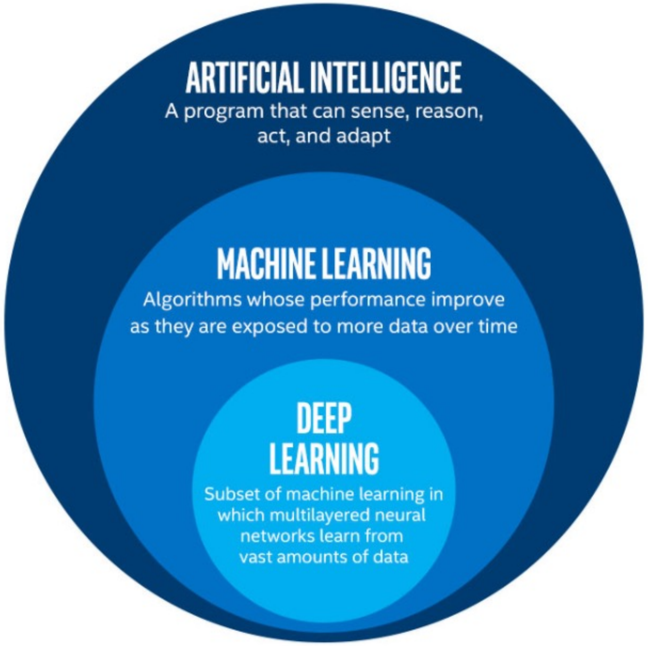
CST/ENP: Envisioning Meeting



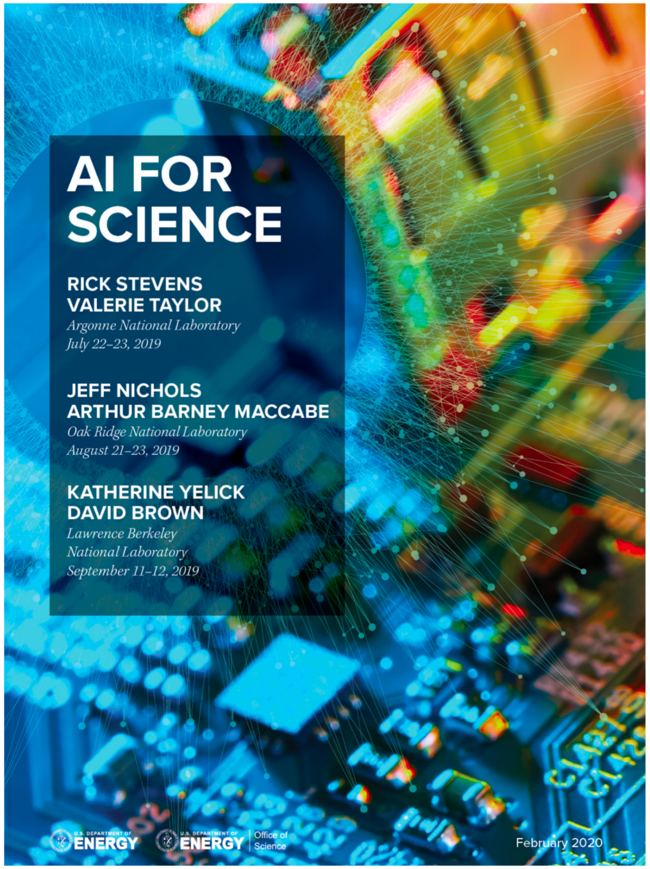
Malachi Schram, Ph.D.
Department of Data Science
Thomas Jefferson National Accelerator Facility

AI for Science Report

"New Deep Learning methods are required to ***detect anomalies*** and ***optimize operating parameters***…"

"... move from ***human-in-the-loop to AI-driven*** design, discovery, and evaluation also manifests across the ***design of scientific workflows***, *optimization of large-scale simulation codes*, and ***operation of next generation instruments***."

- Excerpts from the Executive Summary



Recent Activities

AI Town Hall:

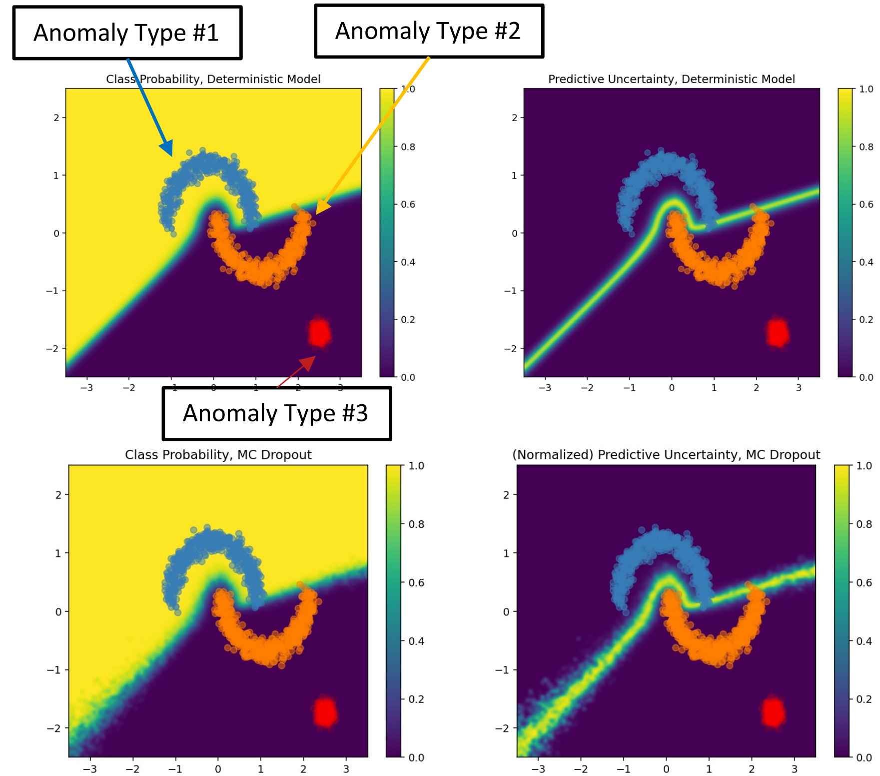
- The recent AI/ML work in the experimental Halls, *specifically Halls B and D*, have shown promising results in areas such as anomaly detection, reconstruction, data driven generators, etc.
  - Augmenting these techniques with *uncertainty quantification* would provide a robust and actionable techniques with a quantifiable levels of confidence
  - Including know *physics constraints and PDEs* would also provide additional robustness in the model predictions
- Unclear how the models are shared across experimental Halls
- Side meeting with Marco on Hall B provide invaluable information for AI/ML report
- Side meeting with Thia and Ole for Hall A resulted in some interesting opportunities for ML
- Work across the experimental Halls, when possible, to maximize scientific productivity

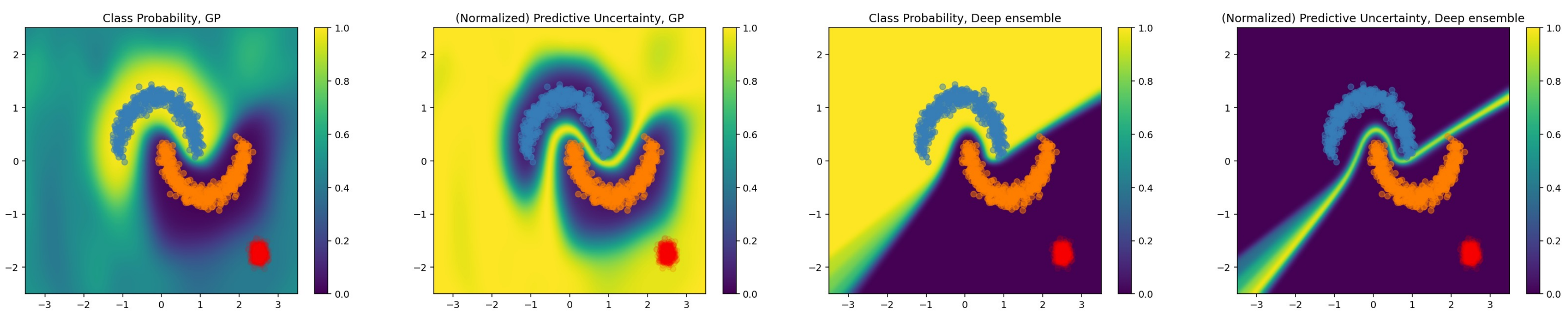
Understand what your model knows and doesn't know

Different method yield vastly different classification predictions, some examples:

- Deterministic
- MC Dropout
- Deep Ensemble
- Gaussian Processes
- Bayesian Neural Networks

- Different models architectures can yield better results if you do not know all classifications





Jefferson Lab

Talking points

Topics	Easy	Medium	Complicated
Datasets repository	X		
Machine learning model repository	X		
Machine learning provenance		X	
Data science workflow tools		X	
Digital Twin			X
Scaling AI/ML workflow			X
Additional considerations			X

An example of a data science pipeline

- What questions are we trying to answer with the data?
- Do we have the right data?
- What do we know about the data?
- Can we learn something from the data before using machine learning (ML) techniques?

Dataset

Data Source

- Real or synthetic
- Quality
- Dimensionality
- Format
- Density
- Size

Data Preparation

- Data cleaning
- Data restructuring
- Correlations
- Dynamics
- Visualization

Models

ML Applications

- Classification
- Regression
- Clustering
- Feature extraction

Workflow/Tools

Training Tools

- Cross-validation
- HPO

Results

- Predictions
- Confidence Level
- Explainability

Across the Halls: Dataset

kaggle

- JLab has a *precious* datasets that can be used for algorithm development
- Similarly, the ongoing efforts in AI/ML at JLab can be leverage to accelerate the science in *other* Halls
- Other national laboratories are developing frameworks to capture elements of ML such as the dataset
- We should develop a *private* collection JLab specific datasets that will allow us to easily collaborate and quickly evaluate algorithms

Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks

Pajau Vangay, Benjamin M Hillmann, Dan Knights

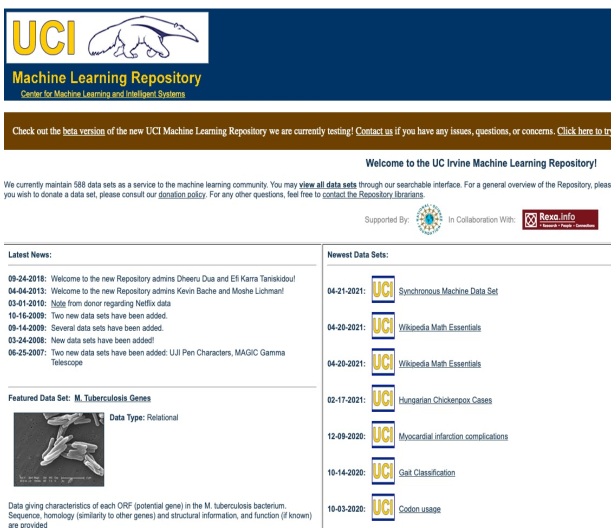
GigaScience, Volume 8, Issue 5, May 2019, giz042, https://doi.org/10.1093/gigascience/giz042

Published: 26 April 2019 **Article history** ▾

PDF | Split View | Annotate | Cite | Permissions | Share ▾

Abstract

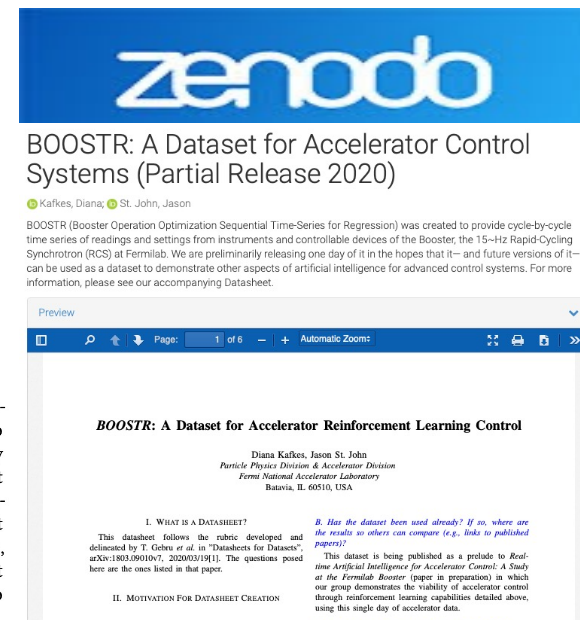
The use of machine learning in high-dimensional biological applications, such as the human microbiome, has grown exponentially in recent years, but algorithm developers often lack the domain expertise required for interpretation and curation of the heterogeneous microbiome datasets. We present Microbiome Learning Repo (ML Repo, available at https://knights-lab.github.io/MLRepo/), a public, web-based repository of 33 curated classification and regression tasks from 15 published human microbiome datasets. We highlight the use of ML Repo in several use cases to demonstrate its wide application, and we expect it to be an important resource for algorithm developers.



Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets.* In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.



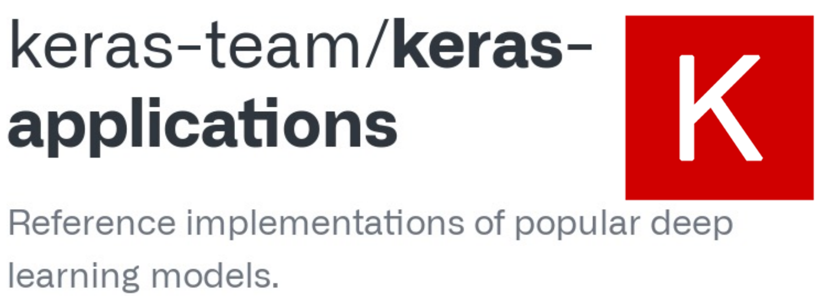
Jefferson Lab

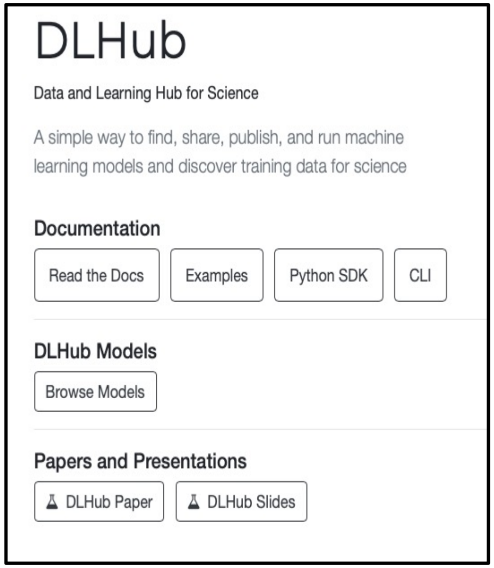
Example questionnaire

- How was the data collected and labeled?
  - Real world data is messy!
  - It will have missing/noisy data that you will need to account for.
- How was it curated?
  - Data curation is the organization and integration of data collected from various sources.
  - Do the various sources of data need to be temporally aligned?
- What are the data formats for your study?
  - Images (tracks and noise), temporal (time series data), categorical (ex: labels A-Z), ordinal (ex: ranking between 1-5)
- What is the dimensionality of the data sources?
  - High dimensional (ex: images)
  - Low dimensional (ex: single variable sensor)
- How many samples do you have?
  - Large number of samples (>10k): Google images or large time series data
  - Limited: A few experimental measurements and/or simulation samples
- Does the data capture the dynamics (physics) of interest or are they distinct samples?
- What are the input and output features of interest?

Across the Halls: ML repositories

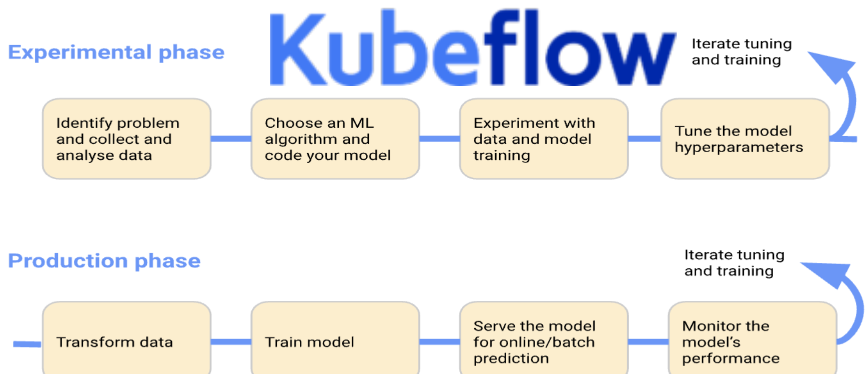
- JLab scientist are developing a lot of ML models from scratch
- There is *no* central system to find exiting models and to share a newly developed models
- For example, Hydra leverages an pre-designed model from keras-applications for anomaly detection
  - Additional changes/extensions to the model is possible to improve performance
- Other national laboratories are developing frameworks to capture elements of ML models and associated meta-data
- We *need* a common repository to capture the ML provenance for all models used for operations
- We can use use the datasets repository to validate the ML model and provide a simple interface for visualization





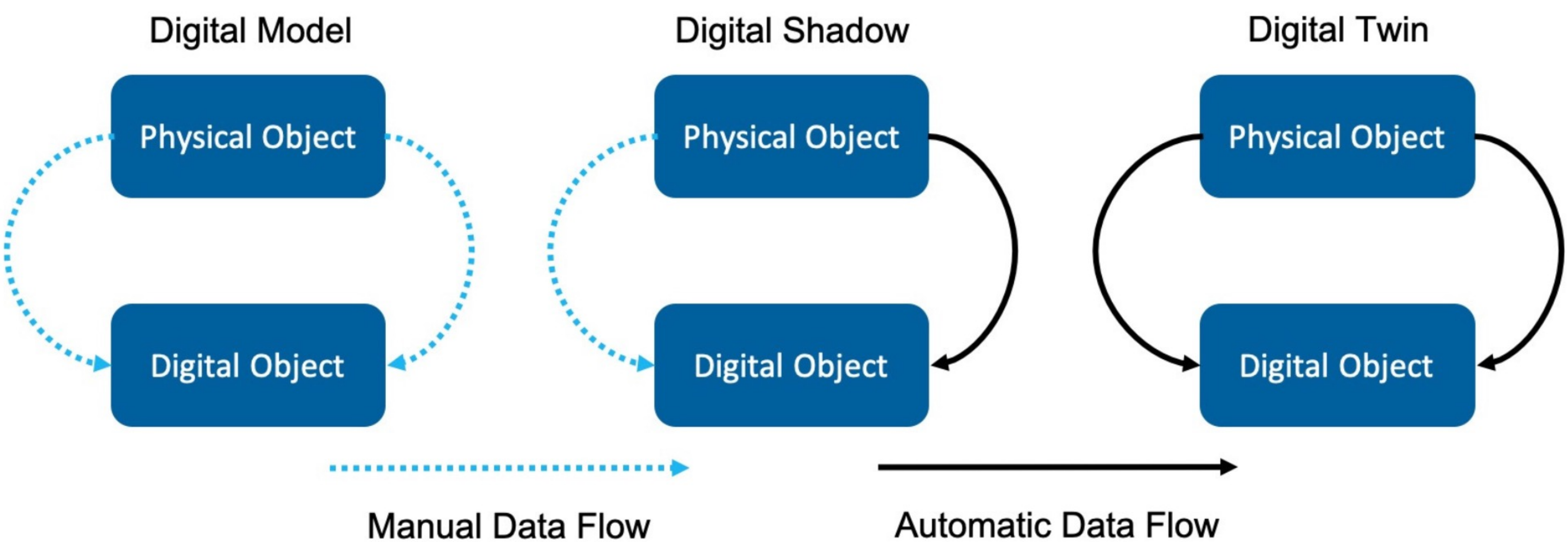
Data science workflows and services

- Develop a dev/ops workflow to allow exploration of SOTA ML packages
- Provide a workflow to perform hyper-parameter scans, model stability tests, etc.
- Leverage the GPU farm with near real-time model building (eye to HPDF)
  - We don't want to do this manually!!!
- Extend workflows to include edge computing and new computing architectures



Moving towards automation

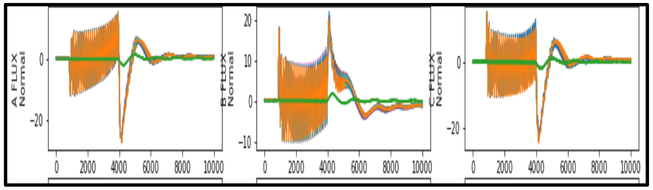
- **Digital Model:** a digital version of a pre-existing or planned physical object
- **Digital Shadow:** digital representation of a physical object with a one-way data flow from the physical to digital object
- **Digital Twin:** data flows between a physical object and a digital object are fully integrated and bilateral

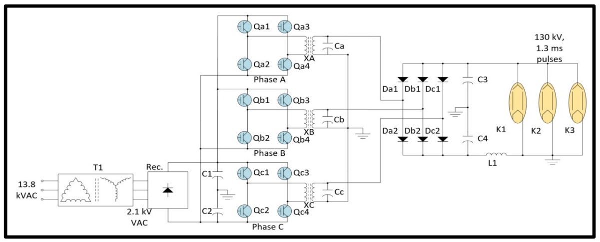


Motivation: Digital twin application examples

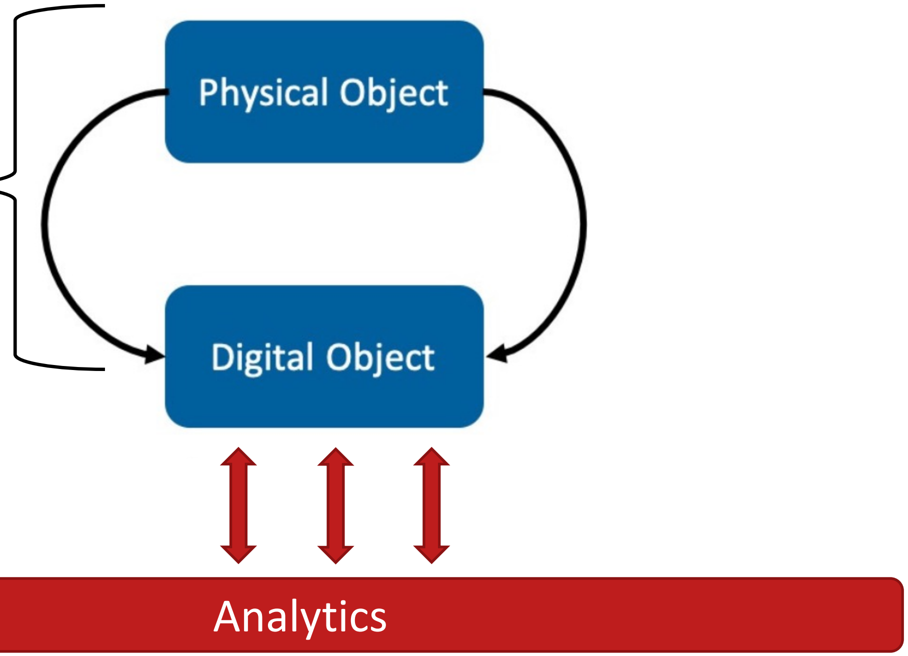
- Digital Twin provides the ability to conduct analytical studies without impacting the physical system, for example:
  - Statistical analysis:
    - Box plots (mean, median, quantiles, etc.)
    - Threads
  - Time series forecasting:
    - Gaussian Processes
    - Quantile Models
    - Recurrent Neural Networks
  - Anomaly detection and classification
    - Random Forest
    - Deep Neural Network
    - Siamese Networks
  - Forecasting component fatigue and failures
    - Physics based models
- Depending on time budget for actionable responses, these studies can be performed on the edge (FPGAs) or on HPC systems

Integrating digital twin into the analytics workflow

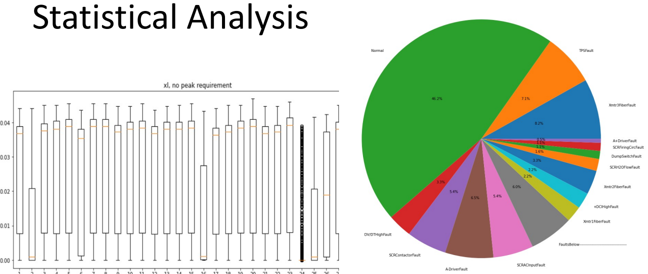




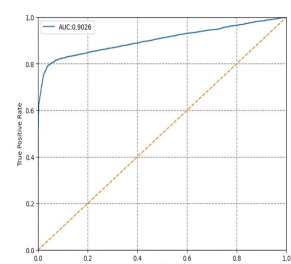
Digital Twin



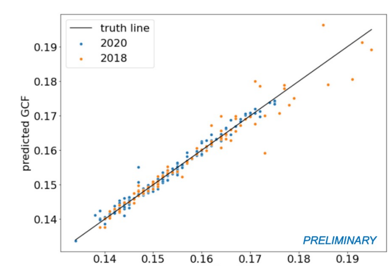
Statistical Analysis



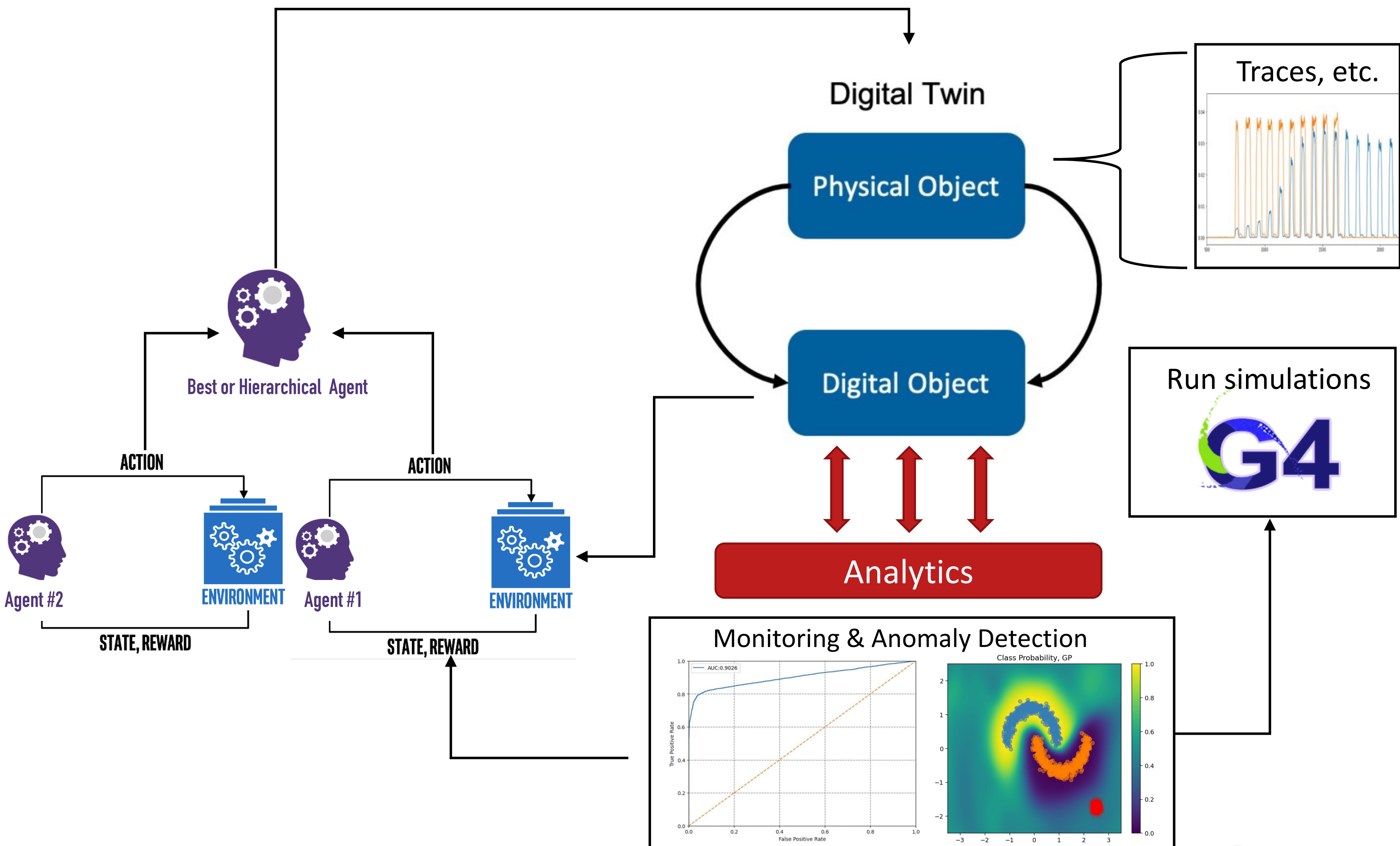
Monitoring & Anomaly Detection



Calibration

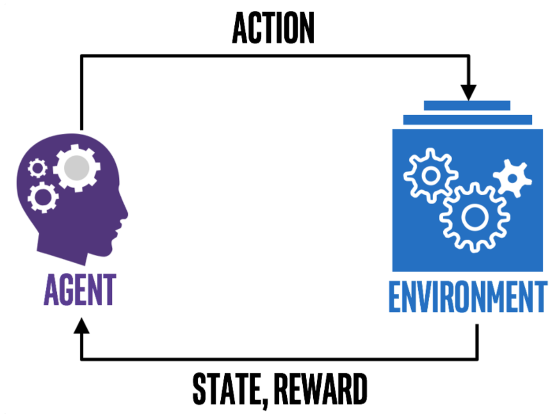


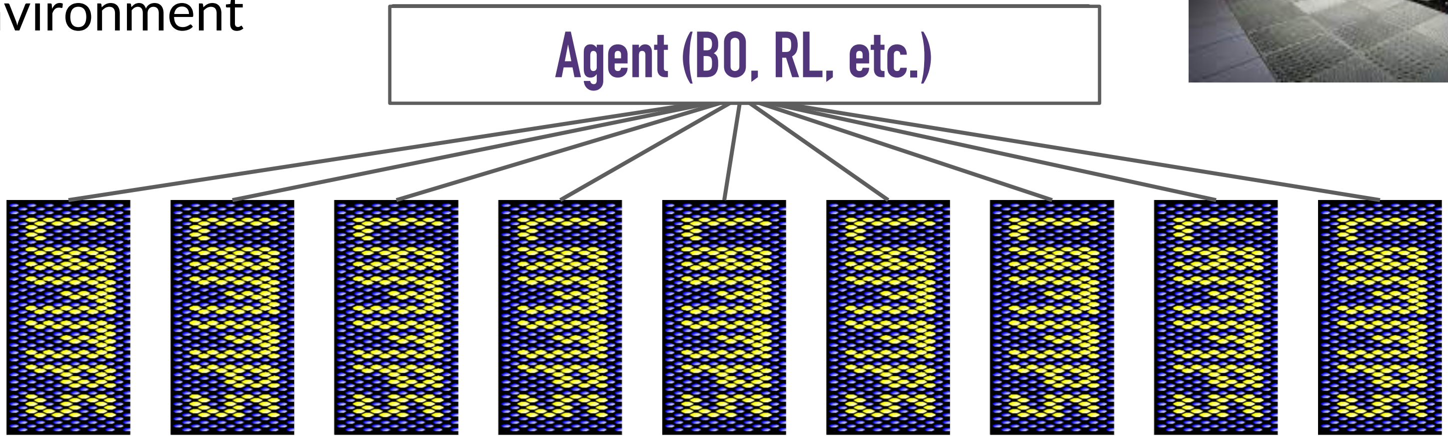
Extending digital twin and analytics for control workflow



Scaling workflow on HPC system

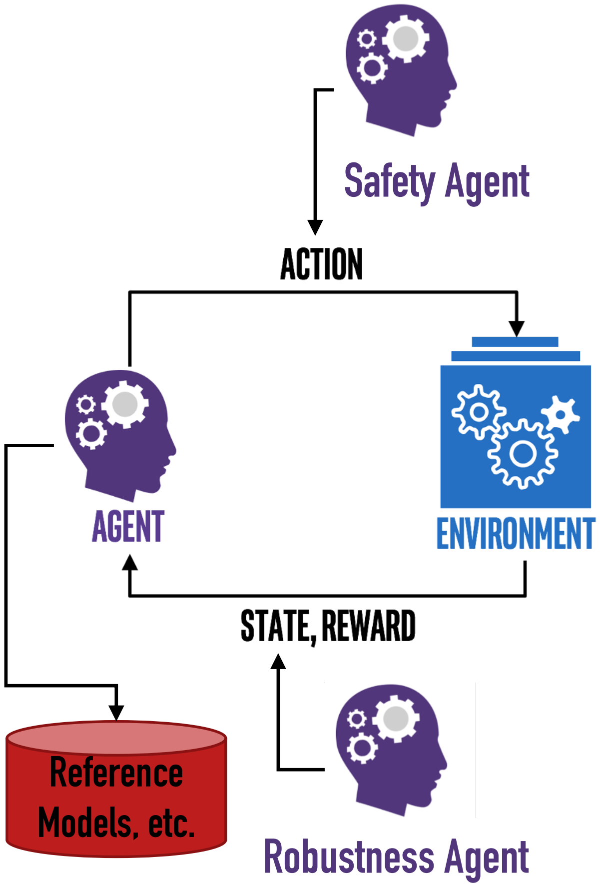
- Digital Object is interfaced with industry standard OpenAI gym environment
- To accelerate the data generation we developed a MPI-based framework
- We created an agent that maps the action-reward for all simulations
- A production job split the MPI communications between the agent and each Digital Object environment





Additional considerations

- **Safety:** We need to ensure that the actions provided by the AI-based controller are within “safe” parameters.
- **Robustness:** Understanding how the AI-based control behaves in the presence of unexpected changes in the input state
  - Models robustness - loss landscape, etc.
  - Impact from noisy and/or dead sensors
  - Adversarial techniques
- **Explainability:** With all AI/ML models we need to understand why the model made a given prediction:
  - Saliency maps
  - Hierarchical models
- **Continuous learning:** The underlying system dynamics can change over time. We need to evaluate the current states to previous states to determine if there has been any notable change that would require the model to be updated
- **Computing infrastructure:** Data-intensive workflow, data movement (DTNs, Wired/Wireless), processing architecture (FPGA, GPU, etc.)



Grand Challenges

- **Incorporate domain knowledge and Uncertainty quantification (UQ)**
- **Transform the operation of accelerators, detector, compute systems:**
  - Fully realized digital twin to provide continuous accelerator and detector monitoring, fault detection, and optimization
- **Scaling Current ML tools:**
  - Scalable high dimension optimization problem (ex: EIC work)
  - Parallelize GAN workflows
- **Data Analysis**
  - Edge analysis and improved reconstruction (tracking, PID, etc.)
  - Integrated exp/theory ML models to solve challenging inverse problems
- **Data management:**
  - Develop ML-based data discovery (similarity score and clustering)
- **Incorporate new techniques, such as Graph NN, to ML workflow**