LA-UR-24-23544

Embracing AI/ML for Mission

Nathan DeBardeleben, Ph.D.
ASC AI/ML Lead for Enabling Manufacturing
Co-Executive Director Ultrascale Systems Research Center

HIGH PERFORMANCE COMPUTING

2024 Salishan Conference on High-Speed Computing, April 2024

AI is sucking all the air out of the room

- Feels like another hype bubble
- Fundamentally, ML is just function approximation
  - Deep resemblance to statistics – Bayesian Optimization, Gaussian Processes, etc.
  - Sometimes highly non-linear function approximation
- Novelty of modern ML:
  - Advanced algorithms often taking advantage of . . .
  - Increased computing power
  - Availability of big data
  - Improved software and tools (e.g. PyTorch, Tensorflow)

HIGH PERFORMANCE COMPUTING

HPC has been here before

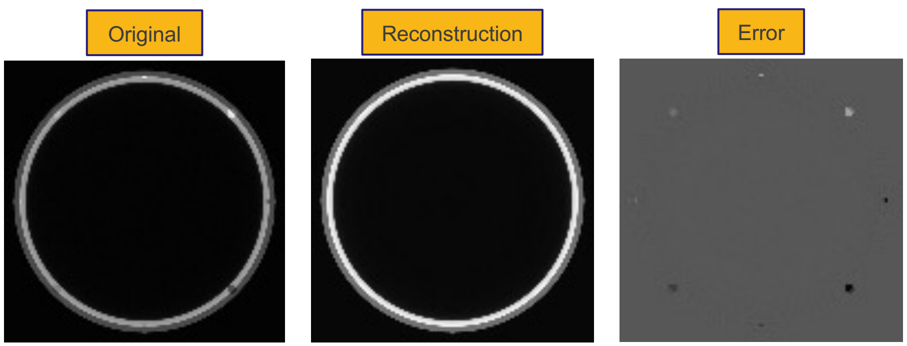
- We have experiences to share
- If you look carefully, we've seen this before
- Obviously the cloud community learned from HPC, and the AI community is as well
- Similarities in scaling challenges:
  - Ease of Entry vs. Scaling Complexity
    - Sequential code and scikit-learn
    - Real challenge comes when you scale these solutions
    - HPC: MPI / OpenMP / Accelerators (Phi, Cell, etc.) –
    - ML: tensor / pipeline / data / model parallelism
  - Optimization requires deep understanding of underlying hardware for performance
- Shared learning curve – Steep!
- Tools are not yet stabilized
  - Settling on PyTorch and Tensorflow . . . but optimization and tuning landscape very much still evolving

AI is / will be useful for mission challenges

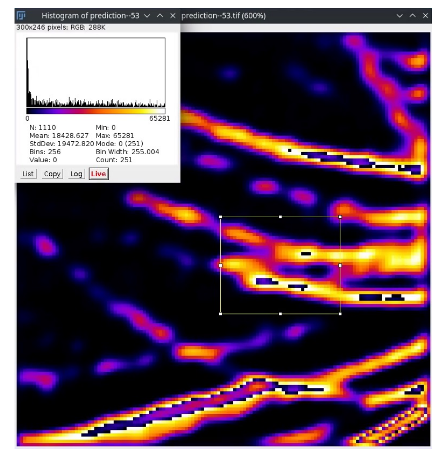
- Operations / production agency / manufacturing / smart factory
- Surrogate models and inverse problems for numerical simulations
- Experiment analysis
- HPC datacenter operations and user-facing tools
- Scientific inquiry / hypothesis generation

Manufacturing process analysis via in-situ and post-mortem studies

- Applying ML to certification and qualification:
  - Weld process analysis
  - CT, radiography, microscopy
  - Sensor telemetry analysis to detect changes / anomalies
  - Assisting inspectors with human-in-the-loop tools



Reconstruction of E6 test object
Garret Kenyon: CCS-3
Matt Sheats: E-6

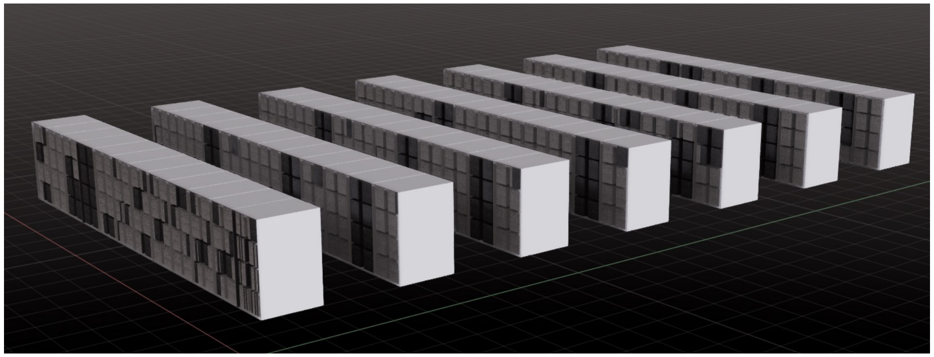


LPBF thermal camera image segmentation
Sean Tronsen: HPC-DES

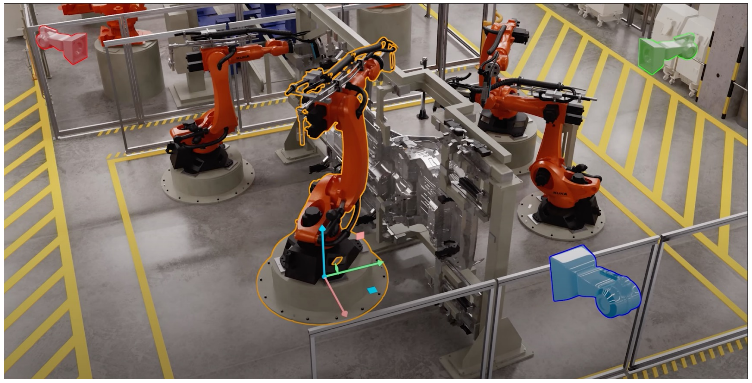
HIGH PERFORMANCE COMPUTING

Operations / PA / Smart Factory

- Digital Twins of manufacturing processes and products
- Manufacturing through to surveillance of stockpile parts – digital twins would empower NNSA mission
- LIDAR and sensor data from factories
- Telemetry empowers reinforcement learning for optimization

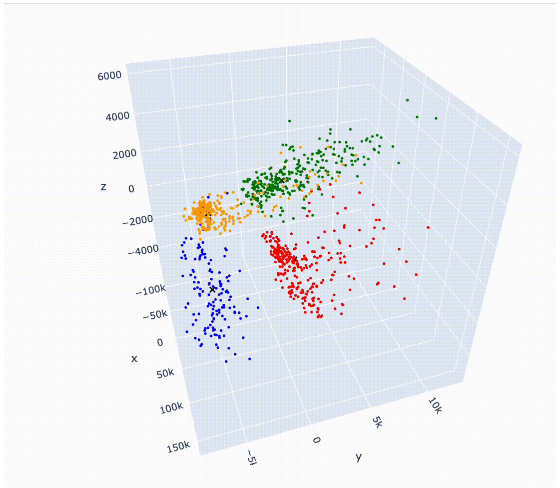


Wes Brewer @ ORNL
Digital Twin of Frontier Supercomputer in Omniverse
AIRES Workshop 2023
Thermo-Fluid Cooling and Power Utilization Models

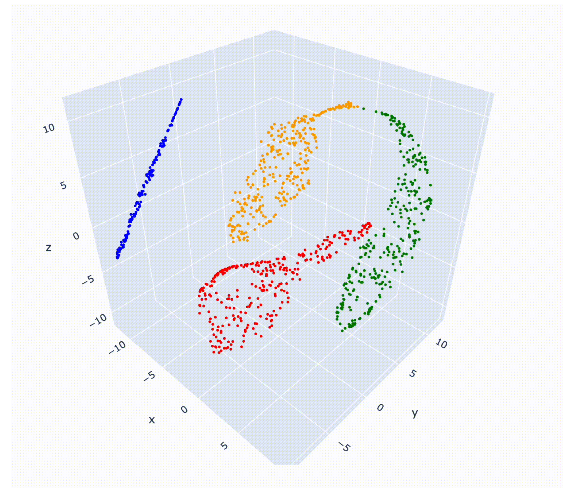


BMW Virtual Factory
NVIDIA Omniverse

Micro-Destructive analysis using laser induced breakdown spectroscopy to study alloy composition



Dimensionality reduction separates the samples somewhat, but there is still some mixing



Samples are clearly separated after Whitaker baseline removal followed by t-SNE using cosine distance

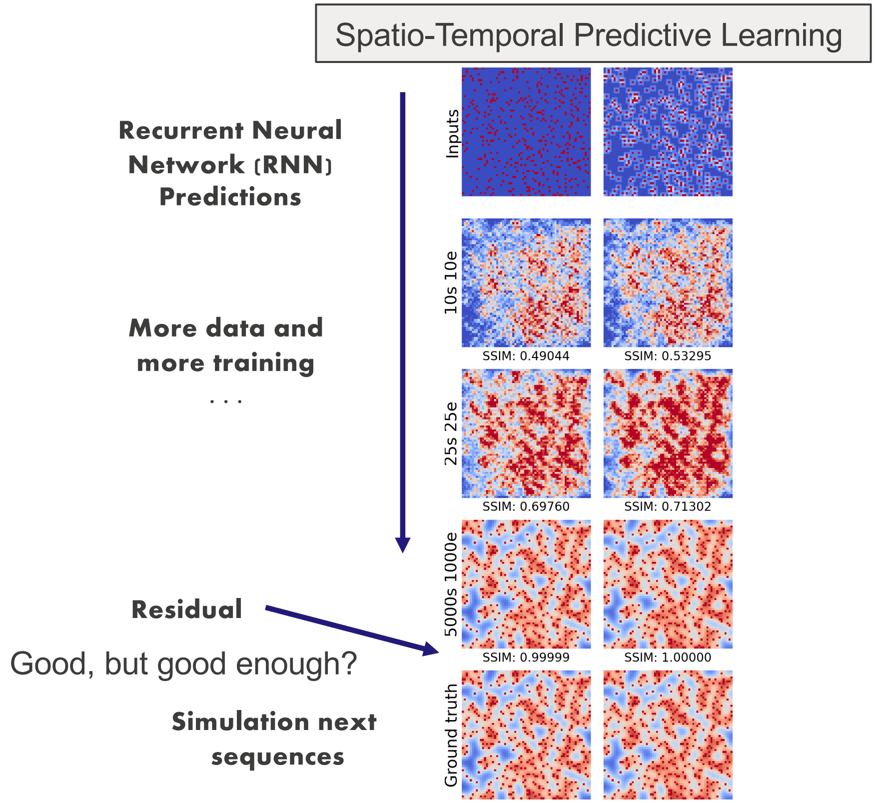
Unsupervised learning for anomaly detection on 700-dimensional data

- Technology used on the Mars Rover to determine presence of elements in rocks
- Used for food analysis in manufacturing

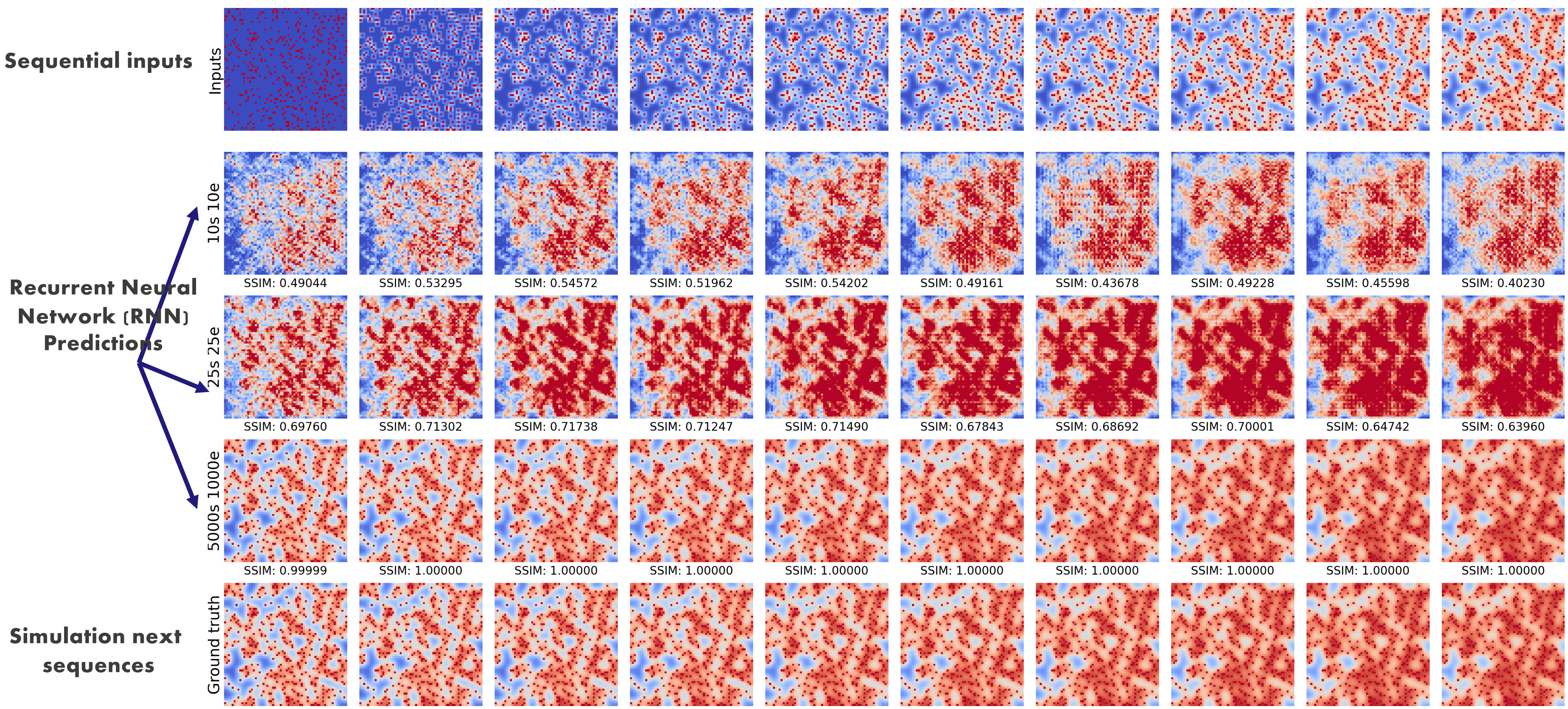
HIGH PERFORMANCE COMPUTING

Data generated from simulation codes used to train ML surrogate models

- Byproduct of running simulations generate training data for models
- (given enough data) Models used to replace, augment, or cross-check simulations
- Surrogates enable inverse problems
- Exploration of design space offline to determine parameters to study



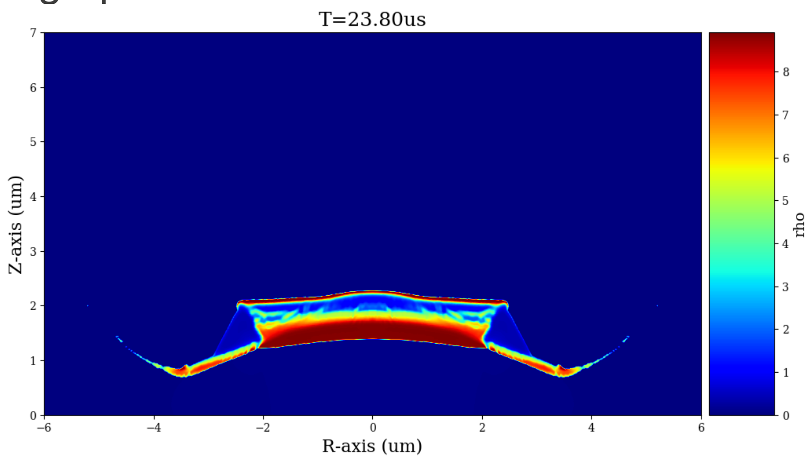
Spatio-Temporal Predictive Learning

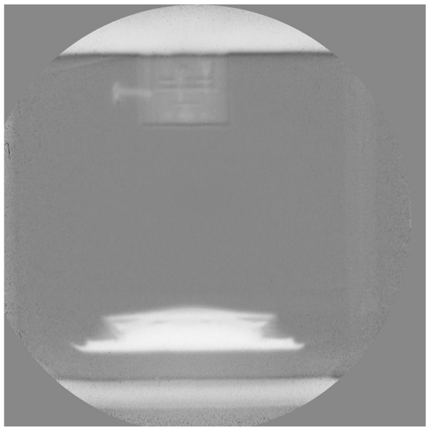


HIGH PERFORMANCE COMPUTING

Neural learning for radiographic analysis

- Allows extraction of detailed, quantitative information from radiographs
- Extract material properties such as damage and strength directly from radiographs
- Use high-fidelity, multi-physics ASC code to generate synthetic radiographs over a range of material properties
- Neural networks trained in synthetic data and tested on experimentally observed radiographs



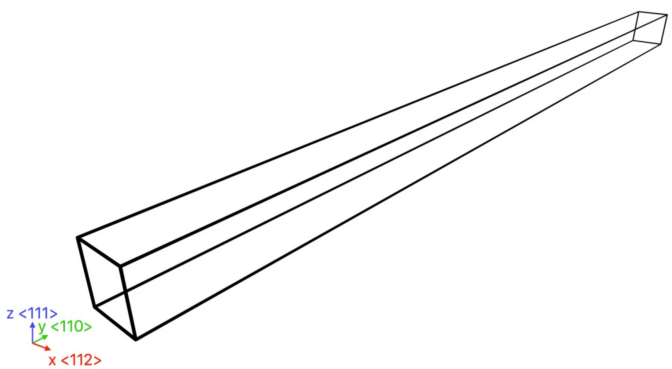


Kyle Hickman, XCP-8
Extracted from: LA-UR-23-22076
TRAINING AND INTERPRETABILITY OF DEEP-NEURAL METHODS FOR DAMAGE CALIBRATION IN COPPER

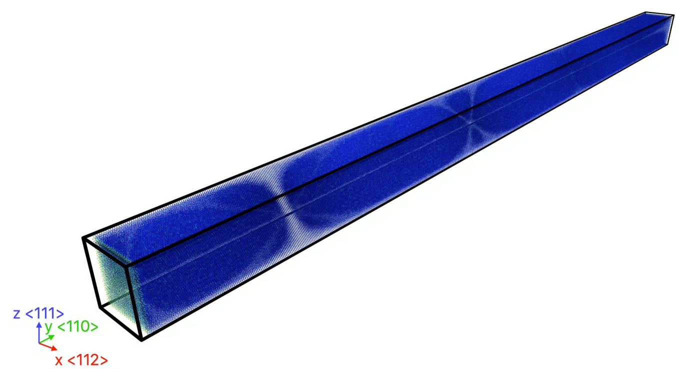
HIGH PERFORMANCE COMPUTING

Machine Learning enables studies that are computationally infeasible other ways

- Big molecular dynamics simulations
- MPI with novel on-node neural network
- ML is ~12 orders of magnitude faster than ground truth simulations



Ben Negben, T-1

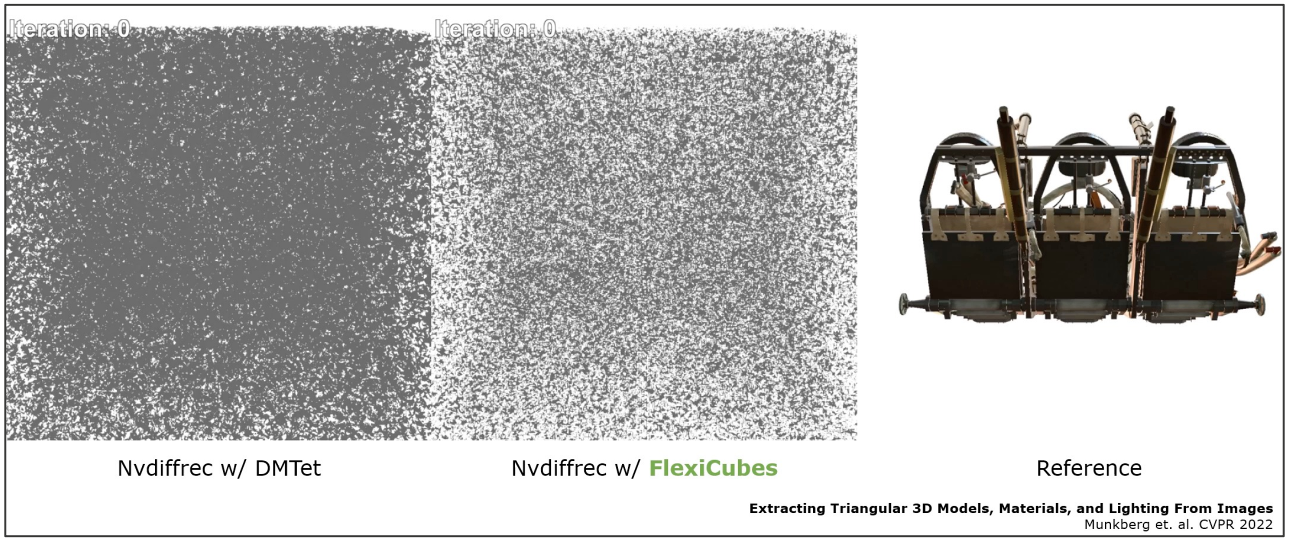


Extracted from: LA-UR-23-22400
Training Machine learned Interatomic Potentials to EXAFS Data for Simulations Under Extreme Conditions

HIGH PERFORMANCE COMPUTING

AI will make improve simulation setup

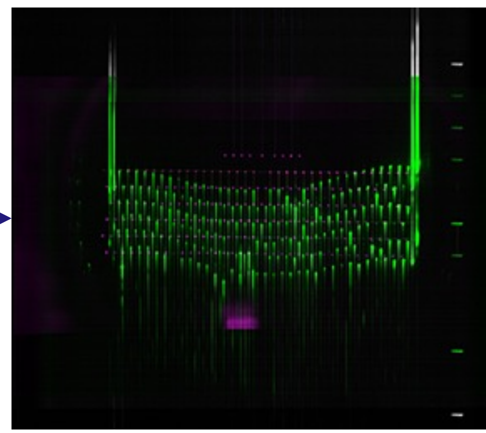
- Meshing setup constitutes an out-sized amount of required effort in a designer's workflow
- LLMs will assist with parameter studies / setup – improving Bayes Opt currently

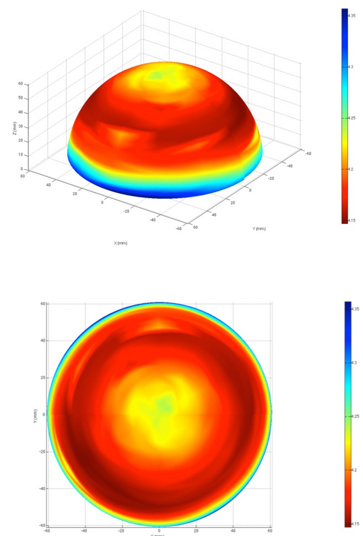


FlexiCubes
NVIDIA

Automating data analysis from experiments using AI

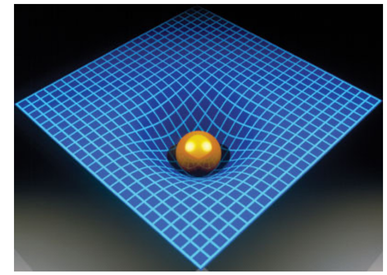
- Explosive (e.g. PBX 9701, 9501, etc.) experiments: fiber optics to a series of streak cameras used to calculate break out curves
- Problem: analysis is labor intensive, measuring pixel “streaks” and entering into a spreadsheet

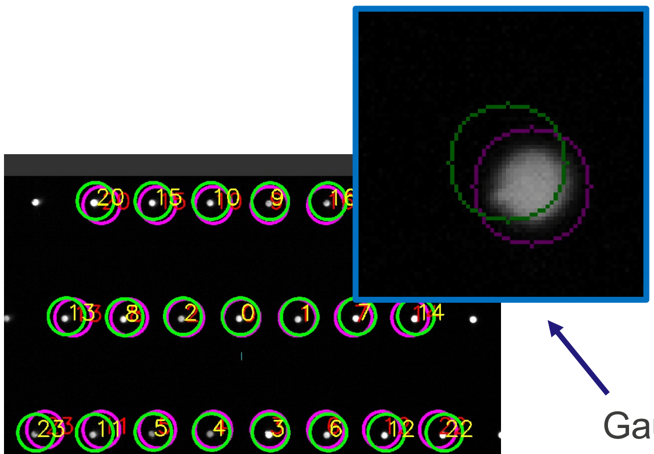




Automating data analysis from experiments using AI

- Approach: combining computer vision techniques with machine learning trained algorithms → user assistance
- Automatic detection of points and measurement of streaks → data file
- Images are highly distorted and vary from shot to shot (different setup, small experimental differences)



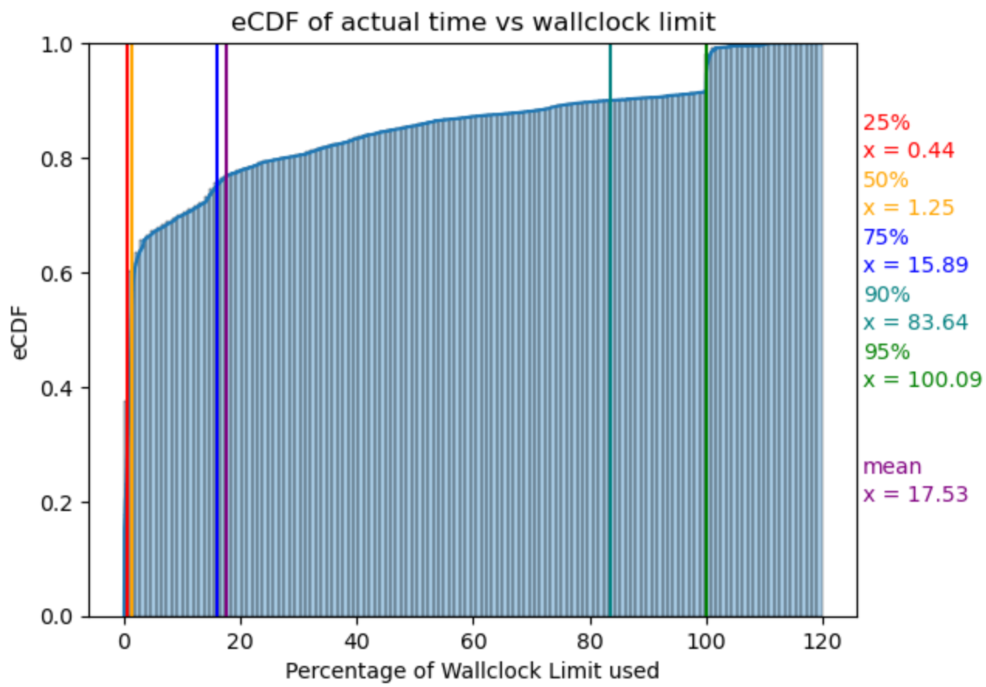


Gaussian Mixture Models trained from datasets assist an AI algorithm in identifying where the points of light are and calculating start points

HIGH PERFORMANCE COMPUTING

HPC Operations

- Changing SysOps is morphing to include MLOps
- AI scheduler prediction:
  - Users are notoriously bad at predicting how long their jobs will run
  - They also have essentially no incentive to predict right
  - Impacts scheduler ability to make informed predictions on queue times
  - ~80+% accuracy in predicting <u>actual</u> job run time (and therefore start time) via ML models



HIGH PERFORMANCE COMPUTING

LLMs will improve business practices and enhance scientific discovery

- Surprising amount of areas in operations that LLMs can assist:
  - Contract writing/review
  - Travel reimbursement
  - Policies
- Hypothesis generation – already happening now:
  - *Harnessing the Power of Adversarial Prompting and Large Language Models for Robust Hypothesis Generation in Astronomy*
  - *Hypothesis Generation with Large Language Models*
  - etc . . .

Why does this matter to HPC folks?

- Datacenters are changing – and many have already changed
- Opportunity to engage with teams deploying inference hardware "at the edge"
- Data captured at the edge needs to be processed in datacenters – it's unclear to me how that is supposed to work
  - Generally, we want the flow that data is captured at experimental device
  - Transferred to the datacenter
  - Used to train a ML model
  - The model is shipped back out to the edge for inference
  - And used on smaller AI/ML hardware at the edge for inference
- Always-on services is not "normal" for HPC batch-like organization

But we have a massive data challenge

- Data is spread around the complex
  - Analysis becomes very complicated
  - Data has to be moved to appropriate compute resources
  - Usually requires deep domain knowledge (e.g. little/no metadata)
  - Most of the data stores do not mount on supercomputer resources (e.g. compatibility problems)
  - Users want global accessibility – laptop to supercomputer

Takeaways

- AI and ML are here – HPC needs to figure out how to adapt
- Don’t be scared, ML is just an enhancement on something we’re already comfortable with – statistics
- AI and ML tools will impact all areas of DOE mission:
  - Manufacturing
  - HPC operations
  - Simulation codes
  - Experimental analysis
- But data is a huge problem and we must address this

HIGH PERFORMANCE COMPUTING

HIGH PERFORMANCE COMPUTING

Over 70 years at the forefront of supercomputing