SDSC SAN DIEGO SUPERCOMPUTER CENTER

NTNU | Norwegian University of Science and Technology

ML Reproducibility: Sources of Algorithmic, Implementation, and Observational Variability

Kevin Coakley - 10/29/2024

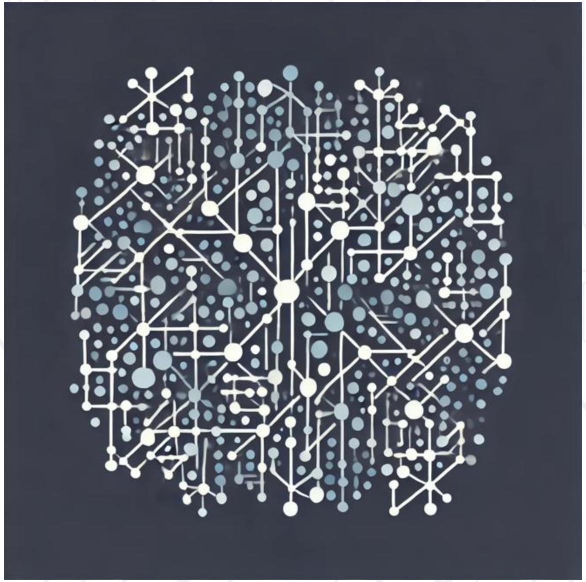
Download Slides Here

Reproducibility Crisis in ML

- Reproducibility is critical for trust in scientific findings, particularly in fields with high-stakes applications like healthcare, autonomous systems, and finance[1].
- In one study, the accuracy of models from 16 identical training runs varied by as much as 10.8%, even after removing weak models[2].
- Machine learning models that produce high variance in results challenge the reliability of findings[1].
- A survey of 901 researchers and practitioners found many respondents were unaware of (31.9%) or unsure about (21.8%) any variance and 83.8% were unaware or uncertain of variance caused by implementation choices[2].

1. Gundersen, Odd Erik, et al. "Sources of Irreproducibility in Machine Learning: A Review." arXiv e-prints (2022): arXiv-2024.
2. H. V. Pham *et al.*, “Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance,” p. 13, 2020.

Reproducibility Crisis in ML



- When developing or evaluating ML models it is critical to understand the sources of variation that can cause ML results to be irreproducible.

What is Reproducibility?

What is Reproducibility?

- Confusion between:
  - Repeatability
  - Replicability
  - Reproducibility

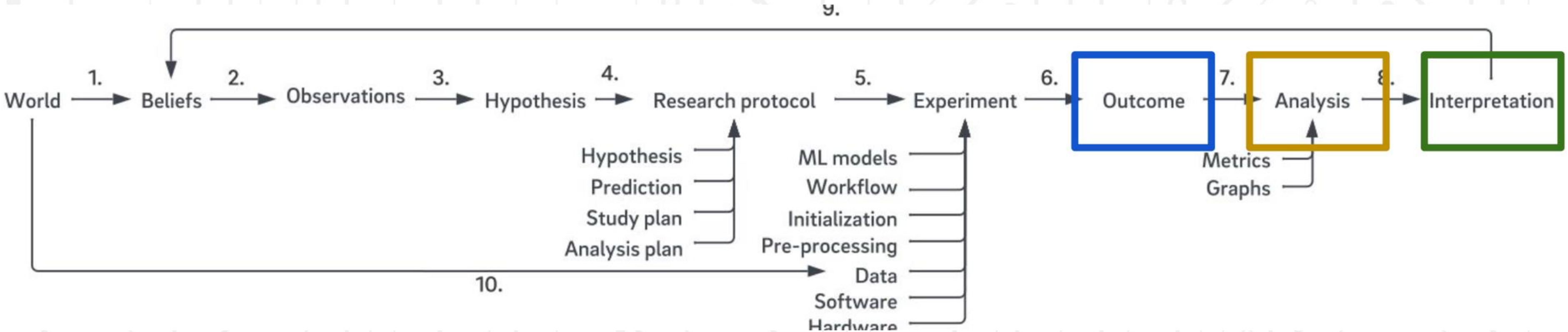
Definitions can differ between scientific disciplines

Definitions can change over time as the literature evolves

Illustration: ACM Definitions

- Artifact Review and Badging – Version 1.0
  - Repeatability: Same team, same experimental setup
  - Reproducibility: Different team, different experimental setup
  - Replicability: Different team, same experimental setup
- Artifact Review and Badging Version 1.1
  - Repeatability: Same team, same experimental setup
  - Reproducibility: Different team, same experimental setup
  - Replicability: Different team, different experimental setup
- 1.1 was updated to match the National Information Standards Organization (NISO) definitions
  - https://www.acm.org/publications/policies/artifact-review-and-badging-current

The Scientific Method in Machine Learning

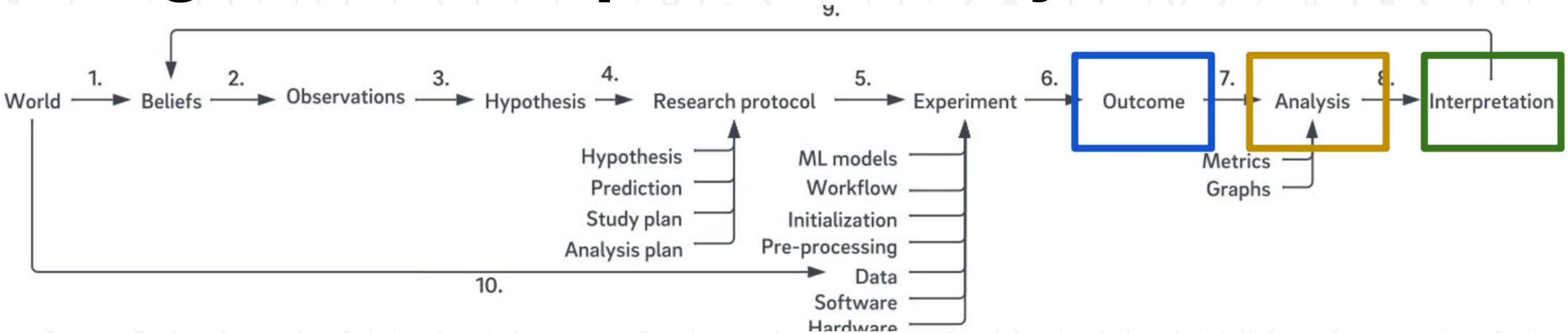


Reproducibility requires independent investigators to draw same conclusions

Three degrees of reproducibility: Outcome, Analysis, and Interpretation

Gundersen, Odd Erik. “The Fundamental Principles of Reproducibility.” ArXiv:2011.10098 [Cs], Nov. 2020. arXiv.org

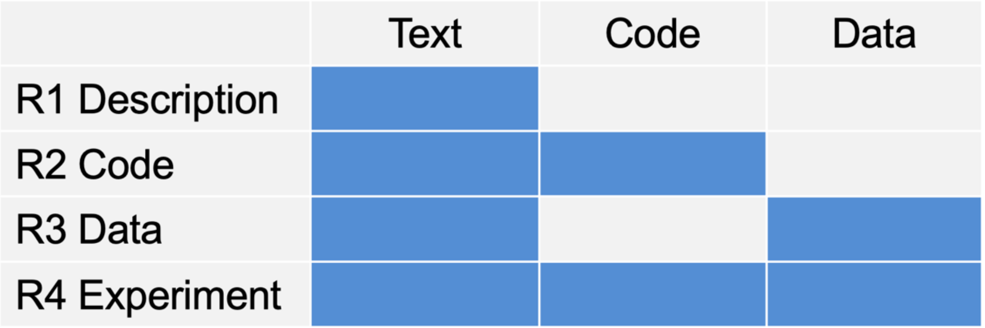
Degrees of Reproducibility



- Outcome: The variability doesn't cause the outcomes to differ. Presumably, the analysis and interpretation won't differ.
- Analysis: The variability causes the outcomes to differ but doesn't change the experiment analysis. Presumably, the interpretation won't differ.
- Interpretation: The variability causes the results and the analysis to differ but doesn't change the experiment interpretation.

Gundersen, Odd Erik. "The Fundamental Principles of Reproducibility." ArXiv:2011.10098 [Cs], Nov. 2020. arXiv.org

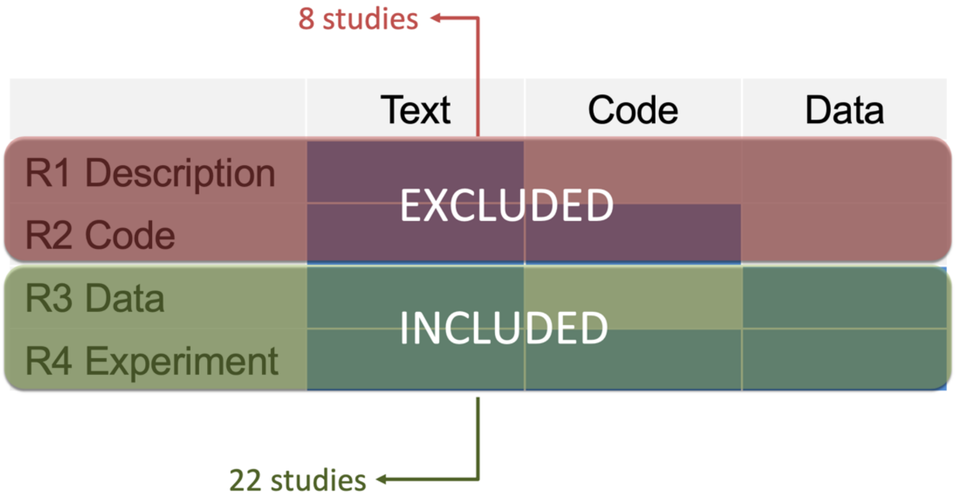
Types of Reproducibility Experiments



- **Text**: Description of the AI method implemented by the AI program, the experiment being conducted and the analysis of the results as well as the hardware and ancillary software used for conducting the experiment.
- **Code**: AI Program code, code for setup and configuration, code controlling workflow, code for analysis of results and visualization.
- **Data**: All data used for conducting the experiment. Are the samples used for training, validation and test specified? What about the results?

Gundersen, Odd Erik. “The Fundamental Principles of Reproducibility.” ArXiv:2011.10098 [Cs], Nov. 2020. arXiv.org

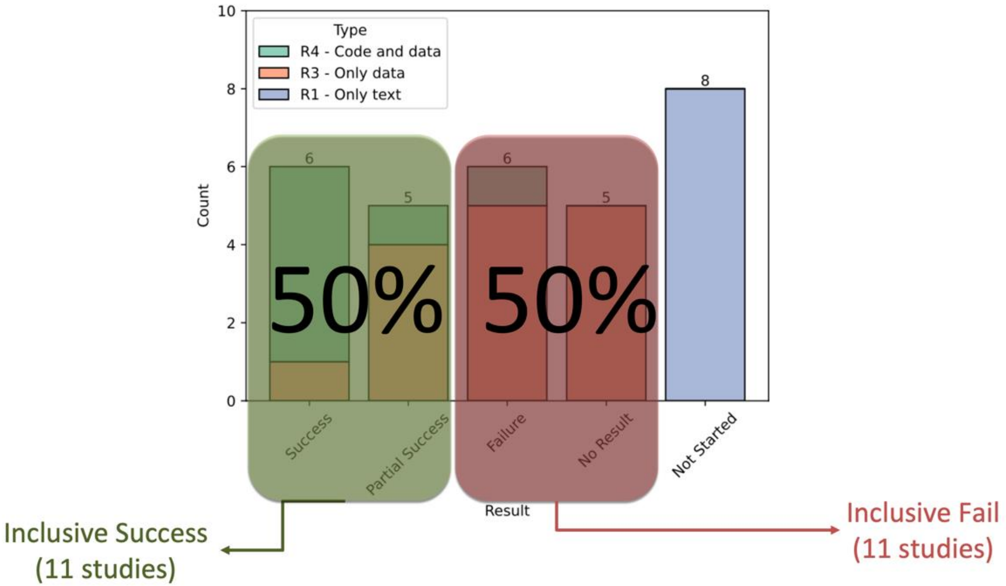
How Documentation Affects Reproducibility

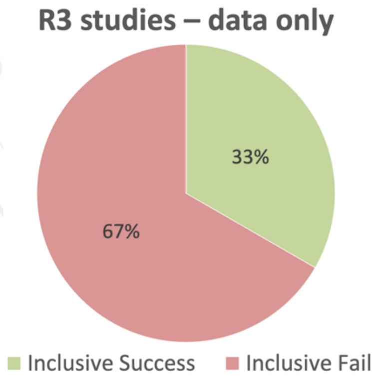


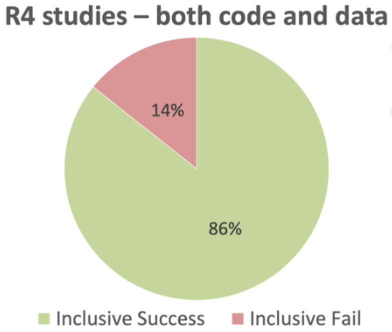
- 30 highly cited AI papers
- 8 papers were excluded because data was not shared

Gundersen, Odd Erik. “The Unreasonable Effectiveness of Open Science in AI: A Replication Study” Under Review

How Documentation Affects Reproducibility







- Overall reproducibility is 50%:
  - Decreases to 33% if only data shared.
  - Increases to 86% if both code and data shared.

How Documentation Affects Reproducibility

- Additional findings:
  - Code documentation and quality is not important. Poor and undocumented code is better than no code.
  - Data and data documentation quality is important for reproducibility.

Reproducibility versus Portability

Reproducibility versus Portability

To avoid confusion with the terms Repeatability, Replicability, and Reproducibility.

Prefer the term **Portability**

- **Reproducibility** focuses on the reliability of the results across different conditions.
- **Portability** focuses on replicating the experiment setup on different systems.

.

Reproducibility versus Portability

- **Reproducibility**:
  - Involves verifying results, analysis, and conclusions beyond just replicating the experimental setup.
  - Can be affected by variations in both hardware and software environments.
- **Portability**:
  - Refers to the ability to transfer and run experiments across different hardware or computing systems.
  - Simplifies recreating the experiment environment but does not guarantee the same experimental results.
  - May still carry over biases and does not address variations due to different hardware setups.

Portability of Experiments

	Text	Code	Data
R1 Description	■		
R2 Code	■	■	
R3 Data	■		■
R4 Experiment	■	■	■

- R1 Description: Hosted by the publisher or a site like arxiv.
- R2 Code: Public version control (GitHub, GitLab), Open research repositories (Zenodo), Domain specific research repositories. Code should include extract, transform, and load (ETL) of data and code to analyze the results.

Portability of Experiments

	Text	Code	Data
R1 Description	■		
R2 Code	■	■	
R3 Data	■		■
R4 Experiment	■	■	■

- R3 Data: Difficult for large datasets. Storage hardware is expensive and difficult to maintain. Cloud storage providers usually charge high fees for data downloads. Free Open Storage Network allocations through Access CI.

https://www.openstoragenetwork.org - https://www.access-ci.org

Portability of Experiments

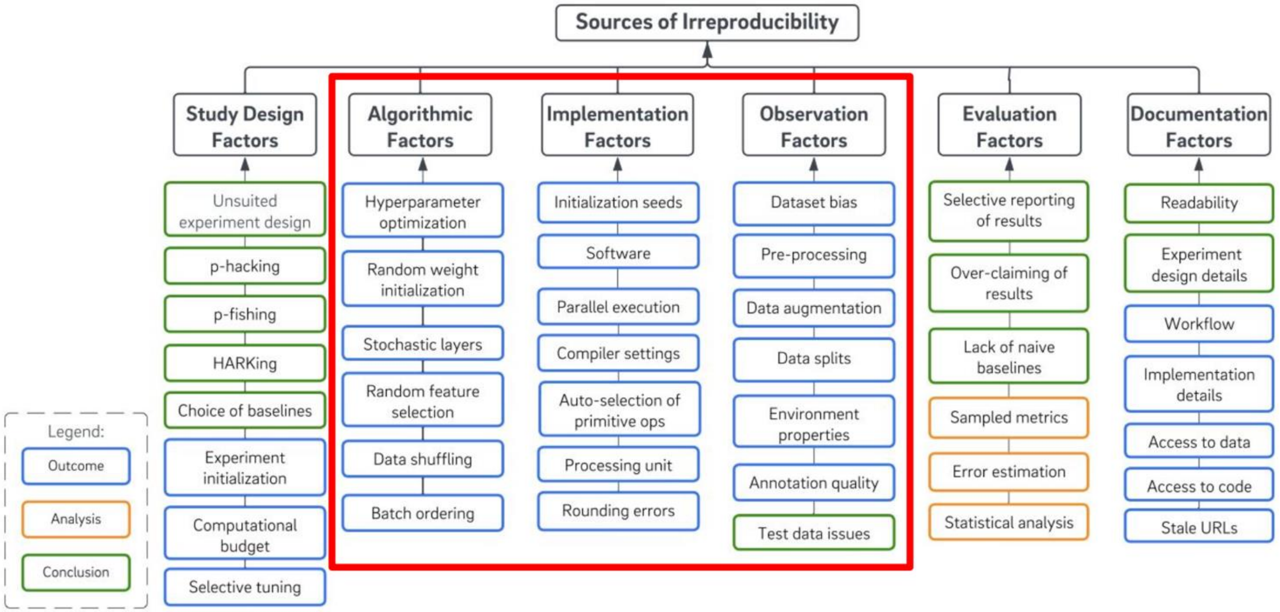
	Text	Code	Data
R1 Description			
R2 Code			
R3 Data			
R4 Experiment			

- R4 Experiment: Pip requirements/Conda environment (Python), Packrat (R), or Docker for portable software environments. Very detailed text descriptions of software used, with version information and details about the hardware environment.

Docker and Portability

Sources of Irreproducibility

Sources of Irreproducibility - Overview



Gundersen, Odd Erik, et al. "Sources of Irreproducibility in Machine Learning: A Review." arXiv e-prints (2022): arXiv-2024.

Note on Variability, it isn't Bad

- Randomness for regularization
  - Helps models generalize by preventing reliance on specific patterns in training data.
- Randomness for speed
  - Techniques like data shuffling accelerate training convergence, helping models reach a stable solution faster.
- Randomness due to design decisions
  - Different libraries or frameworks may introduce slight variability in outputs.

Algorithmic Factors

Algorithmic Factors Causing Irreproducibility - Part 1

AF - Hyperparameter Optimization

Different hyperparameter optimization methods (random, grid, Bayesian optimization, intuition) and optimization budgets (study design factor) will affect outcome.

AF - Random Weight Initialization

The random initialization of weights in neural networks can lead to the model to converge to different local minima.

Algorithmic Factors Causing Irreproducibility - Part 3

AF - Data Shuffling

Random data shuffling done during training so learning converges faster can cause outcomes to differ.

AF - Batch Ordering

Due to memory limitations, data samples are fed into DL algorithms in batches. Randomizing batch order between epochs results in different outcomes between training runs.

Algorithmic Factors Causing Irreproducibility - Part 2

AF - Stochastic Layers

Stochastic model layers, like Dropout, intended to make deep neural networks more robust, affect their outcome.

AF - Random Feature Selection

Many learning algorithms rely on selecting features at random during training, like Random Forests. Which randomly selected features are chosen will influence the outcome.

Algorithmic Factors - Conclusions

- **Stochasticity** in deep learning inherently leads to different outcomes across runs.
- **Significant performance variations** between runs can affect conclusions.
- **Consistent outcomes don't guarantee robustness** — variability must be considered.
- **Report performance variation** over multiple runs to ensure transparency and reliability.

Implementation Factors

SDSC NTNU

Implementation Factors Causing Irreproducibility - Part 1

IF - Initialization Seeds

Different seeds used to initialize the pseudo-random number generator produce different outcomes. The same seed on different platforms produces different outcomes.

IF - Software

Outcomes across DL frameworks (TensorFlow, PyTorch) can vary significantly. Different software (libraries, operating systems) or versions may implement the same algorithm differently, causing different outcomes.

Implementation Factors Causing Irreproducibility - Part 2

IF - Parallel Execution

Random completion order of parallel tasks introduces variation. Truncation error of floating-point calculations introduces variability as $A + B + C \neq C + B + A$ when calculated in parallel.

IF - Compiler Settings

Hong et al[1], found severe sensitivity to Intel compiler optimization levels for weather simulations that rely on floating-point calculations.

1 S.-Y. Hong *et al.*, "An Evaluation of the Software System Dependency of a Global Atmospheric Model," *Monthly Weather Review*, vol. 141, no. 11, pp. 4165–4172, Nov. 2013, doi: 10.1175/MWR-D-12-00352.1.

Implementation Factors Causing Irreproducibility - Part 3

IF - Auto-selection of Primitive Ops

High level libraries implement DL algorithms using GPU-optimized DL primitives from low-level libraries (cuDNN and CUDA). Autotune in cuDNN automatically benchmarks several modes of operation which might change between runs.

IF - Processing Unit

Changing the processor can affect results. The same GPU chip on hardware from different manufacturers can produce different outcomes when running deterministically.

Implementation Factors Causing Irreproducibility - Part 4

IF - Rounding Errors

Different hardware architectures and software implement the rounding of floating-point numbers in different ways. These rounding errors accumulate during long-running calculations, particularly when using GPUs.

Implementation Factors - Conclusions

- Variations in **software and hardware** mirror the inconsistencies seen in physical labs.
- Treat the software and hardware environment as a **calibrated scientific instrument** for ML experiments.
- Consistent results require controlling for differences in **software libraries, hardware configurations, and parallelization**.
- Always **document and share** all configurations to support reproducibility.

Observation Factors

Observation Factors Causing Irreproducibility - Part 1

OF - Dataset Bias

The methods used to gather data (manual or automated) and the way data is captured introduce biases to datasets.

OF - Data Pre-processing

Differences in data pre-processing will change outcomes, so the applied pre-processing techniques must be well documented to facilitate reproducibility.

Observation Factors Causing Irreproducibility - Part 2

OF - Data Augmentation

Stochastic data augmentation procedures are influenced by both algorithmic and implementation factors, which leads to differences in training data and outcomes.

OF - Data Splits

Differences in data splits cause a difference in outcomes.

Observation Factors Causing Irreproducibility - Part 3

OF - Environment Properties

Stochasticity and different dynamic properties of the testing environment could affect the outcome, especially in continuous control simulators such as those used in deep reinforcement learning.

OF - Annotation Quality

Differences in annotations made by humans will affect the target value and the outcome.

Observation Factors Causing Irreproducibility - Part 4

OF - Test Data Issues

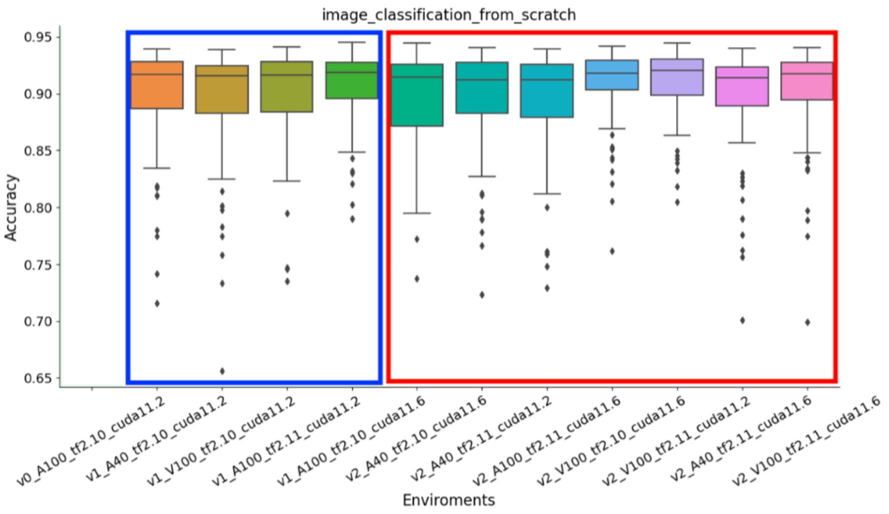
Model performance is overestimated when models are trained on data that should only be available at test time (data leakage).

Observation Factors - Conclusions

- Observation factors might affect the **outcome and interpretation** of an experiment.
- **Dataset bias and pre-processing** significantly impact model outcomes and interpretations.
- Mitigate these effects by setting **random seeds** and thoroughly documenting **data pre-processing** and **provenance**.
- **Careful handling** of duplicate data, outliers, and missing values is essential to avoid **bias**.
- **Dataset shifts** over time may cause models to become outdated—**regularly reassess and update** them.

Why Noise Control Isn’t Enough

- “Simply removing noise from one part of the technical stack is not a robust way to improve training stability[1]”
- The effect of these sources of irreproducibility doesn't appear to be cumulative. Blue one source changed, red two sources changed (Sources: GPU, TensorFlow version, CUDA/CUDNN version)[2]



1 Zhuang, D., Zhang, X., Song, S., Hooker, S.: Randomness in neural network training: Characterizing the impact of tooling. Proceedings of Machine Learning and Systems 4, 316–336 (2022)

2 Coakley Unpublished

Conclusions

- **Irreproducibility is a Complex Challenge**
  - Arises from various factors across algorithms, implementations, and data handling.
- **Interconnectedness Across the Technical Stack**
  - Algorithmic, implementation, and observational factors all contribute to variability in results.
- **No Single Fix**
  - Controlling one aspect, such as random seeds, is insufficient for ensuring training stability.
  - Addressing irreproducibility requires attention to the entire technical pipeline.
- **Share Your Code and Data!**

Questions?

Contact Information:

Kevin Coakley

kcoakley@sdsc.edu

Download Slides Here