Machine learning & Deep learning at ALCF

Elise Jennings

Data Science Group

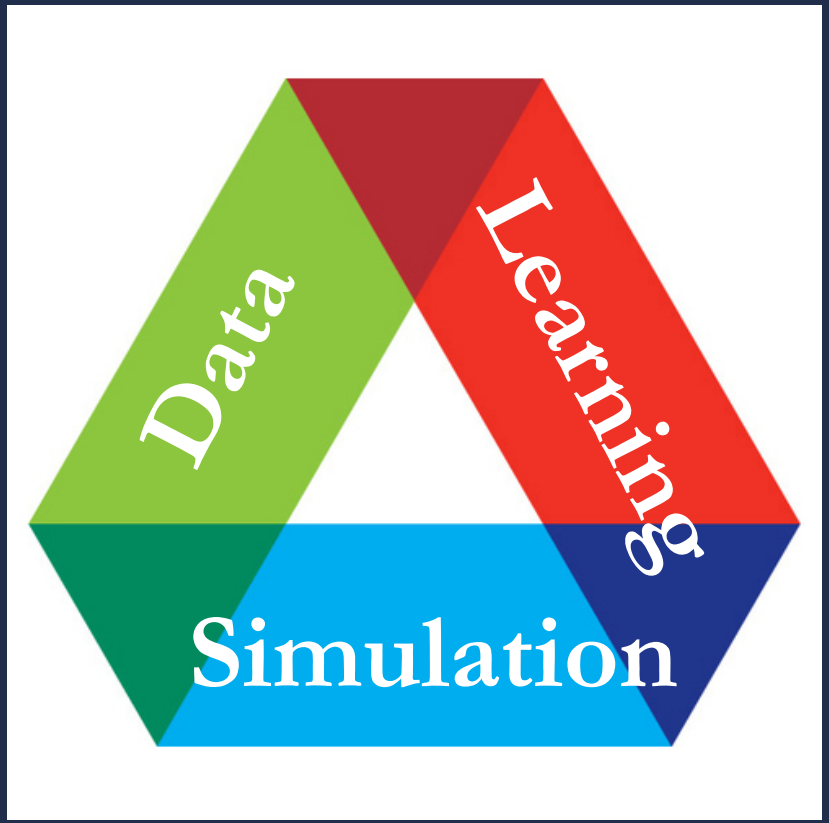
ejennings@anl.gov

Outline

- Data Science Program
- Highlight ADSP projects
- ML, DL & workflow software
- Other research projects

ALCF Data Science Program (ADSP)

- "Big Data" science that require the scale and performance of leadership computing
- Projects cover a wide variety of application domains that span computational, experimental and observational sciences
- Focus on data science techniques including but not limited to statistics, machine learning, deep learning, Uncertainty Quantification, image processing, graph analytics, complex and interactive workflows



ALCF Data Science Program (ADSP)

- Two-year proposal period. PIs required to fill out a renewal application for each allocation period of the award.
- Proposals will target science and software technology scaling for data science
- Projects need high potential impact, data scale readiness, diversity of science domains and algorithms, emphasis on projects that can use the architectural features of Theta

Data

- Experimental/observational data
- Image analysis
- Multidimensional structure discovery
- Complex and interactive workflows
- On-demand HPC
- Persistent data techniques
- Object store
- Databases
- Streaming/real-time data
- Uncertainty quantification
- Statistical methods
- Graph analytics

Learning

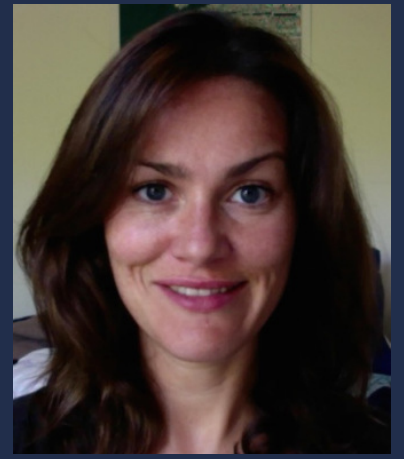
- Deep learning
- Machine learning steering simulations
- Parameter scans
- Materials design
- Observational signatures
- Data-driven models and refinement for science using ML/DL
- Hyperparameter optimization
- Pattern recognition
- Bridging gaps in theory


Tom Uram


Venkat Vishwanath


Taylor Childers


Murat Keceli

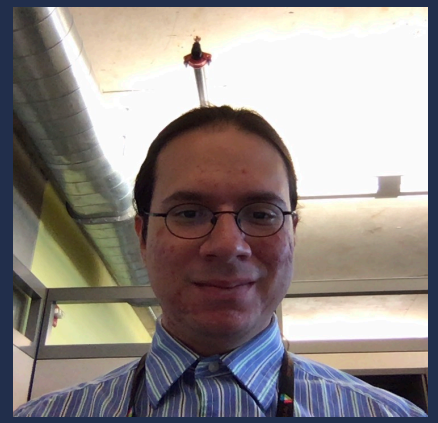

Elise Jennings


Alvaro Vazquez Mayagoitia


Adrian Pope


Misha Salim


Prasanna Balaprakash


Antonio Villarreal


William Scullin


Huihuo Zheng


Richard Zamora


Xiao-Yong Jin

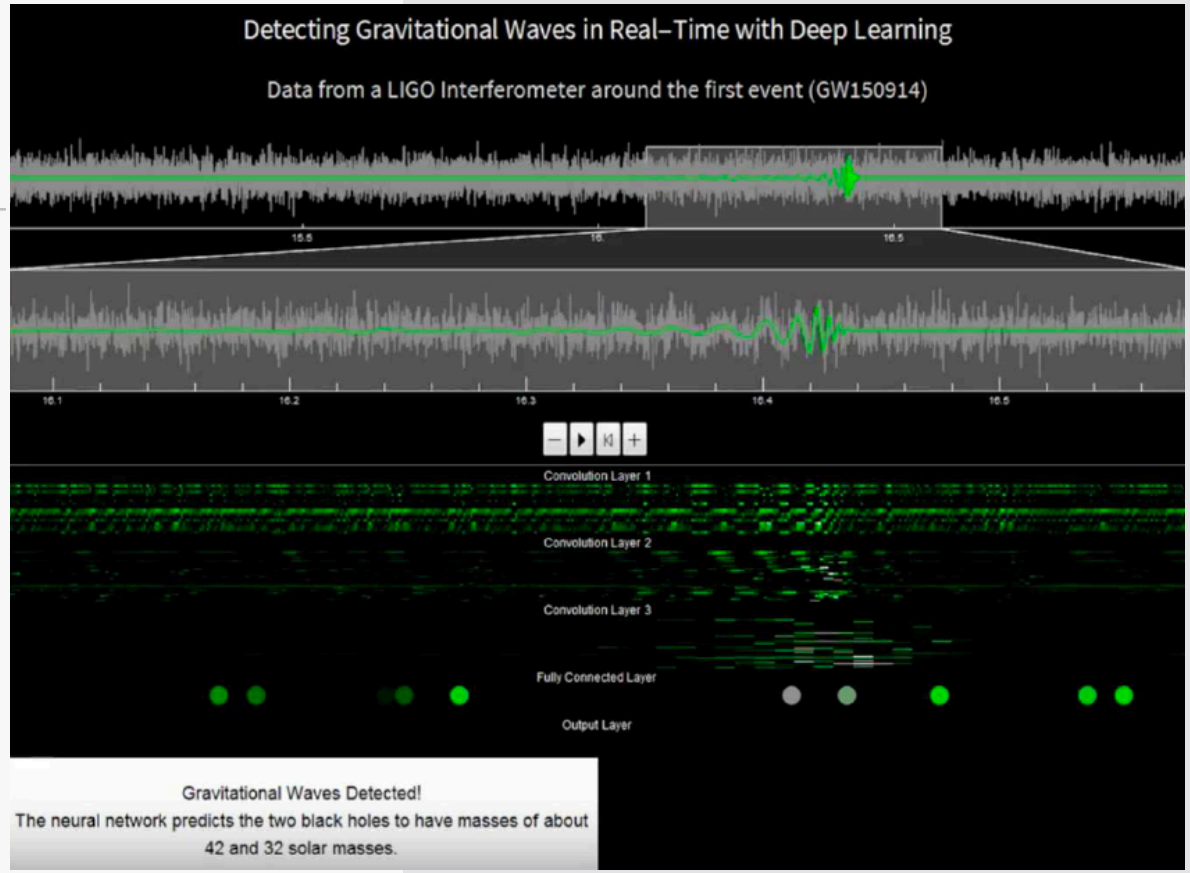

Ganesh Sivaraman

datascience@alcf.anl.gov

New ADSP projects starting Oct 2018 !

Cosmology: Deep Learning at Scale for Multimessenger Astrophysics through the NCSA-Argonne Collaboration

PIs: Huerta, Zhao, Haas, Saxton (NCSA)



Novel data-parallel deep learning to fuse HPC and AI for MultiMessenger Astrophysics (MMA).

Very important topic in modern Cosmology.

Update methods on both the gravitational wave (LIGO) and optical transient (LSST) domains.

Novel visualization of Neural Networks for interpretability.

An overarching goal is to enable the discovery of multi-messenger sources in real-time and at scale.

Materials: Machine learning magnetic properties of van der Waals heterostructures

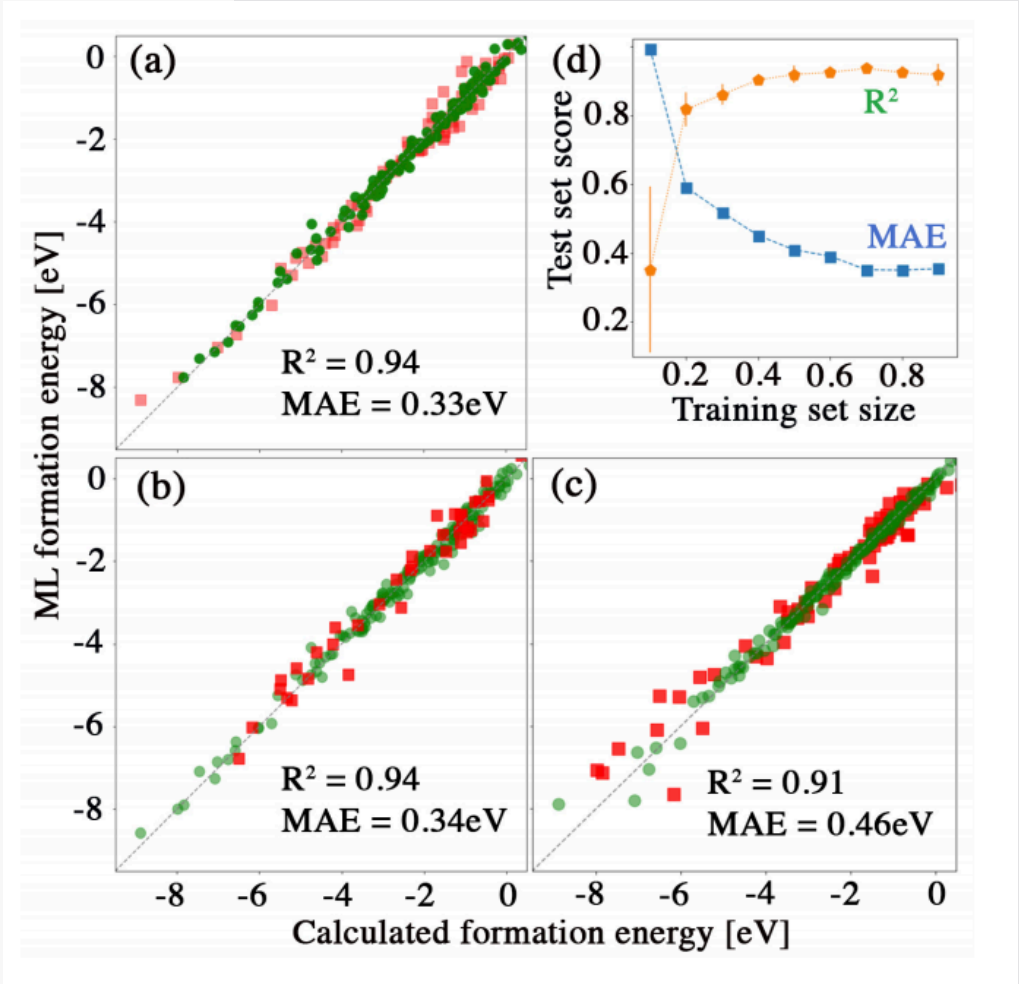
PIs: Kaxiras, Rhone (Harvard)

Use of high-throughput VASP calculations on HPC resources to train ML models for predicting magnetic properties in novel 2D materials.

Will improve understanding of magnetism in 2D materials, high impact on spintronics, data storage devices.

Novel training ML models on-the-fly useful for coupled simulation-learning workflows across all science domains, materials community.

Potentially accelerate material discovery & research in electronics industry



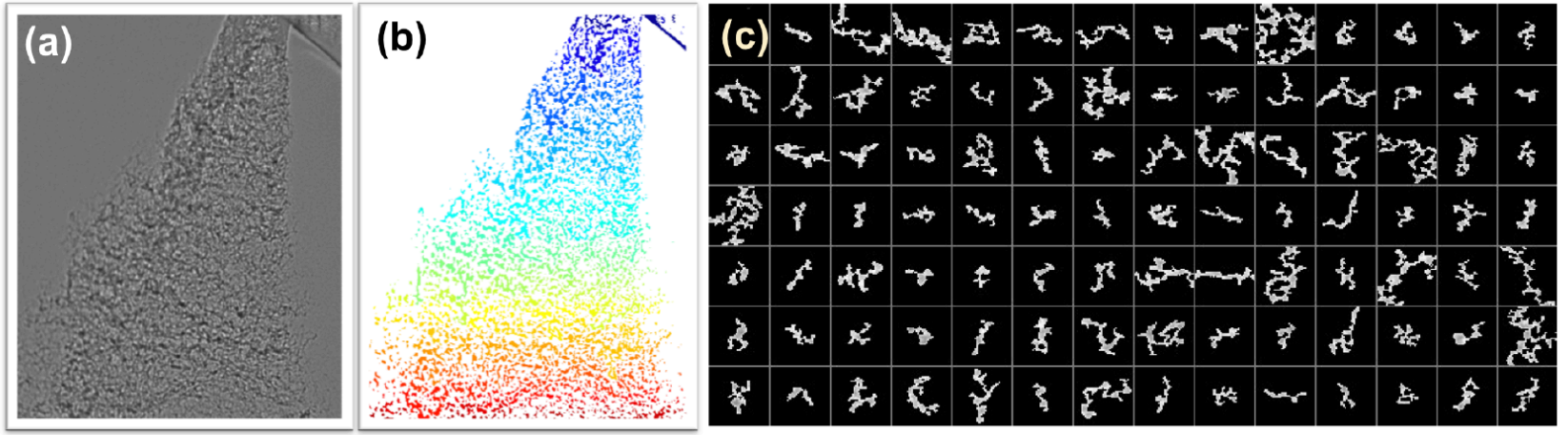
Imaging: Developing High-Fidelity Dynamic and Ultrafast X-Ray Imaging Tools for APS-Upgrade

PIs: Jin Wang et al(APS, ANL)

Real time analysis of APS data at ALCF, CFD simulations & development of deep NN.

Suite of tools will enable real time imaging analysis and improve our understanding of structure, kinetics, dynamics of materials and highly transient processes.

High impact and beneficial to APS users as well as the general scattering community



Initial processing of cavitating flow visualized by ultrafast X-ray imaging (a) normalized image. (b) Image segmentation (c). Isolated patterns.

Imaging: X-ray microscopy of extended 3D objects: scaling towards the future

PIs: Chris Jacobsen, Wild, Nashed (ANL, Northwestern)

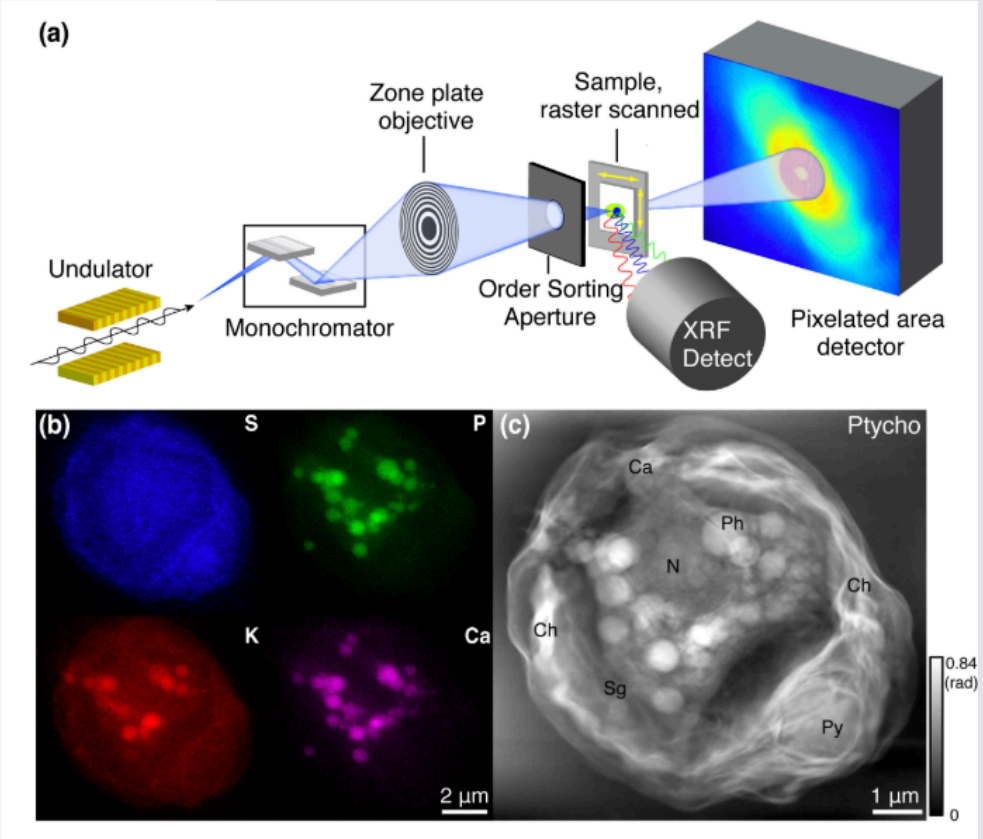
Enable high resolution X-ray imaging for thick specimens going beyond pure projection approximation using novel modeling/optimization techniques.

Apply this reconstruction to eukaryotic cells (sub 20nm) and brain tissue imaging (synapse-level detail at the 10-40nm).

APS-U data sizes require leadership class systems.

Improve our understanding of integrated circuits fabrication and defects, and neuroanatomical structures.

Open source code to users at the APS and other lightsources



Data Science software @ ALCF

- ML/DL, Data analysis, python, containers
- Workflows: Balsam
- Data transfer and management: Globus and Petrel
- Hyperparameter optimization: Deep Hyper
- Interpretability and Uncertainty Quantification in DL/ML: TFP

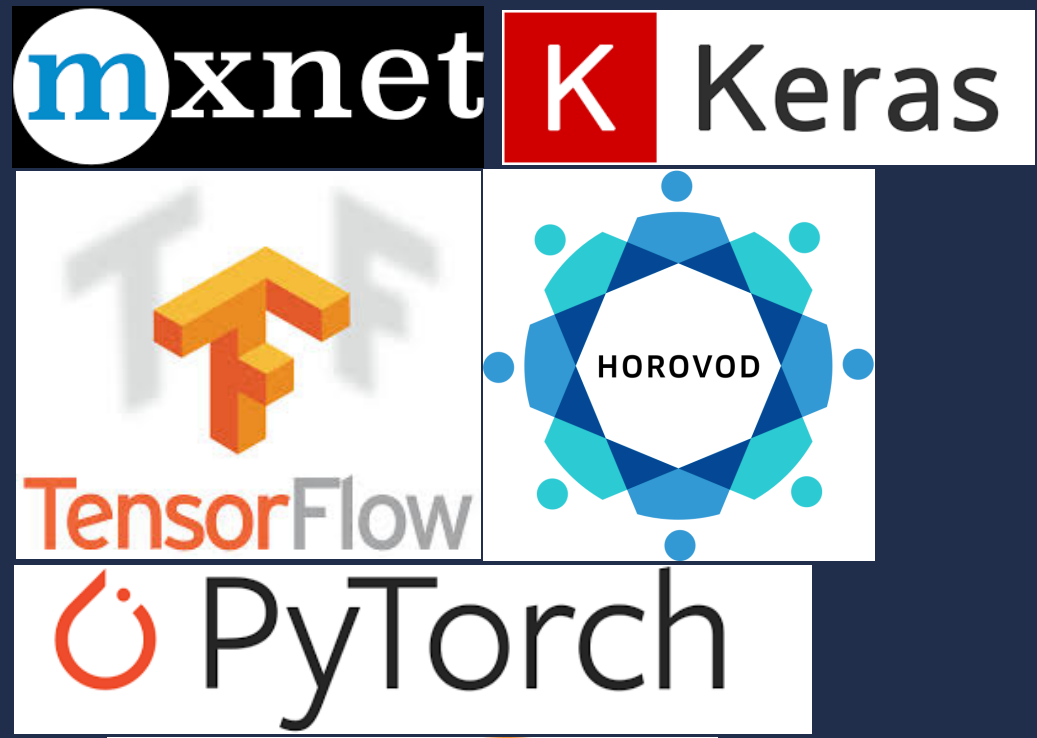
Machine Learning, Deep Learning & Workflow software

ML/DL:

- **TensorFlow, Keras, Neon, MXNet, Caffe2, Theano, CNTK, PyTorch, Sci-kit Learn, Graph Analytics (Cray Graph Engine), Horovod**
- **With performance libraries e.g. Intel MKL, MKL-DNN, LibXSMM enabled**
- **Intel optimized Tensorflow**
- **Conda package on Theta**
- **Intel Distribution for Python's optimized numpy**

```
module load
```

```
datascience/horovod-0.13.11
datascience/keras-2.2.2
datascience/pytorch-0.5.0-mkldnn
datascience/tensorflow-1.4
datascience/tensorflow-1.6
datascience/tensorflow-1.10
```



Machine Learning, Deep Learning & Workflow software

Containers

- Singularity container solution for application science workloads.
- Environment imported into container

Data Analysis

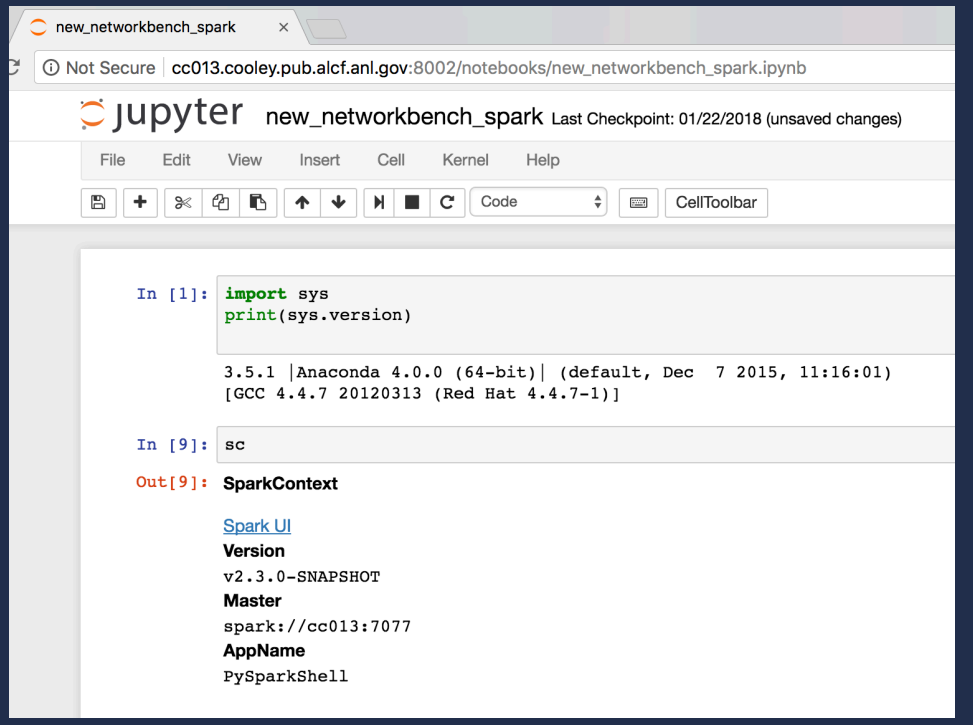
- MongoDB, Apache Spark, R

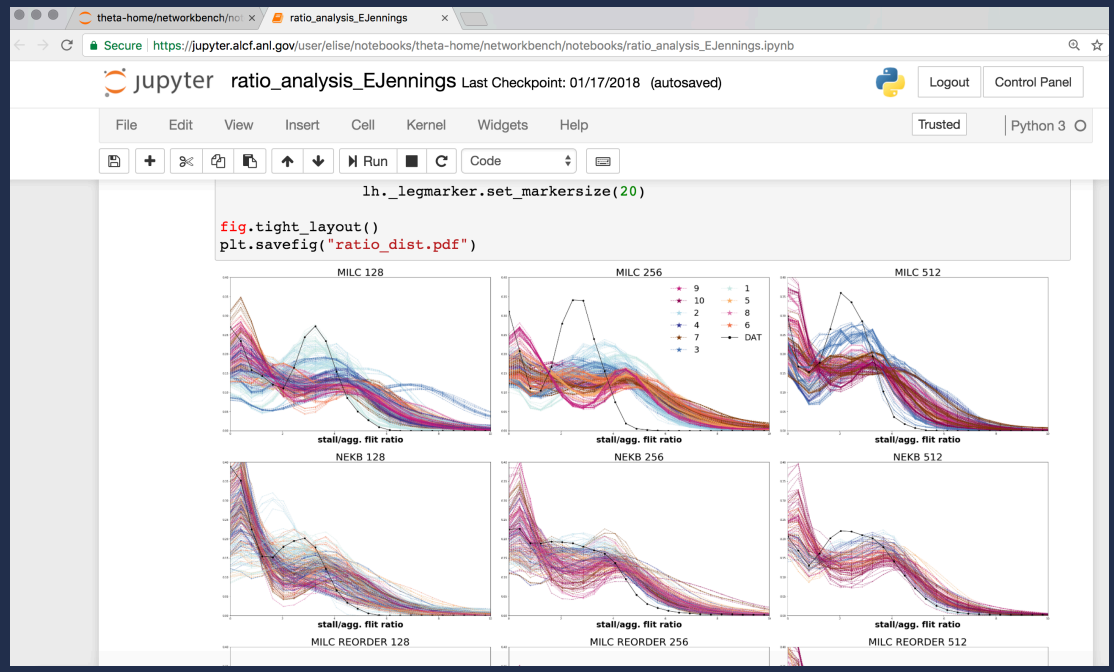
Python

- Intel and Cray modules on Theta
- ALCF alcfpython/2.7.14-20180131
- Conda modules
- Jupyter Hub

Jupyter Hub

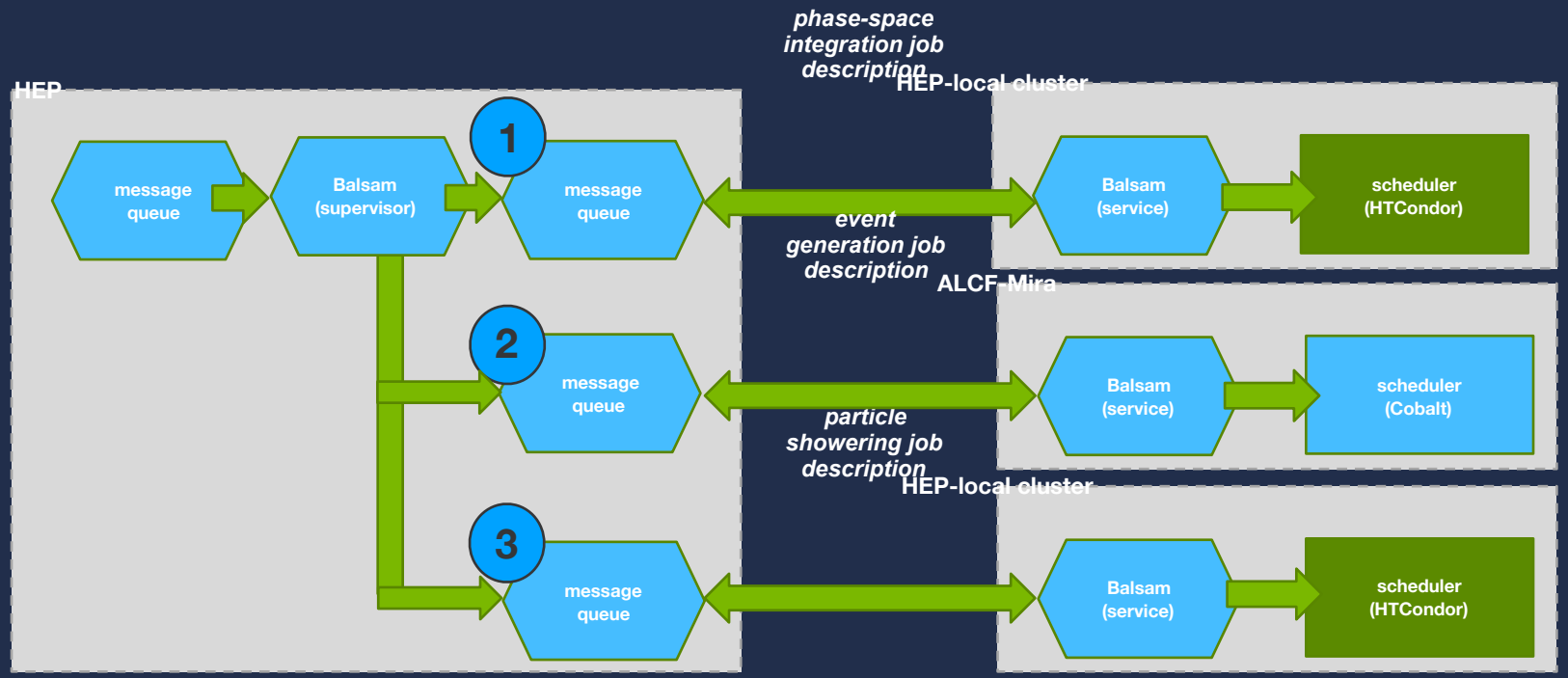
- Interactive computing environment for Python and R users.
- Login jupyter.alcf.anl.gov/theta
- Access to Theta /home and /projects folders
- You can submit jobs to Theta queues. (Run !qsub myjob.sh on a cell)
- Default Jupyter kernel is Python 3.
- Create custom conda enviroments to install any Python module and use custom kernels. Sample Jupyter notebooks /projects/datascience/jupyter
- Note: if you login to jupyter.alcf.anl.gov (without /theta), then you get access to Cooley projects folder not Theta and your jobs will run on Cooley. (This will change in the future)





Workflow manager: Balsam

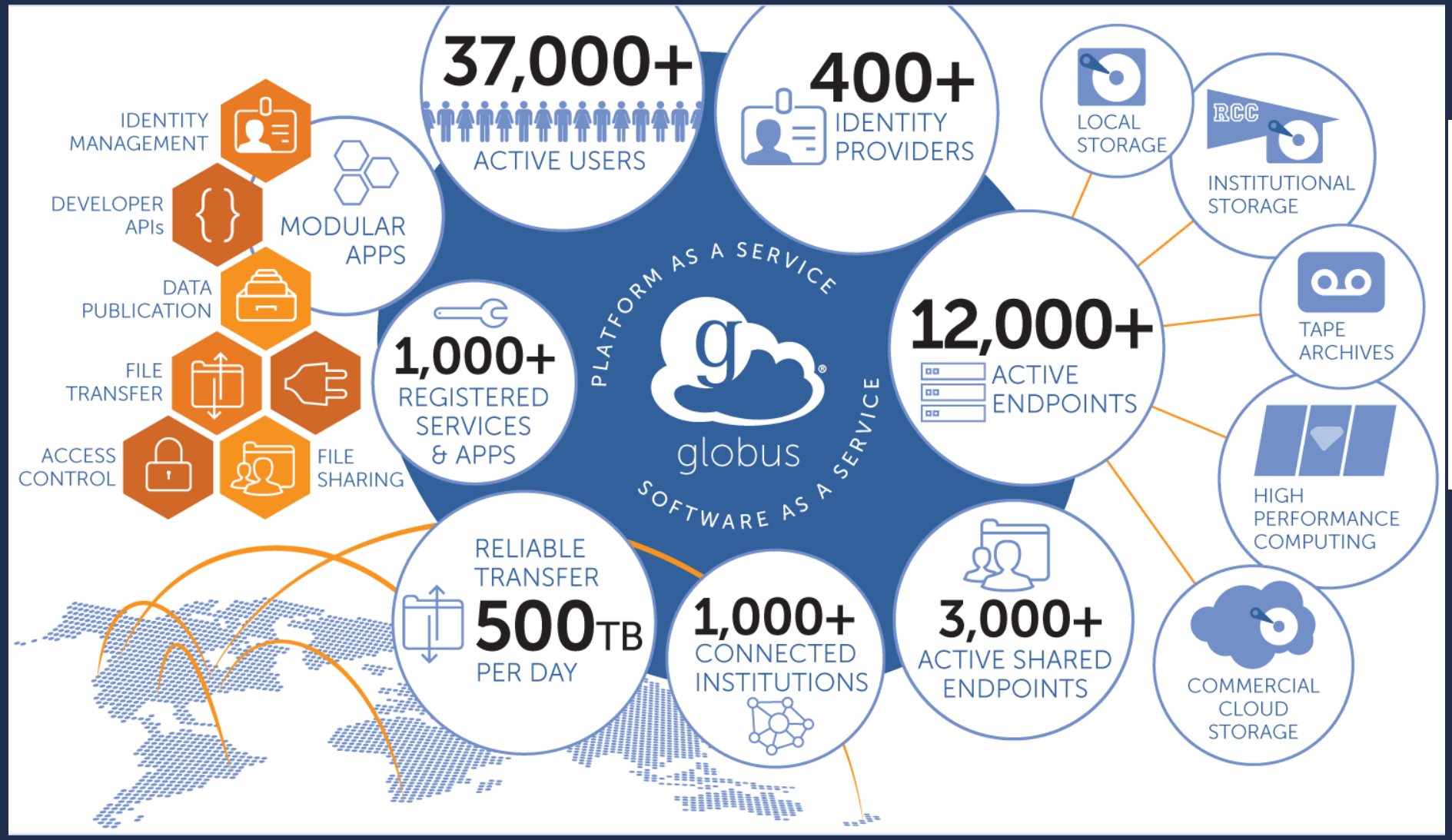
- Project run by the ALCF Data Science group, to optimize workflow execution on ALCF systems
- Users can easily describe a large campaign of jobs to be run, with varying sizes and interdependencies, and let Balsam manage flowing the jobs out to the target systems
- Delivered >150M Core hours for production science on ALCF systems
- The ATLAS experiment has used Balsam to run hundreds of millions of compute hours of event generation jobs on ALCF systems.



https://www.alcf.anl.gov/balsam

Balsam is a workflow manager that simplifies the task of running large-scale job campaigns on ALCF resources while minimizing user involvement and improving productivity.

Data transfer: Globus

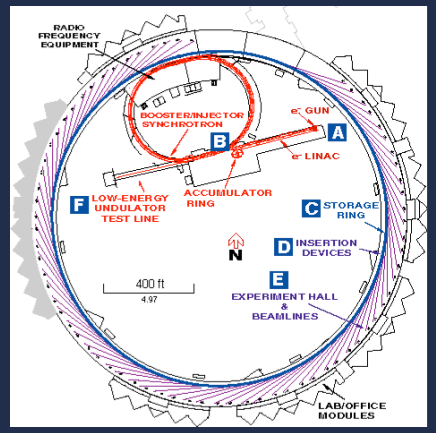


ALCF Globus Endpoints

- Theta: alcf#dtn_theta
- Mira: alcf#dtn_mira
- Cetus: alcf#dtn_mira
- Tukey: alcf#dtn_mira
- Vesta: alcf#dtn_vesta
- HPSS: alcf#dtn_hpss

Data management and sharing: Petrel

Search & discovery via portal



Instrument facility (APS)

Modern Research Data Portal

PETREL

Data Management and Sharing Pilot

Share with collaborators

Compute (ALCF, LCRC)

Manage permissions

Publicly available

http://petrel.alcf.anl.gov

Argonne NATIONAL LABORATORY

Petrel by the numbers (2017)

31 projects	106 Million files moved	1.4 PB data stored
320 total users	2.2 PB transferred	372 TB largest project
110 users of single project	12,000 transfer tasks	100TB allocation/ project

Hyperparameter Optimization: Deep Hyper

- Hyperparameter optimization is of paramount importance for deep learning for science. Key data science workload on exascale systems

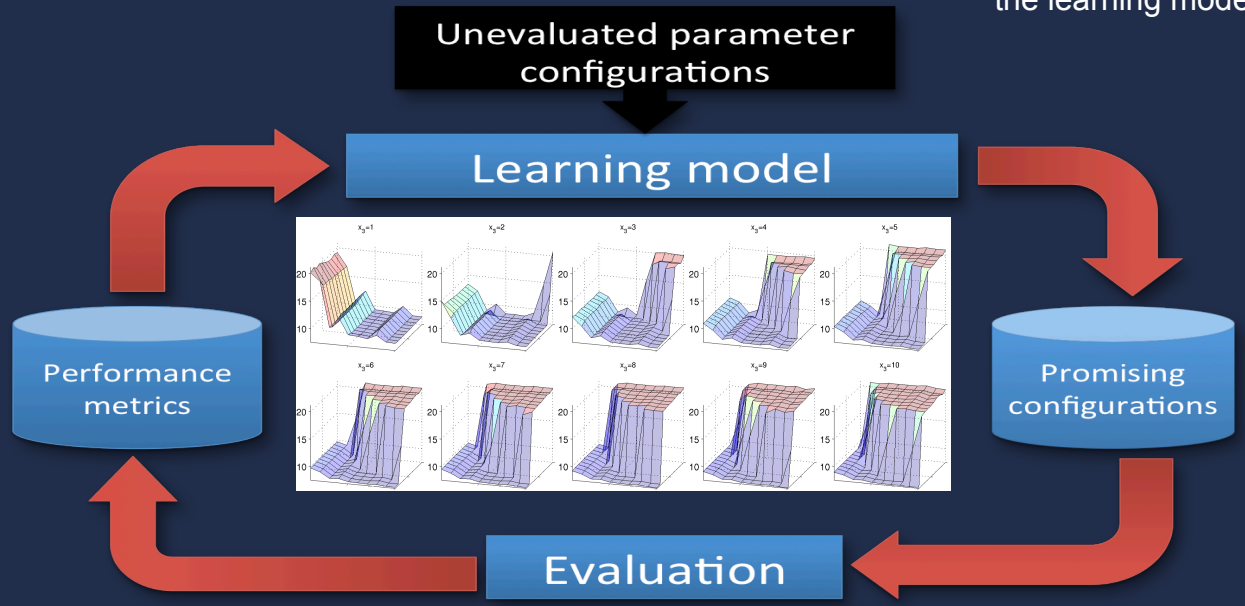
- Search iteratively refines a model in promising input region (based on expected improvement)

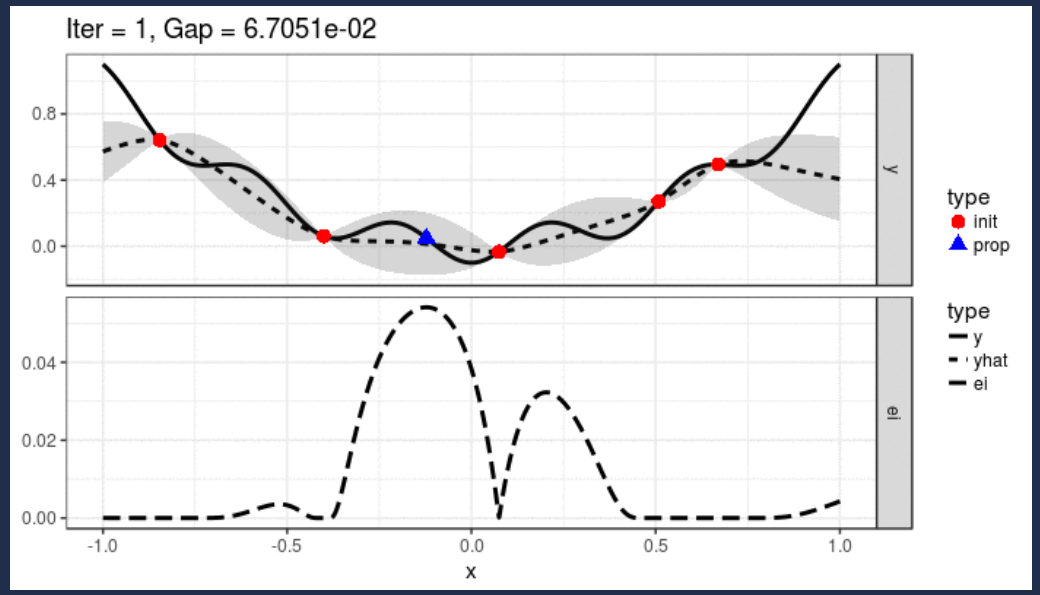
General framework:

- Initialization phase using Random or Latin hypercube sampling
- Iterative phase wherein we fit a model and sample using this model

- Genetic Algorithm, Gaussian processes, random forest and xgboost algorithms are used for building models.

Example Surrogate Model Fitted to Sampled Performance (iterative refinement improves the learning model)

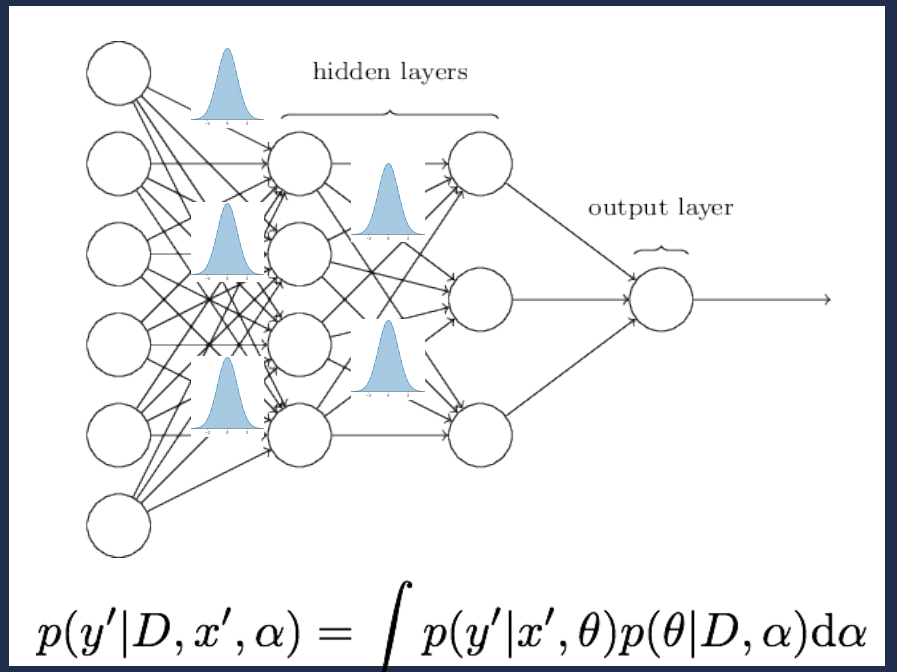




Uncertainty Quantification in deep learning:

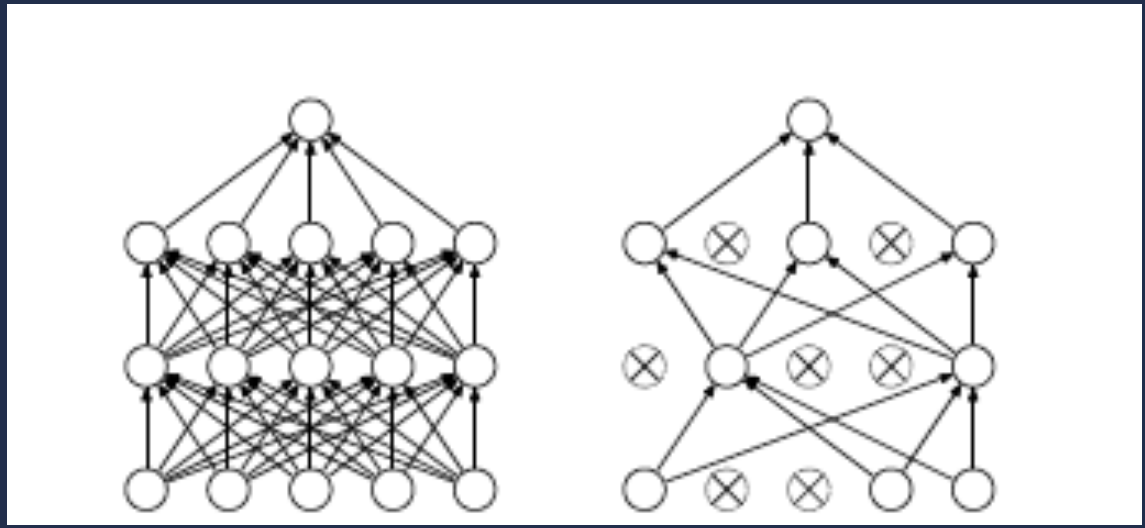
Bayesian Neural Networks

- Marginalization over hyperparameters
- More robust to overfitting
- Regularization = choice of prior
- Model comparison via Bayesian Evidence



Dropout

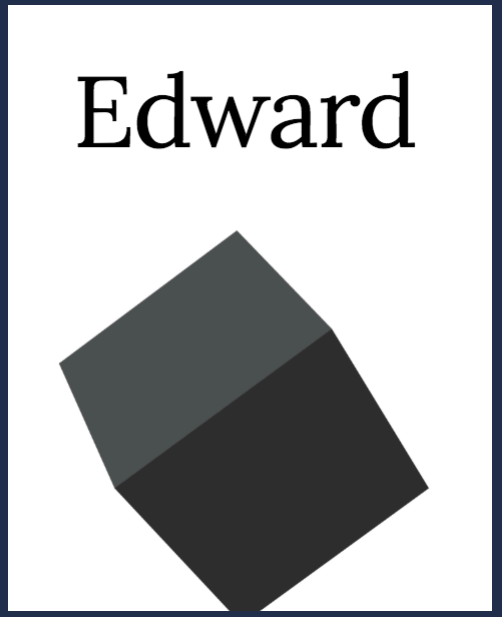
- Prob p to drop weights from network at training time
- Averaging exponentially many models with shared weights
- Each model is equally weighted
- Fast to use at train and test time



Uncertainty Quantification in deep learning:

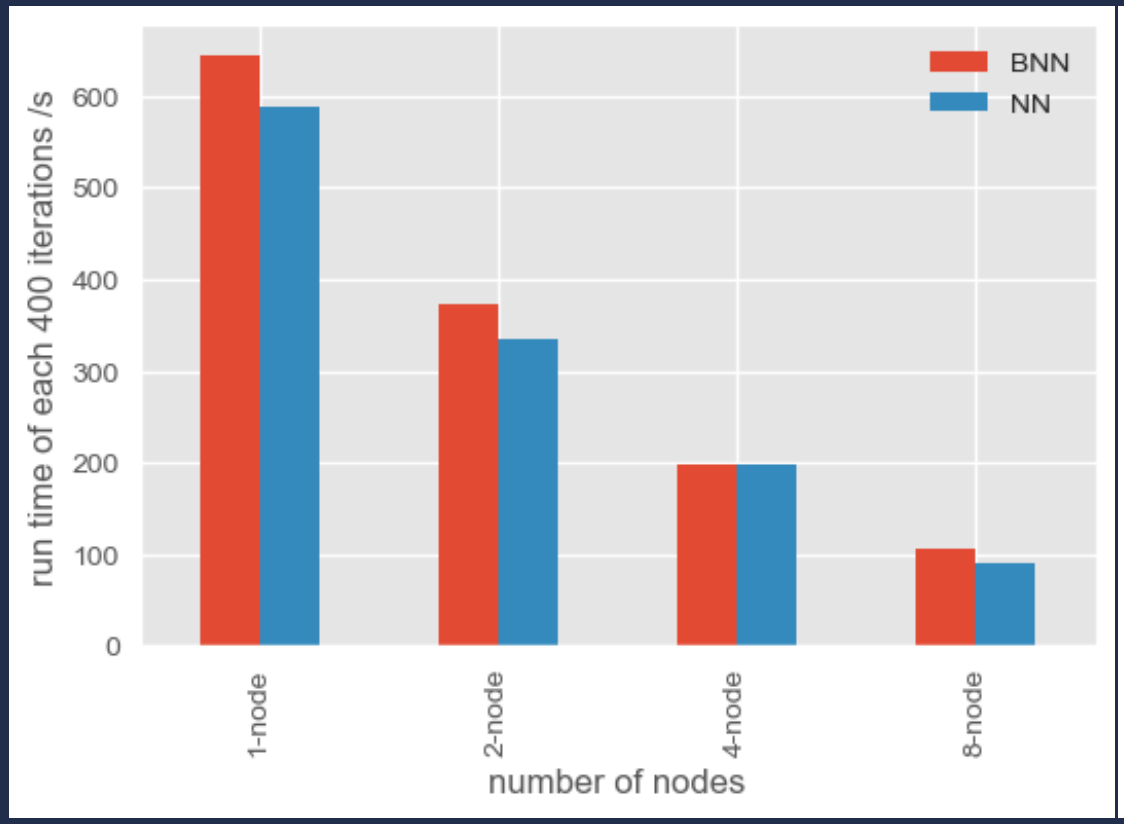
Software:

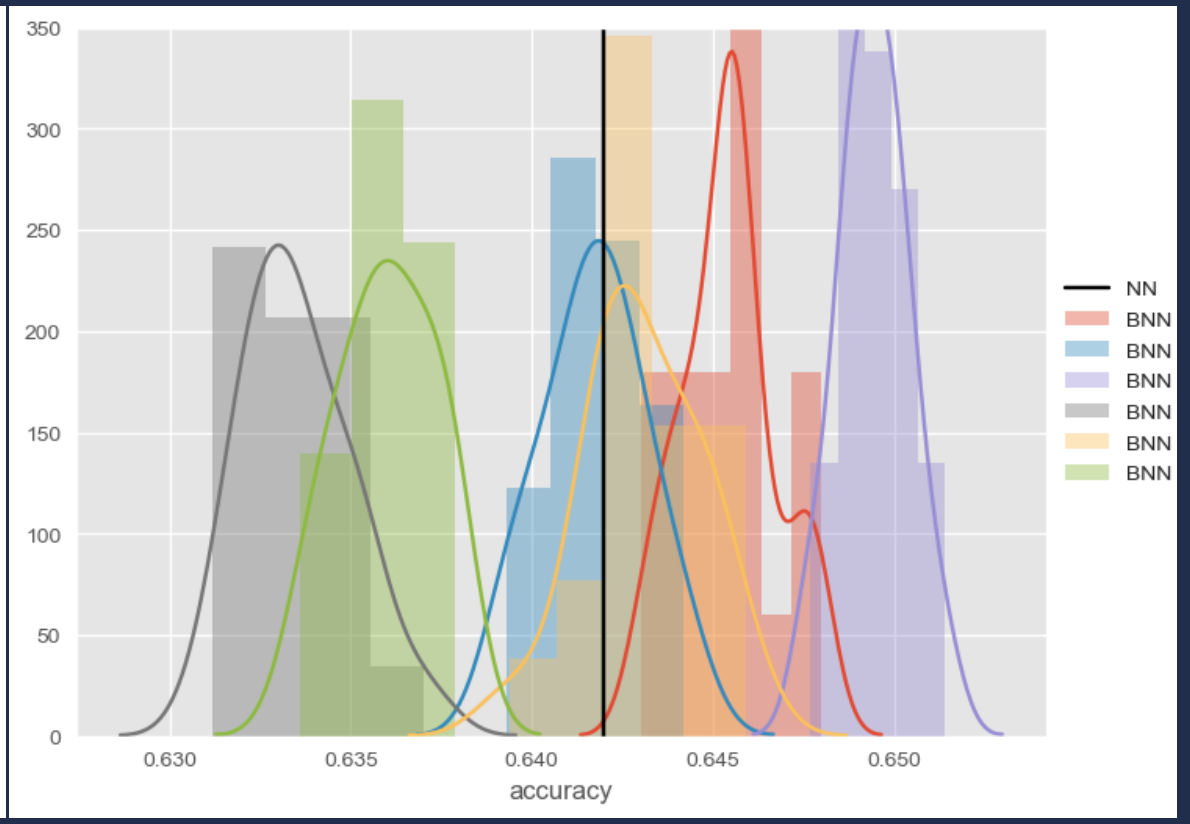
- Tensorflow Probability for probabilistic reasoning and statistical analysis in TensorFlow.
- Tensorflow Distributions
- tf.layers extended to probabilistic layers
- Loss function=log Prob of Likelihood + log Prob Prior
- Many methods for inference: MCMC, VI, HMC…



Results on Theta:

- ***Data:*** *CIFAR10*
- ***Architecture:*** *Convolutional layer (32 kernels w/ size 3 by 3) + Max pooling (size 2 by 2) + fully connected*





Thank you !

datascience@alcf.anl.gov