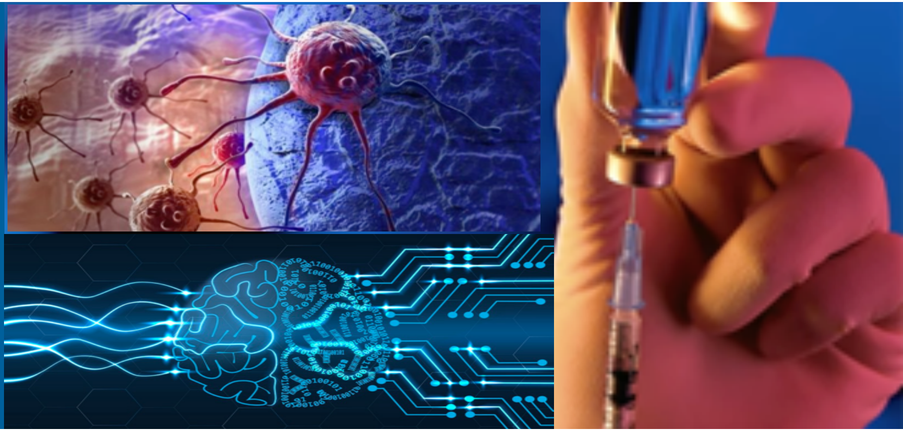
USING THE MOST ADVANCED COMPUTERS TO SOLVE THE HARDEST PROBLEMS

Argonne NATIONAL LABORATORY

EXASCALE DEEP LEARNING ENABLED PRECISION MEDICINE FOR CANCER



THOMAS BRETTIN
Strategic Program Manager
Argonne National Laboratory
University of Chicago

Thursday June 20th, 9:00 am – 4:00 pm.
Deep Learning for Science, ISC2019

ACKNOWLEDGEMENTS

Materials in this presentation comes from **Argonne, Oak Ridge, Livermore and Los Alamos National Laboratories, National Cancer Institute, Frederick National Laboratory for Cancer Research, and the University of Chicago**

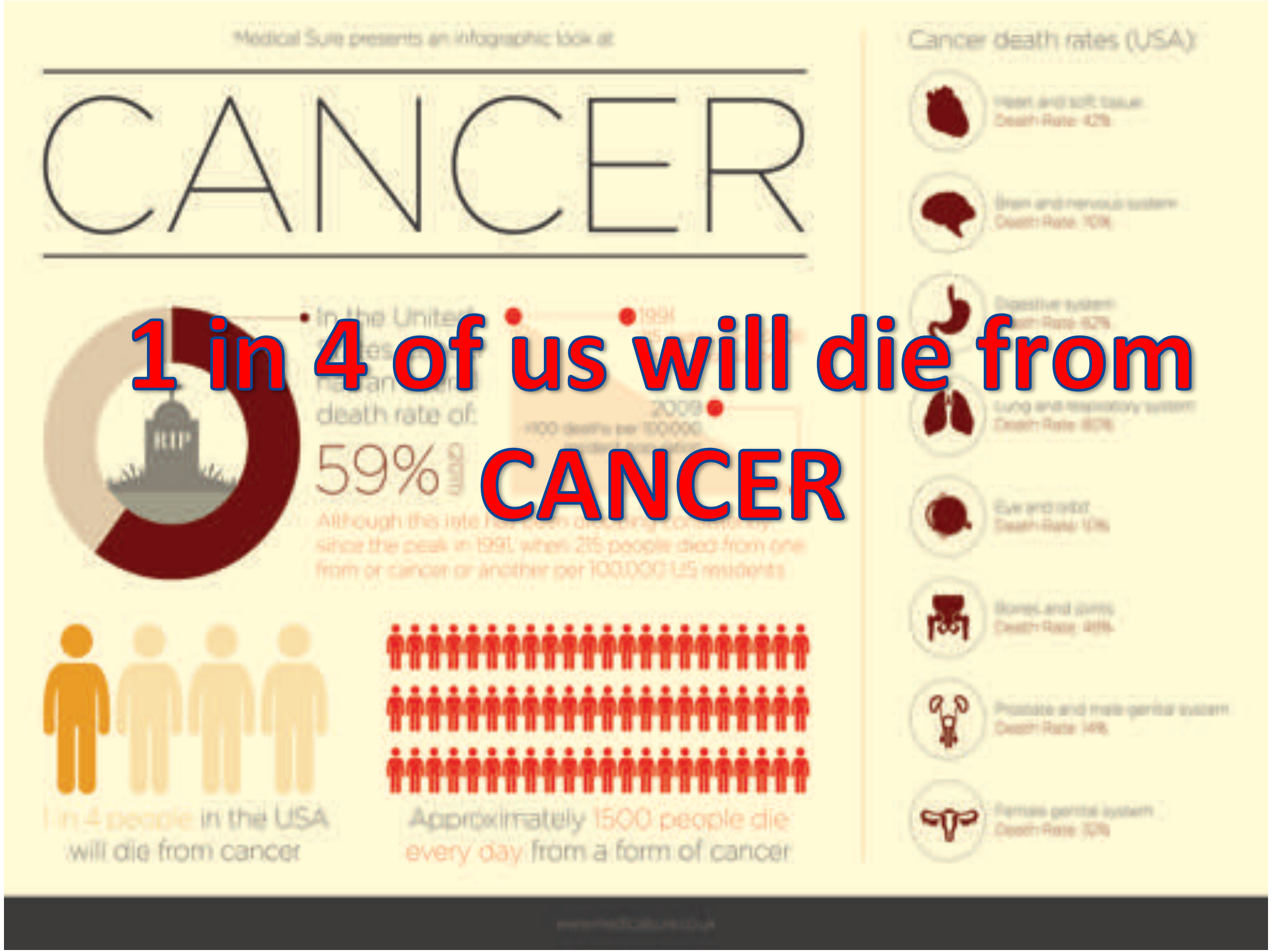
Special thanks to:

Gina Tourassi (ORNL). Fred Streitz and Brian Van Essen (LLNL). Tanmoy Bhattacharya and Jamal Mohd-Yusof (LANL). Eric Stahlberg and George Zaki (NCI). Rick Stevens, Fangfang Xia and Arvind Ramanathan (ANL)

THE 15 SECOND OVERVIEW

- What is the motivating problem
  – Cancer
- What is CANDLE
  – It is an exascale computing project application
  – It is a framework for executing computational (DL) experiments
- Results
  – It scales and is enabling discovery
- Next steps
  – Bigger challenge problems

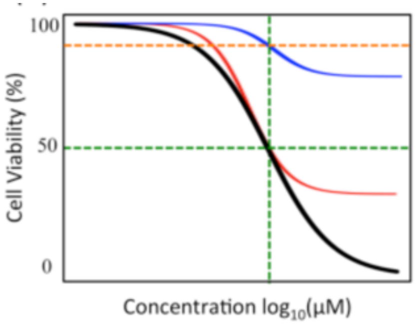
1 in 4 of us will die from CANCER



CANDLE (CANCER DISTRIBUTED LEARNING ENVIRONMENT)

The National Cancer Institute challenges which CANDLE addresses

- understanding the molecular basis of key Ras protein interactions



- developing predictive models for tumor response to drug treatments

- extracting information from cancer patient records to determine optimal cancer treatment strategies

CANDLE IS A DOE EXASCALE COMPUTING PROJECT APPLICATION

EXASCALE COMPUTING PROJECT

ECP's work encompasses the development of an entire exascale ecosystem: **<u>applications</u>**, system software, hardware technologies and architectures, along with critical workforce development.

Exascale Systems 2021 - 2022

ANL

Frontier

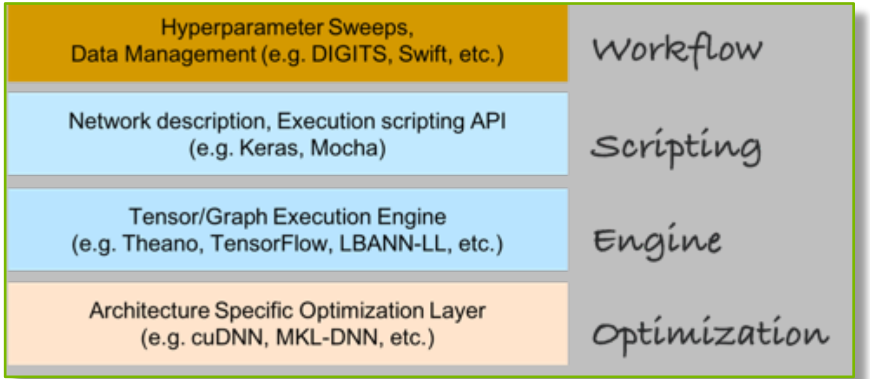
ORNL

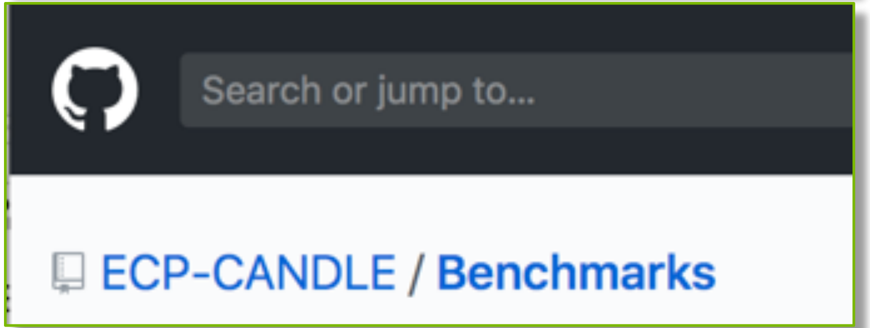
El Capitan

LLNL

PROGRESS

Updates since SC18, Dallas TX





- **Milestone 10** – CANDLE library released
- **Milestone 11** – New benchmarks released
- **Milestone 12** – CANDLE v0.2.0 released

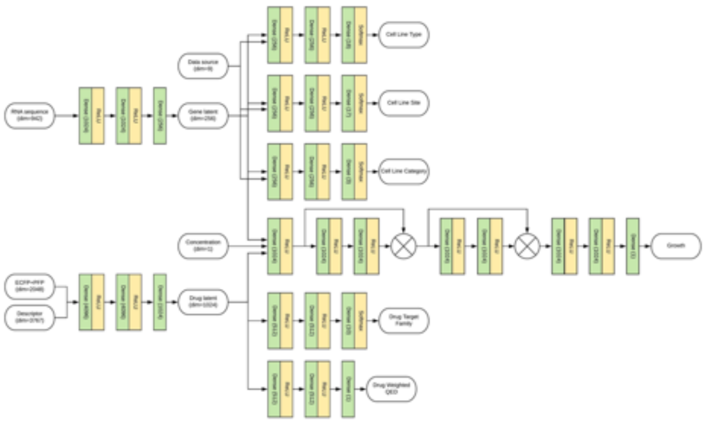
MULTI MODAL MULTI TASK DRUG RESPONSE MODEL

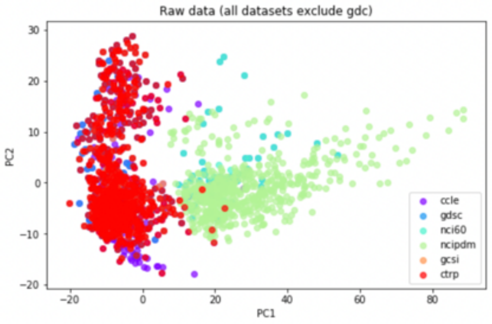
New CANDLE Benchmark

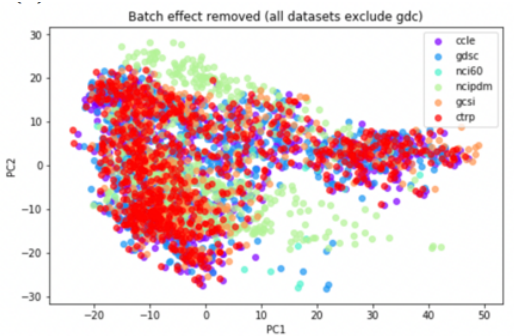
we proposed a new formulation of the dose response prediction problem that unified single and paired drug screening data

Data Source	# Tumor Samples	# Drugs	# Dose Response Samples	Treatment Type
NCI-ALMANAC	60	104	3,686,475	Drug pair
CCLE	504	24	93,251	Single drug
CTRPv2	887	544	6,171,005	Single drug
gCSI	409	16	58,094	Single drug
GDSC	1,075	249	1,894,212	Single drug
NCI	60	52,671	18,862,308	Single drug
GDC	11,081	N/A	N/A	N/A
NCI-PDM	1,198	12	518*	Single and paired drugs

* PDM drug response were measured differently from cell line dose response data.

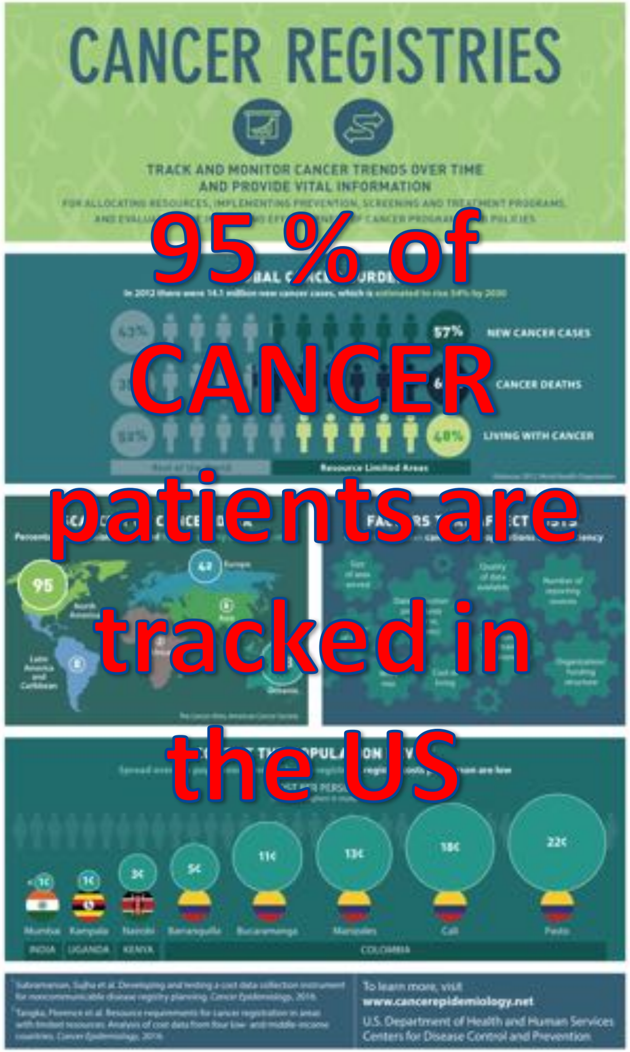






MODELING POPULATIONS

- Cancer registries present unique NLP challenges
- Pathology reports generated by hundreds of pathology labs and thousands of different pathologists
- Capturing accurate information for approximately
  - 60 cancer sites
  - 330 cancer topographies
  - 400 histological types
  - 9 histological grades
  - 5 laterality codes



MULTI-TASK HIERARCHICAL CNN WITH ATTENTION FOR INFORMATION EXTRACTION

New CANDLE Benchmark

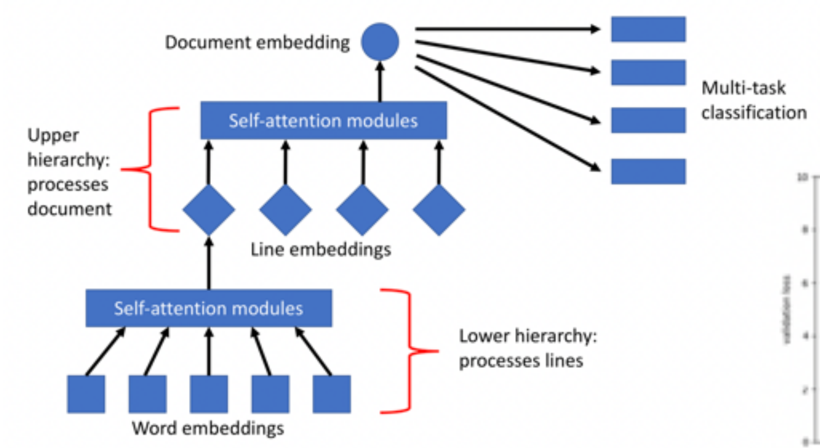
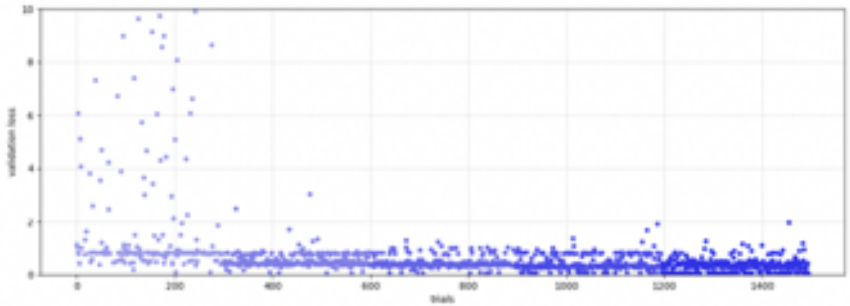


Figure 1: Multi-task hierarchical convolutional attention network (MT-HCAN) architecture



Clinical Task	Micro-F1 (MT-CNN)	Micro-F1 (MT-HCAN)	Macro-F1 (MT-CNN)	Macro-F1 (MT-HCAN)
Subsite	0.975	0.986	0.963	0.980
Laterality	0.966	0.981	0.965	0.980
Behavior	0.988	0.992	0.971	0.980
Grade	0.966	0.970	0.963	0.967

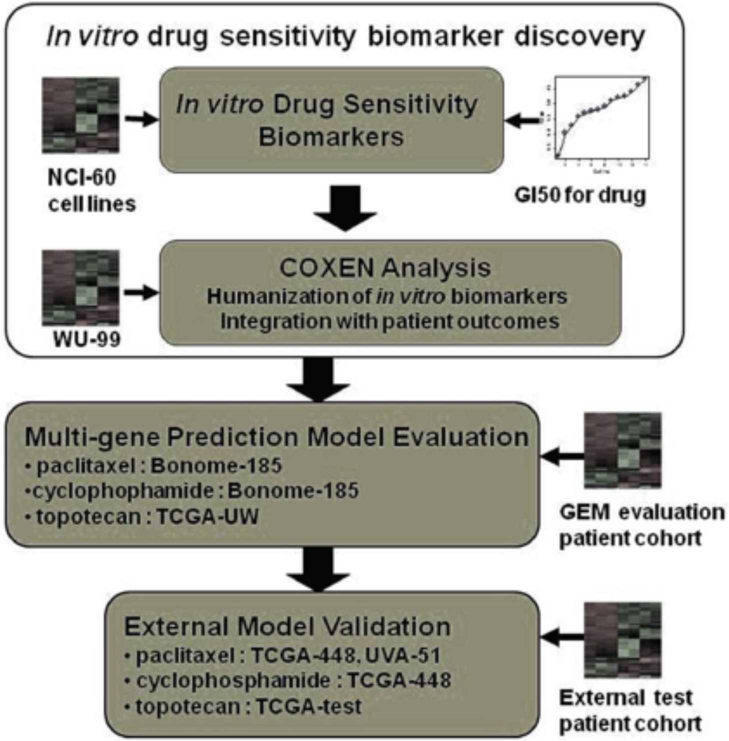
DRUG RESPONSE CROSS STUDY VALIDATION

Candle Demonstration, Release v0.2

- Train on one study, predict on the others
- Perform feature selection based on cross-correlation
- Use biological knowledge derived features as control
- Perform hyperparameter optimization for different feature sets
- Train N best models
- Infer on other studies
- Investigate model uncertainty

FEATURE SELECTION

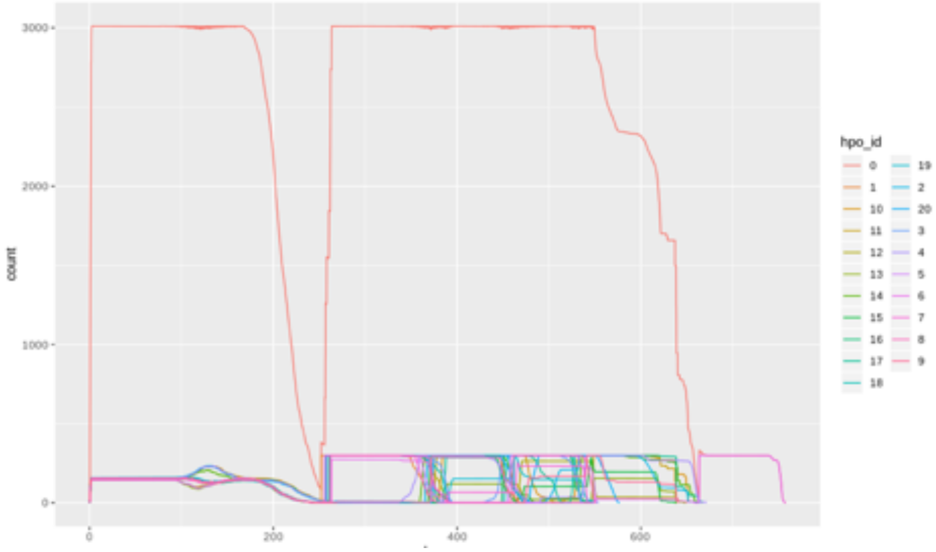
- CO-eXpression ExtrapolatioN (COXEN) algorithm
  - identify the genes whose expression in one study was related to drug sensitivity and then determined which of these genes maintained *concordant expression* in a second study.
- Lincs 1000
  - 1,000 landmarks genes sufficient to impute 82% of the remaining gene expression levels



HYPERPARAMETER OPTIMIZATION

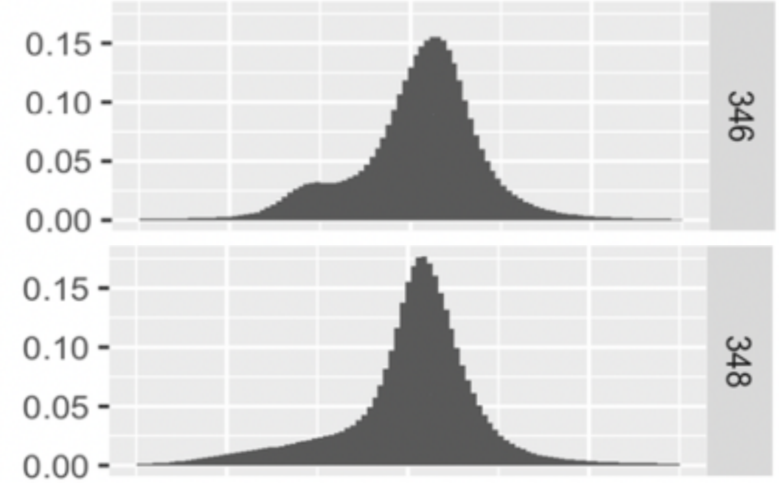
Concurrent experiments

- Perform 25 hyperparameter optimization experiments
  - 20 COXEN selected feature sets
  - F5 Lincs1000 selected feature sets
- Structural and learning parameters examined
  - network depth
  - network width
  - optimizers
  - activations
  - etc



INFERENCING

		Testing set				
		NCI60	CTRP	GDSC	CCLE	gCSI
Training set	NCI60	R2 = 0.81 MAE = 17.1	R2 = 0.38 MAE = 32.2	R2 = 0.24 MAE = 35.3	R2 = 0.48 MAE = 33.4	R2 = 0.46 MAE = 33.4
	CTRP	R2 = 0.44 MAE = 29.8	R2 = 0.68 MAE = 22.7	R2 = 0.23 MAE = 34.4	R2 = 0.61 MAE = 28.3	R2 = 0.60 MAE = 28.5
	GDSC	R2 = 0.32 MAE = 34.0	R2 = 0.25 MAE = 36.7	R2 = 0.53 MAE = 27.2	R2 = 0.50 MAE = 32.6	R2 = 0.60 MAE = 29.2
	CCLE	R2 = 0.27 MAE = 36.9	R2 = 0.20 MAE = 39.2	R2 = 0.11 MAE = 38.9	R2 = 0.68 MAE = 25.4	R2 = 0.39 MAE = 34.2
	gCSI	R2 = 0.00 MAE = 44.9	R2 = 0.11 MAE = 43.1	R2 = 0.05 MAE = 42.8	R2 = 0.33 MAE = 40.6	R2 = 0.80 MAE = 192



CROSS STUDY VALIDATION

- Best results to date
- Room for improvement

UNCERTAINTY QUANTIFICATION

- Normalized error for 450 models
- Exploring how to compare models

CANDLE DEMONSTRATION

Milestone 12 – CANDLE v0.2.0 released

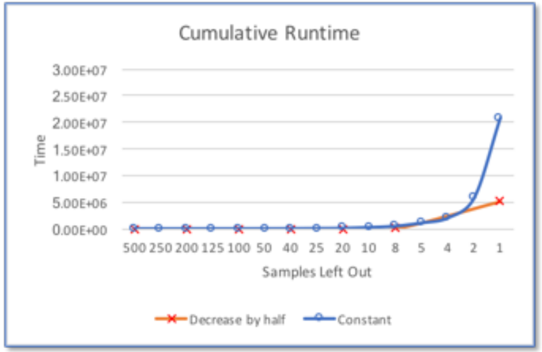
Stage	Quantity Description
Data Pre-Processing	20 Feature selected data sets, one control (LINCS 1000)
HPO Searches	25 HPOS (21,300 models searched)
Model Training	450 Models trained
Cross-Study Validation	450 Models screened

Summit	
Inferences/node	82,782,891
Inferences/hr/node	6,898,574
Inferences/sec/node	1,916
Nodes	450
Total Inferences	37,252,301,160
Total Inferences/hr	3,104,358,430

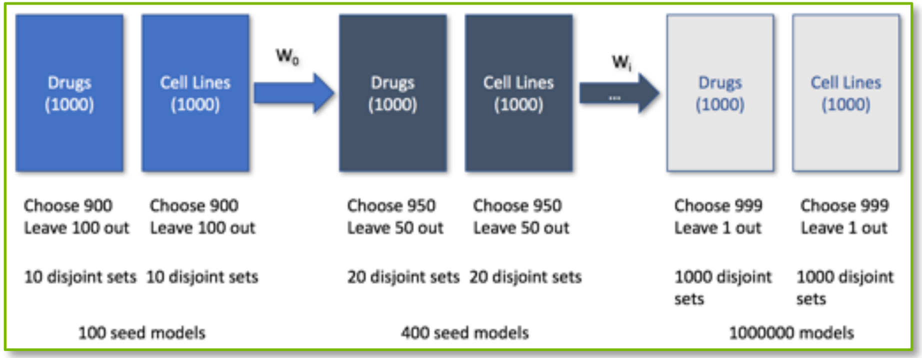
ACCELERATED DNN TRAINING METHODOLOGY

August 2019

- Develop accelerated training methodology
  - small portion of predictors will be trained from scratch
  - utilize pre-trained weights to speed-up model convergence for large number of models
  - explore recent advancements such as parameter sharing and pruning, and low-rank factorization



- Design learning rate schedulers
  - learning rate schedulers that allow models to converge when trained with pre-trained weights
- Explore DNN architectures
  - test pre-trained weights with various architectures that require long time to converge
  - include attention-based neural networks i.e. dot-product and multi-head attention



GITHUB AND FTP

- **ECP-CANDLE GitHub Organization:**
- https://github.com/ECP-CANDLE

- **ECP-CANDLE FTP Site:**
- The FTP site hosts all the public datasets for the benchmarks from three pilots
- http://ftp.mcs.anl.gov/pub/candle/public/

Thank You

Dear DLS speakers,

Thanks again for your coming presentation at the Deep Learning for Science workshop at ISC'19, June 20th, Frankfurt Germany.

We have posted a tentative program at https://dlonsc.github.io/, please let me know if you need to present at a different slot.

Best,
Zhao, Vali, Ian