WORKSHOP NOTES

Artificial Intelligence and the Accessibility and Analysis of Geospatial Data: A SCINet Workshop

Wooton Hall, Jornada Exp. Range ARS, 2995 Knox St, Las Cruces, NM
September 10-11, 2019

Tuesday, September 10

8:00	Sign In: Wooton Hall (enter thru front door at corner of Knox and Frenger)	
8:15	Opening Remarks: Dr. Deb Peters	
8:30	Participant Introductions – research area, experience with SCINet/HPC, experience with AI/ML; Workshop goals and products	
9:30	Geospatial successes on the HPC	Rowan Gaffney: Big Data & Machine Learning: Mapping Grassland Vegetation
9:50	Break	
10:10	Geospatial Challenges and Opportunities on the HPC	Dr. Alisa Coffin: "HPC systems and AI in the Long-Term Agroecosystem Research Network–status, challenges, and potential for network level modeling and geospatial research"
10:30		Dr. Dave Fleisher: "Mapping Crop Yields in the Northeastern Seaboard Region: There Must be an Easier Way!"
10:50		Dr. Scott Havens (remote presentation): "Challenges of spatial modeling in the cloud during the era of big data"
11:10		Dr. Feng Gao: "Large area crop phenology and water use mapping using satellite data: opportunities and challenges"
11:30	Working lunch: Common issues to be solved among geospatial ag problems for using the HPC	
1:00	SCINet Basics, Introduction to SCINet resources for geospatial data Dr. Andrew Severin and Jim Coyle, Iowa State University (zoom)	
2:00	Small groups: Identifying SCINet Issues for Geospatial Researchers	
3:00	Break	
3:15	Small Groups continue	
4:00	Report Outs from groups	
5:00	Poster session	
6:00	Adjourn – dinner on your own	

Wednesday, September 11

8:00	Opening Remarks and Summary of Day 1	
8:30	AI/ML in Geospatial Research	Dr. Laura Boucheron (NMSU): "From Rules to ML to DL" "Convolution Neural Networks: Basic Structure" "Flavors of DL" "Convolution Neural Networks: Epic Fails"
9:15		
10:00	Break	

10:30	AI/ML in Geospatial Research, continued	Dr. Dawn Browning (Jornada ARS): "Applications of ML in natural resources w/geospatial data"
11:00		Dr. Niall Hanan (NMSU): "Machine learning: friend and foe of geospatial and ecological science"
11:30	Discussion	
12:00	Lunch Break	
1:30	Small working groups (3): integrating ML/DL and the HPC potential and challenges for solving geospatial problems	
3:00	Break	
3:30	Presentations by working groups	
4:00	Development of a SCINet Geospatial Research Working Group: Goals, Roles & Responsibilities; outcomes and products	
5:30	Wrap-up, Closing Remarks and Collection of Participant Feedback	
6:00	Adjourn	

9/10/2019

8:15am - Deb Peters - Opening Remarks

Link to TOC

Workshop Goals

To create a geospatial working group for improving SCINet for geospatial researchers
To communicate researcher computational needs (training, software, etc)

Questions

Can you talk about the other SCINet workshops that have been going on?
Dawn: Phenology WG (August 2019), exploring options for overcoming computational bottlenecks
Adam Rivers, Gainesville FL, hands-on ML training
This workshop - geospatial SCINet needs + AI exposure
Beltsville AI Conference for RL, exposure to AI methods to inspire research ideas
There's also a new SCINet website under development where you'll be able see past and future opportunities, will share link when up and running

How do I get SCINet funding for a workshop or meeting?
The type of event is flexible but it has to have a SCINet component or focus. You will have to work with Deb to ensure the agenda is approved

9:30am - Geospatial Successes on the HPC - Rowan Gaffney: Big Data & ML: Mapping Grassland Vegetation

Link to TOC

Working with NEON Hyperspectral Data
Uses Jupyter Lab on SCINet for processing large data

File format that is helpful on HPC/cloud computing: Zarr, netcdf is transitioning to using zarr under the hood
Parallelizing code using Dask. Can build a cluster and visualize the workers processing on a dashboard - helpful for seeing if the cluster is working properly and also see how long the processing will take
For python programmers working with gridded data the xarray package is very helpful. This package helps you hold onto metadata, for example dimension names, sizes, data units, more
For machine learning: SciKit Learn python package, Support Vector Machine model

Questions
How do you integrate process physical based models with ML techniques that don't have any underlying physics or biology mechanisms?
How does computing on Ceres work with clusters and nodes and submitting a compute job?
How did you get your data onto SCINet? Globus, ~1Tb took less than 1 hr

10:10am - Geospatial Challenges & Opportunities on the HPC - Alisa Coffin: HPC systems and AI in the Long-Term Agroecosystem Research Network–status, challenges, and potential for network level modeling and geospatial research
Link to TOC

LTAR network of 18 sites looking at "food for the future: understanding and enhancing the sustainability of agriculture"
Focuses on sustainable intensification of Ag, integrating question driven research projects with common measurements on multiple ecosystems, coordinating research across scales/sites
Working on developing network data management capabilities
Past: Working on local storage, local machines including lab servers
Challenge: how to do network level computational research, there are many labs, some linkages, some have no geospatial capability, LTAR network not fully connected yet, HPC system can be critical for better connecting LTAR
Recent Developments:
- Integration through Communication, data harmonization, data sharing
- Cooperation through clearer leadership, proj management thru coordinated working groups, identification of network level research questions (phenology, regionalization-understanding the regions that the sites represent for modeling purposes, manuresheds, etc)

Next Steps:many labs, many linkages, more have geospatial capability, network more connected. But what about network computing?
Computational Needs (every site requires the ability to):
- Harmonize data from multiple sites and contribute it to the network
- Share lrg research data files that aren't publicly accessible (flat, DBs, metadata, gridded)
- Access common pool resources (software, baseline data)
- Communicate easily across distances (zoom)
- Data provenance - tracking changes
- Collaborate on model and code development in real time

Explore and visualize results of very large datasets
Securely store and rapidly access stored data

Challenges for the coming months/years:

Dev network computing resources with LTAR
Leadership needs to clarify data sharing policies in harmony with USDA
What's the minimum computing "standard" needed for sites to work in the network? (people, expertise)
Dev clear picture of needs and assets of each site wrt connectivity, storage, expertise
Building expertise and capacity for using SCINet/HPC
Documentation procedures for data, methods clarified
Networked visualization of very large datasets
Instant and easy visual and voice communication

Vision for the future (5 years from now):

LTAR experiments will have published results
Finalized geospatial datasets
Common measurement and automated routines for updating DBs quickly
Working datasets that are easily accessible to researchers
LTAR using HPC systems regularly

There are Tutorials for using SCINet and GEE in the Remote Sensing & GIS Working Group Basecamp Docs&Files

Questions

Between the LTAR sites are there common measurements and experiments? Yes, but the challenge is integrating the data from multiple sites into a larger framework that can be shared throughout the network

10:30am - Geospatial Challenges & Opportunities on the HPC - Dave Fleisher: Mapping Crop Yields in the Northeastern Seaboard Region: There Must be an Easier Way!

Link to TOC

Scaling up point models or model intercomparisons

US Northeastern Seaboard Region modeling study (food security based)

Imports 65-80%% of the fresh fruits and veg
Multiple food security concerns
Can re-regionalizing the food system in the area address the food security concerns?

Quantify current and potential production capacity by using process based crop and soil models integrated with multiple geospatial databases available on the public domain

Producing geospatial yield maps to show what can be grown where

Crop/Soil Modeling: inputs meteorological vars, soil info, cultivar and management parameters, outputs - yield, model processes - crop growth, soil processes, etc.

The plant model is point based, the soil model is 2 dimensional

Running the models many many times (more than 10000 times, is computationally heavy)

Moving forward - want to look at adaptation responses (shifts in planting dates, production on marginal land, land use re-allocation based on optimizing crop and climate interactions) will require more modeling
Using 5 different computers - each computer has all of the input data and they had to manually manage what computer was running which simulation in a spreadsheet. Each model run takes 2-3 minutes
Challenges:

- expertise/domain knowledge (how to access, do I require HPC, how to revise scripts for parallel computing)
- intimidation/unfamiliarity (will learning this be efficient use of my time, 'language' barrier)
- Rapid changes in technology (it took me a year to learn it and the system changed!, backwards compatibility)

Workflow summary

- Common types of input data (soil, weather, climate, veg, etc)
- Prepare the data for model runs
- Run the model
- Analyze the model output

Simplifying the learning curve for using HPC, reduction of compute time could enable research on many more science questions

Questions/Comments
Reproducing the modeling workflow is important - are singularity or docker containers available on SCINet? Yes, this can also help with backward compatibiliy concerns
Not all of our modeling processes are so simple, often there are serious issues with pre-processing of the input data and integration issues when we're working with data from different sites

10:50 - Geospatial Challenges & Opportunities on the HPC - Scott Havens: Challenges of spatial modeling in the cloud during the era of big data
Link to TOC

Water supply forecasting in the western US for supply management by water managers in CA
Stakeholders CADWR, NRCS, US Bureau of Reclamation, CA water management agencies
Model: iSnobal; 54k square kilometers, 21+million grids
Model input: HRRR 3km hourly
Input data size for WY2017 = 50TB and will only grow
It used to take a couple days to run 1 year, but now down to a couple of hours on HPC
They are fully automated: input data pre-processing, model runs, post-processing of output
Using Docker for portability and reproducibility - any user can replicate the model results and publication results
Scott was an original tester of Ceres, but data was too big because required 500TB and 6 solid weeks of computing
Shortcomings of HPC environment:

- Shared resource, the queue is a problem when near real time model results are required

Docker wasn't supported before on Ceres
Need public access to the data
Felt that Ceres was Meant for 1-time jobs that are brief, not projects like theirs that run all of the time

They determined their project was better suited for Amazon Web Services - where they are hosting their model results

Geoserver (GIS) for model results (stakeholders are building web apps based on data from the geoserver)
S3 bucket linked to their website for snowpack summary reports

Cloud environment is providing their project needs:

Infinite and on-demand resources
Built for docker
Public access
Resources 100% of the time
HPCs can be built on the cloud

Take Aways:

AWS cloud computing meeting their project needs, but SCINet/Ceres cannot
Your stakeholders don't need to know much about HPC systems to use AWS cloud computing

How do you get the massive input data into the computational environment?

Requirements:

1 basin, 1 year = 1.5 TB/year model output
30 year forecasts (50TB/basin)
All 5 basins = 250TB for a single 40 year run
Ensemble runs (say 100) to address uncertainty for 5 basins = 25Pb → only the cloud can handle this → cloud cost would be 150k/month- not doable

Have to rethink how they use/access/store large spatial datasets

Stop the store-it-all mindset
The cost of running the model is almost nothing compared to storage
Input data access → using THREDDS data server for netcdf for allowing multiple connections to data file

Questions

What's the current AWS cost? 150k/month but that doesn't include the compute costs, they've got their own in house system

Where does the THREDDS server run and is it a good option? Their files are relatively small so they can run the server locally. When you access files through THREDDS, you are accessing it not transferring the data. The data is transferred on demand when your code calls for it and you can call for just a small subset of a large data file. Also many people can be accessing the same file simultaneously.

11:10 - Geospatial Challenges & Opportunities on the HPC - Feng Gao: Large area crop phenology and water use mapping using satellite data: opportunities and challenges

Link to TOC

Crop water use and phenology mapping
Multi-Source Agricultural Monitoring
    Many sources of input data
    Pre-processing (reconciling different spatial and temporal resolutions of input data)
    Model runs
    Analysis of model output
Near real time crop phenology mapping using high temporal and spatial resolution VENUS data over BARC (2019)
He's taking 16 day MODIS products and creating daily data "data fusion"
Detecting green up dates
Application to variable irrigation
Opportunities:
    methods/algorithm are becoming mature, employed over multiple LTAR sites, moving from research to operational
    High temporal/spatial resolution
    High performance computing
Use of Google Earth Engine (GEE)
    Evaluating yield variability of corn and soybean using landsat-8, sentinel-2, and MODIS in GEE - all this data already exists in GEE, don't need to move data, only write a small Script. This wouldn't be possible on the lab server due to data size
    Monitoring water demand and Use: OpenET - using GEE again
Challenges:
    Data Storage - daily 30m res data - 1 layer, 1 year, 1 variable = 10TB (large input data)
    Data transfer - would need to download from NASA/USGS/NOAA to lab server then to Ceres? Can we go from the agencies directly to SCINet?
    Product distribution - long term data archive and distributions (to ag data commons or other repository) (analysis output data is smaller)
    Personnel - need help from multi-disciplinary background (comp sci, GIS, Remote Sensing, agro-informatics, agronomy/ecology/geography) - he could use a postdoc to port their analysis over to SCINet/Ceres and automate/parallelize

Questions
Have you used the AgRee tools that can crop mask Landsat8 and Sentinel2 data? Feng hasn't used this
What does your lab server look like? 20 nodes
Are you trying to predict yields? No, just filling in data gaps to capture variability

1pm - SCINet Basics, Intro to SCINet Resources for Geospatial Data - Andrew Severin and Jim Coyle
Link to TOC

Andrew Severin VRSC
SCINet = VRSC, high speed network, high-performance computer

VRSC - Virtual Research Support Core - manage Ceres, install software, troubleshoot software issues, Manage Rstudio and Jupyter Notebooks, develop best practices and tutorials for computing
For bioinformatics they have developed a “bioinformatics workbook” that steps people through some analysis processes
Enable researchers to translate big data into informative data
Capture the collective knowledge of the USDA and connect those with the knowledge to those that need it
We should all be thinking about how we can contribute/share our knowledge to the benefit of others at ARS
Ceres is the name of the high-performance computer
What is an HPC cluster?
    - Collection of multiple separate servers/computers - called nodes with multiple computing cores
    - Computing on 1 core on SCINet may not be faster than computing on 1 core on a newish laptop
    - The power is in using multiple cores - parallel computing
Types of nodes: login, data transfer, compute
Ceres also has storage
Currently: 65 community nodes with 40 cores, 2 newer nodes, with 80 cores and high RAM, private nodes include a GPU node
There is a job scheduler with Queues for computing on Ceres: queues include brief for less than 2 hrs, short for less than 2 days, medium for less than 7 days, and many more
To see the programs/software that are installed - type module avail at the command line once you’ve logged into SCINet
Software of interest that are on Ceres now: RStudio server, Jupyter Notebooks, ENVI/IDL 1 license
VRSC could install an ArcGIS server on Ceres as well- you can use your local ArcGIS desktop and connect it to the remote server on Ceres - this way you can have all your data on Ceres in one place and process data through the GUI interface - in order to be truly effective on Ceres would need to use certain python services/packages that enable ArcGIS to run in parallel
On basecamp there are a lot of documents on how to work on SCINet (example, how to use RStudio, Jupyter Notebook)
For help contact scinet_vrsc@iastate.edu
Containers - singularity ← docker
    - Using singularity to avoid security issues that come with docker
    - Creates a static environment for running programs, can export/import this environment to other computers/operating systems so that you can run your codes in the same environment
When would you use SCINet/Ceres?
    - If you have large datasets
    - Sharing data within the same group
    - Collab on a project
    - Data integration from multiple researchers

Build a container/environment for running codes

Project management is an important aspect when your projects get big or there are many collaborators - Andrew has ideas for project management located at www.bioinformaticsworkbook.org "Introduction to Project Management"

Ceres not meant for long-term storage, but long-term storage does exist on SCINet

In your Ceres Home Directory - smaller quota - request a project directory for larger space

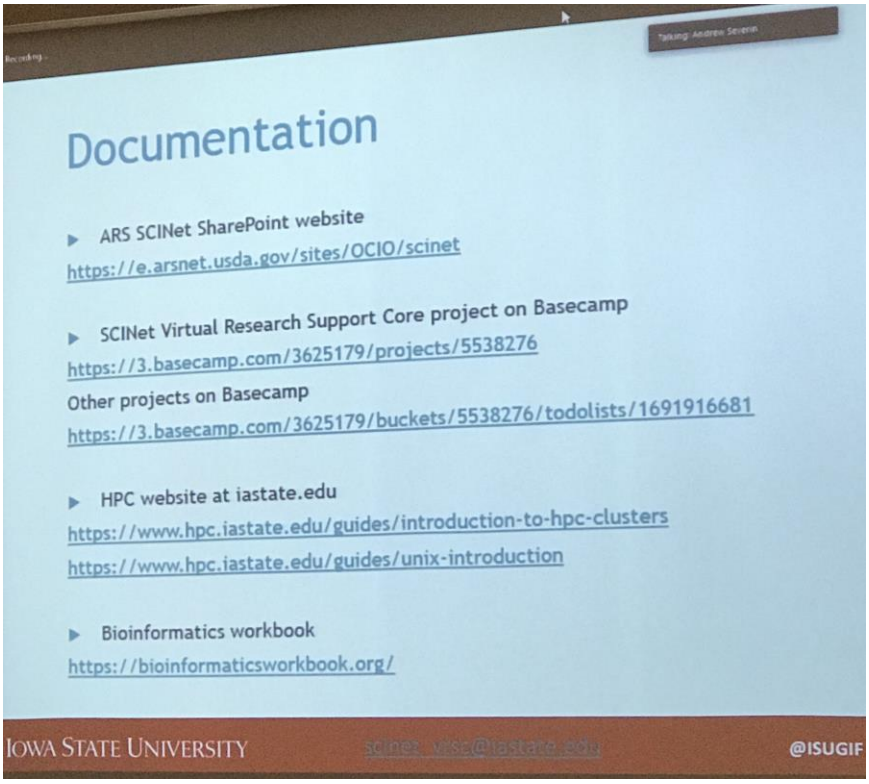
Only 1/10th of a project space is backed up on Ceres, users should back up important data elsewhere off of Ceres

Data transfer

Large data transfer for high speed sites - globus, ftp, more - see basecamp docs

For slow sites - physically send iowa state a hard drive of your large data until there is better connectivity for more sites

SCINet Information can be found at:



Don't post to basecamp your individual user issues, instead email the VRSC scinet_vrsc@iastate.edu

Roadblocks identified by the workshop participants are mostly about the learning curve

Use basecamp resources and VRSC help to get going on SCINet/Ceres

Getting software onto Ceres:

for things that require a license or are universally useful go thru the software request process that requires approval. Example ArcGIS server

The other way is to install software yourself in your project directory if it's only going to be you/your group using.

Experimental Design questions: how do I approach a certain type of analysis

Post to basecamp

Workflow exposure/parallelization/tutorials

like the bioinformatics workbook, there could be a similar resource for the geospatial Community

Questions/Comments

Are we able to push and pull from Github on SCINet? SCINet is a lower security network than ARSnet, so yes

How to keep track of all the model versions?
?

How to access log files on ArcGIS server vs desktop?
?

How to overcome the project management issue of not being able to access the GIS user accounts of our technicians? How to manage projects so that if a technician leaves, we don't lose all of their work?
?

What about public facing web hosting on SCINet? The snow group has an AWS site, the other option you would need a globus end point. Ceres isn't really a mechanism for serving data

Tifton recently set up a Next (?) server which is a cloud service and connect to local ArcGIS computing (Alisa).

Bruce has AgCROS in the Azure Cloud which will have an image server, geoevents server, and more. It will not require eAuthenticate. How to integrate AgCROS with SCINet/Ceres?
?

2pm - Small Group Breakouts

What is the issue? Do we all have unique solutions or is there a common solution? THink about short, mid, and long-term goals

Groups

1. How to deal with Large input/output files - should there be a library somewhere, etc.?
2. How to deal with the storage issues, not long-term storage but "longer term" storage
3. Products for stakeholders - what would we have to do to have an outward facing part of SCINet. Box, AWS, etc.
4. Vision Group - Workflow development - we're talking about Ceres and Cloud Computing - what's the workflow and are there other things we should be thinking about.
5. What this group needs (further training specifics for example) to be able to use SCINet/Ceres? Practical next steps for training/workshops

4pm - Report Outs from Breakout Groups

BREAKOUT GROUP - Large input/output files

Need: The Remote Sensing and GIS community are requesting a repository/library of commonly used data on Ceres/SciNet. This will reduce the duplication of popular datasets on Scinet as well as reduce the barrier of adoption for many in the spatial community.

Recommendations: Common datasets can be outlined by the RS/GIS working group, but will likely include continental or global scale data from Landsat, MODIS, Sentinel, PRISM, SSurgo, Polaris, etc… In building the centralized library on Scinet, the following aspects need to be explored:

1. What is the best method for serving large data to users – flat files or an imagery server (ie Thredds, OpenDAP, ESRI Imagery server, GeoServer)? The file type or server should meet the following criteria:
    a. High IO for distributed reads
    b. Able to serve data to a wide variety of platforms (R, Python, GDAL, ESRI, etc…)
2. Who will build and maintain (update with new data) the library? We suspect this effort will be too large to outsource to the wider ARS community, and will need a designated person to build and manage/update.

An additional aspect we would like to explore is setting up I2 connections to the NASA DAAC data repository network. If possible, this would allow access (via http protocol or OpenDAP) to a massive collection of valuable spatial data.

BREAKOUT GROUP - Data Storage

1. Talked about the beginning of Scinet and the purpose.
2. We talked about the learning curve and most scientist don't use Scinet
3. Should Scinet store data 2 months to 6 months then permeant storeage
4. Should Scinet only store datasets that will be used by more than one scientist more than one time

5. Who is responsible for the data on the NAS at each hub
6. From a research stand point when will data be placed in an archival location and who makes that decision?
7. Should data be stored on Scinet for the length of time as a project plan for 5yrs.
8. Does Scinet have a project tool that allows a scientist to view the time limit on their data and when it has to be removed?

BREAKOUT GROUP - Practical Next Steps for Computing Using HPC

Computational needs of the group:

image processing and how to make analysis totally reproducible
collaborative infrastructure for reproducible science, on the edge of needing HPC, upcoming process intense techniques
Physical process based modeling

AgCROS/SCINet integration: how to download large data from AgCROS directly to SCINet
process intensive techniques for predictive disease

There are a handful of specific projects that could use help porting to SCInet

Envisioning of future hands-on trainings

1) Data Carpentry or similar style training
    a) Logging in
    b) Command line comfort
    c) Building containers for reproducibility
    d) File Mgmt issues
    e) Parallel processing
        i) How to use SLURM
        ii) How to set up input scripts and programs
        iii) software for optimization of codes to run in parallel (i.e. DASK)
        iv) Accelerating the speed of science! Parallelization and embarassingly parallel
    f) Work through (e) with examples together with geospatial data
        i) Point data with thousands of points - serial or sequential execution
        ii) Executables from various operating systems
        iii) Spatial analysis - imagery over time or fused with point data
        iv) Large homogeneous data from sensors
        v) Smaller data where processing time is taxing RAM
2) Hands-on Training on Reproducibility and Collaboration Tools
    a) Docker/singularity contained environments for running programs
    b) Git for version control
    c) GitHub for collaborative scientific programming with provenance for version control
3) Repeat of the AI Training that Adam just held in Gainesville (AI, ML and DL)
    a) What are they and when to use them?
    b) What are the techniques and when would you use them
    c) What's on SciNET?

What is keeping you from taking steps to using SciNET

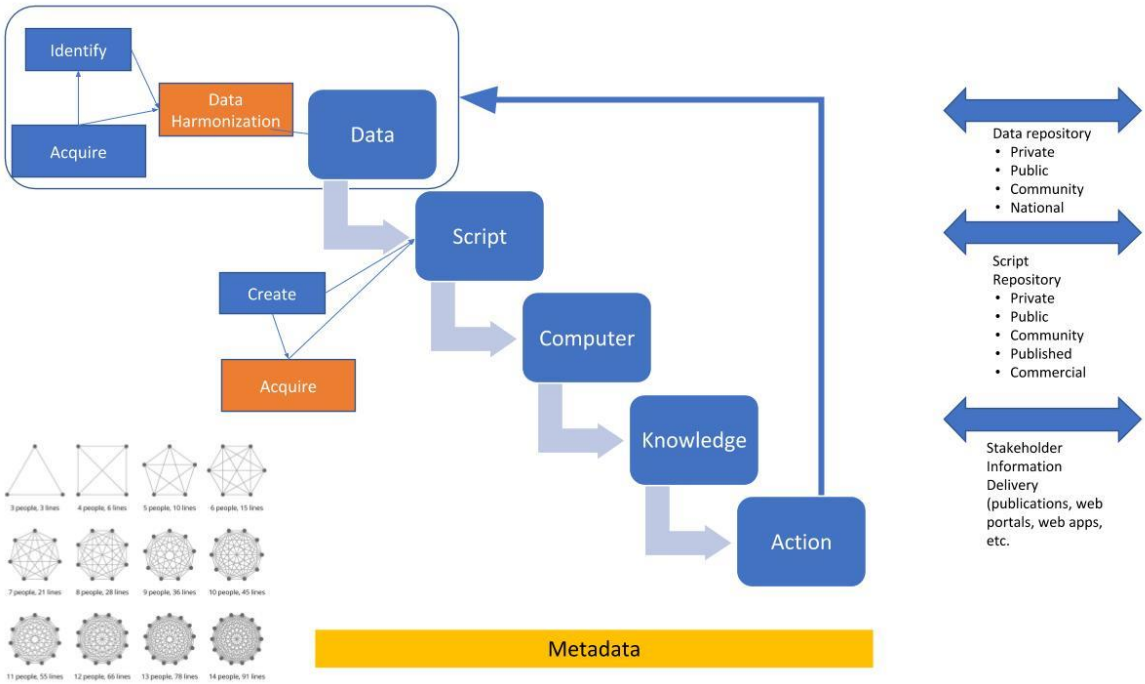
1. When is the time to migrate?
    1. Do you just have a feeling you can be doing things differently?
        1. More data than Excel can handle or complexity of the models
            1. Bird abundance simulation models (processing is intense, although data are small)
            2. Cattle tracking data
            3. Sensor network data
        2. Create library of datasets from other networks to utilize data
        3. Need shared computing space for cross-site collaborations

4. Want collaborative cyber-infrastructure environment for conducting reproducible science
5. Working with Big Databases from AgCROS to get a better understanding of how to approach working with a very large heterogeneous data sets could work on SciNET
    1. Download directly fom AgCROS to SciNET
    2. Instrumented watershed data can it be loaded to AgCROS?

2. How to make a request for software, and know are my tools on SciNET?
3. Use as a means to become more familiar with the SCInet framework and learn how to collaborate and manage a project on it, so that when you do have the modeling needs with HPC, we will be in a better place for using Ceres and working within the SCInet framework
4. Helpout with making containers or other data management tools, if you don't need HPC
5. Linux vs. windows based model - Ask that the compilers are available for your code (VRSC)
6. Library of software and compilers are available
7. Where are the best HPC resources that are searchable - any university….MI, FL, Iowa
8. module avail at command line

Existing resources at other agencies that we might look to:
Cyber Carpentry Training through NSF - reproducibility and workflow
Here's the link to the course's github page
Also check out NSF XSEDE

BREAKOUT GROUP - vision



BREAKOUT GROUP - stakeholders

Possible products and outreach for Stake-holders using ARS-Sci-net. 9-10-2019

As prepared by Dan Long, Merle Vigil, Jorge DelGado and Anapalli Saseendran

Most of the following potential research/technology outcomes are related to satellite remotely sensed data that would be used in decision support by farmers, farm managers, university extension professionals, NRCS and Agricultural consultants on a regional and national scale.

1. Change detection: identification of anomalies in production fields caused by diseases, insects, weeds (including glyphosate resistant weeds), and other factors within the growing season and among seasons.
2. Identify source areas for tumble weed infestations (rights of way, fence rows, borrow pits, etc.) and provide recommendations for control while the weeds are young to minimize spread.
3. Monitor spread of new pest wheat stem sawfly across wheat growing areas that has had a change in its ecology …. and other pests that may have moved to new areas due to climate change.
4. Monitor crops for drought and other stresses using an NDVI trend analysis approach.
5. Development of decision support products for planting decisions and crop rotations including cover crops based on GDD, ET, annual precipitation and weed management.
6. Yellowness index to identify canola production acreage and growing regions for procurement by oil crushing plants.
7. Develop regional expectations for P index and N index and for their loss to the environment through water erosion and run-off, based on university soil tests, weather, soil type, manure or nutrient application and NDVI.
8. Identify best locations for implementation of conservation practices that include buffer strips, terracing, and other approaches of precision conservation.
9. Growth analysis of crop development using low altitude, high resolution imagery and photogrammetric methods.
10. Near surface air flow inversion modeling and prediction of cold air pockets (winter kill, frost damage) real time prediction. Sulfanyl urea damage prediction.
11. Bioinformatics of soil microbial communities in relation to soil management as affected by landscape position, hydrology soil type and management (My-Philo DB).

    Side bar ideas:

    A. Supplement ag industry interest in proprietary farm data collection (tractor fuel use, hybrid yield, on the go yields all going to the cloud) by providing real data on environmental impacts on soil carbon, soil erosion, N leaching, and other ecosystem services. The importance of the applications listed above lies in calculation of value of carbon markets, 3rd party certification of good conservation practices, conservation programs, and water quality market credits.

    B. Ag-CROS agricultural Conservation research outcome systems, Data preservation, for posterity and future analysis, legacy data sets.

C. Using broadband capabilities to promote teamwork across ARS locations including those on main campus near LTARs and remote places, increasing geographic accessibility and data sharing, providing for an ability to work in a virtual environment, and enhancing possibilities for involving farmers in true interdisciplinary research and linking learning groups of farmers across regions.

Recommendations:

1. Help with funding of high-speed fiber for connecting remote locations to SciNet that would need it in cross locational bioinformatics and geoinformatics research.
2. Training on how to manage long-term data sets and archival of data of retired scientists that have historical significance.
3. Training on how to access Sci-net.

9/11/2019

8:30am - Laura Boucheron (NMSU)

Link to TOC

From Rules to ML to DL

How might humans explain the difference between hand written digits?

- Number of line segments, but this gets complicated quickly due to the variation how people write digits
- rule-based learning- leverage human to provide labeled training data - ground truth
- Leverage human to work with specific examples- select features that are expected to be discriminatory - feature space
- Leverage human to discriminate between digits - decision boundary

How can we better leverage the computer to do this?

Classical ML

- Leverage human for labeled training data - ground truth
- Leverage human to work with specific examples - feature space
- Leverage computer for the decision boundary

Supervised Classification

- The computer draws the boundary between classifications given the human ground truth and feature space
- Is the feature space not discriptive enough?
- Is the decision boundary not appropriate for the space?
- Is there not enough training data?
- Are they just difficult samples to classify?

You want to be careful not to overfit ML classification because the solution may then only be very specific to the training dataset

Deep Learning - feature extraction and classification

Human ground truth
Computer feature space - the computer decides what the discriminatory features are
Computer decision boundary

Neural network is a deep learning technique

Convolution Neural Networks (CNN): Basic Structure

Convolution means filtering of a signal
The filter slides across the raster image pixels and comes up with pixel weights
The filter is used multiple times, “layers”
The first layer in almost any image processing is edge recognition
Second layer can combine edges from layer 1 to recognize corners, circles, shapes
Third layer can combine shapes and learn to represent more complicated structures
Pooling layers - reduce the spatial res via subsampling
In the CNN example shown she uses multiple convolution layers, then maxpool
Activations -
Must define a loss function - backpropagation- rectifying the predicted answer with the expected answer

Flavors of Deep Learning (not comprehensive)

Spatial (spatiospectral) Image Classification/Regression
    The CNN example
Object Detection
    2or3D image in
    Bounding box with label/confidence
    Region based CNNs
Image Segmentation
    2or3D image in
    Image of delineated objects with labels
    Mask R-CNN
Image Translation - given a number of features/description, generates the image, inferring better images from “cheaper” data
    2or3D image in
    2or3D image out
    Paired images ground truth
    Adversarial networks
Temporal Classification/Regression
    Vector input (temporal)
    Discrete class label (classification) or continuous label (reg)
    Recurrent neural network
Spatiotemporal Classification/regression
    Image sequence 3or4D in - images over time
    Discrete or continuous class label
Image Captioning
    Image 2or3D in

Natural text description out
Ground truth images w captions
CNN + recurrent neural network

Transfer learning - leverage something someone else has done on a completely different dataset

She has applied DL to predicting solar flares based on the sun's magnetosphere images

Unsupervised learning

The computer learns to identify enough features of an image that it could decompose the image into features and then successfully recompose the image - and if it can't do it then re-identify the feature until it can work

Convolution Neural Networks: Epic Fails

Questions

How do you encapsulate what you learned? Once trained, the CNN is a model that can be applied to other data

Are these techniques and data open sourced? The DL community supports open source and there are free packages that can be used for example with Python or R, etc

Are these techniques used to recognize pests in agriculture? Yes, likely, also for recognizing plant health

Will these methods ever be appropriate for smaller imagery dataset - size ~200 images? All of these methods are very data hungry. There a possibility to apply transfer learning where you only have to tweak the last couple of layers of the model. You will most likely need many more images to capture as much of the variation as possible, which likely isn't possible with a couple hundred of images.

Does spatial or temporal autocorrelation in data cause overfit in the models? The model overfit is usually cause by high variation in data not autocorrelation. You don't necessarily have to worry about autocorrelation with these techniques like you would with traditional statistical techniques

The needs to be some physically based decision metric built into these DL techniques. Usually on of the last layers of the model is a "softmax" where bounds can be applied to the decision

Can the DL models learn from their error and then correct for it? Yes, she didn't know the name for this technique but says it's possible and requires even more computational power

Why is the kappa statistic (used in RS) not used for understanding the uncertainty of ML/DL classifications? In ML/DL we usually look at a confusion matrix. There's no reason that you couldn't use some other method to understand accuracy. Accuracy might not always be an

appropriate thing to look at though especially if you're working with very unbalanced data for example the data to predict solar flares where 5% or less of the images are flares.

In imagenet data are there a lot of attributes/metadata attached to each image? Different image dataset have a different amount and level of detail of annotation for each image. You need to find an image dataset that's closely related to where you want to go for example using cluttered images if that more applicable than using images of an object on a white background.

Ecology/RS doesn't have a standard image dataset, would it be beneficial for the community to create one versus focusing on new ML/DL methods? It's definitely a good idea to create new image datasets that are different than the ones that exist currently, but note that it's easy to create datasets that are missing something that the computer really needs for successful classification

10:30am - Dawn Browning: Applications of ML in natural resources w/geospatial data
Link to TOC

Knowledge learning analysis system (KLAS)
- Started from the need to automate the processing of remote meteorology sensors
- Previous approach:
  - Someone physically driving to the sensors to grab the raw data (L0)
  - Someone QA/QC'd manually to yield L1, L2
  - But # of met stations has increased dramatically- there's now 100
  - Needed automation
- Current approach:
  - Stations transmit the data to a cloud server
  - Some QA/QC is automated to L1
  - More QA/QC - human rule-guided ML process to get to L2, the ML algorithm will flag anomalous data that requires human attention, still refining the algorithm for improvements or when new weather extremes are hit to tell the algorithm not to flag
  - Data is put on a server for researchers to access

Vesicular stomatitis virus (VSV) case study
- Grand challenge: what are the environmental factors in the spread/predicted location of VSV
- The vector is black flies, sand midges, and sand flies. The disease affects horses
- Multivariate analysis: Input variables tied to the vectors, for example if the vector is tied to streams then an input variable could be distance to stream
- Using ML (maximum entropy technique MAXENT), 5+ environmental variables emerge as drivers of the disease spread/location

Take aways:
- Modeling human behaviour with ML increased efficiency of data handling and QA/QC
- ML can distill complex environmental relationships to yield novel insights
- Env characteristics are more important that viral characteristics in determining spatial

patterns in occurrence

Questions/Comments

Is mimicking human behaviour with ML the best way to QA/QC data? Shouldn't we be letting the computer detect things that a human may not see? By modeling human behaviour she meant having the human establishing some bounds and rules for the ML algorithms

11am - Niall Hanan (NMSU): Machine learning: friend and foe of geospatial and ecological science

Link to TOC

Looking at the predictive capabilities of ML and where/how to derive ecological insights
It's great that ML can predict where the forest is, but we also want to understand why the forest is there- ecological insight hidden in the ML black box
We could try to develop physically based non-linear models but ML helps us do this much more quickly
ML: friend of geospatial prediction, foe of ecological insights
Success: mapping woody cover at regional scale using ML with satellite radar and optical data
    ML shows large improvement for predicting tree cover over non-ML methods
Success: predicting future veg structure and carbon stocks with ML and climate forecasts
    Random forest technique out performs non-ML model
ML issues: Woody cover in African savannahs: the role of resources, fire, and herbivory
    Ranking of the relative importance of different predictor variables
    But many of the predictors even though the model fits the data well, the researchers still dont understand the ecological relationships
ML issues: analysis of stable states in global savannas: is the CART pulling the horse?
    Tree cover estimated using CART methods
    Are we detecting discontinuities (bifurcations) and alternate stable states because of the use of the CART model?
    The created dummy data (pseudo tree cover) where there were no bifurcations to test the CART model
    Compare distributions of pseudo tree cover and CART predictions - CART is changing the distribution and adding features/modes in the data
    Using CART model plus smoothing out of the nodes, results could still be mistaken for bifurcations
Now testing the ability of various random forest models to reproduce known ecological relationships by looking at partial dependence plots

Questions/Comments

Should colinearity of variables be addressed before applying these ML methods? Some ML methods deal well with colinearity but others do not, so it may be better to deal with this before applying ML techniques in some cases

How do you engage in these ML approaches in a reproducible way? We could write our own R packages so others could reproduce the results
Why is the CART model changing the distributions of data? The CART model uses a regression tree to break down a dataset into discontinuous nodes and despite smoothing these nodes can remain in the data which is the problem. The issue is a statistical artifact

1:30pm - Afternoon Workgroups
Link to TOC

Groups:
1. How would we do ___ (geospatial +AI) on Ceres?
2. Deep learning and HPC with Laura Boucheron
3. Next steps for SCINet group

3:30pm - Breakout Group Report Outs and Notes
Link to TOC

BREAKOUT GROUP - How would we do ___ (geospatial +AI) on Ceres?
SCINet questions the group has
- how to facilitate scientist using scinet, do they need it, etc
- where are the instructions, how do I use it, has at least 3 projects that will require the use
- how do i get the projects set up
- how to get data there
- what scripts do I need to write
- what else do i need to think about
- how to I use a container, port env for reproducibility
- how to collaborate
- using globus to move data
- should I set up a linux box in my lab? no you can interface to
- how to get common input files on the system and access ones that are already there
- what DL packages are available for a specific type of analysis, can we get them on Ceres

Things we want to learn in this session
- globus data transfer
- logging in
- Working in linux and navigating to your project space
- building/using a container
- accessing rstudio and jupyter lab

Demonstration session by Rowan
- Logging in
- Navigating in linux
- Seeing available software modules

Accessing rstudio

Wish list for SCINet support / recommendations

Paste Dawn's notes here

Establishment of a GeoTeam - computer folks with some domain expertise

Programmers to help parallelize and/or translate languages

Helping with workflow and code optimization

Contractor for 3-5 years

Documentation of all the work

SHared code repository

Could have a process for applying to get the geoteam type services

Searchable forum of scinet questions/answers

Shared Library of data on the system

Purchase of a GPU for remote sensing and UAV processing on SCINet

Prioritized recommendations from the Practical Next Steps group (continued, Day 2)

1. GEOTEAM
    a. Programmers to parllelize existing code or translate to new programs
    b. Workflow and code optimization strategies (computer people with domain experience)
    c. 3 to 5 years to document trainings to facilitate the transfer of expertise and knowledge
    d. Apply for use of those resources (code that needs to be translated or parallelized)
2. Common input files and data resources. What are the input files you'd want?
3. Searchable forum on the new website
4. Pix4D Engine software would be a huge asset that would facilitate use of HPC for drone image processing.
5. Additional GPU note to facilitate the use of CERES for image processing when we get to a point of it being a limited resource.
6. (From Feng Gao) Facilitate ML on CERES via these two steps
    a. ML uses MODIS LAI to get to LANDSAT LAI
    b. ML to sharpen the thermal band (Rulequest –old name, Cubist-paid version)

BREAKOUT GROUP - Deep Learning and HPC with Laura Boucheron

How does DL relate to HPC?

Need HPC for large training set s but can test smaller on local machines

It would be nice to have sample scripts that use DL for an agricultural application

Deep Learning (DL) with High-Performance Computing (HPC) on CERES

- There are different approaches for parallelizing applications, and the choice of approach depends on the algorithm (i.e. complexity of operations to be repeated) and hardware availability (GPU or CPU). The way that DL applications can significantly benefit from

HPC is GPU -based approaches by the nature of the algorithms. If we want to pursue DL research and want HPC support to facilitate that research, there will likely need to be GPU resources.

- An option to assess the benefits of CPU- vs GPU-based approaches for DL, we could run a test on a cluster with both, e.g. Discovery at NMSU, with a test example. CPU-based approaches can also be tested on CERES for completeness.
- Most DL interests in geospatial group likely about segmentation, region-based, or mask-based DL methods. These are more computationally expensive than whole image classifications, increasing the need for HPC support for feasible run times.
- Scripts implementing DL libraries, e.g. in python, can be tested locally then transferred to CERES. Only modification that should be needed is turning on parallel options (e.g. go look for GPUs) within the script.
- **Recommendation: Do a cost-benefit analysis of adding more GPUs in terms of adopting DL methods in geospatial research.**

UAV Imagery Classification with DL

- Knowledge transfer can be used to reduce the workload of training a Convolutional Neural Network (CNN) for classifying scientific imagery like UAV imagery. It has been similarly done for astronomy and medical image analysis. Even so, HPC will likely be useful.
- Biggest hurdle will be labeling what we want classified in the UAV imagery, e.g. vegetation type.
- Although we often view UAV imagery as mosaics, the individual images can be used for trainings. This way, one flight will provide 100s of images, not one. Alternatively, a mosaic can be tiled into smaller images to increase the sample size. Labeling can be done on individual images or mosaic, whichever is easier for the person labeling. A way to speed up labeling is doing an alternative classification method, then having a human correct it.
- **Recommendations: A collection of tutorials from agriculture/geospatial-relevant applications and example scripts for using DL (on CERES) be created.**

BREAKOUT GROUP - Next steps for SCINet group

GPUs
Support for running models on Ceres and parallelizing them
Optimizing code for parallel processing
Getting Software onto Ceres (paste the list here)
Detailed software carpentry training to move from excel to R or python

Idea from the larger group: hands-on group work sessions in R to teach those who don't know R the skills they need to process incoming data with R instead of excel
The jornada has a similar working group based around R and evolving beyond R, with presentations and code sharing

Recommendations

- Need More GPU's
- Capable of hierarchical or multiple models running continuous at the same time with a way to receive output results
- Transform modeling code to enable parallel processing
- Make machine learning tools available on SciNet examples, open cv, KERARS, Lidar tools, Pytorch, tensorflow, GrassGIS, QGIS,
- Provide detailed software carpentry training

4pm - Development of a SCINet Geospatial Research Working Group: Goals, Roles, & Responsibilities; outcomes and products

Link to TOC

Deb: there will be a quarterly news letter to all ARS that can advertise things like software/computational tips to make your life easier, and info on how to sign up/request a data carpentry training

The working group could provide continued input to the direction of SCINet
A place where the group decides their needs and pushes to have them met
Participation in the group could help facilitate your research through connections to details of the research of other geospatial researchers in the working group

There will also be a postdoc associated with this workshop to help carry out the recommendations of this group

for example the postdoc could create the library of input files on SCINet
Show up at follow on working group sessions and working with individual research groups/projects

This group should send in their recommendations prioritized for this postdoc to work on
The hope from the administrators is that the postdocs will drink the koolaid and be hired on as permanent scientists
Set up a SCINet Geospatial Working Group on Basecamp - done Alisa started it
If this group wants to get together again, we should decide and let Deb know

How can the LTAR working groups better interface with SCINet? Incorporate learning sessions and demos into LTAR gitweeds, serve on the SCINet Advisory Committee

If there are people that should be included in this working group that aren't, check with these people if they want to be involved first and send their names to Kerrie. Kerrie to send a blurb for people to use.
Kerrie to send out the link to the shared drive
Kerrie to forward info on how to sign up for the SCINet Advisory Committee

Kerrie to send out the post-workshop survey
Ask who wants to be involved in the working group from the participants, keep a tracking list, send to Alisa to get them added to Basecamp