FROM HPC TO THE EDGE TO ENABLE ACCELERATED AND REPRODUCIBLE AI DISCOVERY

ELIU HUERTA
Lead for Translational AI
Computational Scientist
Data Science and Learning Division
Department of Computer Science, University of Chicago

Machine Learning for Industry Forum
August 10-12, 2021

BIO

Theoretical Astrophysicist, Mathematician, Computer Scientist

Lead for Translational AI
Data Science and Learning Division
Argonne National Laboratory

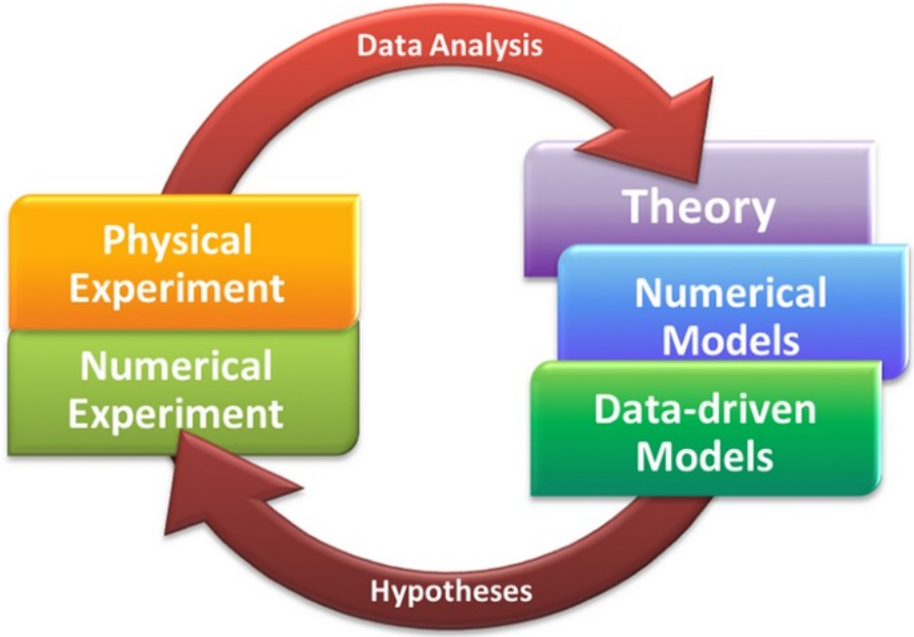
Department of Computer Science, University of Chicago
Department of Physics, University of Illinois at Urbana-Champaign

PhD in Theoretical Astrophysics
Master of Advanced Study in Mathematics
University of Cambridge, UK

DATA-DRIVEN DISCOVERY

WHAT

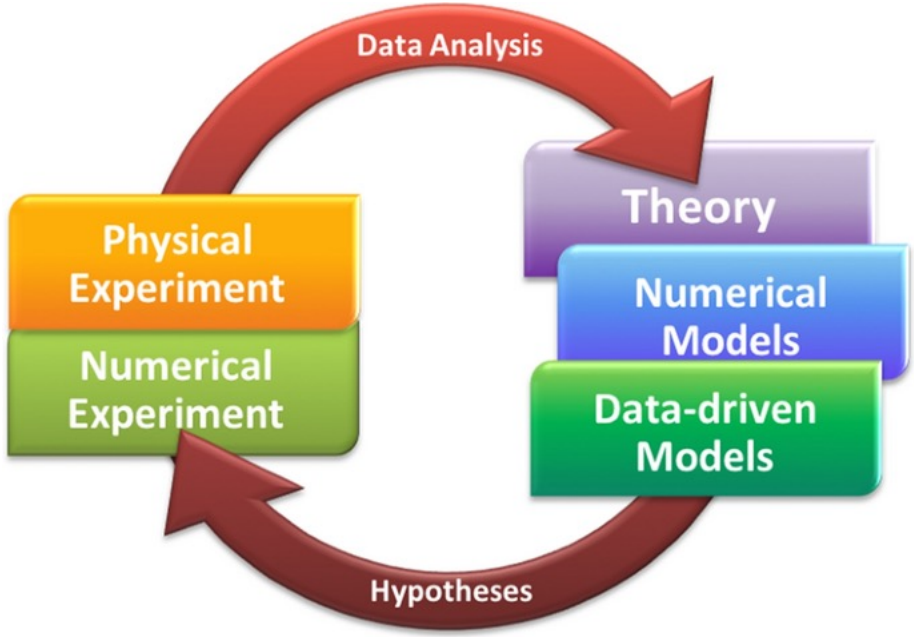
Transition from **first principles** modeling and large-scale simulation to domain-informed, interpretable, accelerated and reproducible **AI**



DATA-DRIVEN DISCOVERY

HOW

Bridge the gap between AI based on first principles & simulated data and AI that captures the complexity and non-linearity of experimental data



DATA-DRIVEN DISCOVERY

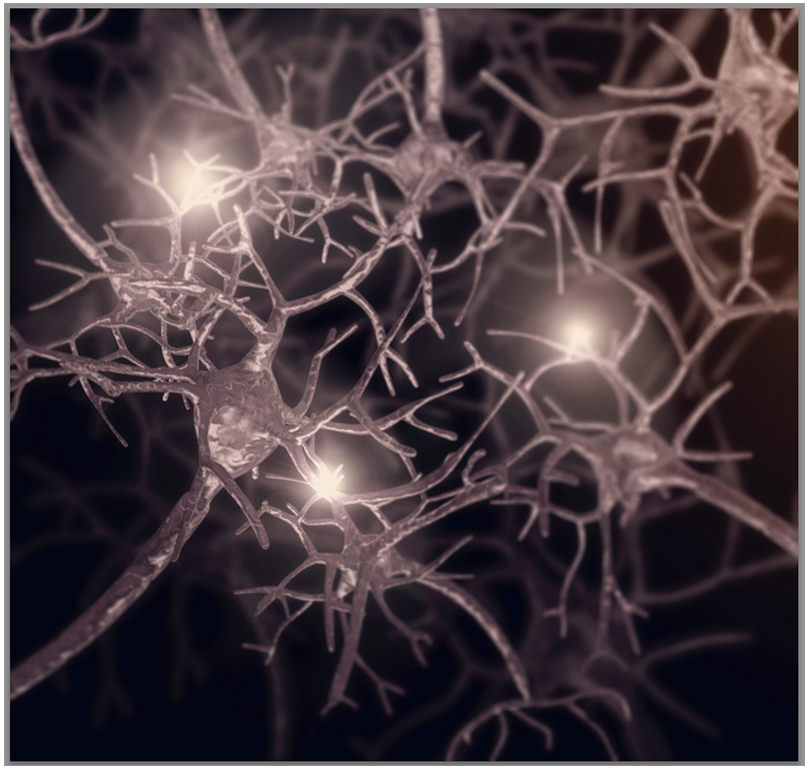
DO's & DON'Ts

DO's: Translational Research

Who's been there and done that?

Learn from success and avoid pitfalls

Awareness: Open Source Software for Data-driven discovery
[NVIDIA, Argonne, …]



DATA-DRIVEN DISCOVERY

DO's & DON'Ts

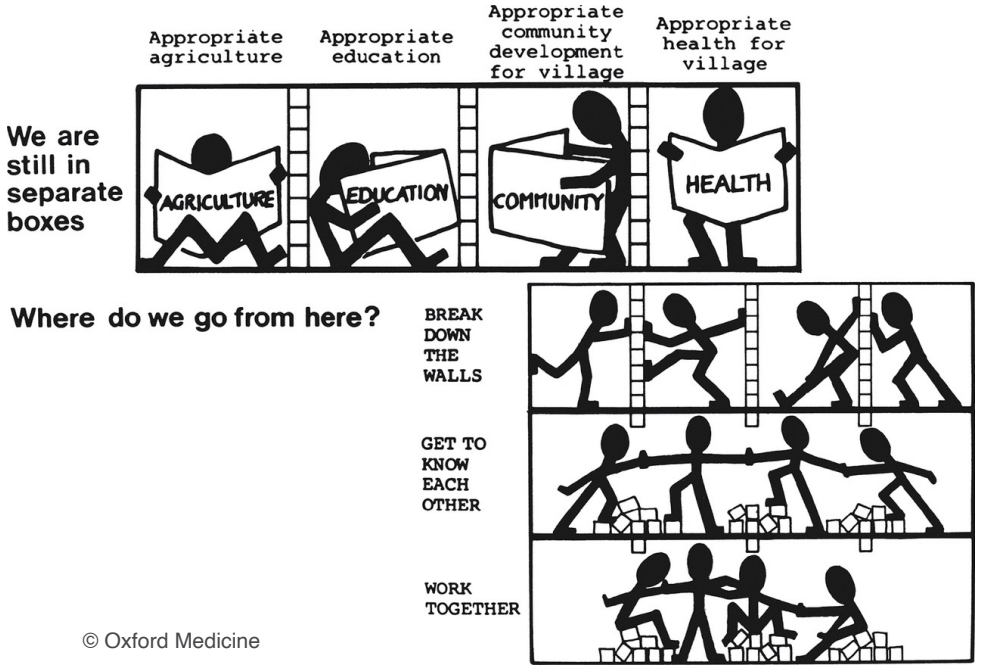
DONT's:

Work in silos

Dump your data on AI practitioners and hope for the best

Ignore AI advances [geometric deep learning, edge computing …]

Wait for mathmos to figure all out



© Oxford Medicine

AT ARGONNE

Data Science and Learning Division

https://www.anl.gov/dsl

IN THE NEWS

AI Detects Gravitational Waves Faster than Real Time

NVIDIA

IN THE NEWS

Detecting gravitational waves using AI

Tech Explorist

IN THE NEWS

Scientists develops AI model to detect gravitational waves

Sify

CONTACT US

Data Science and Learning General Inquiries

+1-630-252-2000

AI Distinguished Lecture Series

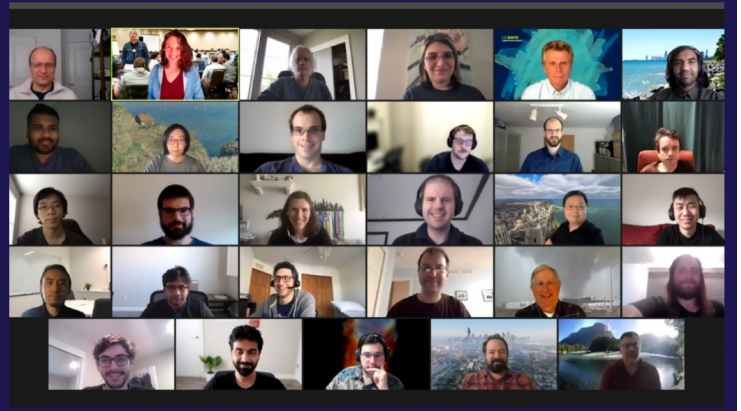
Argonne's AI Distinguished Lecture Series feature pioneers and innovators from around the world conducting research in foundational and applied AI. The lectures cover a wide variety of topics in academia, industry, finance, and technology.

LEARN MORE

Argonne Leadership Computing Facility

https://www.alcf.anl.gov

NEWS OUTREACH



ALCF training events help prepare researchers for current and future supercomputers

AT ARGONNE

Be nimble and agile

Harness extensive expertise in applied AI and advanced computing

Identify critical areas of development

Enhance & develop AI skills within your company

David Martin

Manager, Industry Partnerships and Outreach

+1 630-252-0929
dem@alcf.anl.gov

Argonne National Laboratory
9700 South Cass Avenue
Building 240 - Rm. 3126
Argonne, IL 60439

SAMPLE CASES

facta, non verba

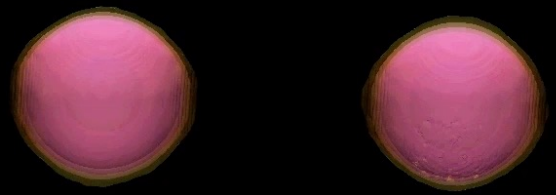
AI SURROGATES

Multi-scale and multi-physics simulations

© Never underestimate a drone: deep learning for turbulence
Astrobites blog, 2 Jan 2020

Turbulence: approximate, first principles models; highly non-linear; complex mathematical formulation

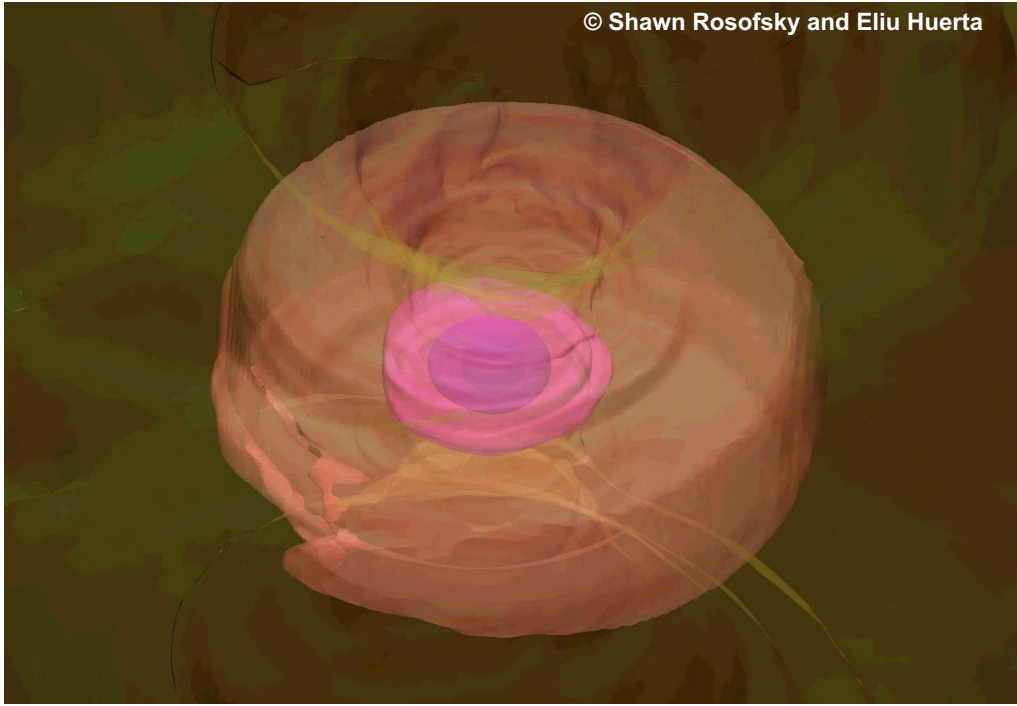
May AI be capable of learning and accurately describing the physics of turbulence?



GRAVITATIONAL WAVE ASTRONOMY

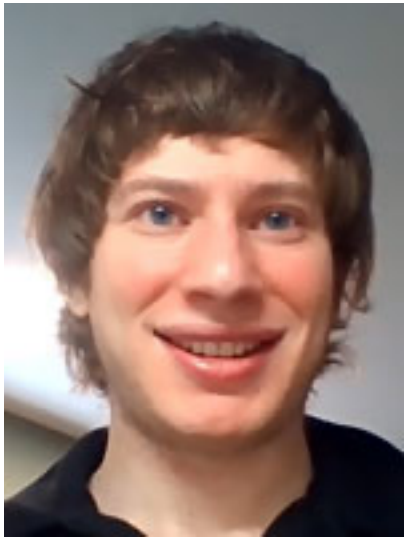
WHAT

Gravitational (and electromagnetic) wave observations of neutron star mergers



AI FOR TURBULENCE

Data-driven discovery
Cross-pollinate expertise between academia and industry



Shawn Rosofsky

PHYSICAL REVIEW D

covering particles, fields, gravitation, and cosmology

Highlights Recent Accepted Collections Authors Referees Search Press

Artificial neural network subgrid models of 2D compressible magnetohydrodynamic turbulence

Shawn G. Rosofsky and E. A. Huerta
Phys. Rev. D **101**, 084024 – Published 9 April 2020

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD® ABOUT CENTERS CAPABILITIES RESEARCH

Focal Points

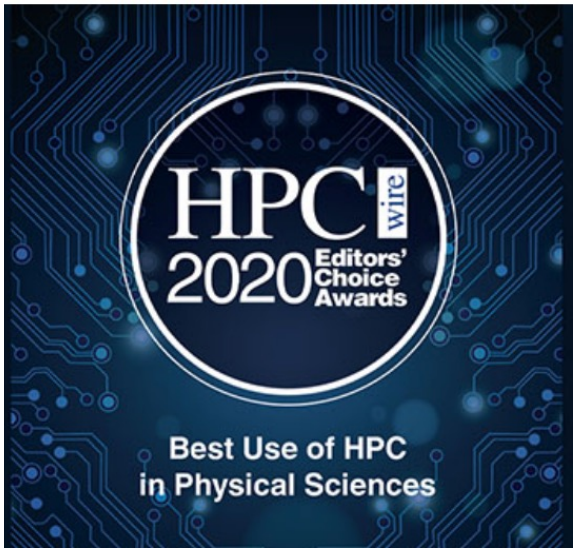
AI AND MACHINE LEARNING

We're harnessing the power of artificial intelligence and machine learning in ways that benefit our nation, with an emphasis on ethics and safeguarding privacy.

CONVERGENCE OF AI AND LARGE SCALE SIMULATIONS

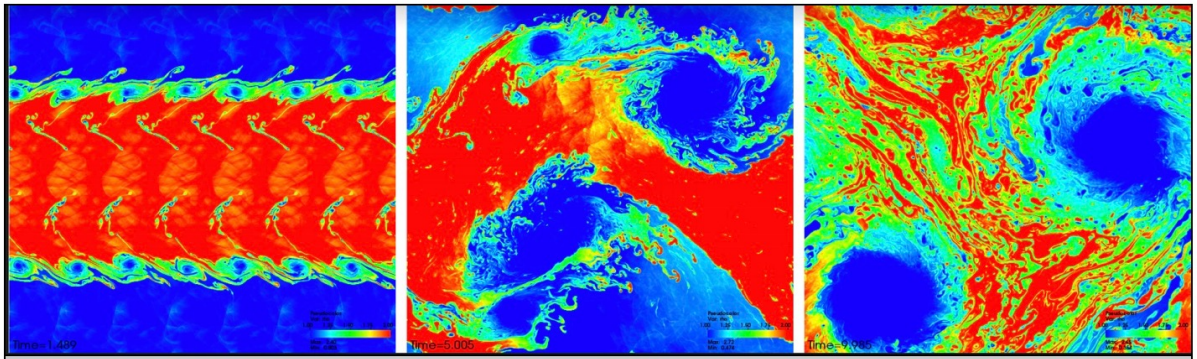
Star Crash

HOW



Artificial Intelligence on XSEDE Systems Is Key to Speeding Simulations of Neutron Star Mergers

By Ken Chiacchia, Pittsburgh Supercomputing Center



The intense magnetic fields accompanying movement of matter from neutron-stars past each other causes increasingly complicated turbulence that is computationally expensive with standard simulation methods. In this time series, a deep learning AI provides a simulation of this process at a fraction of the computing time.

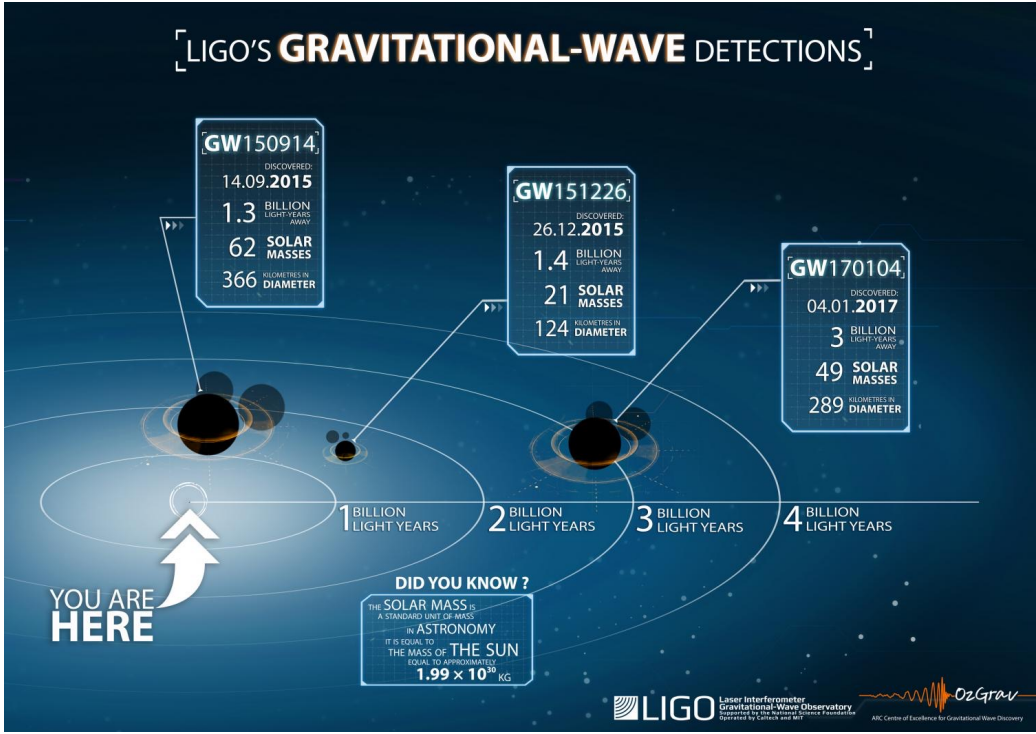
SAMPLE CASES

facta, non verba

GRAVITATIONAL WAVE ASTRONOMY

WHAT

Ground-based detectors continue to improve their sensitivity to gravitational wave sources



EXTRACTING WEAK SIGNALS IN NOISY BACKGROUNDS

Realistic datasets

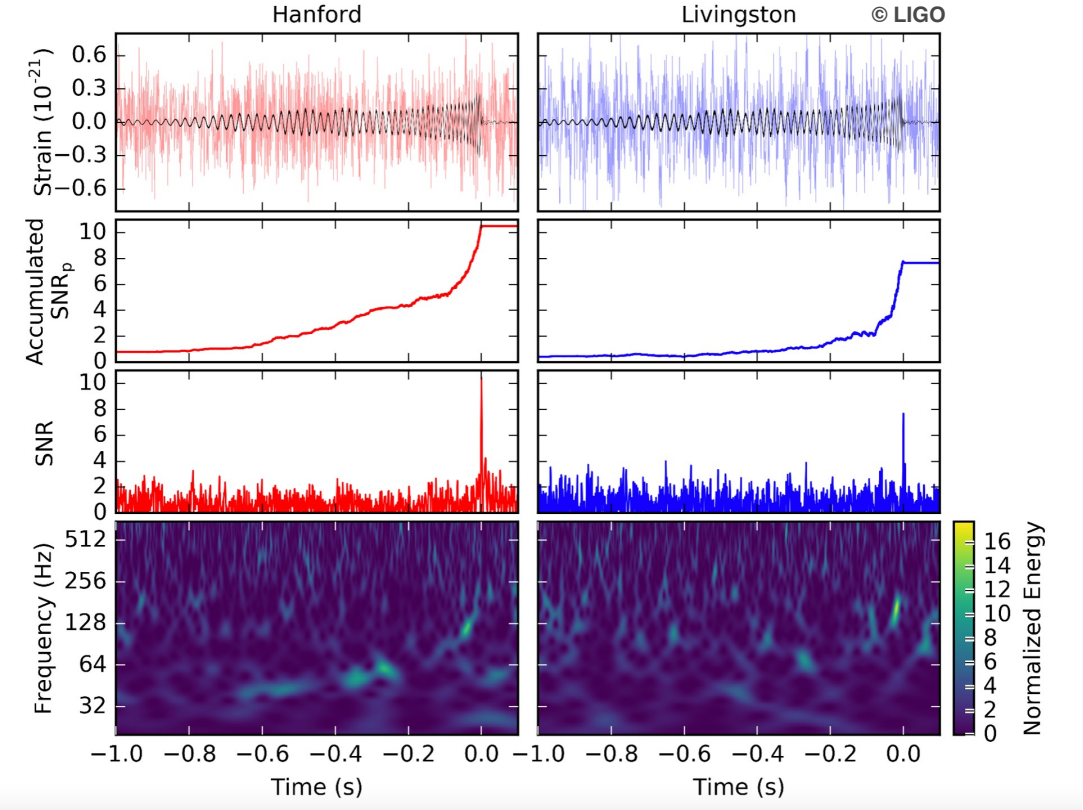
Challenge:

High-dimensional signal manifold

Lightweight, high speed data production

Non-Gaussian and non-stationary noise

Noise contamination

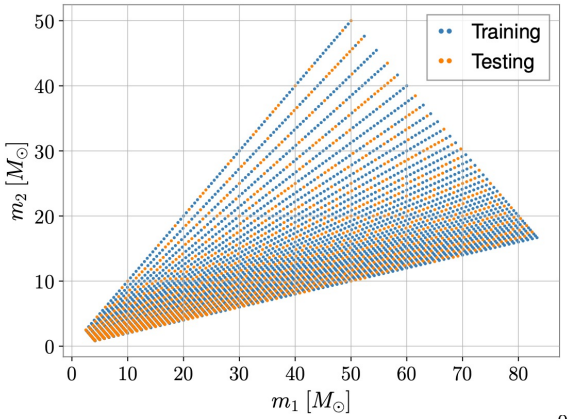


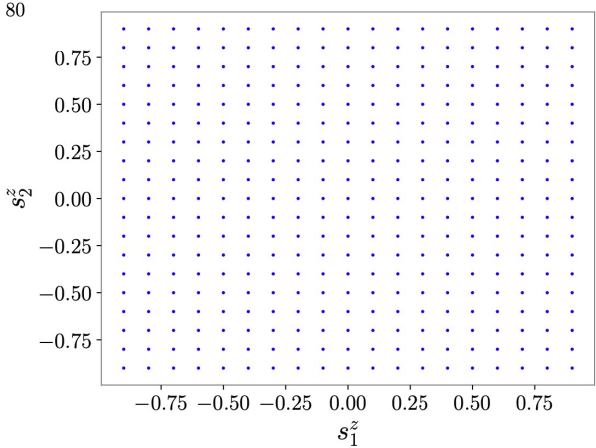
GRAVITATIONAL WAVE ASTRONOMY

WHAT

Demonstrate that AI + HPC provide a novel solution for production scale AI-driven gravitational wave detection

Consider 4-D signal manifold of real-time gravitational wave detection algorithms





GRAVITATIONAL WAVE ASTRONOMY

WHY

Number of detections continues to grow

Available computational resources remain finite and oversubscribed

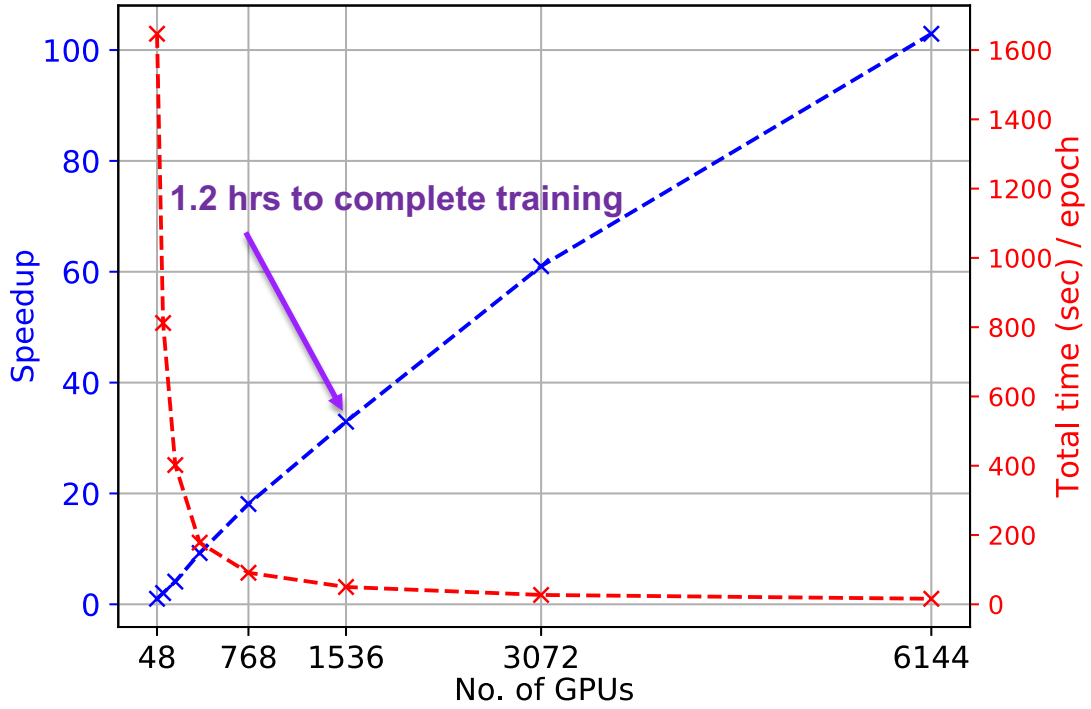
Radical re-thinking of computational methods for gravitational wave discovery

GRAVITATIONAL WAVE ASTRONOMY

WHAT

Densely sampling this 4-D signal manifold requires millions of modeled waveforms

Training stage: 1 month with a single NVIDIA V100 GPU



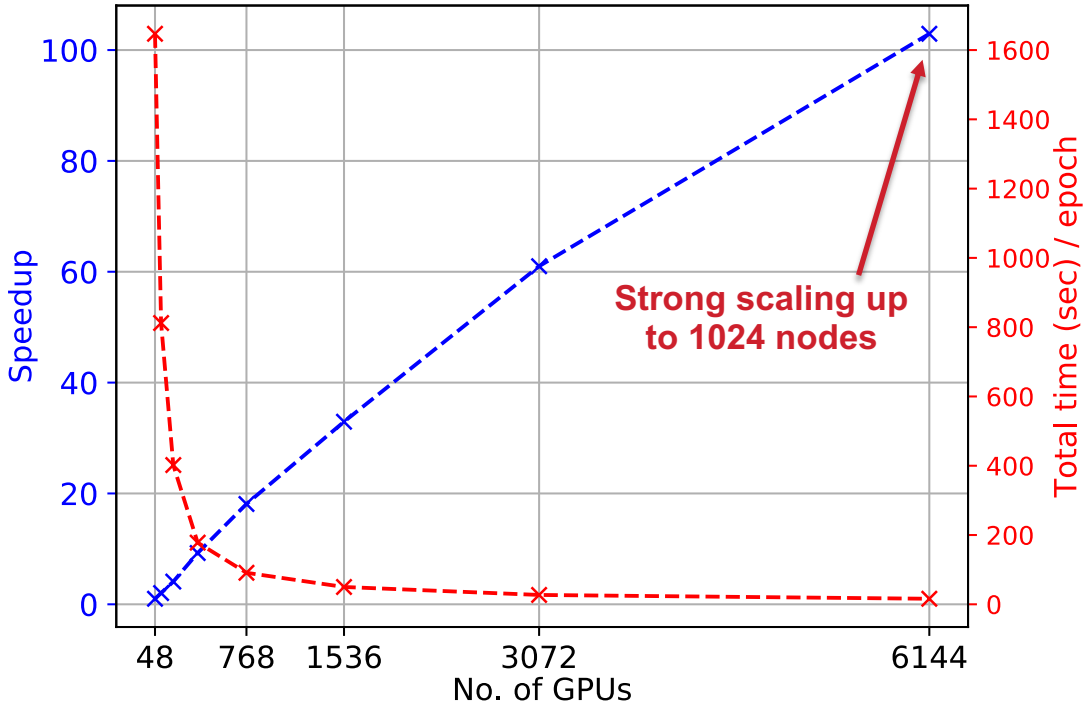
GRAVITATIONAL WAVE ASTRONOMY

HOW

Deployed and used new optimizers in Summit to reach optimal classification performance

600-fold speed up in training

Developed AI ensemble for real-time gravitational wave detection



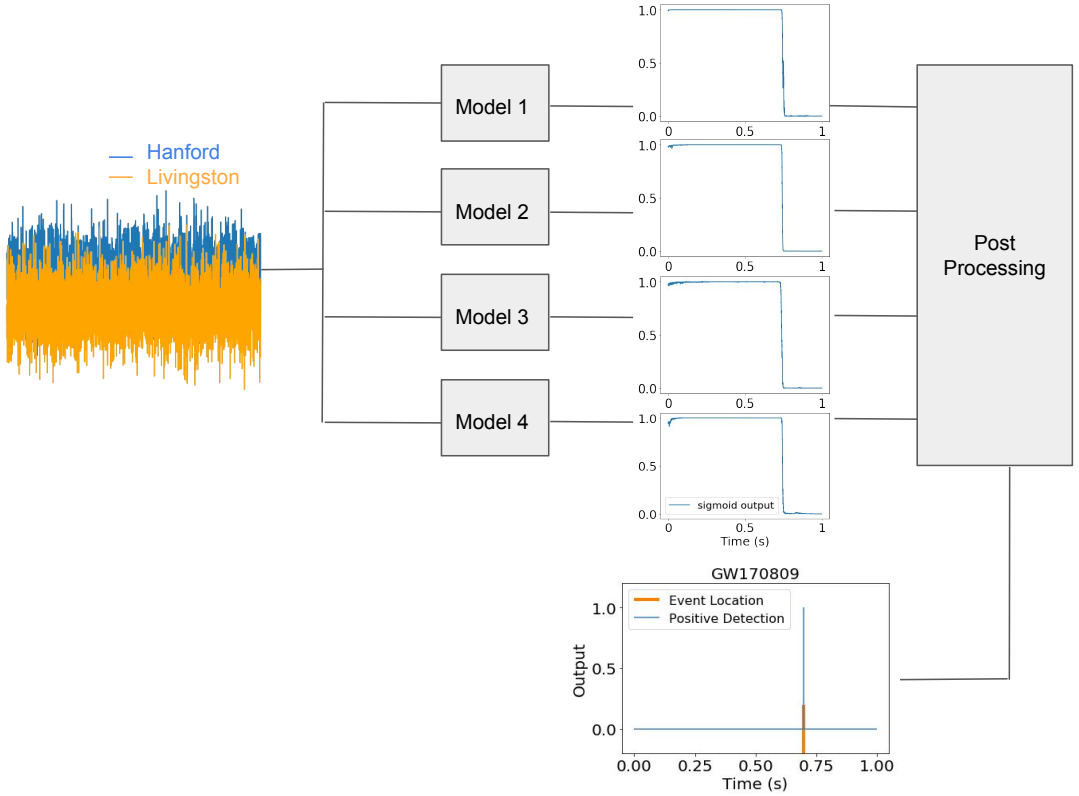
GRAVITATIONAL WAVE ASTRONOMY

HOW

4 AI models processing advanced LIGO data in tandem

Each model processes data faster than real time

Combine output of all models to increase confidence of detection



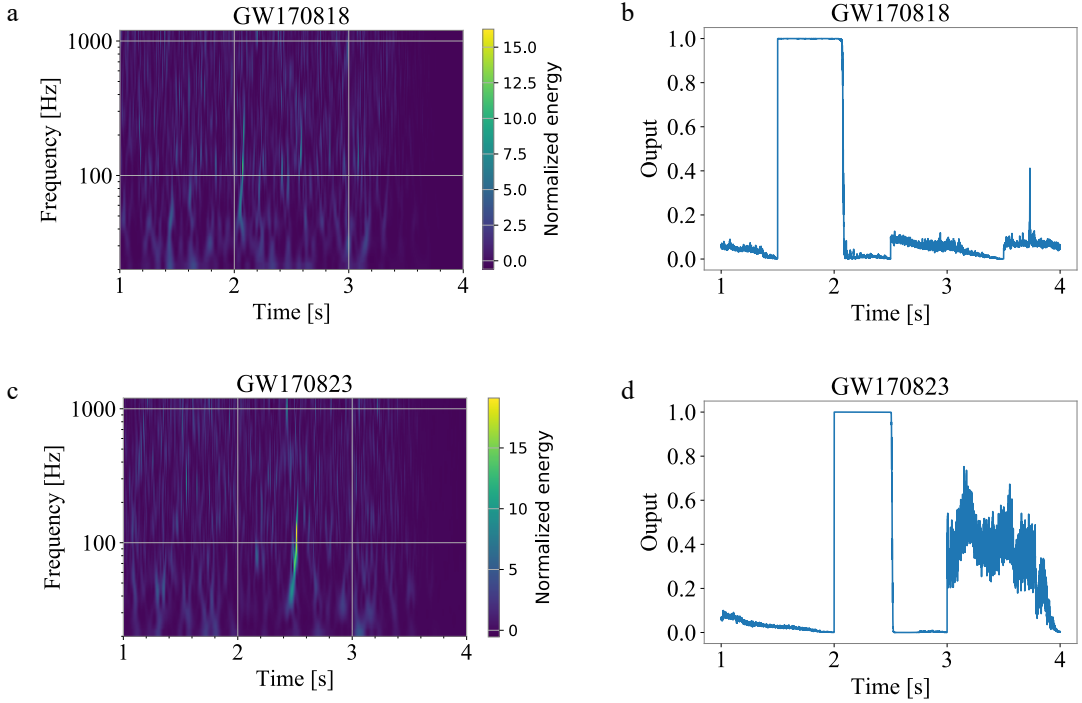
GRAVITATIONAL WAVE ASTRONOMY

HOW

4 AI models processing advanced LIGO data in tandem

Each model processes data faster than real time

Target: identify real events while reducing # of misclassifications



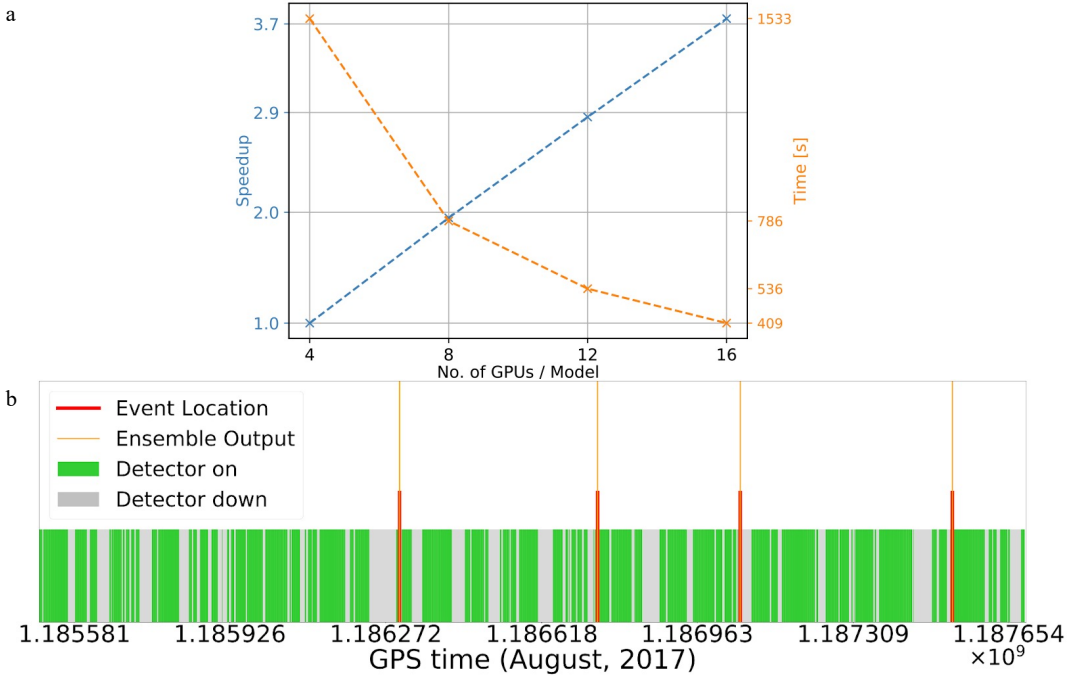
GRAVITATIONAL WAVE ASTRONOMY

HOW

Use AI ensemble to process one month of advanced LIGO data

Quantify sensitivity, inference speed and scalability

Distribute inference over the entire Hardware-Accelerated Learning (HAL) cluster at NCSA
[IBMPower9 system with 64 NVIDIA V100 GPUs]



GRAVITATIONAL WAVE ASTRONOMY

REALLY?

Establish reproducibility, scalability and performance of results

Make AI ensemble and post-processing pipeline open source and containerized at the Data and Learning Hub for Science (DLHub)

DLHub

Data and Learning Hub for Science

A simple way to find, share, publish, and run machine learning models and discover training data for science

Documentation

Read the Docs | Examples | Python SDK | CLI

DLHub Models

Browse Models

Papers and Presentations

DLHub Paper | DLHub Slides

GRAVITATIONAL WAVE ASTRONOMY

REALLY?

Open source + containerized is great

Can we do better than that?

DLHub

Data and Learning Hub for Science

A simple way to find, share, publish, and run machine learning models and discover training data for science

Documentation

Read the Docs | Examples | Python SDK | CLI

DLHub Models

Browse Models

Papers and Presentations

DLHub Paper | DLHub Slides

PRESENT – STATIC APPROACH

DLHub+funcX: reproducible, scalable and accelerated AI-discovery at the edge

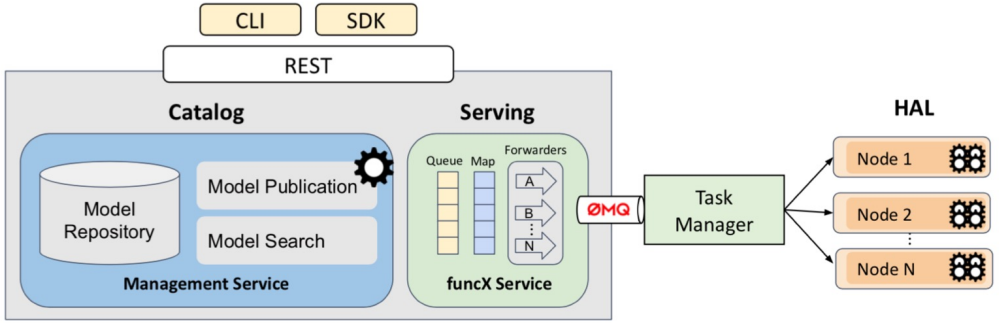
Reduce time-to-insight with HPC platforms
Optimal distributed training

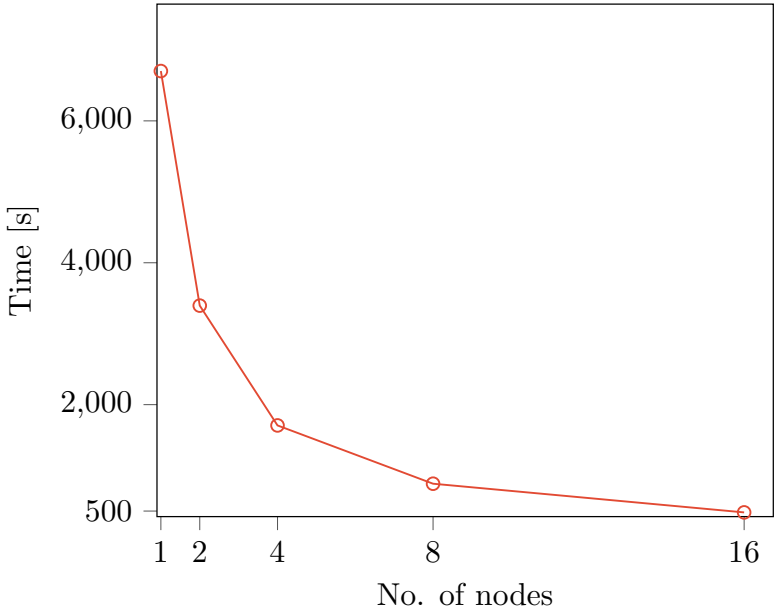
Deploy AI models in DLHub

Already used at scale!

GRAVITATIONAL WAVE ASTRONOMY

funcX + DLHub





Deploy funcX end-point at HAL

Call AI models hosted at DLHub

Optimal scalability, reproducibility established

PRESENT – STATIC APPROACH

DLHub+funcX: reproducible, scalable and accelerated AI-discovery at the edge

Reduce time-to-insight with HPC platforms
Optimal distributed training

Deploy AI models in DLHub

TensorRT

Already used at scale!

TensorRT further reduced the analysis to just 2 minutes!

Contributor | Nature Astronomy

BEHIND THE PAPER

From Disruption to Sustained Innovation: Artificial Intelligence for Gravitational Wave Astrophysics

Eliu Huerta
Lead for Translational AI, Argonne National Laboratory

Follow

Published Jul 06, 2021

Open source, accelerated, reproducible AI

Extreme scale computing

Edge computing

Article | Published: 05 July 2021

Accelerated, scalable and reproducible AI-driven gravitational wave detection

E. A. Huerta, Asad Khan, Xiaobo Huang, Minyang Tian, Maksim Levental, Ryan Chard, Wei Wei, Maeve Heflin, Daniel S. Katz, Volodymyr Kindratenko, Dawei Mu, Ben Blaiszik & Ian Foster

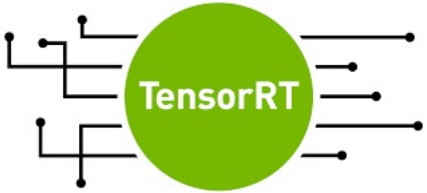
Nature Astronomy (2021) | Cite this article

DYNAMIC AI

DLHub+funcX: reproducible, scalable and accelerated AI-discovery at the edge

Summit
Theta-G AURORA …

Reduce time-to-insight with HPC platforms



Deploy dynamic AI models in DLHub

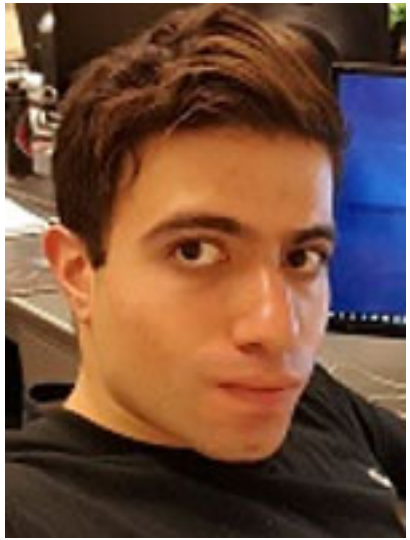
Edge Distributed Computing TensorRT …

Major upgrade of AI models

Active/Transfer/Reinforcement

Burst training

CROSS-POLLINATION OF EXPERTISE

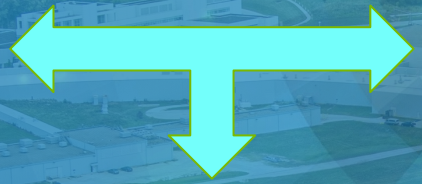


Asad Khan
Goldman Sachs
2021 Summer Intern

Wei Wei
Goldman Sachs
Associate

Minyang Tian
ByteDance AI Lab
2021 Summer Intern

AI-ready datasets



Innovative computing

FAIR, interpretable, physics-inspired, accelerated AI models

Data fusion & new modes of data-driven discovery & smart cyberinfrastructure

ACKNOWLEDGEMENTS

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725

We acknowledge support from NSF OAC-1931561, OAC-1934757, OAC-2004894 and DLHub: Argonne LDRD Project

We acknowledge support from NVIDIA and IBM


Argonne
NATIONAL LABORATORY
75
1946–2021