Accelerating Physics With AI at MIT

EM Breakout session - white paper drafting

Attendance:

Jeroen Audenaert, Michael Coughlin, Matthew Graham, Konstantin Malanchev, Daniel Muthukrishna, Gautham Narayan, Josh Peterson, Benedikt Riedel, Kate Scholberg, Ashley Villar

White paper - notes/questions:

- Key missions in EM that benefit from ML acceleration? Timeline of missions
    - *2025 LSST*
    - *ZTF, YSE, TESS - ongoing*
    - *Roman Telescope - 2027*
    - *SKA, radio facilities*

- *Transient vs Variable Astronomy*
    - *Time criticality for fast phenomena (e.g. shock breakout, and fast events) - quick followup - currently involves human in the loop - need for automating the distribution of resources*

- What latency is required for LSST/other EM? - How does this fit on the A3D3 graphic from Phil's talk (streaming data rate (B/s) vs latency requirement (s)?
    - Kilonovae/GRBs < 1 day
    - *60second Latency from LSST*
    - *30 - 40 million alerts per night 500Hz inference rate*
    - Single alert is 80kB, but final object going to ML algo is larger and may include: light curve, catalog cross-matches, etc
    - Latency requirement depends on a target: fast phenomena requires ~minutes latency, slower phenomena could exchange latency for passthrough and use in-bulk alert processing.
    - Space telescopes rely on the DSN. Limited bandwidth. Need to store data onboard with donwlinks ~once per week
        - Potential for onboard ML
            - Currently cosmic-ray rejection algos exist onboard spacecraft. More advanced methods in future?

- *Existing Resources*
    - *Alert Brokers*
        - ANTARES (NOIRLab, running in GCP),
        - GooglePitt (Google resources),
        - AlerCE (Chile),
        - Fink (CNRS),
        - LASAIR (Edinburgh compute (not much ML yet)),
        - AMPEL

    - Broker capabilities
        - Take in alerts from multiple surveys, store in database, build data lake, provide access to community in digestible way.
        - Brokers have no requirement for compute back end - only listening and filtering capacity.
    - Hermes - stores photometry and makes data public
        - Only aggregates data does not allow compute time
    - LINCC funded to develop infrastructure
        - Team to help community develop Rubin relevant infrastructure

- Brokers/HPC
    - Where are ML models being run? Brokers or HPC?
        - **We need a standardized platform for compute resources**
            - **Standard interface to the platform (that's running e.g. Kubernetes)**
    - HPC is for data releases, while broker is for online/live access and archival availability
    - **Brokers can be used to trigger algorithms on other facilities**
        - Brokers have limited resources, but can offer some small model running requests from the community
        - What are the heavy models we have to run, is this a problem for brokers?
            - Snoopy, KN models take minutes to run per object, Reinforcement learning models depend on slow kilonova grids etc.. REFITT
            - ZTF currently runs on every single data point coming through alert stream (on Minnesota cluster)
                - Not possible for LSST.
    - Standardized API to make requests outside of firewall. Need a simple way (e.g. API tokens) to trigger cluster jobs.
        - And need prioritization at cost
    - Not clear which broker is responsible for different applications

- *How can HPC resources help?*
    - *Run our marshalls on NERSC with easier access - current firewalls limit access*
        - *Better way to authenticate our marshalls*
    - *TOM toolkit being used for LSST*
    - *NERSC Kubernetes might be more useful for us to take advantage of?*
    - *We require recomputing of models as new data come in in an automated way*
    - *Infrastructure for easy access and uploading of models - currently barrier to entry for eg. grad students*
    - *Cron tasks: model retraining, re-running new or existing models on "old" alerts*
    - *Bayesian model, ML models to inform followup*

- What are the roadblocks to the integration of Machine Learning in scientific computing and the large-scale adoption of these systems?
    - Where do we run our models?

    - Educational gap
        - How to use these in scientific computing is not well-understood in the community
        - No common tooling for LSST / transient data, could LINCC help?
        - Data science knowledge is limited
        - Grad student resources to learn condor/slurm, HPC resource running
        - We fall short compared with GW/HEP
        - The simpler we can make the standardized interface, the smaller this gap will be.
    - TVS could be responsible for this

- What synergies exist between EM, GW, and HEP?
    - What can we learn from other disciplines?
        - Stadardized model to train and run models - already happening on GW side and is allowing models to be run at scale
        - Simulated/fake alert stream to practice run models
            - Like ELASTICC++
            - Stream some data by request
        - GW already take care of distributed computing. Lower barrier to entry to run community models
    - *Benchmarks for ML models*
        - *PLAsTiCC, ELAsTiCC*
            - *All simulated datasets. Not clear how we're going to perform on real datasets*
        - *ZTF as precursor to LSST*
    - Hardware?
        - Could space telescopes benefit from onboard machine learning with FPGAs?
        - Telescopes currently have some cosmic-ray rejection
        - Looking for computing infrastructure for ML science deployment
        - *Would be really nice to have a community computing resource*

- How can we be prepared for LIGO O4?
    - Currently multiple groups running the same KN models
    - Could overlap with LSST commissioning

- Neutrinos - DUNE/Icecube
    - Neutrinos could point to a supernova and inform EM follow-up
    - Need real-time neutrino supernova classification and localization to trigger EM follow-up
        - Need better coordination

White paper questions

- Summary of key points of the workshop
- Outline
    - Discussion of computing tools and software:
        - Path to aligning these across domains
    - List of critical models in the field
    - One plot to rule them all and bind these sections
- What can we offer other disciplines?
- Computing demands
    - ‣ We can assemble a list of common hardware(+tools)
        - GPU request
        - Real-time models on space telescopes?
- Software stack,
    - ‣ With all ML algorithms aim for a set of core software tools
    - ‣ Need for good tools to validate and deploy algorithms
        - *Would be really nice to have a community computing resource*
- ML problems
    - ‣ Across the domains similar ML problems exist
    - ‣ Highlighting the similarity is critical
- Inference-as-a-service for EM?
    -
- We are training a lot of models, but how many are actually being actively used on real-time datasets? (same point from Alec Gunny talk on GW)
    -
- What ML systems exist?
- *Simulation-based inference, Variational Inference*
- Can we learn new physics from ML?
- Multimessenger Astrophysics requirements
- Classification, anomaly detection, data generation/augmentation
- What are the benchmarks of success for anomaly detection etc.?
- What role do brokers play?
- How necessary is ML for upcoming surveys?
    - What are the pros/cons?
- Real data vs simulations
-