



AGC Workshop 2023

User Experience for Machine Learning

May 3, 2023
Elliott Kauffman - Princeton University





Goal: Add machine learning component to AGC CMS Open Data ttbar analysis

- Provide options for physicists who want to incorporate ML into their analyses on analysis facilities
- Factorized training and inference

▷ AGC Demo Day #1

First steps using inference server at coffea-casa facility

<https://indico.cern.ch/event/1218004/#1-first-steps-using-inference>

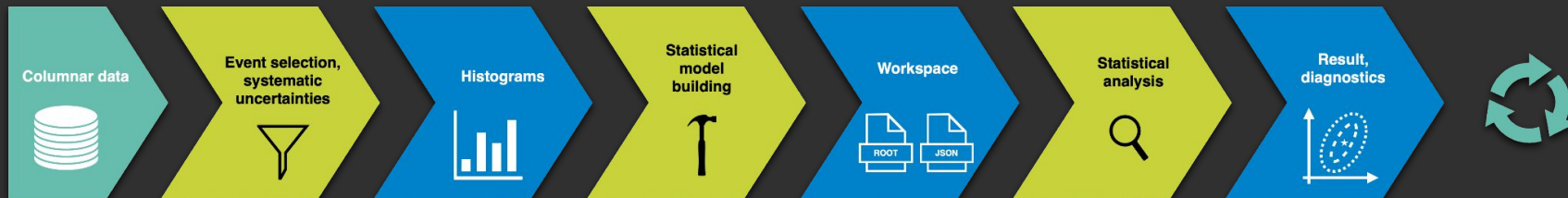
▷ AGC Demo Day #2

Integrating MLFlow into AGC Workflow

<https://indico.cern.ch/event/1232470/#5-integrating-mlflow-in-agc-wo>

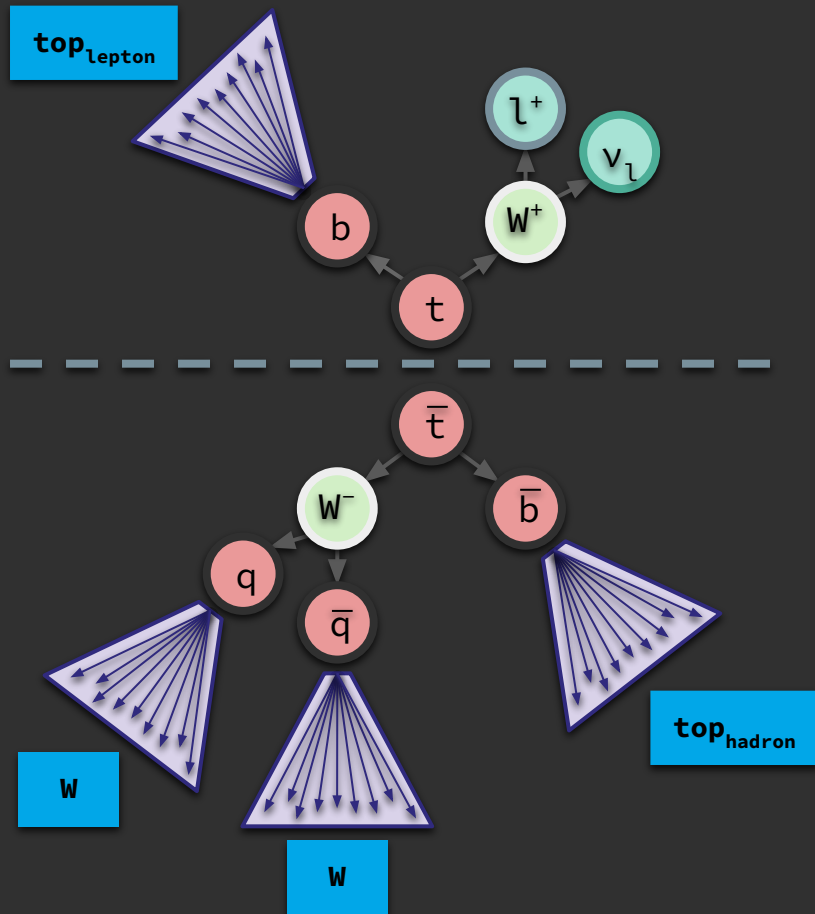
Previous Demonstrations

Previous Analysis Pipeline



Analysis Pipeline with ML



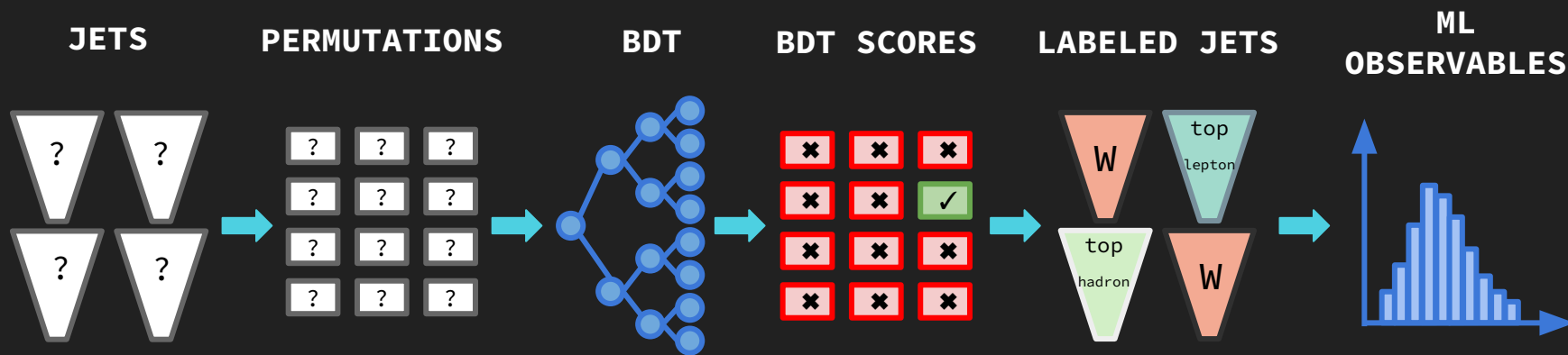


Machine Learning Task

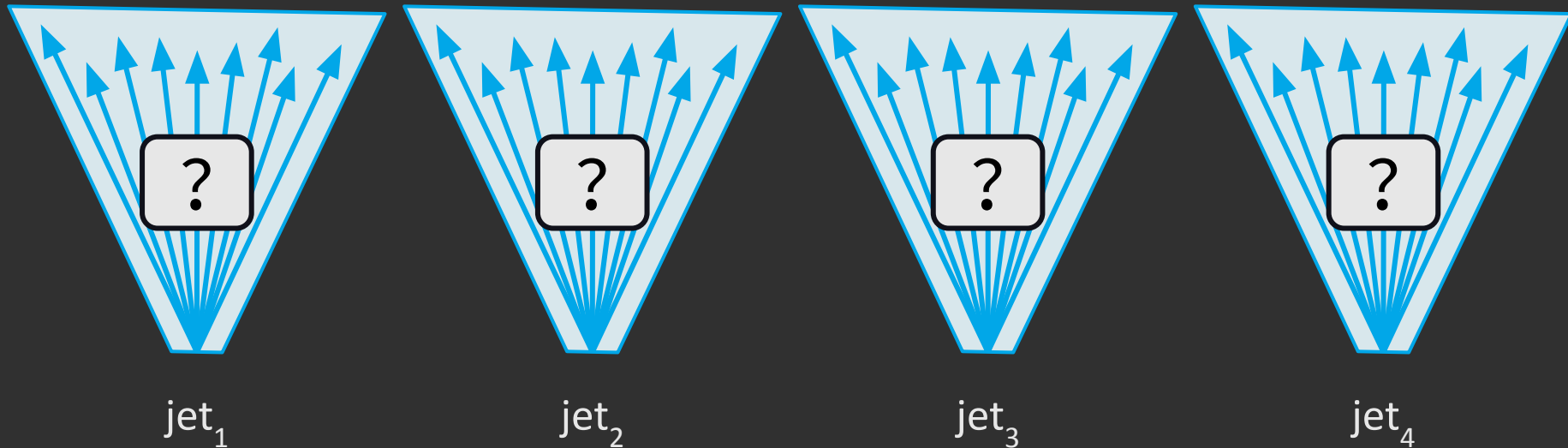
- Assign jets to their parent partons
- Allows us to approximate observables such as
 - ◆ Mass of top quark (combined mass of top_{hadron} and two W jets)
 - ◆ Angle between top_{lepton} jet and lepton ($\Delta\phi$)

Approach to ML Task

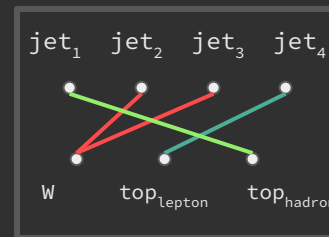
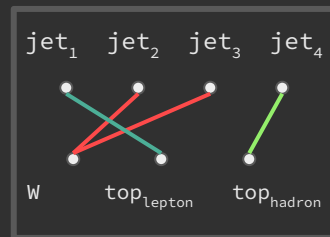
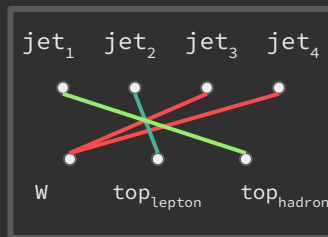
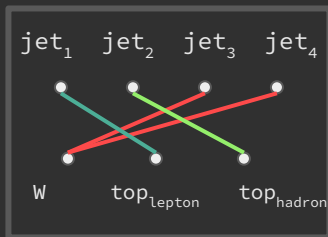
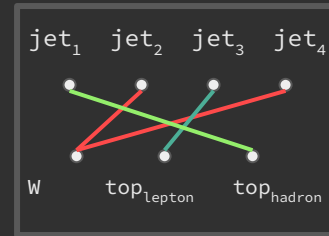
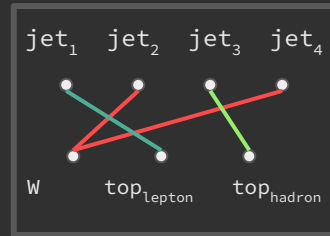
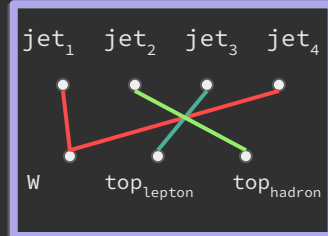
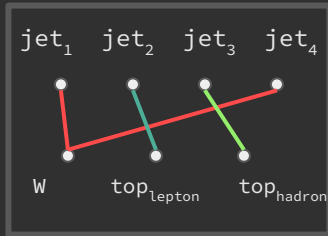
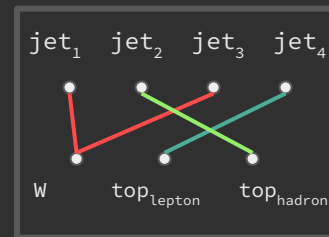
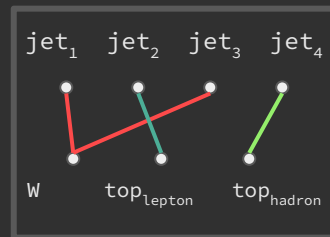
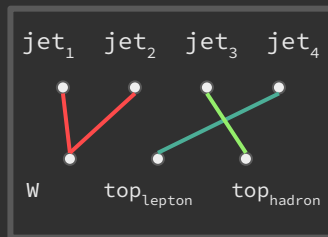
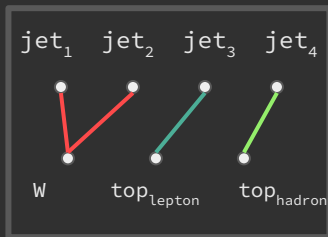
1. Consider **leading N jets** in each event
2. Find **all possible permutations** of parton assignments of these N jets (two W , one top_{hadron} , one top_{lepton})
3. **Calculate features** for each set of permutations and **feed into BDT**
4. Select permutation with **highest BDT score**
5. Use selection to **label jets**
6. Calculate **ML observables**



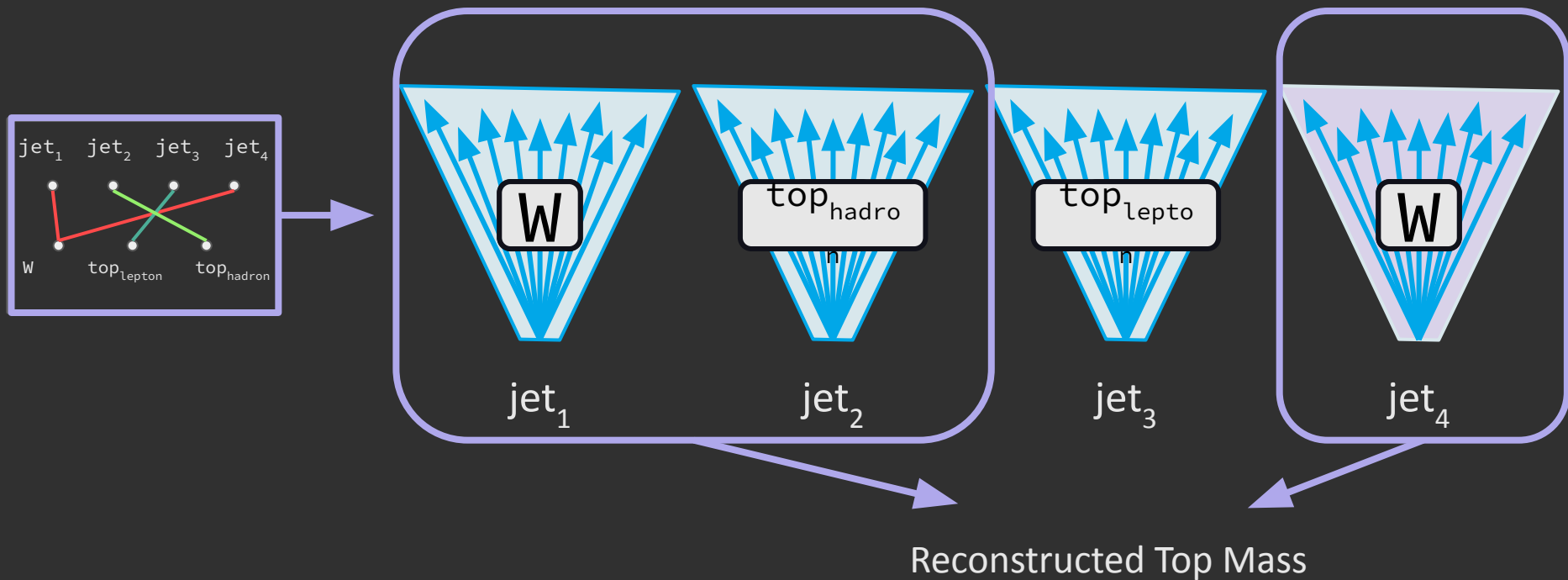
ML Task Example



ML Task Example

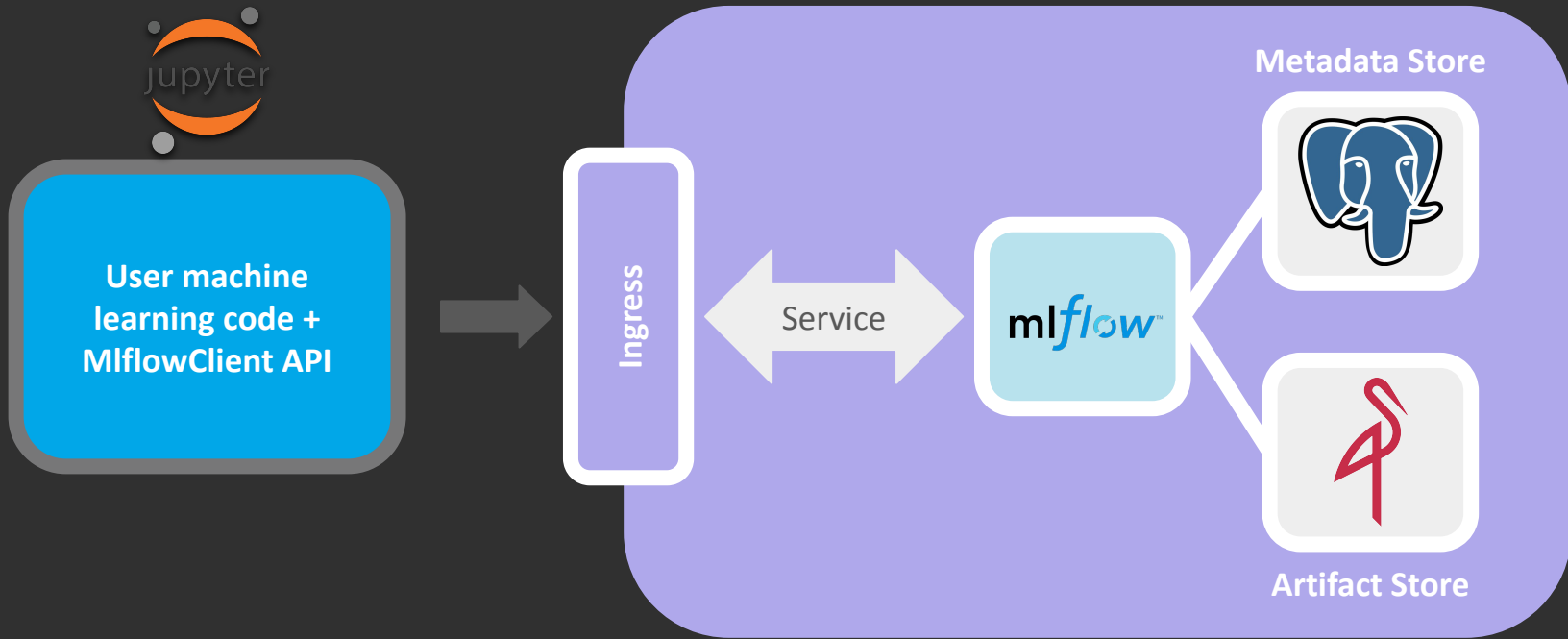


ML Task Example



Training





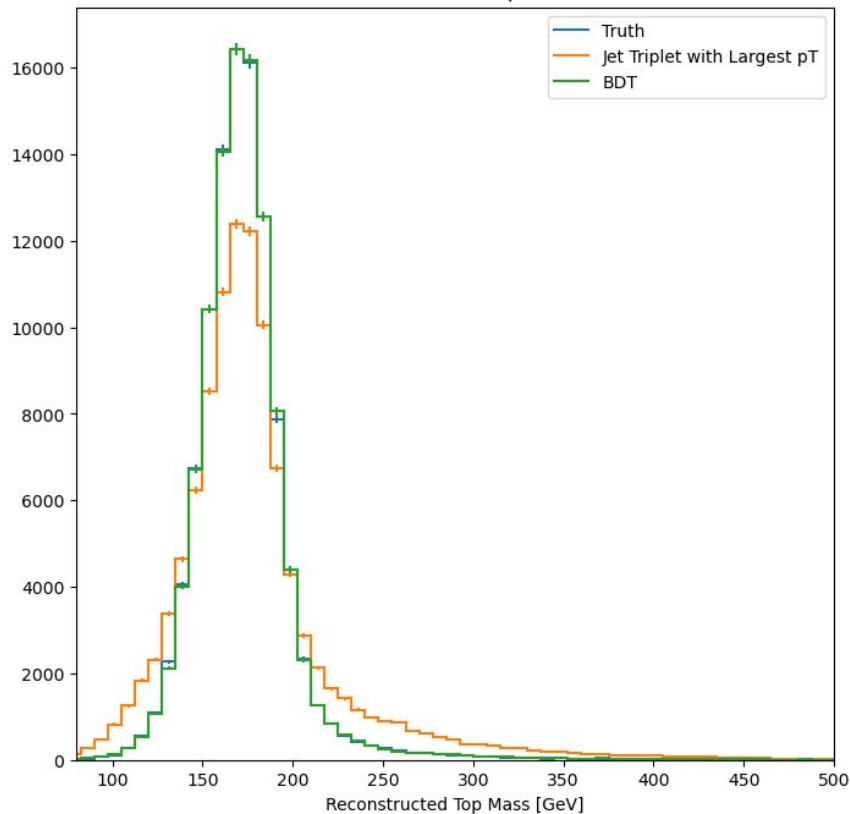
mlflow Architecture

DEMO

GitHub

Temporarily using NCSA
mlflow instance

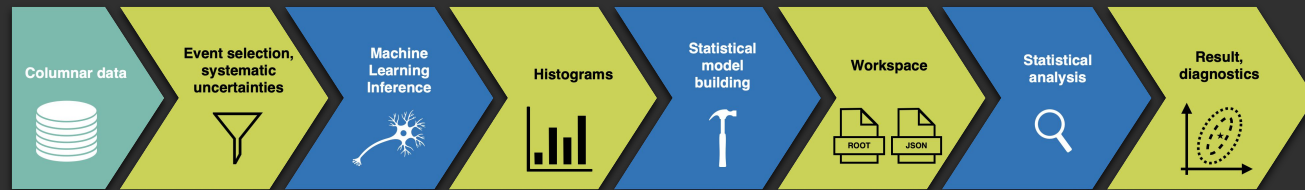
Reconstructed Top Mass

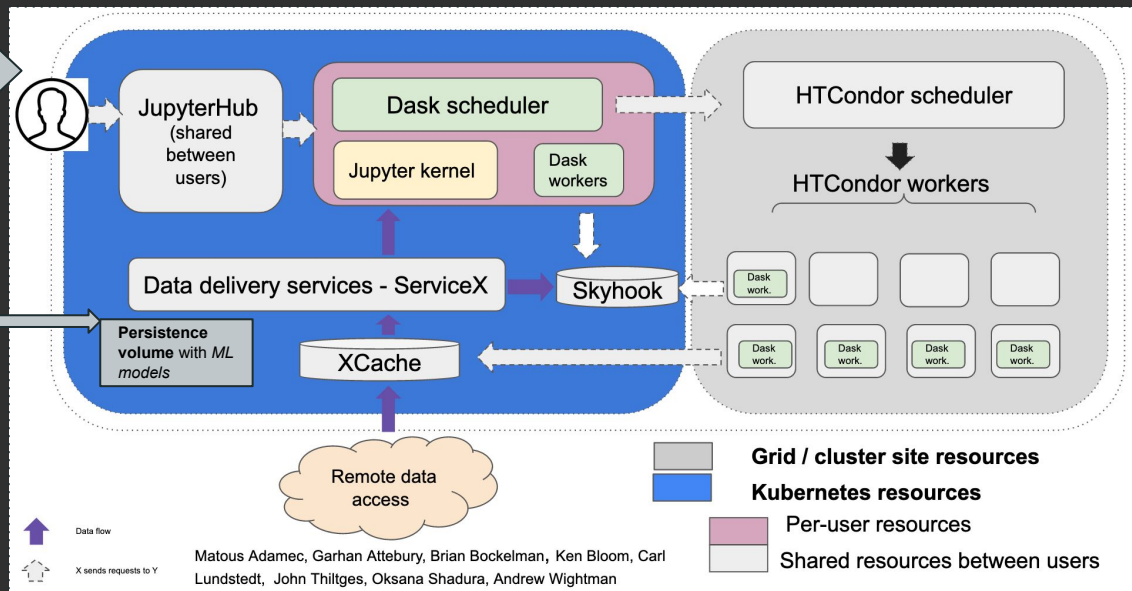
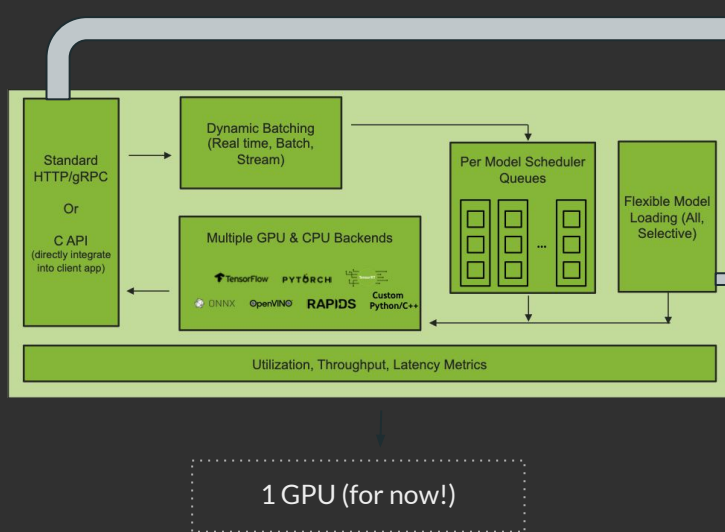


User Experience

- Right now we are using custom methods for hyperparameter optimization then mapping to Dask workers
- This is to be able to utilize mlflow
- Want to instead use dask-ml, but it is difficult to use with mlflow
- Can use mlflow with local setup, but have to define environment variables

Inference





NVIDIA Triton Architecture

DEMO

GitHub

General User Experience

→ xgboost json format doesn't seem to work well with NVIDIA Triton

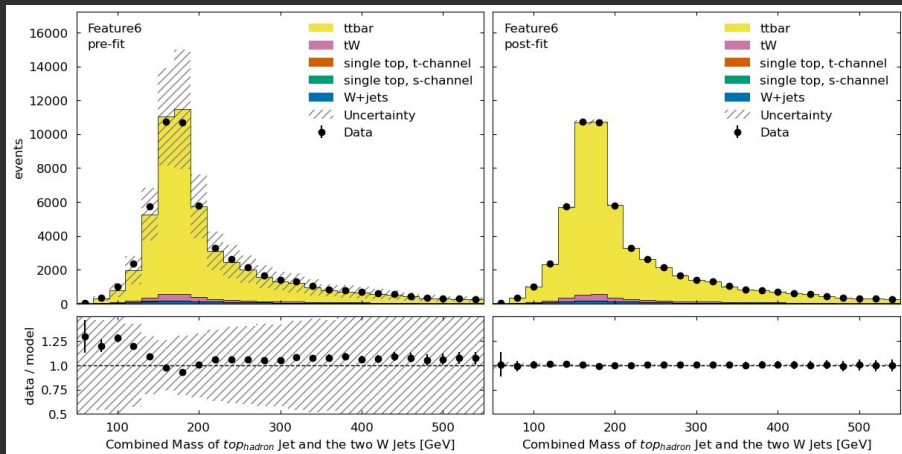
- ◆ Want to switch to torch or tensorflow for next iteration of AGC ML task, which will resolve this issue

→ Triton has somewhat unclear instructions for writing config files

→ If `USE_TRITON = False`

- ◆ If `USE_DASK=False`, need to pass model paths to coffea processor constructor and load models in processor.
- ◆ If `USE_DASK=True`, need to pass loaded models to constructor

→ Having method definitions in notebook is a bit too much information, would be nice to have dask dependency management





Future Goals:

- More complex machine learning task (looking into neural network-based approach)
 - Avoids dealing with each permutation of jets separately
- Integrate Triton + mlflow
- Improve user experience by utilizing dependency management for distributed dask workers