SDSC SAN DIEGO SUPERCOMPUTER CENTER

NTNU | Norwegian University of Science and Technology

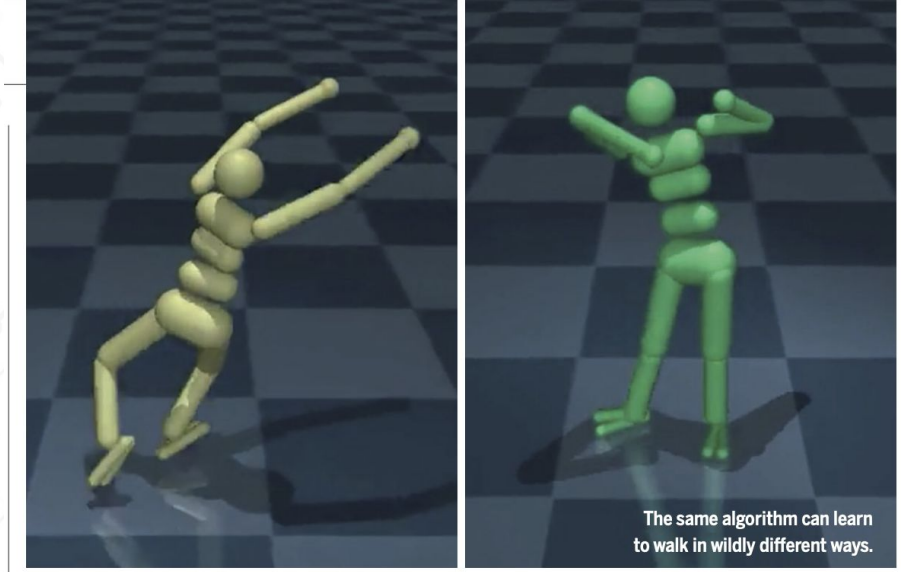
AI Reproducibility

Health AI Exploration: Virtual Meeting #5

Kevin Coakley - SDSC/NTNU
Christine Kirkpatrick - SDSC

Overview

- Reproducibility and reproducibility factors
- Case study: reproducibility exploration using the Open Science Grid
- Implications of variance
- Takeaways



The same algorithm can learn to walk in wildly different ways.

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

Hutson, Matthew. "Artificial Intelligence Faces Reproducibility Crisis." *Science*, vol. 359, no. 6377, Feb. 2018, pp. 725–26. *DOI.org*

What is Reproducibility?

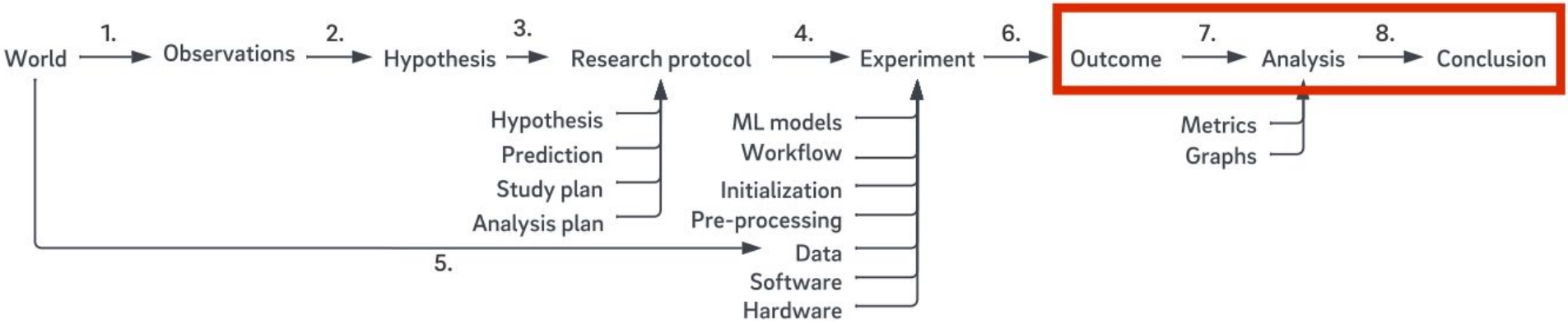
- Confusion between:
  - Repeatability (10x)
  - Replicability
  - Reproducibility

← repeat on same infra yourself / reproduce independently →

- Definitions can differ between scientific disciplines
- Definitions can change over time as the literature evolves

The Scientific Method

Perhaps a better way to think of reproducibility to examine the scientific method:



Reproducibility requires independent team of investigators have to conduct the same experiment and achieve the same conclusions. There are different requirements needed to achieve reproducibility at the outcome, analysis and conclusion steps.

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

Implementation Reproducibility Factors

- Initialization seeds
- Software
  - Software version
  - Bugs in software
- Non-deterministic ordering of floating-point operations
- Parallel execution
- Compiler settings
- Auto-selection of primitive operations
- Processing unit
- Rounding errors

Gundersen, Odd Erik, Kevin Coakley, and Christine Kirkpatrick. "Sources of Irreproducibility in Machine Learning: A Review." *arXiv preprint arXiv:2204.07610* (2022)

What Can Be Done?

Not all sources of irreproducibility can be controlled for.

However, many of the sources of irreproducibility can be examined by performing multiple runs in heterogeneous computing environments consisting of different hardware and software environments.

The biggest challenge for researchers can be getting access to multiple heterogeneous computing environments.

Comparing AI Results on Different Hardware

- Used Open Science Grid for access to many clusters
  - Different CPUS & GPUS, different versions of tensorflow
- 3 use cases from Keras (running deterministically)
  - Computer Vision - Simple CNN classifying handwritten digits using MNIST
  - NLP - A 2-layer bidirectional LSTM classifying sentiments using IMDB
  - Structured Data - Dense NN performing binary classification on the Kaggle Credit Card Fraud Detection dataset
- 780 runs, each example 5x (10 CPUs & 4 GPUs)

Varying Hardware Test (CPU)

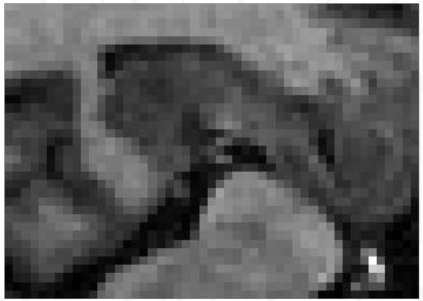
Processing Unit:Software Environment	Bidirectional LSTM	Binary Classification	mnist
AMD EPYC 7513:tensorflow_2.8.0-gpu	0.8616	0.9972	0.9919
AMD EPYC 7F52:tensorflow_2.8.0-gpu	0.8616	0.9935	0.9915
Intel Xeon E5-2650 v3:tensorflow_2.8.0-gpu	0.8463	0.9923	0.9918
Intel Xeon Gold 6148:tensorflow_2.8.0-gpu	0.8425	0.9987	0.9916
AMD EPYC 7302:tensorflow_2.9.1-gpu	0.8506	0.9912	0.9919
AMD EPYC 7443:tensorflow_2.9.1-gpu	0.8506	0.9933	0.9921
AMD EPYC 7713:tensorflow_2.9.1-gpu	0.8506	0.9961	0.9919
AMD EPYC 7F52:tensorflow_2.9.1-gpu	0.8506	0.9934	0.9916
Intel Xeon ES-2650 v3:tensorflow_2.9.1-gpu	0.8563	0.9931	0.9919
Intel Xeon Gold 6148:tensorflow_2.9.1-gpu	0.8664	0.9639	0.9918
Intel Xeon Gold 6242:tensorflow_2.9.1-gpu	0.8026	0.9941	0.9916

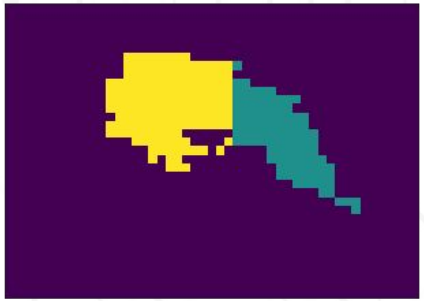
Varying Hardware Test (Software Bug)

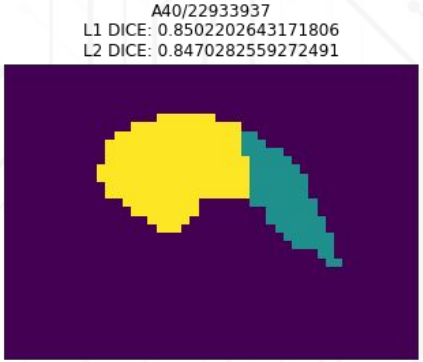
Processing Unit:Software Environment	0	1	2	3	4
NVIDIA A40:tensorflow_22.03	0.86584	0.86584	0.86584	0.86584	0.86584
NVIDIA A100-40GB:tensorflow_22.03	0.86580	0.86580	0.86580	0.86580	0.86580
Tesla V100-16GB:tensorflow_22.03	0.85203	0.85203	0.85203	0.85203	0.85203
NVIDIA A40:tensorflow_22.06	0.86584	0.86584	0.86584	0.86584	0.86584
NVIDIA A100-40GB:tensorflow_22.06	0.86580	0.86580	0.86580	0.86580	0.86580
Tesla V100-16GB:tensorflow_22.06	0.85203	0.85203	0.85203	0.85203	0.85203
NVIDIA A40:tensorflow_2.8.0-gpu	0.86571	0.86555	0.85680	0.85927	0.86147
NVIDIA A100-40GB:tensorflow_2.8.0-gpu	0.86511	0.84516	0.86360	0.85895	0.86479
Tesla V100-32GB:tensorflow_2.8.0-gpu	0.85983	0.78835	0.85764	0.86031	0.86908
NVIDIA A40:tensorflow_2.9.1-gpu	0.83631	0.86287	0.83891	0.85724	0.85276
NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.85979	0.86619	0.83156	0.86412	0.84564
Tesla V100-16GB:tensorflow_2.9.1-gpu	0.84447	0.85320	0.85996	0.86511	0.86779
NVIDIA A40:tensorflow_2.9.1-gpu-cuda11.3-cudnn8.2	0.86184	0.86184	0.86184	0.86184	0.86184
NVIDIA A100-40GB:tensorflow_2.9.1-gpu-cuda11.3-cudnn8.2	0.84631	0.84631	0.84631	0.84631	0.84631

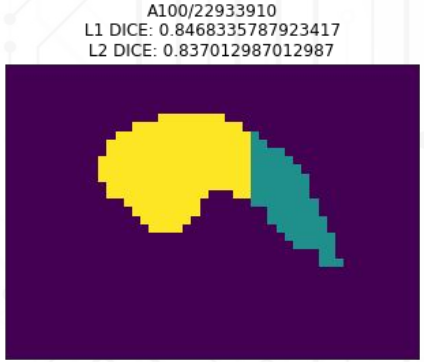
Tolerance for Erroneous Conclusions: Low

Running nnUNet with Medical Imaging



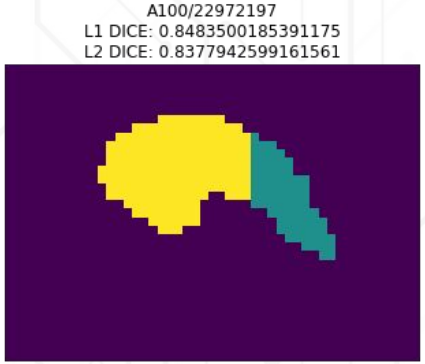


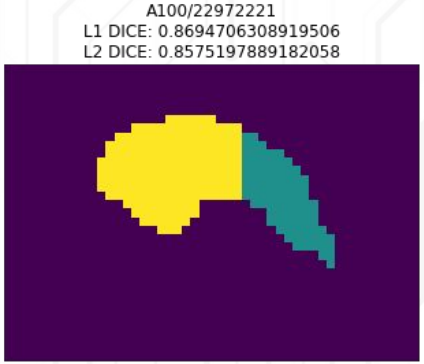


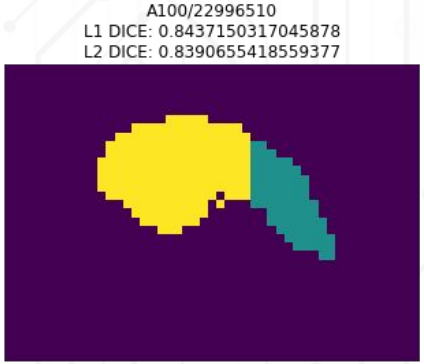


What amount of error is tolerable?

Need to quantify the variance







Experimentation Conclusions

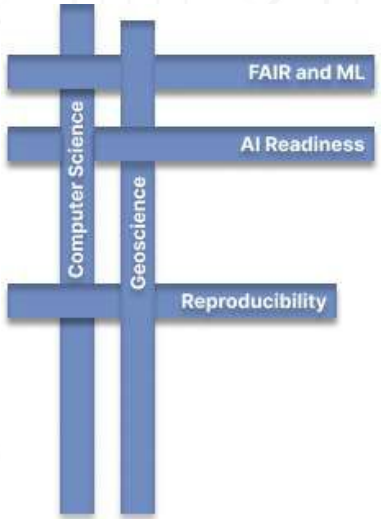
- There are **reproducibility factors that cannot be controlled for** when dealing with **limited research time and budgets**.
- Running the same Deep Learning (DL) experiment in different environments can produce results that are **varied enough that a researcher may come to a different conclusion** than they would have if they had only ran the experiment in one environment.
- **Running DL experiments in multiple environments** will expose how sensitive the results are to the implementation factors.
- ML researchers should not focus on the variation in the outcomes but ***how the variation affects the analysis and conclusion***.

Takeaways

- AI/ML under constant development
  → Consider usage carefully for application with impact
- Commitment to repeatability needed with AI
- Document the applicable ‘implementation factors’
  - Publishers play a role
  - Nanopublications could help
- Awareness building needed
- Research and tools needed for prioritizing reproducibility work/documentation
  → Realistic goals vs. perfection

FARR – FAIR AI Readiness & Reproducibility

- Building communities to
  - promote better practices for AI
  - harness community efforts
  - improve efficiency and reproducibility
  - stimulate and enhance new research
- Activities will include
  - workshops
  - assessing community needs
  - fostering new collaborations (proposals)
  - setting research agendas
  - community-led reports
- For more info, contact us



Contact

Kevin Coakley - kcoakley@sdsc.edu
Christine Kirkpatrick - christine@sdsc.edu

Sources of Irreproducibility in Machine Learning: A Review

https://arxiv.org/abs/2204.07610