Conclusion
Conclusion: Inferential Reproducibility

Stefan Riezler and Michael Hagmann
Inferential Reproducibility

- Validity, reliability, and significance are methodological pillars of empirical science.
- Easily neglected in race for improved state-of-the-art results on benchmark data.
- Old-fashioned statistical methods come the rescue to analyze inferential reproducibility!
  - Enter **interpretable GAMs and LMEMs** as analysis tools.
  - **Statistical tests like GLRT, VCA, or circularity test** are **justified by identifiability and consistency** of maximum likelihood estimators for GAMs and LMEMs.
  - **Wide applicability, well established software.**
Focus of our work

- **Significance:**
  - Related to partial conjunction testing for multiple datasets
    [Dror et al., 2017],
  - and to score distribution comparison for multiple models
    [Dror et al., 2019].
  - **Our focus:** **Unified approach** for significance testing under **meta-parameter and data variation**, using likelihood ratio tests.
Focus of our work

- **Reliability:**
  - Related to approaches that analyze meta-parameter importance in model prediction [Hutter et al., 2014, Bergstra and Bengio, 2012],
  - or report expected validation performance w.r.t. computational budget [Dodge et al., 2019, Tang et al., 2020].
  - **Our focus:** Explain variability by LMEM variance component analysis and justify reliability by ICC-like coefficient.
Focus of our work

- **Validity:**
  - Related to descriptive statistics to detect dataset bias
    [Poliak et al., 2018, Gururangan et al., 2018],
  - with goal of using machine learning to reduce influence of bias features
    [Clark et al., 2019, Kim et al., 2019].
  - **Our focus:** GAM-based test to **detect** validity-violating features and **remove** them from datasets.
Open Questions, Comments, Suggestions

- Towards **inferential reproducibility** as a **new standard in machine learning evaluation**?
  - How to get there?
  - Would you go the extra mile?
  - What did we forget?

- Please tell us in Q&A or by email to
  {riezler,hagmann}@cl.uni-heidelberg.de
Thank you!

Data, code, and preprint:

https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods/


A systematic review of reproducibility research in natural language processing.
In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

An empirical investigation of statistical significance in NLP.
In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, Korea.

Random search for hyper-parameter optimization.

Some asymptotic theory for the bootstrap.

Cambridge University Press.

Test validity in cognitive assessment.

The concept of validity.


*CoRR*, abs/2204.02311.


What does research reproducibility mean?

We need to talk about standard splits.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy.

A methodology for building a patent test collection for prior art search.

Nonparametric regression and generalized linear models: a roughness penalty approach.
Crc Press.

Annotation artifacts in natural language inference data.

PHS: A toolbox for parallel hyperparameter search.

Generalized additive models.

*Generalized Additive Models.*
Chapman and Hall.

Spline smoothing in a partly linear model.

Reproducibility standards for machine learning in the life sciences.
*Nature Methods, 18*:1122–1144.

Deep reinforcement learning that matters.
In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA.

The large-sample power of tests based on permutations of observations.

Artificial intelligence faces reproducibility crisis.

An efficient approach for assessing hyperparameter importance.
In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China.

*The Growth of Logical Thinking from Childhood to Adolescence.*
Basic Books.

Scaling laws for neural language models.

Generalization in deep learning.
*CoRR*, abs/1710.05468.

Learning not to learn: Training deep neural networks with biased data.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA.

Derivation of new readability formulas for navy enlisted personnel.
Technical report, Technical Report, Naval Air Station, Millington, TN.

A guideline of selecting and reporting intraclass correlations coefficients for reliability research.
Correct me if you can: Learning from error corrections and markings.
In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

*Content Analysis. An Introduction to Its Methodology.*
Sage.

Deep learning reproducibility and explainable AI (XAI).
*CoRR*, abs/2202.11452.

BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Automatic evaluation of summaries using n-gram co-occurrence statistics.

Towards reproducible machine learning research in natural language processing.

**Lucic, M., Kurach, K., Michalski, M., Bousquet, O., and Gelly, S. (2018).**
Are GANs created equal? A large-scale study.

**Magdy, W. and Jones, G. J. F. (2010).**
Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task.
In In *Proceedings of the CLEF 2010 Workshop*, Padua, Italy.

**Manning, C. D., Raghavan, P., and Schütze, H. (2008).**
*Introduction to Information Retrieval.*
Cambridge University Press.

**Marie, B., Fujita, A., and Rubino, R. (2021).**
Scientific credibility of machine translation research: A meta-evaluation of 769 papers.
In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online.

**McCulloch, C. E. and Searle, S. R. (2001).**
*Generalized, Linear, and Mixed Models.*
Wiley.

**Melis, G., Dyer, C., and Blunsom, P. (2018).**
On the state of the art of evaluation in neural language models.


CoRR, abs/2204.06815.


Smoothing parameter and model selection for general smooth models.

Assumptions behind intercoder reliability indices.
*Communication Yearbook, 36*:419–480.

Extractive summarization as text matching.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online.

Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl.
*CoRR*, abs/2006.13799.