

HPC+AI for Science at NERSC



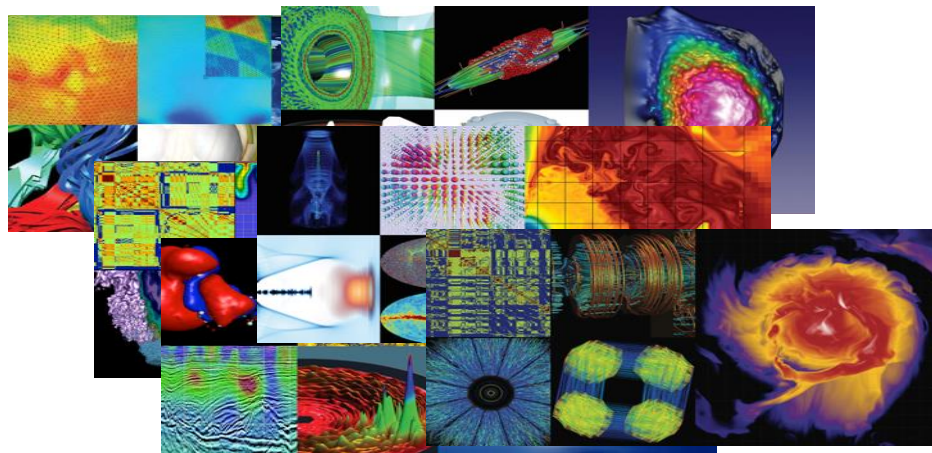
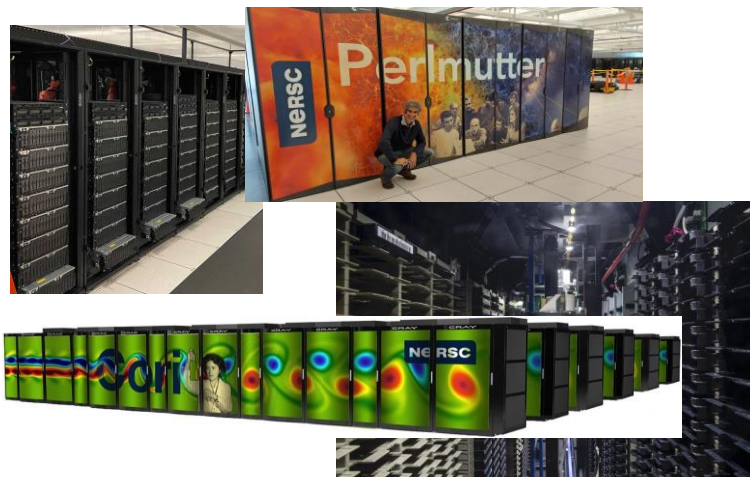
Peter Harrington
Data & AI Services, NERSC
Lawrence Berkeley National Laboratory



Outline

- I. AI for Science @ NERSC
 - Center overview & AI strategy
 - AI workload characteristics
- II. Application Highlight:
 - Data-driven forecasting & HENS

NERSC: Mission HPC for the Dept. of Energy Office of Science



Large compute and data systems

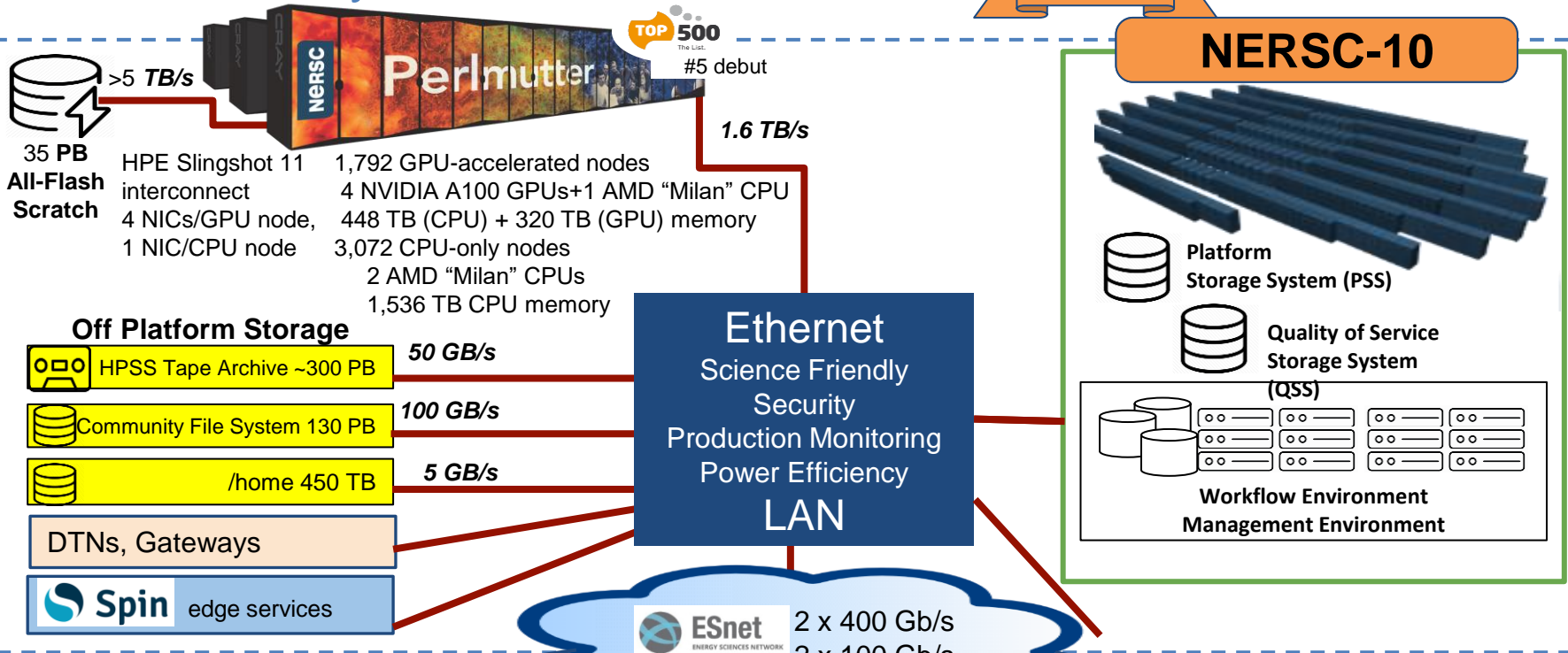
- Perlmutter: ~7k A100 GPUs
- 30 PB all-flash scratch filesystem
- 128PB Community Filesystem

Broad science user base

- > 10,000 users,
- 1000 projects,
- Gov, industry, academic research

NERSC Facility

**Coming
2026/27**



NERSC-10

Experimental Facility

ASCR Facility

Home Institution

Cloud

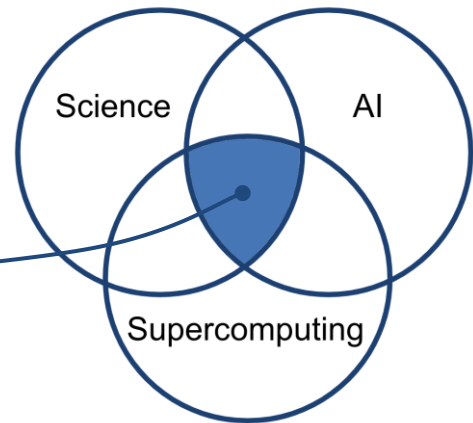
Edge



NERSC AI Strategy

- The **intersection** of HPC, AI, & science

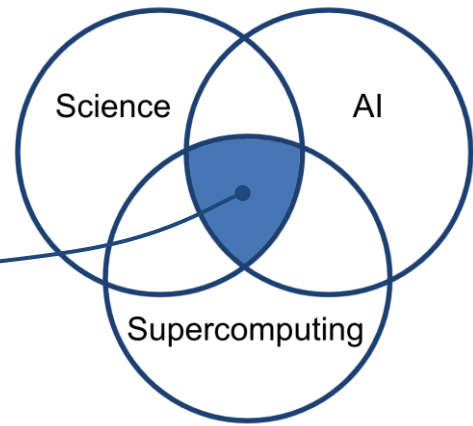
NERSC AI



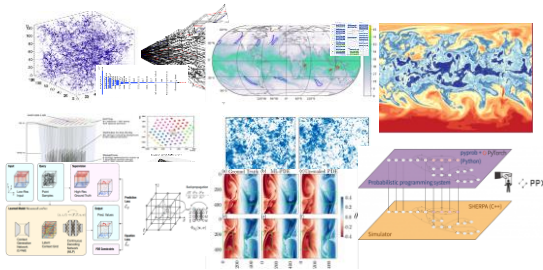
NERSC AI Strategy

- Focus activities in three main areas:

NERSC AI



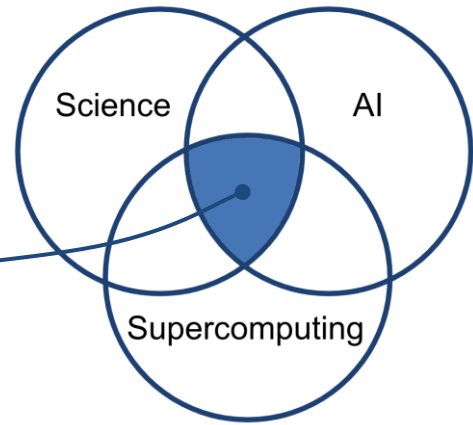
Methods and Applications



- NESAP and strategic projects
- Performance, scaling, tooling in addition to new methods & apps
- Leverage lessons learned for broader community

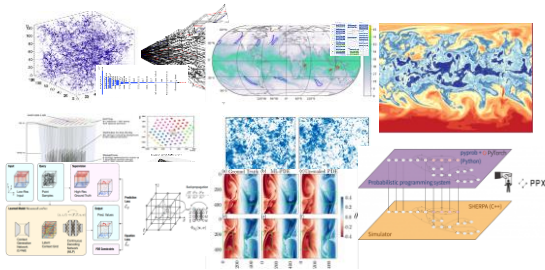
NERSC AI Strategy

- Focus activities in three main areas:



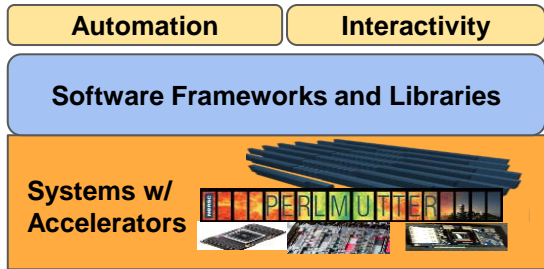
NERSC AI

Methods and Applications



- NESAP and strategic projects
- Performance, scaling, tooling in addition to new methods & apps
- Leverage lessons learned for broader community

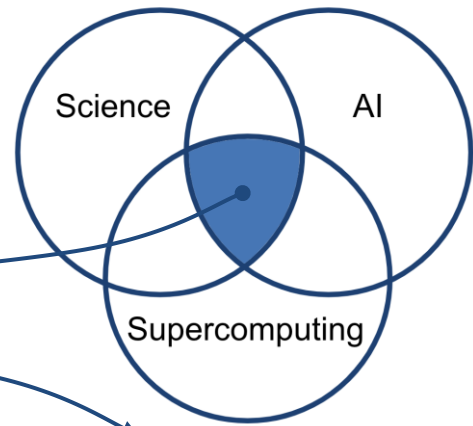
Deployment



- Optimized hardware and software systems for AI & data
- Collaborations on the full AI software stack (e.g. NCCL+SS11)
- Integration of ML tooling ecosystem

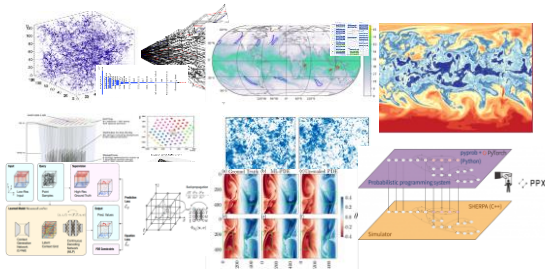
NERSC AI Strategy

- Focus activities in three main areas:



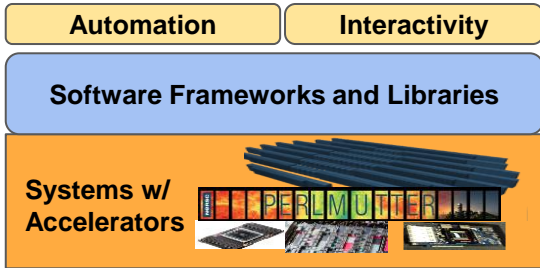
NERSC AI

Methods and Applications



- NESAP and strategic projects
- Performance, scaling, tooling in addition to new methods & apps
- Leverage lessons learned for broader community

Deployment



- Optimized hardware and software systems for AI & data
- Collaborations on the full AI software stack (e.g. NCCL+SS11)
- Integration of ML tooling ecosystem

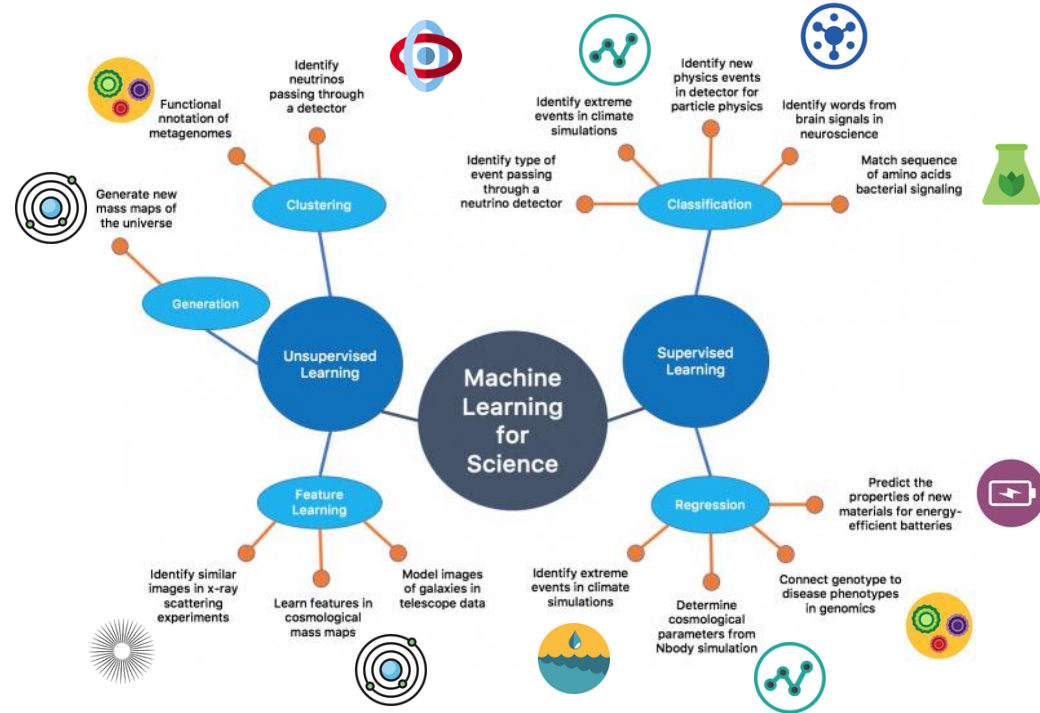
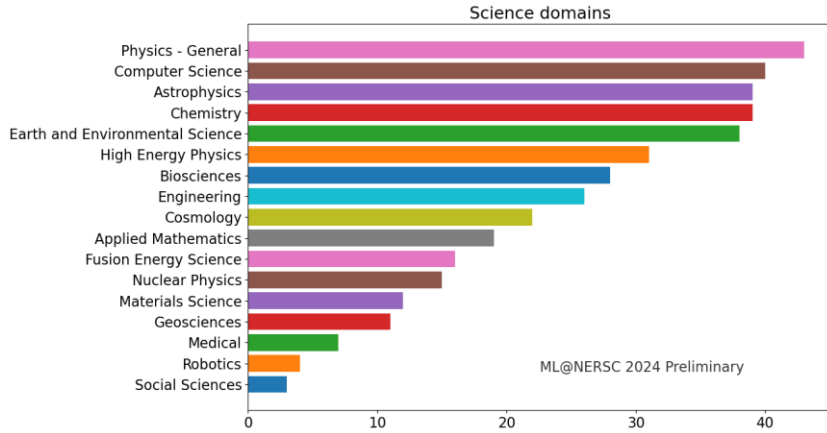
Empowerment



- Seminars, training and schools as well as staff, student intern and postdoctoral programs
- Over 20 DL@Scale tutorials

NERSC AI workload

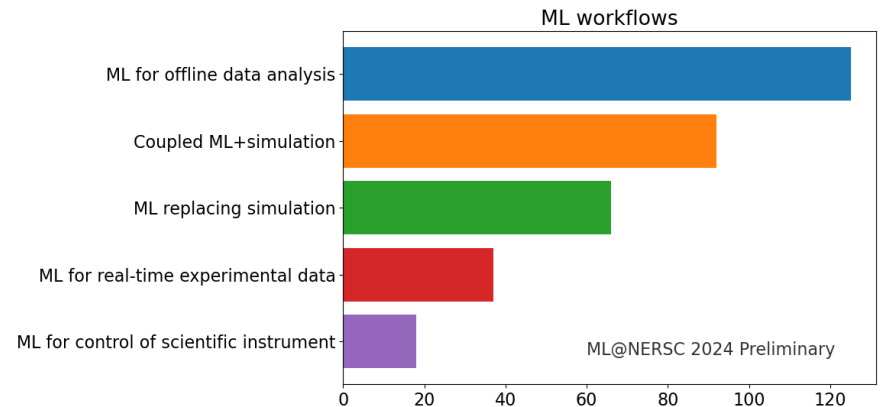
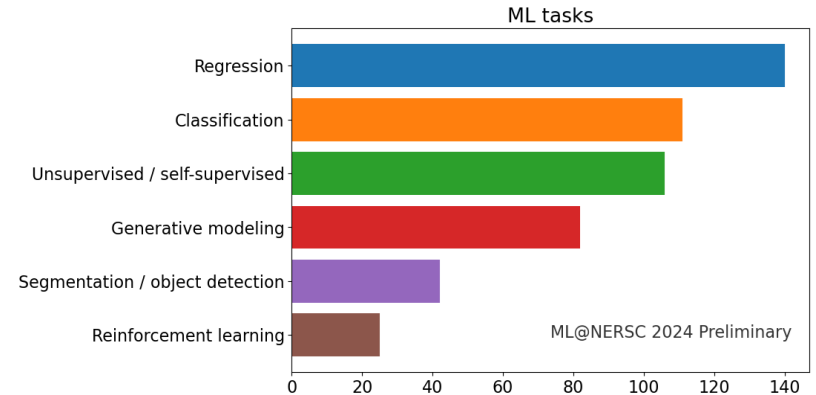
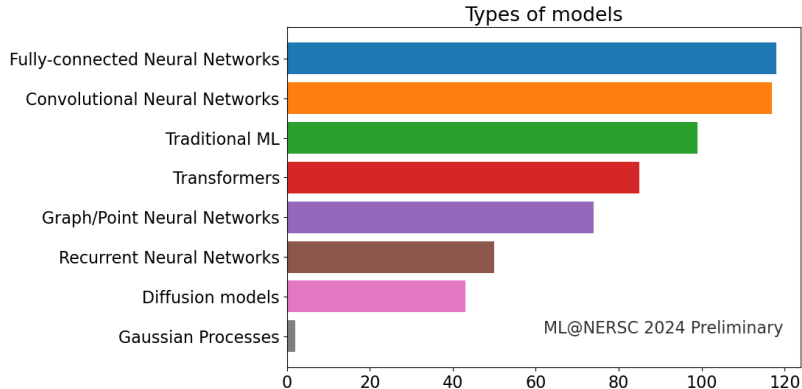
AI for science:
multifaceted
diversity!



NERSC AI workload

ML@NERSC Survey (preliminary results):

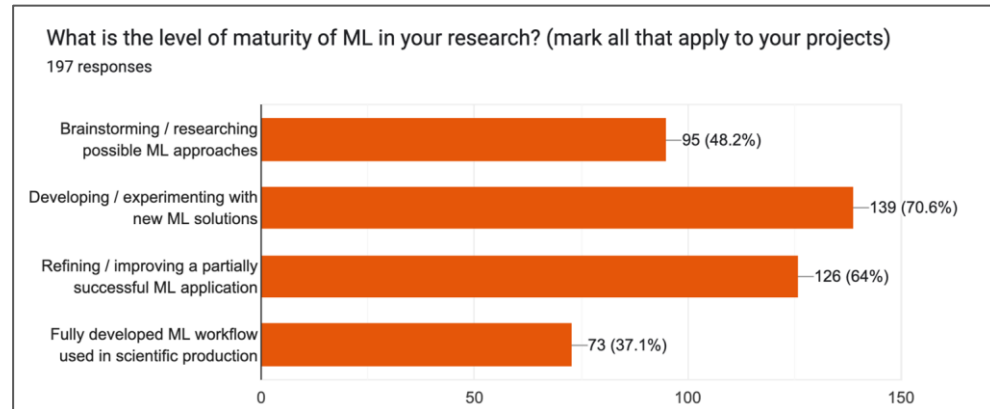
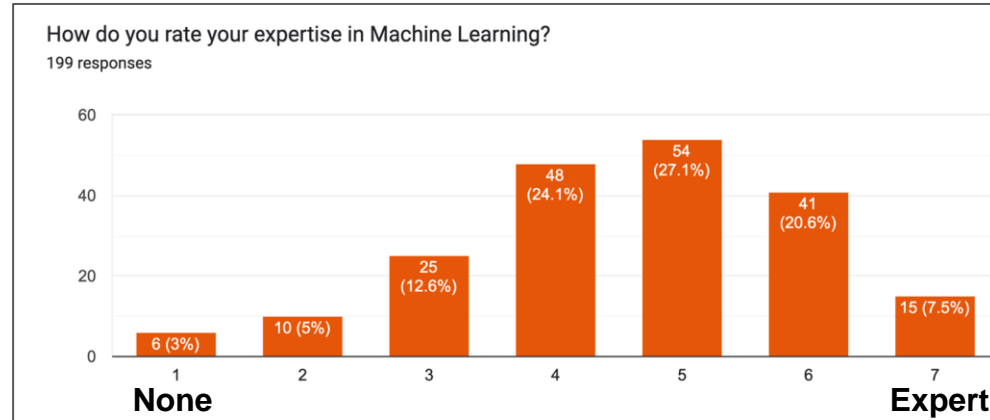
- No one dominant use-case
- Applications spanning full range of sophistication



NERSC AI workload

Range of expertise and maturity:

- Shifted gradually upwards since 2018
- 37% claim to have fully-developed ML “in production” for science



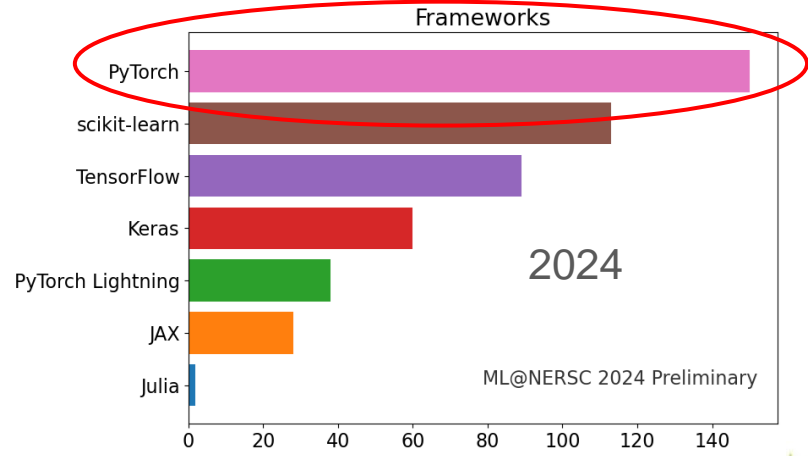
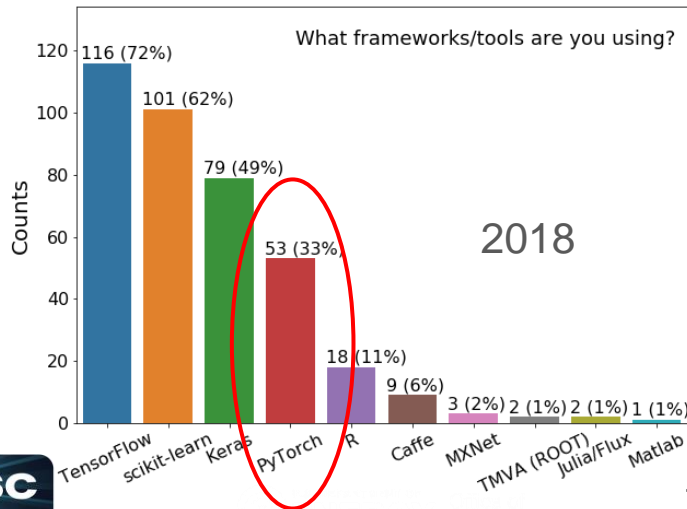
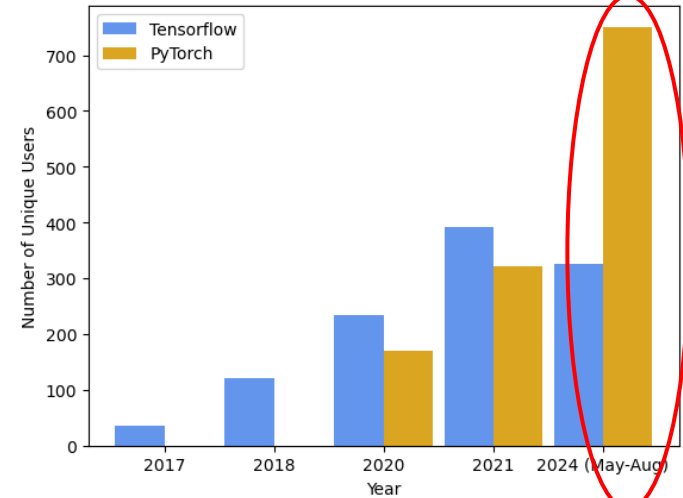
ML@NERSC2024 Preliminary



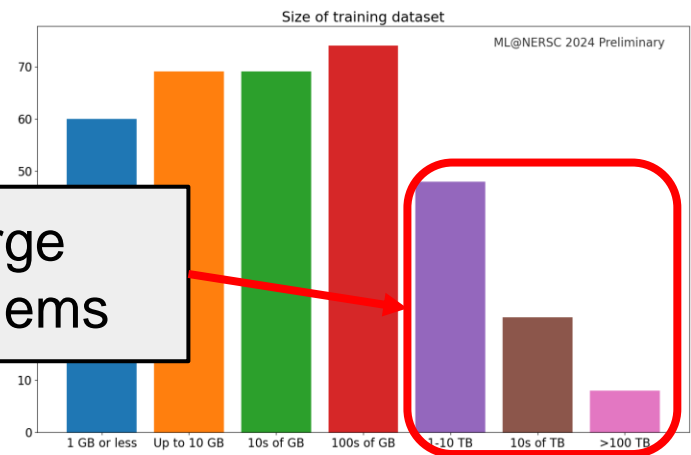
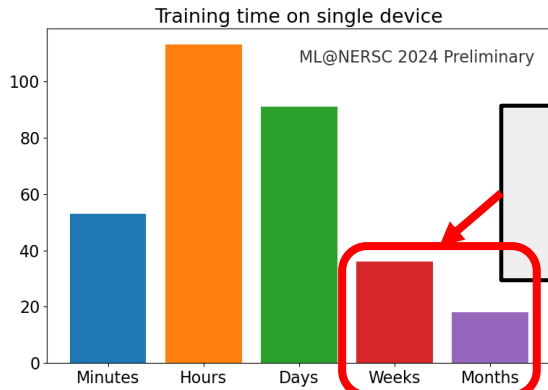
NERSC AI workload

AI software tracking via Python monitoring:

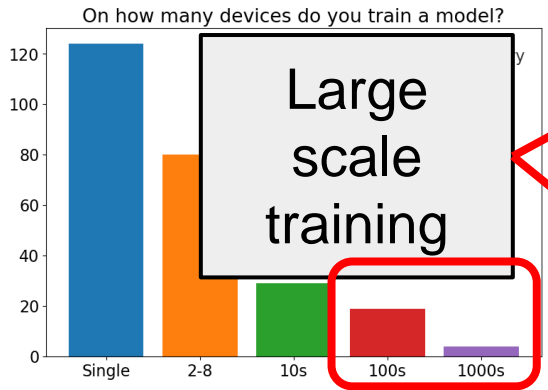
- **20x increase** in deep learning users since 2017
- PyTorch has taken the crown from TensorFlow



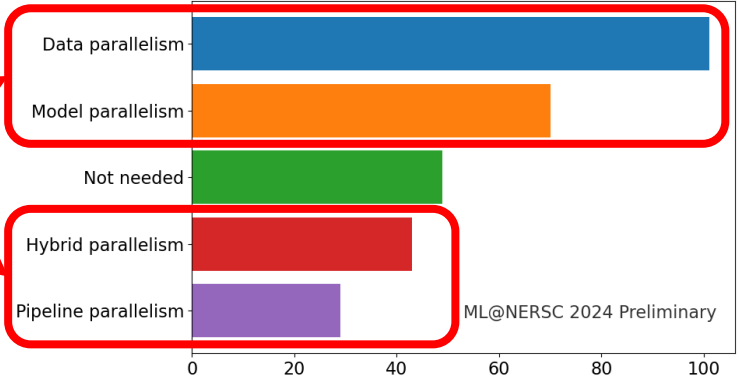
NERSC AI workload



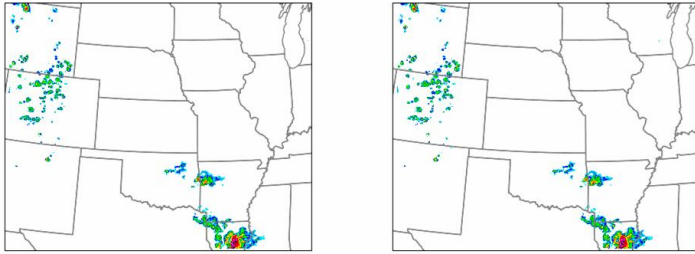
Large problems



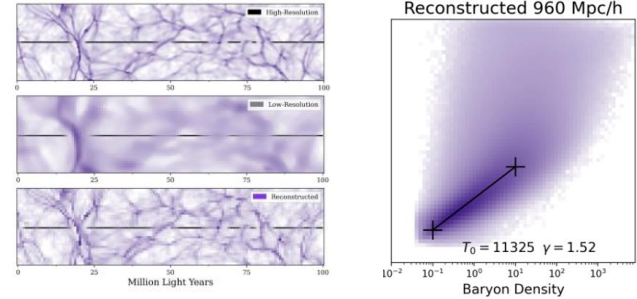
Large scale training



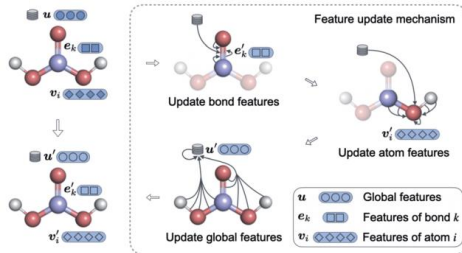
NERSC AI workload examples



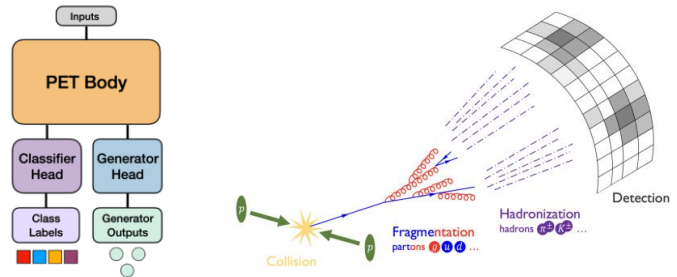
[StormCast: Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling](#) (Pathak et. al. Aug 2024) (led by Nvidia)



[Gigaparsec-scale super-resolution](#) for cosmological simulations (Zarija Lukić, LBNL)



[BonDNet](#): GNNs for reaction network active exploration (Sam Blau, LBNL; Wenbin Xu, NERSC)



[OmniLearn](#): a versatile foundation model for HEP (Vinicius Mikuni, NERSC; Ben Nachmann, LBNL)

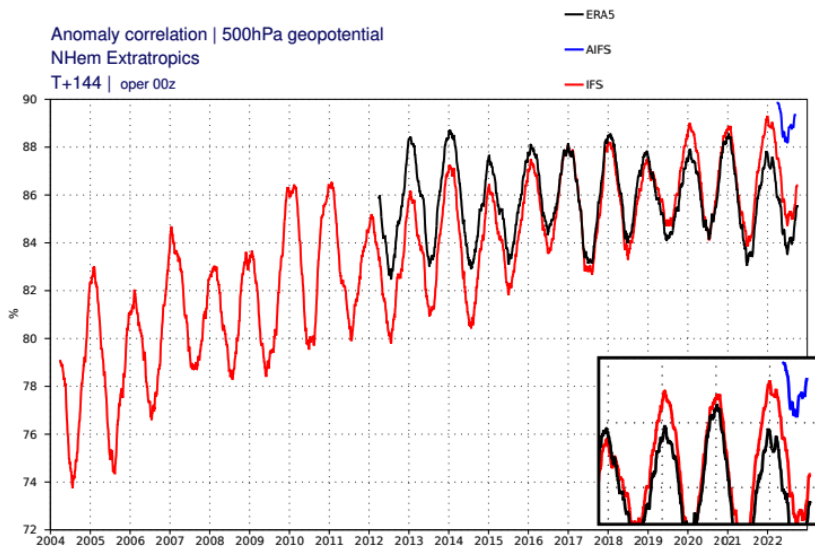


Outline

- I. AI for Science @ NERSC
 - Center overview & AI strategy
 - AI workload characteristics
- II. Application Highlight:
 - Data-driven forecasting & HENS

Data-driven weather & climate forecasting

- Deep learning has **rapidly caught up to the resolution and skill** of production NWP
- State of the art skill (RMSE*) in medium-range weather, with impressive speedups over NWP



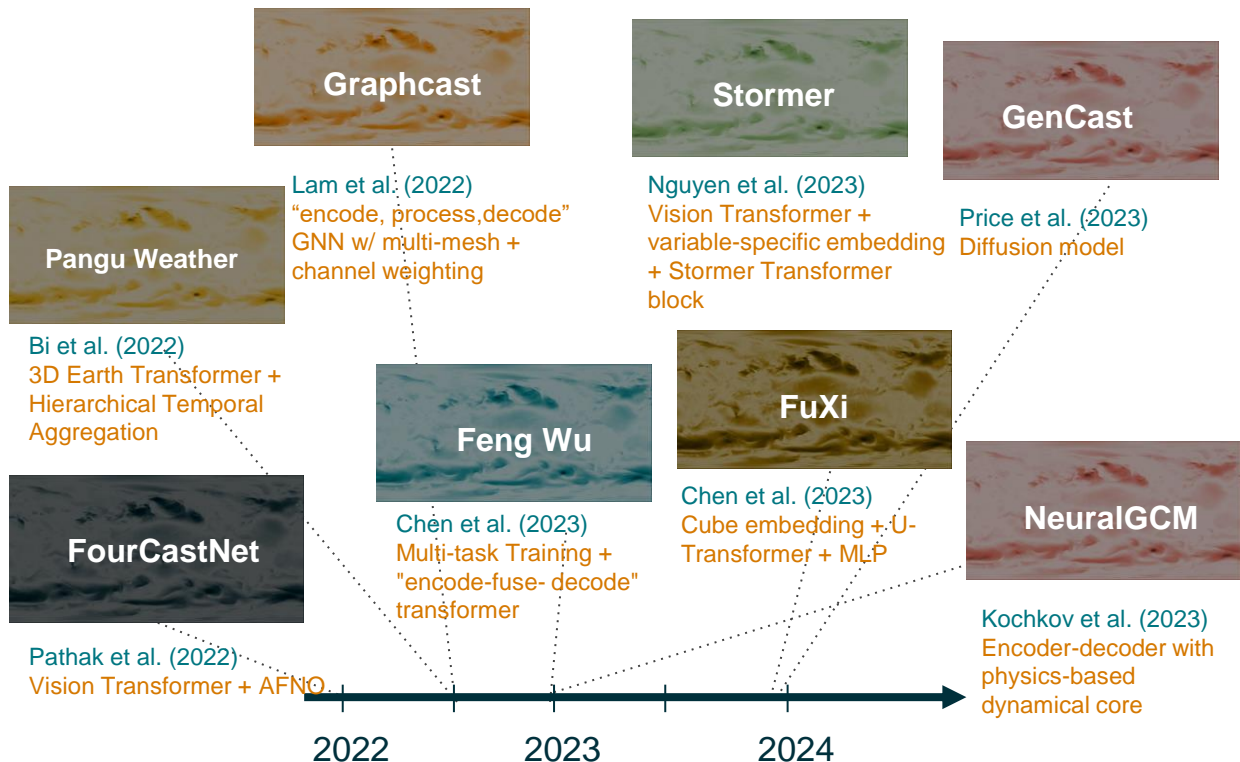
[Lang et al. \(2024\)](#)

<p>Latest forecast</p> <p>Experimental: GraphCast ML model: Mean sea level pressure and 850 hPa wind speed</p> <p>GraphCast (Google DeepMind): a deep learning-based system developed by Google DeepMind. It is initialised with ECMWF HRES analysis. GraphCast operates at 0.25° resolution.</p>	<p>Latest forecast</p> <p>Experimental: Pangu-Weather ML model: Mean sea level pressure and 850 hPa wind speed</p> <p>Pangu-Weather: a deep learning-based system developed by Huawei. It is initialised with ECMWF HRES analysis. Pangu-Weather operates at 0.25° resolution.</p>
<p>Latest forecast</p> <p>Experimental: AIFS (ECMWF) ML model: 500 hPa geopotential height and 850 hPa temperature</p> <p>AIFS (ECMWF): a deep learning-based system developed by ECMWF. It is initialised with ECMWF HRES analysis. AIFS operates at 0.25° resolution.</p>	<p>Latest forecast</p> <p>Experimental: FourCastNet ML model: 500 hPa geopotential height and 850 hPa temperature</p> <p>FourCastNet v2-small: a deep learning-based system developed by NVIDIA in collaboration with researchers at several US universities. It is initialised with ECMWF HRES analysis. FourCastNet operates at 0.25° resolution.</p>

<https://charts.ecmwf.int/>

DL in weather & climate “nowadays”

- Large variety of sophisticated models
- Community moving past singular benchmark of medium-range forecast skill (simple RMSE)
- What are the next frontiers?
 - Probabilistic forecasting
 - Foundation models
 - **Ensemble forecasting**



HENS: Huge ensembles for extreme weather

Actualize the potential of deep learning for massive ensemble forecasts in weather & climate

Collaboration (NESAP program at NERSC) between industry, gov, academia:

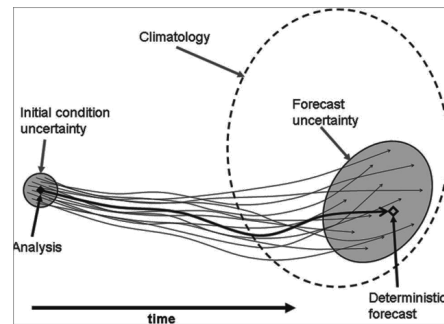
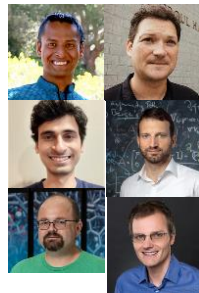
NERSC

Peter Harrington
Jared Willard
Shashank Subramanian



Nvidia

Mike Pritchard
Karthik Kashinath
Boris Bonev
David Pruitt
Thorsten Kurth
....others



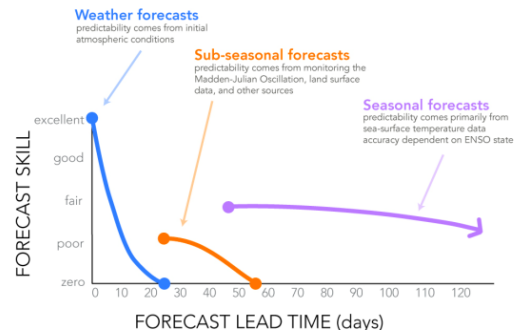
UC Berkeley / LBL Climate and Ecosystem Sciences

William Collins
Ankur Mahesh



Indiana University

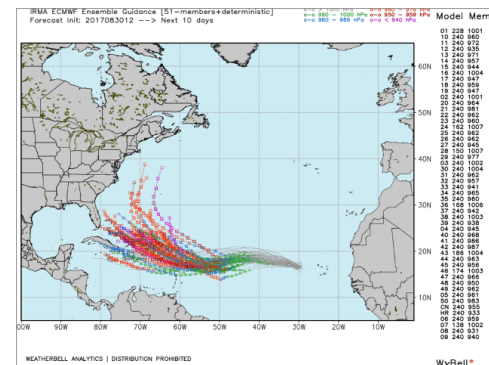
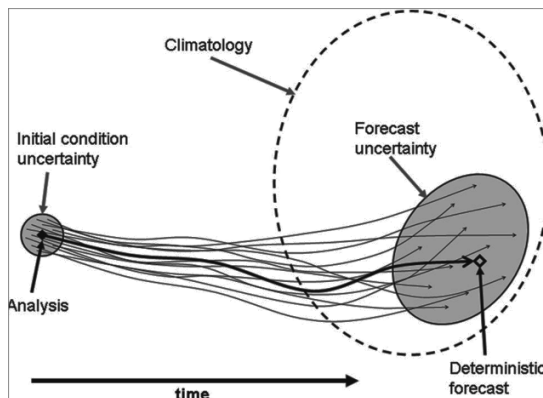
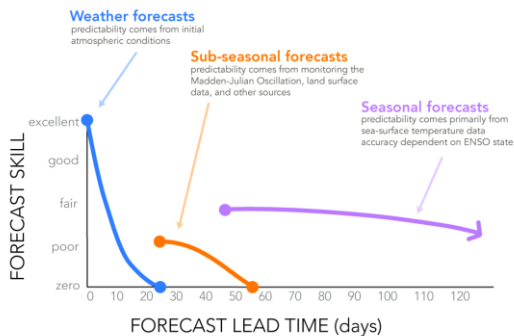
Travis O'Brien



HENS: Huge ensembles for extreme weather

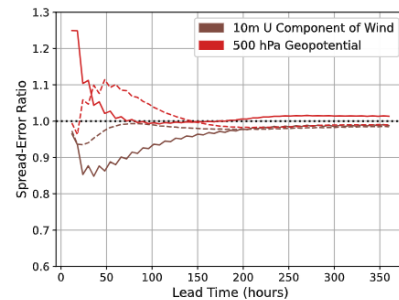
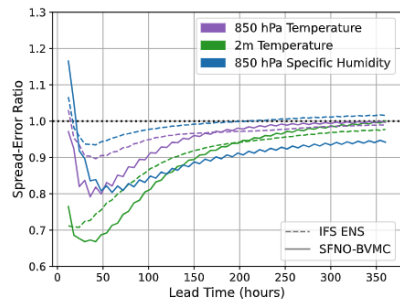
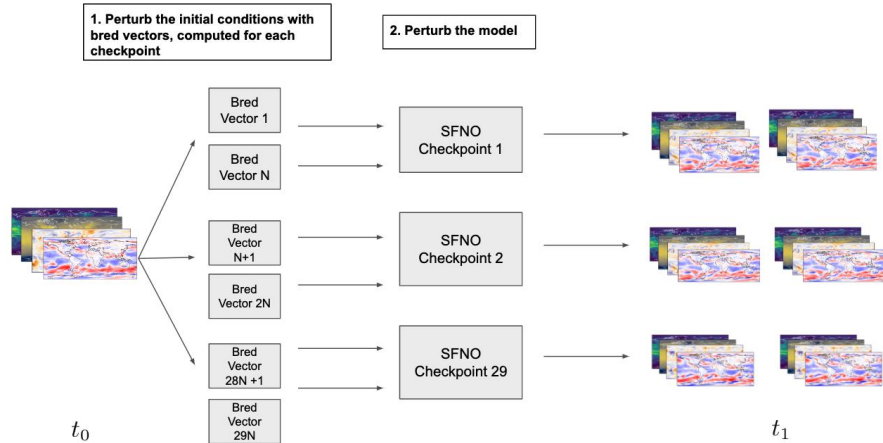
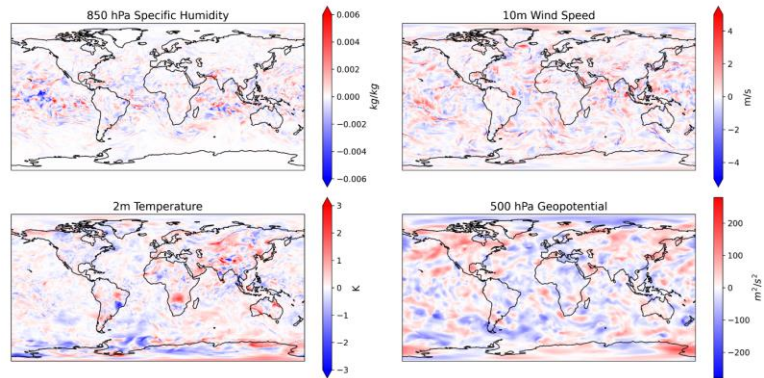
Deep learning enables much larger ensemble forecasts than previously possible. Questions:

- How do we **initialize and propagate** a DL-based large ensemble?
- Does a well-calibrated ensemble better characterize & capture **extreme events**?
- What is the **value add** of a $O(1000)$ member ensemble over conventional $O(100)$?



HENS: Huge ensembles for extreme weather

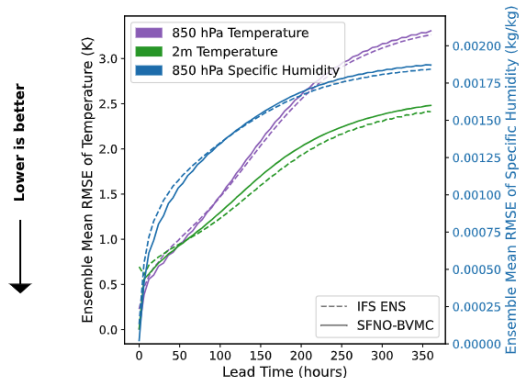
- How do we initialize and propagate a DL-based large ensemble?
 - IC perturbations: **bred vectors**
 - Iteratively breed fast-growing modes of the system for each var
 - Model- and dataset-agnostic
 - Model dispersion: **multi-checkpoint ensemble**



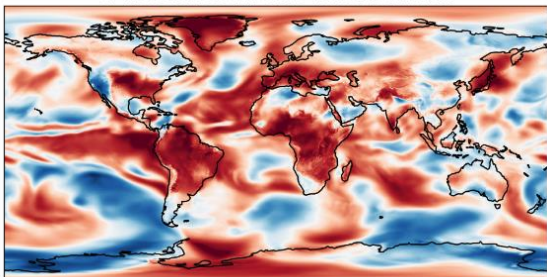
Comparable to IFS ensemble skill, calibration

HENS: Huge ensembles for extreme weather

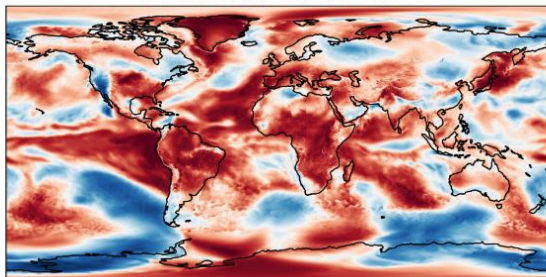
- Does a well-calibrated ensemble better characterize & capture **extreme events**?
 - Ensemble mean and dispersion match IFS on aggregate
 - We also see close agreement in extreme metrics:
 - Extreme Forecast Index (EFI) spatial patterns: capturing hot and cold extremes of varying intensities
 - Reliability diagrams match IFS for extreme percentiles



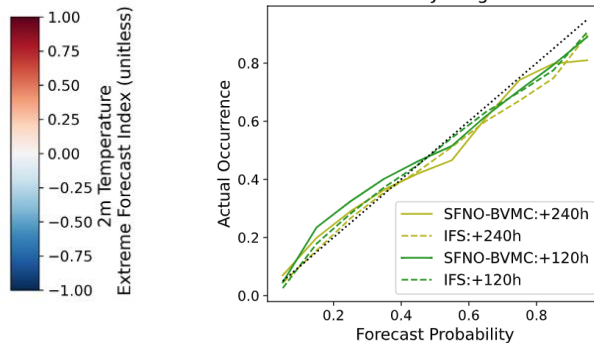
SFNO-BVMC Extreme Forecast Index



IFS Extreme Forecast Index

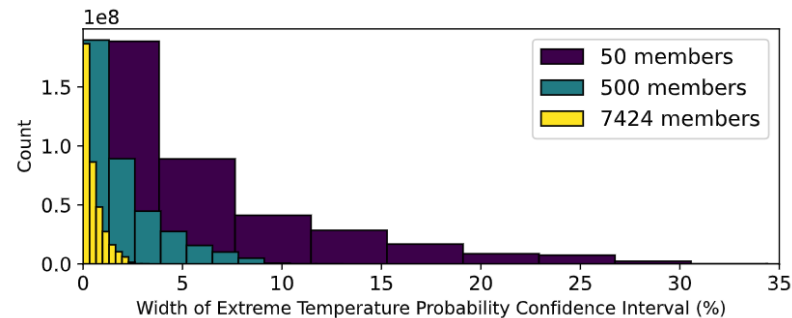
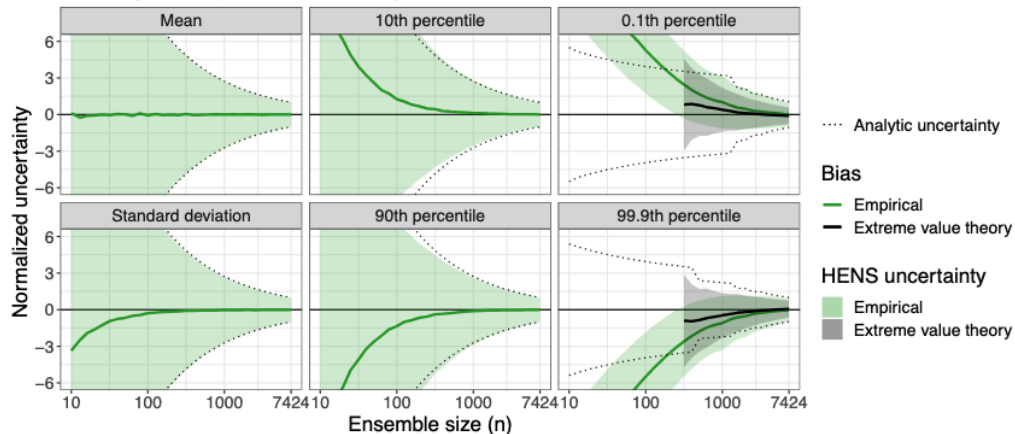


95th Percentile 2m Temperature Reliability Diagram



HENS: Huge ensembles for extreme weather

- What is the **value add** of a $O(1000)$ member ensemble over conventional $O(100)$?
 - Summer 2023 HENS run: 7424 members, 3PB; largest ever ensemble weather forecast at this scale
 - Order of magnitude reduction in uncertainty of extremes!



HENS: Huge ensembles for extreme weather

- **Case study:** 08/2023 Kansas City extreme heatwave
35°C air temperature, 56% relative humidity, heat index of 43°C
- 10-day IFS ensemble: warmer than average, but no members captured both surface heat & humidity.
- HENS samples the tails of the forecast distribution and is able to capture the magnitude of the event, plus contours

Record Breaking August 2023 Heat

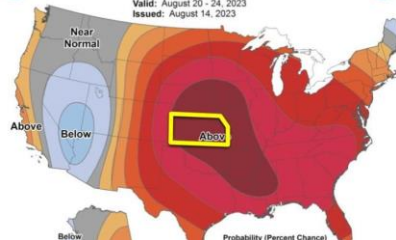
[Weather.gov](#) > [Wichita, Kansas](#) > Record Breaking August 2023 Heat

Wichita, Kansas
Weather Forecast Office

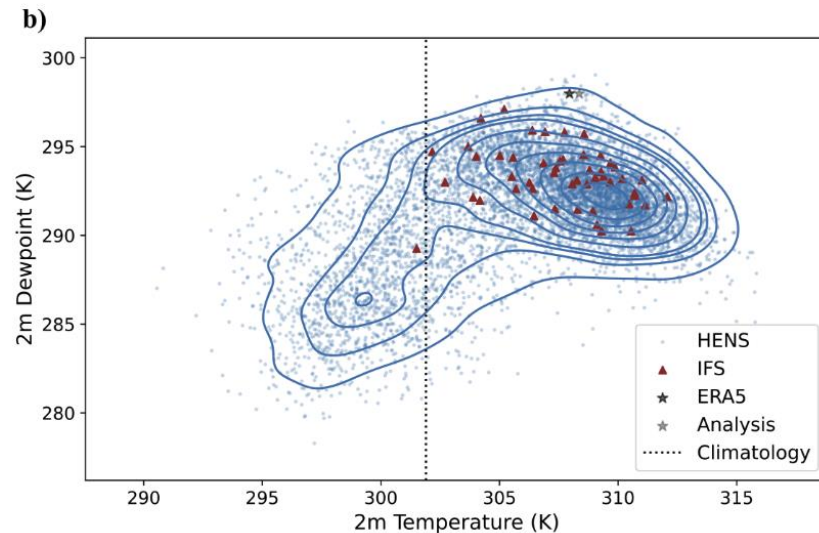
[Current Hazards](#) [Current Conditions](#) [Radar](#) [Forecasts](#) [Rivers and Lakes](#) [Climate and Past Weather](#) [Local Programs](#)

6-10 Day Temperature Outlook

Valid: August 20 - 24, 2023
Issued: August 14, 2023



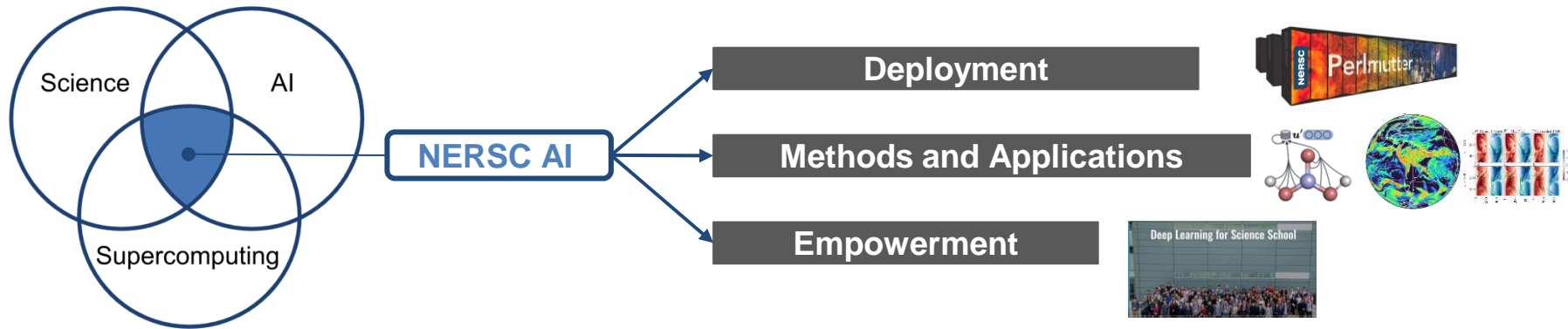
https://www.weather.gov/ict/event_2023AugHeat



More details: HENS [Part I](#), [Part II](#)

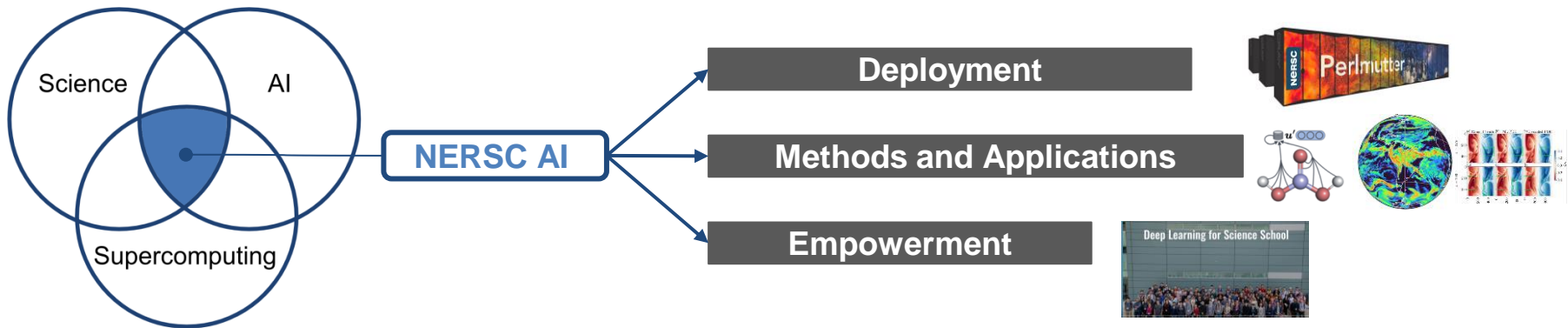
Summary & conclusions

- AI for science is very broad; we see increasing sophistication & scientific impact:
 - Accelerated by interdisciplinary collaborations (engineering teams, domain scientists, and AI expertise)
 - ...and HPC infrastructure!
- Future is bright, and many open questions in scientific AI!



Summary & conclusions

- Catch us at SC24! **Deep Learning at Scale Tutorial** in collaboration with NVIDIA, OLCF
 - Performance optimization
 - Data and model parallel training
 - Hands-on scientific AI examples





Backup



BERKELEY LAB



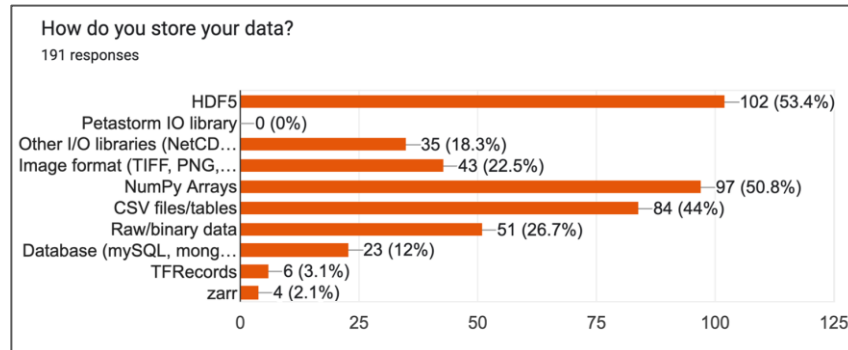
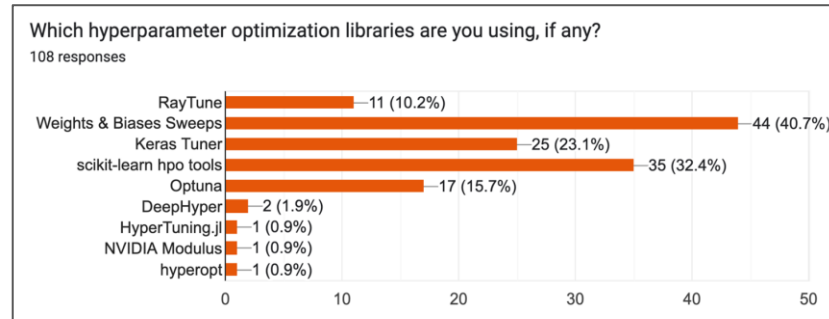
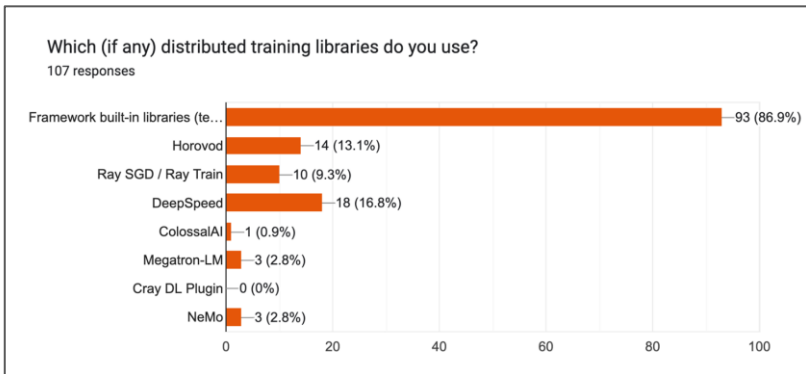
U.S. DEPARTMENT OF
ENERGY

Office of
Science

NERSC AI workload

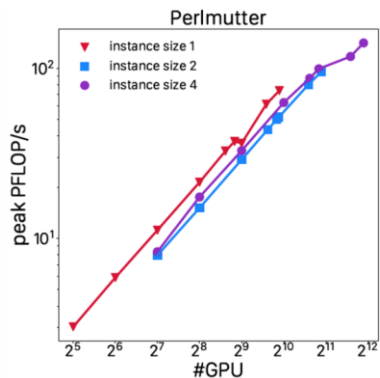
ML software ecosystem usage:

- Native distributed training strategies (e.g. DDP) dominate
- Variety of external tools used for end-to-end ML 'workflow'
- HDF5/NumPy remain popular



Data-driven weather & climate forecasting

- 2022 was the **breakout year**
- Recipe: Scaling up on...
 - Data (resolution, size of variable set)
 - Compute (model complexity)
 - (+ algorithmic improvements)



[Kurth et al. \(2023\)](#)

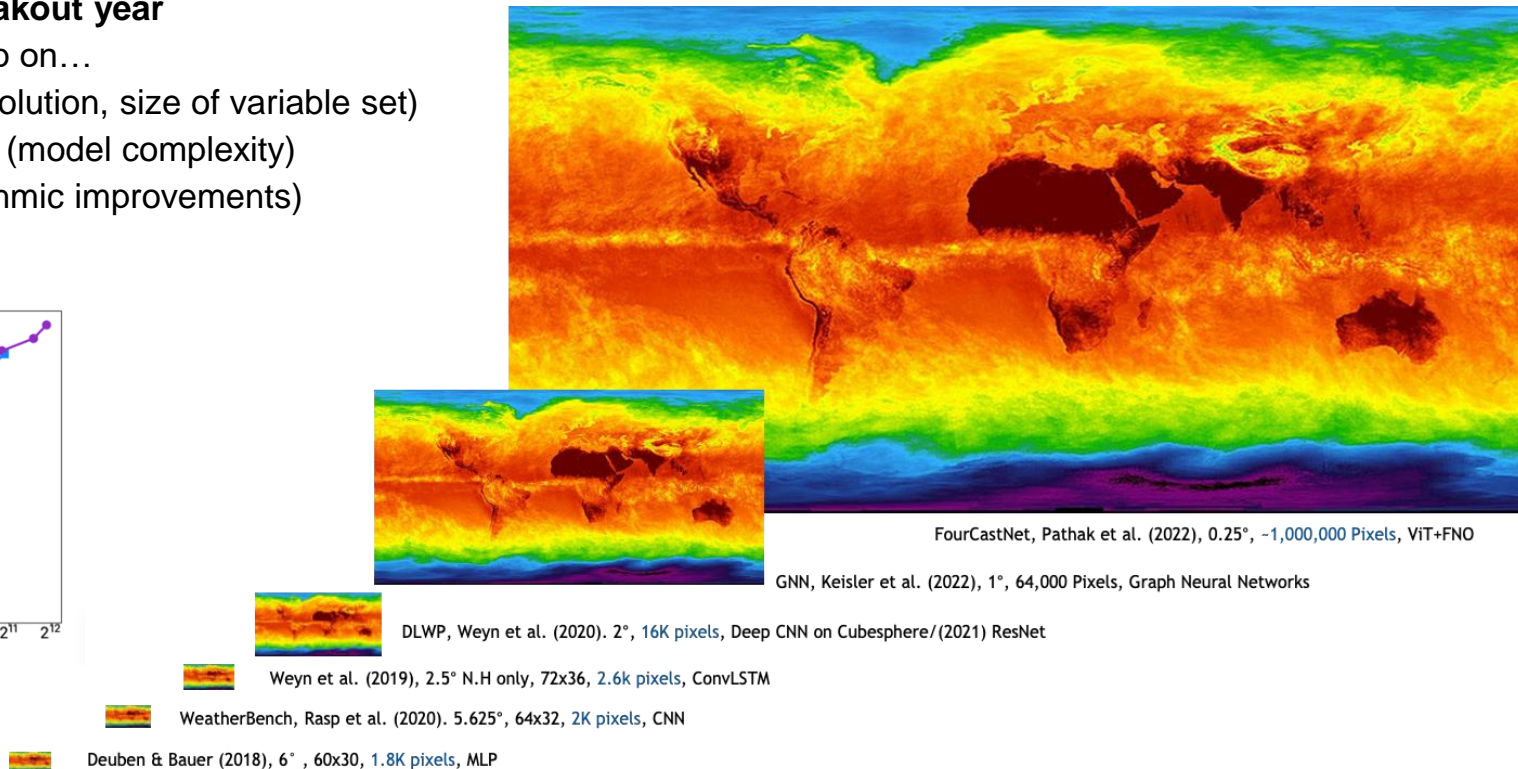


Fig. credit: Karthik Kashinath, NVIDIA



OmniLearn: a foundation model for HEP jet physics



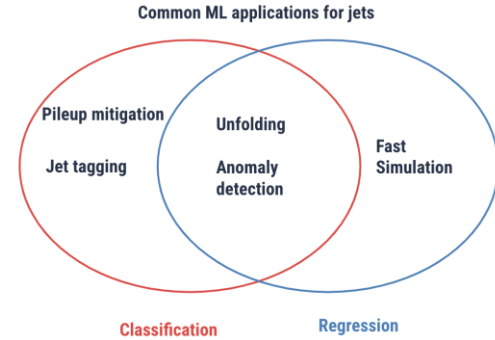
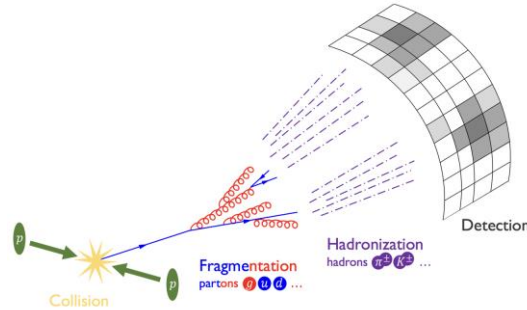
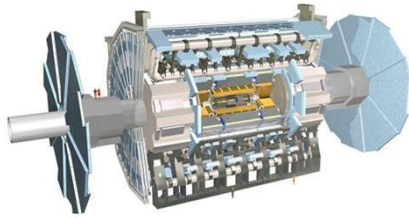
BERKELEY LAB



U.S. DEPARTMENT OF
ENERGY

Office of
Science

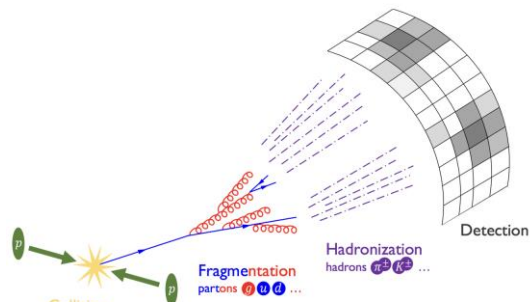
OmniLearn: a foundation model for HEP



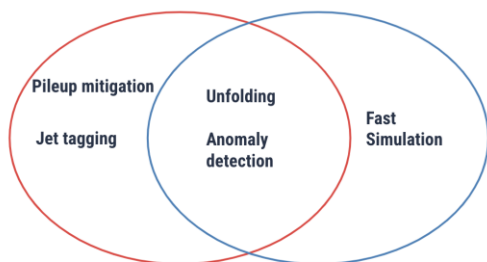
- Particle colliders (and simulations of them) provide a **wealth of rich data**
- High-energy physics (HEP) has been **eager in productionizing & scaling ML** for various scientific analyses

- Most previous work: bespoke models for bespoke tasks
- How to proceed in the **foundation model era?**

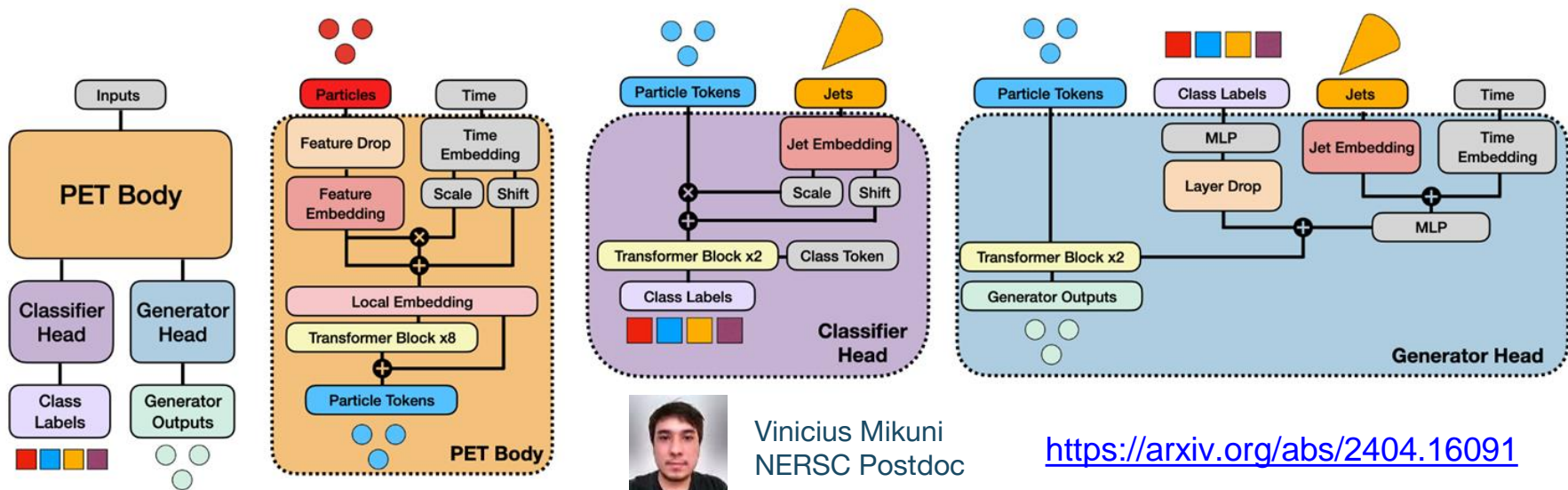
OmniLearn: a foundation model for HEP



Common ML applications for jets



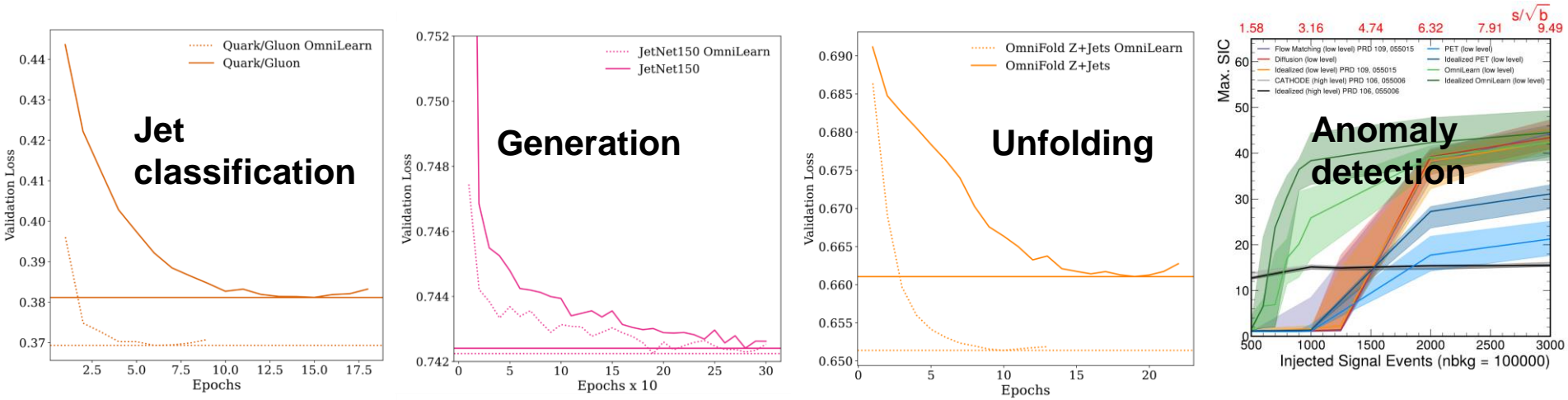
A classifier + generator architecture applicable to many downstream jet physics tasks!



Vinicius Mikuni
NERSC Postdoc

<https://arxiv.org/abs/2404.16091>

OmniLearn: a foundation model for HEP



- Improved classification across different physics experiments
- Faster convergence in generative models for simulation
- Improved detector unfolding (reweighting)
- Improved sensitivity in anomaly detection for unlabeled searches of new physics!