

Breakout Session 2: Track B

Scaling CDE Curation Model Training

Mr. Mark Weston
CEO, Netrias, LLC



FY24 ODSS Cloud Supplement Program PI Meeting

Scaling CDE Curation Model Training

FY2023 Funding Request Notice for Supporting the Exploration of Cloud in NIH
Intramural Research and Contracts

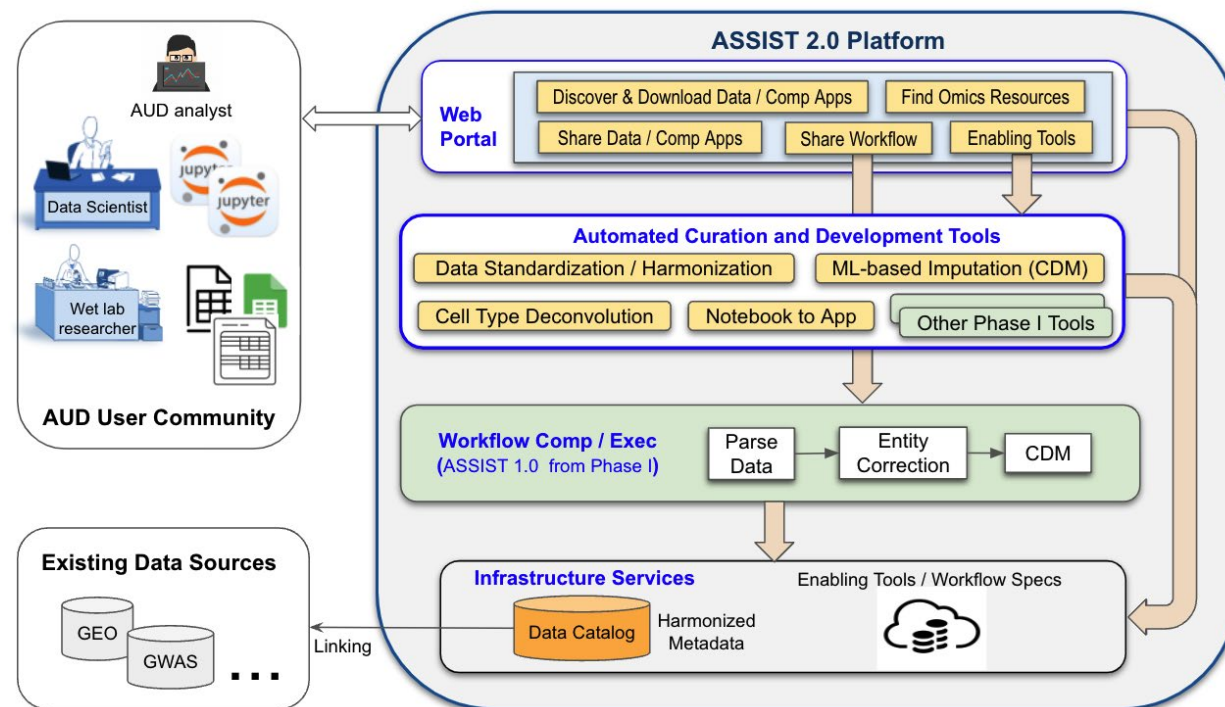
Mark Weston

weston@netrias.com

www.netrias.com



Summary

- Currently performing on SBIR Phase II for NIAAA: Alcoholism Solutions: Synthesizing Information to Support Treatments (ASSIST), Contract #75N94022C00003
- Build a data curation and harmonization toolkit and web portal to help researchers rapidly curate metadata and support data sharing and data archive initiatives



Project Goals

Advance and reduce barrier of entry for:

- **Data harmonization:**
 - Develop an extensible set of Common Data Elements (CDEs) for AUD researchers
 - Create a data curation and harmonization toolkit to ease data access and collaboration
 - AI modeling to harmonize between heterogeneous CDE representations 
- **Imputation:**
 - Integrate data with prior datasets and infer missing conditions 
 - Validate model by holding out some conditions
- **Deployment:**
 - Simplify deployment of apps from Jupyter Notebooks
 - Access tools through an easy-to-use web portal



Cloud Resource Benefits

- Two main areas that benefit from cloud resources:
 - Metadata term curation models to harmonize between heterogeneous CDE representations utilize smaller, bi-directional LSTM natural language processing models - one per type of data.
 - Cloud resources allows training of larger, generalizable models and accelerate multi-model training
 - Machine learning based imputation models (“Combinatorial Design Model”) benefit from larger neural network models with greater numbers of parameters

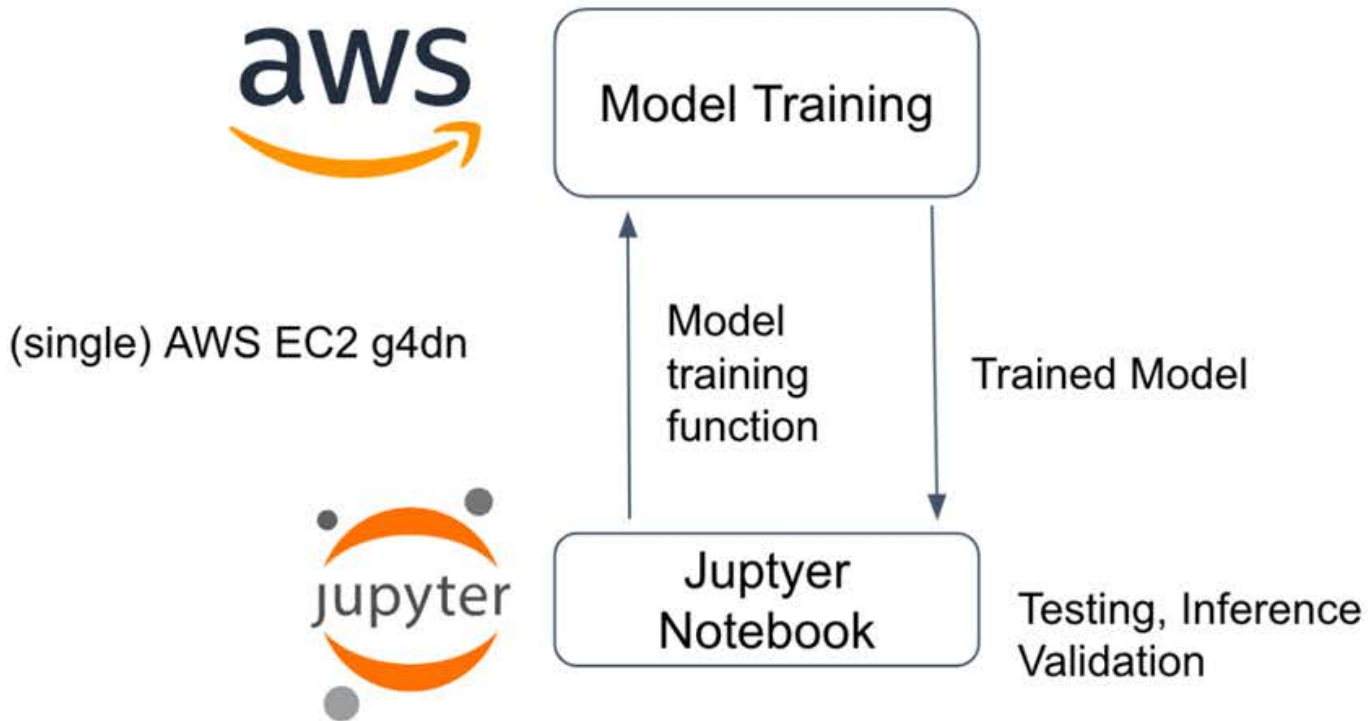


Motivation for Cloud Resources

- Previously utilized Netrias-owned compute and selective, limited use of Cloud services: Amazon Web Services (AWS) GPU instance (g4dn instance type). 1 Nvidia T4 GPU, 16 VCPUs, and 64 GiB of ram
- Under the current resources, limited to running a single instance at a time. Model approaches the available memory during training; can only train one instance of the model at a time
- Given available instance memory, cannot train larger models without utilizing a larger instance type



Previous Process



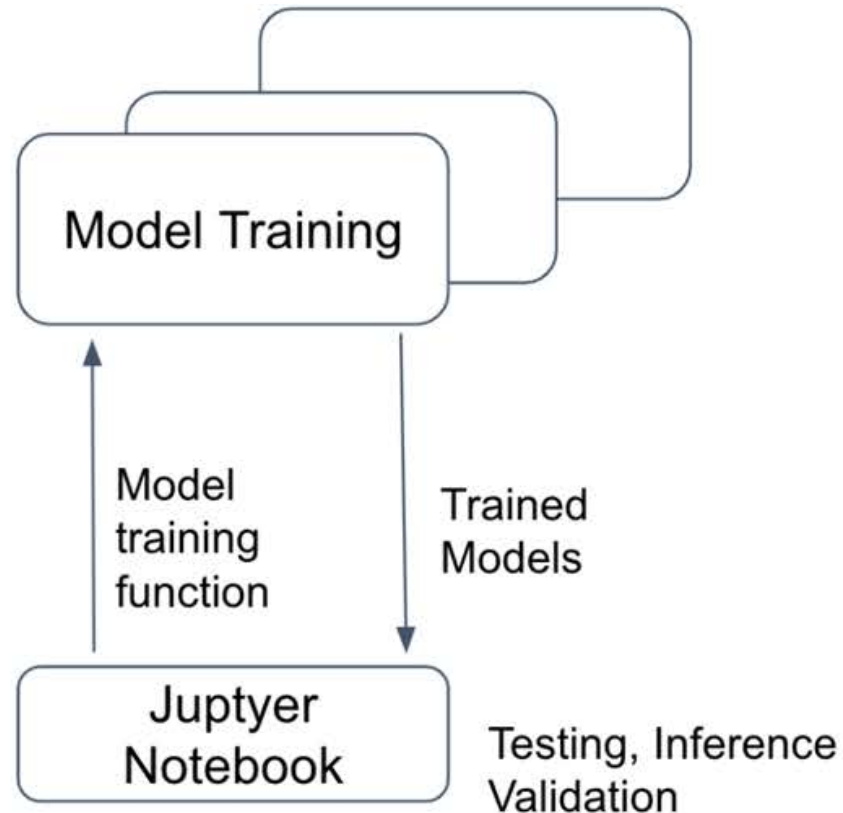
- Single Instance Training
- Limited Memory and Model Size
- Notebooks run locally on instance
- Model export via SCP/SSH



Updated Process



(multiple) AWS EC2 g4dn;
p3



- Multi Instance Training
- Greatly expanded GPU, Instance Memory and Model Size
- Notebooks run locally on instance
- Model export via SCP/SSH



Benefits

- Train multiple models simultaneously
 - Faster exploration of model hyperparameter space, model architecture variations, and model validation and testing
- Train models on larger and more powerful instances with increased GPU, VCPU, and memory amounts
 - Exploration of models that have larger parameter counts that do not fit on our current instance type
- More powerful training instances will improve iteration time, accelerating our overall development timeline



Benefits and Metrics

- Accelerate model training and development work as we make use of these additional instance types
- Facilitates faster iteration and release of models to the larger NIAAA research community
- Metrics
 - Number of model(s) trained before and after the cloud service improvement
 - Size (in parameter space) of the models trained before and after the cloud service improvement
 - Model performance and generalizability in accuracy metrics for both the curation and CDM models



Thank you!

Questions / Discussion

