A unified approach to data science at NASA Goddard Space Flight Center

Mark Carroll, Dan Duffy, Craig Pelissier, Jules Kouatchou, Grey Nearing, Roger Gill, Glenn Tamkin, Jian Li, Thomas Maxwell, Mary Aronne, John Schnase

1 NASA Goddard Space Flight Center, Maryland, USA
2 Science Systems and Applications Incorporated, Maryland, USA
3 University of Alabama, Tuscaloosa, Alabama, USA
4 Arctic Slope Research Corporation

Presented by Liz Hoy NASA/GST, Inc.
Computational & Information Science and Technology Office (CISTO)
Purpose and Goal of the AI/ML CoE:
- The purposes of the CoE is to provide all necessary AI/ML resources needed by the SMD. The goal is to accelerate Science Discoveries through the use of AI/ML.

Scope
The of the AI/ML CoE is represented by this diagram.

- IT Resources
- Examples
- Training
- Partnership
- Support

Accelerating the discovery through the use of AI/ML, enhancing the competitiveness of science and engineering, and creating the next generation workforce for Goddard.
From AI to ML and DL

**ARTIFICIAL INTELLIGENCE**
Programs with the ability to learn and reason like humans

**MACHINE LEARNING**
Algorithms with the ability to learn without being explicitly programmed

**DEEP LEARNING**
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data
Machine Learning for Monsoon Prediction
Using Global Teleconnections
Thomas Maxwell, Jian Li, Thomas Favata

- Predict All-India Monsoon rainfall accumulation one year in advance using a two-layer neural network
- Inputs: First 16 global PCs of surface temperature & 500 mbar height (1 year lag time- August values)

Comparison of predicted to actual monsoon precipitation

Back-projected Activation Patterns
Surface Temperature
500 Mbar Height

Strong Monsoon
Weak Monsoon
Difference
MERRAMax automated variable selection

**MERRA weekly average trend - 490 predictors**
- Random Selection
  - MaxEnt run
    - 2.5 min 1 CPU
  - Batch of results from MaxEnt
    - 10,000 runs
  - Identify top ten percent of predictors
    - 3 batches
  - Identify top ten predictors
    - 10,000 runs

**MERRA monthly average trend - 113 predictors**
- Random Selection
  - MaxEnt run
    - 2.5 min 1 CPU
  - Batch of results from MaxEnt
    - 10,000 runs
  - Identify top ten percent of predictors
    - 3 batches
  - Identify top ten predictors

**Final run**
- Final map and ranking of variables in terms of overall importance
  - 42 hrs per dependent variable
  - 84 hrs total

**Random Selection**
- 3 batches
- 10,000 runs

**MaxEnt run**
- 2.5 min 1 CPU
- 10,000 runs

**Batch of results from MaxEnt**
- 10.5 hours 40 CPU’s
- 3 batches

**Identify top ten predictors**
- 10,000 runs

**Final map and ranking of variables in terms of overall importance**
- 31.5 hrs per dependent variable
- 63 hrs total

**MERRA automated variable selection**
- 42 hrs per dependent variable
- 84 hrs total
Random Forest for Lake Depth

30T Training Data: MSE = 0.41

30T Verification Data: MSE = 1.50
Multiple ML Comparison

Error

Algorithm
ADAPT 2.0 Planned Configuration

Public Access

Data Services
https, GIS, TDS, etc

Jupyterhub
remote vis

ADAPT VMs including
Machine Learning

NCCS User

App Catalog

HPC Allocation

Self-Provisioning

6 PB
Local

34 PB
CSS

Discover HPC
What’s next?

• Continue to offer platform as a service with Discover and ADAPT
  • Explore and test new hardware options to enhance the current hardware options – ADAPT 2.0
  • New GPU cluster with 20 4 GPU systems will be online by early 2020
• Increase data holdings at NCCS to facilitate discovery through AI/ML
  • ICESat-2 data are currently being added to the collection
  • Over 3.5 PB of Digital Globe data available
  • Data buy (Planet labs, expanded DG data) data are being stored and will be available soon
• Develop software collaboratively with scientists to maximize utility
• Training and outreach
Thanks!