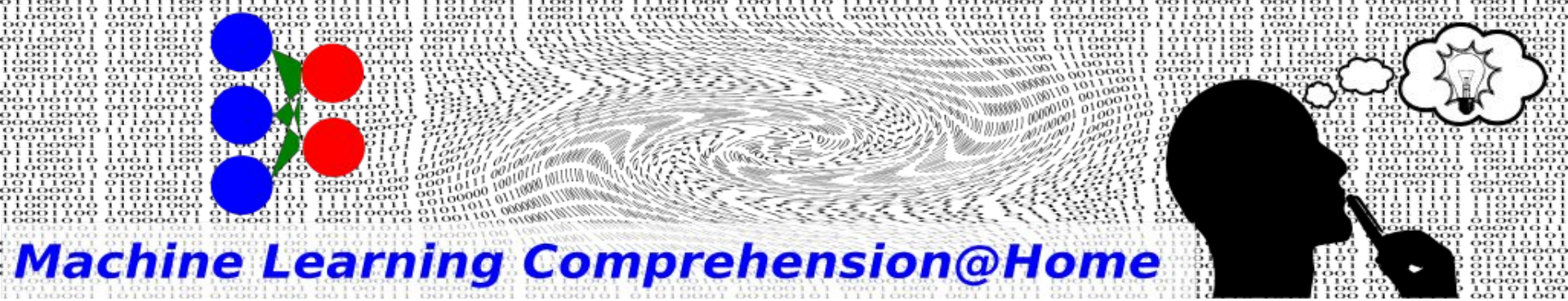


MLC@Home: A Distributed Platform for Studying and Understanding Neural Networks

2021 BOINC Workshop
Apr 14 2021

John Clemens <clemej1@umbc.edu>
Cognition, Robotics, and Learning (CORAL) Lab
University of Maryland, Baltimore County (UMBC)

UMBC

Motivation

Neural networks are ***powerful***, but ***complex***!
Analysis is hard:

- Many possible "right" answers
- GPT-3 has **175 billion** parameters

Network creation still more ***art*** than ***science***

Complexity can mask undesirable behavior:

IBM's "Watson for Oncology" cancelled after unsafe treatment recommendations
Tesla tricked into speeding by researchers using electrical tape
Amazon ditched AI recruiting tool that favored men for technical jobs
TrojanNet: Embedding Hidden Trojan Horse Models in Neural Networks

MLC@Home seeks to understand how neural networks encode knowledge, characterize a model's limits, and develop new methods to evaluate the effectiveness and fitness of a model.

XKCD #1838

MLC@Home

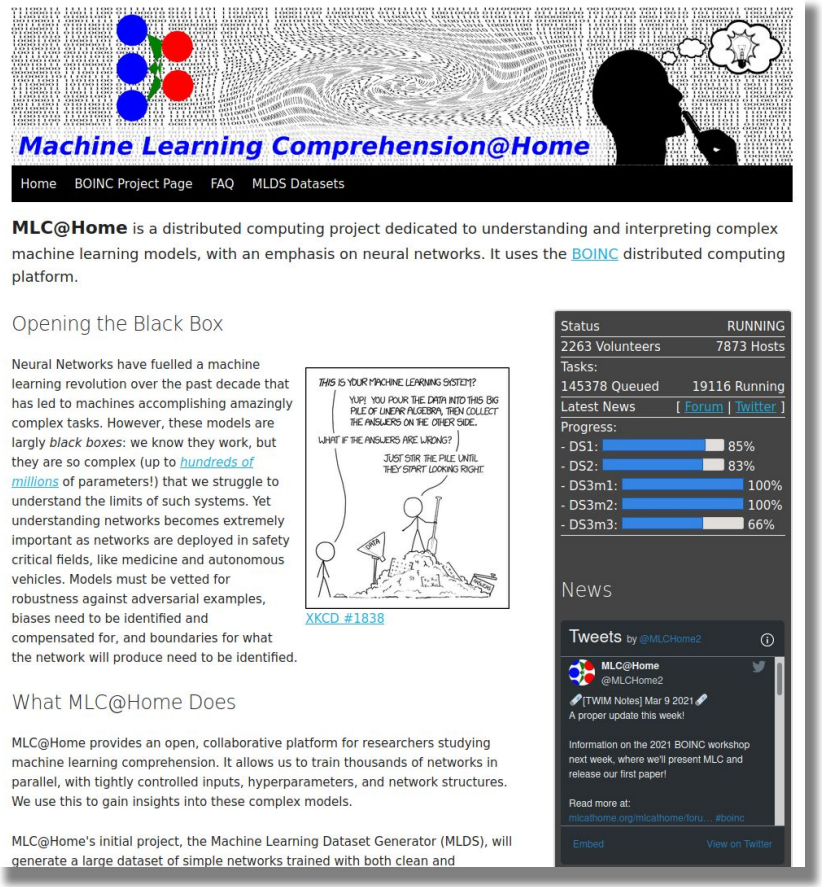
MLC@Home is an umbrella project which harnesses volunteer computing power from around the globe to support projects which study the *how* and *why* of neural networks. Project areas include:

- Model meta analysis
- Reproducibility studies
- Hyperparameter and Architecture search
- Neuro-Evolution

Out of scope: narrowly-applicable point solutions to specific problems, chasing SoTA, etc.

Current status:

- Live running project since July 2020
- 2200+ volunteers, 7800+ computers
- Linux, Windows, amd64, ARM, CUDA, and ROCm support
- Active community on Forums, Discord, and Twitter
- First project, MLDS, releasing results ***today!***



Machine Learning DataSet Generator (MLDS)

UMBC

To understand neural networks, you need *a lot* of examples. Then we can use those examples to do meta-analysis of the resulting models. MLC@Home's first project, ***MLDS***, generates this dataset.

- Volunteers compute thousands of neural networks
- Three different kinds of training data
- Questions:
  - Can we identify the dataset used to train a model?
  - How many examples are needed?
  - Are there differences between CPU and GPU-trained models?
- Capture training process metadata, as well as final output

Dataset 1: 42570/50000

Name	100	500	1000	5000	10000
SingleDirectMachine	100	500	1000	5000	10000
EightBitMachine	100	500	1000	5000	10000
SingleInvertMachine	100	500	1000	5000	10000
SimpleXORMachine	100	500	1000	5000	10000
ParityMachine	100	500	1000	2566	2566

Dataset 2: 41818/50000

Name	100	500	1000	5000	10000
ParityModified	100	500	1000	1583	1583
EightBitModified	100	500	1000	5000	10000
SimpleXORModified	100	500	1000	5000	10000
SingleDirectModified	100	500	1000	5000	10000
SingleInvertModified	100	500	1000	5000	10000

Dataset 3

Overall Completed: : 663581/757338

Milestone 1 (100x100) : COMPLETE (10000/10000)

Milestone 2 (1000x100) : COMPLETE (100000/100000)

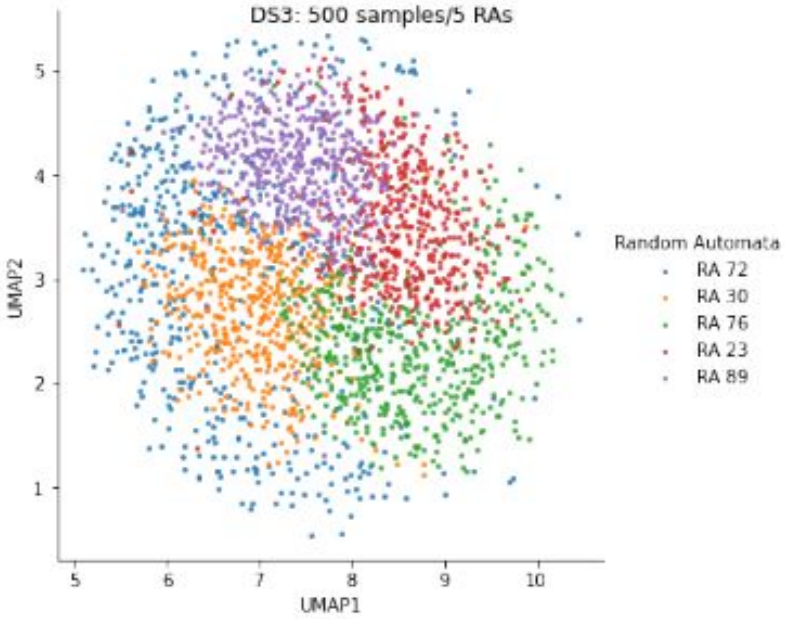
Milestone 3 (10000x100) :

6795	6820	6828	6456	6761	6513	6452	5783	6815	6240
5865	6525	6679	6828	5799	6416	6686	6783	6690	6776
6753	6793	6807	6823	6722	5482	6791	6662	6743	6802
6789	6699	6705	6776	6776	6451	6767	6515	6806	6754
6757	6819	6710	6703	6794	6784	6648	6799	6260	5944
6774	6823	6496	6641	6793	6770	6789	6613	6811	6797
6828	6799	6181	6342	6779	6802	6766	6749	6796	6817
6667	6791	6828	6798	6217	6813	6796	6804	6802	6802
6789	6821	5516	6829	6475	6725	6831	6191	6797	5973
6807	6812	6662	6641	6812	6601	6805	6795	6619	6452

0%	< 25%	25% - 75%	>75%	100%

To date, MLC@Home volunteers have generated over ***750,000*** example neural networks to enable this analysis. To our knowledge, this is ***by far*** the largest dataset of its type in the field.

MLDS Results



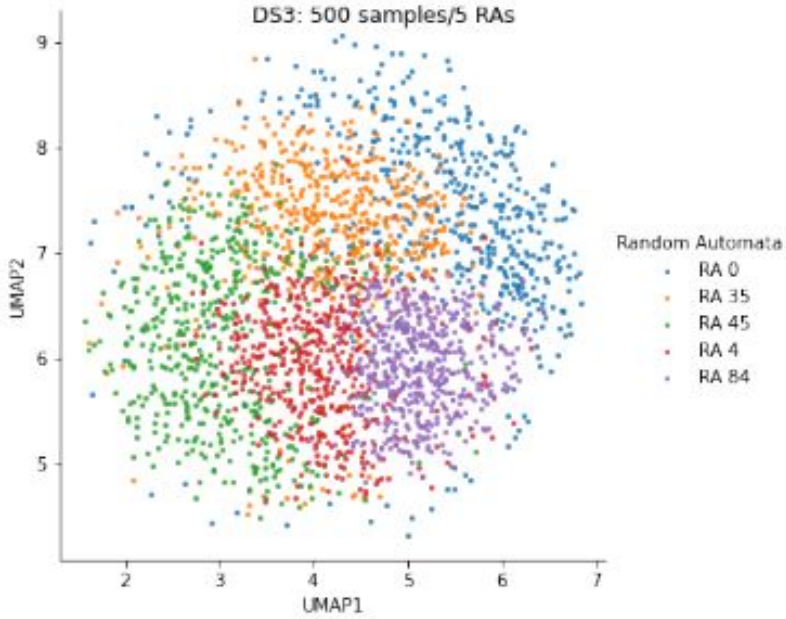


Figure 4: MLDS-DS3 UMAP weight space projections for 5 randomly-selected automata.

MLDS Results

Machine	SVM	DecisionTree	RandomForest	MLP	AdaBoost	NaiveBayes
EightBit	87%	78%	96%	91%	94%	99%
SingleDirect	96%	81%	95%	96%	99%	100%
SingleInvert	52%	49%	52%	53%	48%	55%
SimpleXOR	86%	85%	91%	85%	94%	99%
Parity	63%	55%	71%	66%	69%	61%

Table 2: Classifiers attempting to classify whether a given network is from of DS1 (normal) or DS2 (modified) for N=1000 samples of each, 800 training, 200 validation.

Future Plans

MLC@Home is here for the long term

Plenty of fun and interesting research questions to explore

MLDS

- Current workload to produce ***1.1 million*** networks
- MLDS has plans for CNN, fully connected, and Transformer examples, and more

Needs

- Scientific collaborators
  - Unique and powerful platform
  - Public dataset available ***today***!
- Technical
  - Data scientists / engineers
  - C++ developers
  - System administrators / moderators

Summary

If you want to help advance scientist's understanding of machine learning and artificial intelligence, make these technologies safer and easier to use in the real world, and participate in a dynamic and exciting field, join MLC@Home!

- ***Volunteer***
  - ***CPU and GPU equally useful!***
- ***Collaborate***
  - ***New experiments***
  - ***Technical Help***
- ***Community***
  - ***Home https://www.mlcathome.org/***
  - ***BOINC Forums***
  - ***Twitter @MLCHome2***
  - ***Discord https://discord.gg/ZX75hr27***
  - ***Email mlcathome2020@gmail.com***
  - ***Git https://gitlab.com/mlcathome/***

Thank you!

Questions?

***MLC@Home*: A Distributed Platform for Studying and Understanding Neural Networks**

John Clemens <clemej1@umbc.edu>
Cognition, Robotics, and Learning (CORAL) Lab
University of Maryland, Baltimore County (UMBC)

UMBC

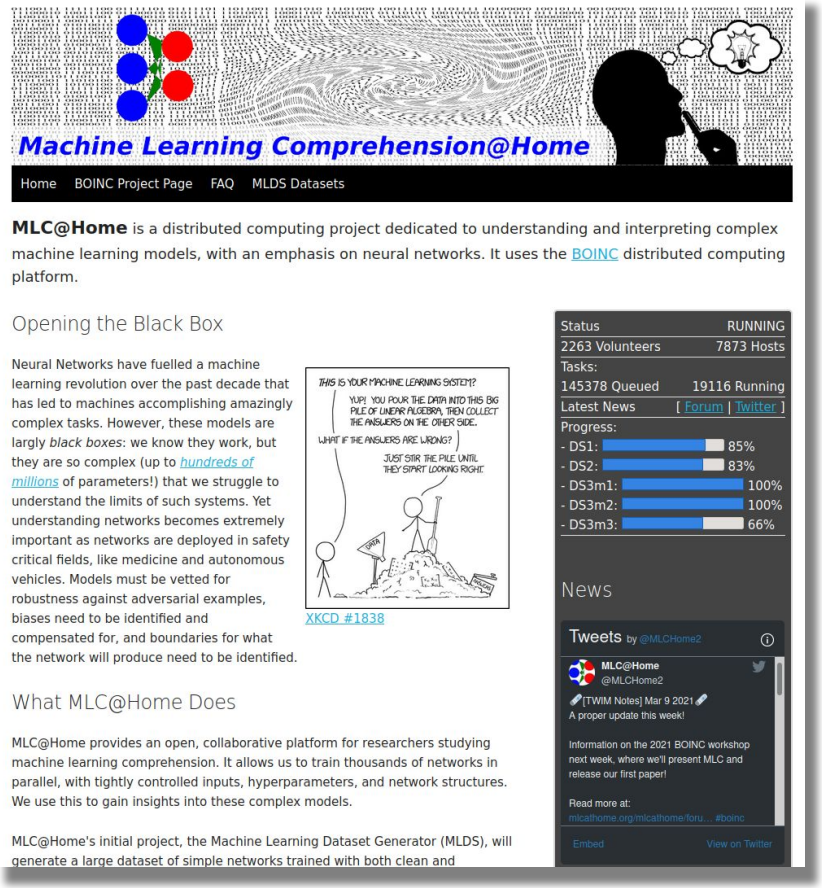
Backup Slides

MLC@Home and BOINC

BOINC provides a solid foundation to build MLC@Home, the community embraced the project from the beginning. Largely positive experience.

Issues:

- Each network trains in a variable amount of time, but BOINC expects WUs to take a similar amount of time
- Validation
  - There's not one "right" answer
  - Two networks that perform well on the training set might have massively different weights
- ML is highly environment dependent
  - VirtualBox means you can't use GPU
  - PyTorch didn't allow static compiling



MLDS Results