results in your work?

References

- Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature, 588(7837), 203-205.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677), 1416-1421.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. arXiv preprint arXiv:2111.15366.
- Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do imagenet classifiers generalize to imagenet?. In International conference on machine learning (pp. 5389-5400). PMLR.
- Nguyen, K., Fookes, C., Ross, A., & Sridharan, S. (2017). Iris recognition with off-the-shelf CNN features: A deep learning perspective. IEEE Access, 6, 18848-18855.
- Salaudeen, O., & Hardt, M. (2024). ImageNot: A contrast with ImageNet preserves model rankings. arXiv preprint arXiv:2404.02112.
- Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the special collection on the fragile families challenge. Socius, 5, 2378023119871580.
- Schlangen, D. (2020). Targeting the benchmark: On methodology in current natural language processing research. arXiv preprint arXiv:2007.04792.