ORNL/TM-2019/1442

The Trillion Pixel GeoAI Challenge Workshop

Dalton Lunga
Hamed Alemohammad
Yan Liu
Shawn Newsam
Fabio Pacifici
Hector Santos-Villalobos
Eric Shook
Robert Stewart
Sophie Voisin
Lexie Yang
Budhu Bhaduri

Approved for public release.
Distribution is unlimited.

December 12, 2019

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2019/1442

The Trillion Pixel GeoAI Challenge Workshop

With contributions from: materials presented at the workshop and notes from volunteers
(Jacob Arndt, Jordan Bowman, Rohan Dhamdhere, Umesh Gupta, Tao Liu, Nikhil Makkar)

Presentations can be downloaded from the workshop website at:
https://geoai.ornl.gov/trillion-pixel/

Date Published: December 12, 2019

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-Battelle, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

ACRONYMS

AI	Artificial Intelligence
ANL	Argonne National Laboratory
API	Application Programming Interface
AWS	Amazon Web Services
BigEarth	BigEarth European Research Council
CADES	Computing and Data Environment for Science
CNCF	Cloud Native Computing Foundation
CV	Computer Vision
CVPR	Conference on Computer Vision and Pattern Recognition
DARPA	Defense Advanced Research Projects Agency
DARPA STO	Defense Advanced Research Projects Agency Strategic Technology Office
DHS	Department of Homeland Security
DL	Deep Learning
DOE	Department of Energy
EO	Earth Observation
FEMA	Federal Emergency Management Agency
FPGA	Field-Programmable Gate Array
FTP	File Transfer Protocol
GCP	Google Cloud Platform
GDAL	Geospatial Data Abstraction Library
GeoAI	Artificial Intelligence for Geographic Knowledge Discovery
GIS	Geographic Information Science
HDD	Hard Disk Drives
HPC	High-performance computing
HTTP	HyperText Transfer Protocol
IoT	Internet of Things
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NAIP	National Agriculture Imagery Program
NASA	National Aeronautics and Space Administration
Navy	Naval Information Warfare Center Pacific
NCCS	National Center for Computation Sciences
NDAAS	NSG Data Analytic Architecture Services
NDWI	Normalized Difference Water Index
NGA	National Geospatial Intelligence Agency
NISR Brazil	National Institute for Space Research Brazil

NLIP	National Land Imaging Program
NVM	Non-Volatile Memory
OGC	Open Geospatial Consortium
ORNL	Oak Ridge National Laboratory
OSM	Open Street Map
PNNL	Pacific Northwest National Laboratory
POST	Prioritizing Operations Support Tool
SaaS	Software as a Service
SAR	Synthetic Aperture Radar
SDK	Software Development Kit
SRS	Spatial Reference System
SSD	Solid-State Drive
STAC	Spatial Temporal Asset Catalog
TPU	Tensor Processing Unit
UAV	Unmanned Aerial Vehicle
UNACORN	Universal AI CORe eNvironment
UQ	Uncertainty Quantification
USAF	US Air Force
USGS	U.S. Geological Survey
VGI	Volunteer Geographic Information
VPU	Vision Processing Unit
WFS	Web Feature Service
WMTS	Web Map Tile Service
WPAF Base	Wright-Patterson Air Force Base

ACKNOWLEDGMENTS

We would like to thank the organizing committee, moderators and workshop participants (names listed in Table 2) for the intellectual contributions towards the workshop. We also would like to thank the administrative support team and many volunteers who helped with the logistics of the workshop.

Trillion Pixel GeoAI Challenge Workshop
Oak Ridge National Laboratory - September 26, 2019

Figure 1. The Trillion Pixel Challenge for GeoAI Workshop Participants

Table 1. The Trillion Pixel GeoAI Challenge Participants

Participant	Affiliation
Arndt, Jacob	ORNL
Batic, Matej	Sinergize Hub
Bhaduri, Budhu	ORNL
Bingham, Phil	ORNL
Borth, Richard	USAF
Bowman, Jordan	ORNL
Brown, Christopher	Google Earth
Chinthavali, Supriya	ORNL
Coletti, Mark	ORNL
Cooper, Frank	NGA
Corcoran-Freelander, Melanie	Analytic Fusion
Crawford, Cary	ORNL
Davis, Chris	ORNL
Demir, Begum	BigEarth Research
Dhamdhere, Rohan	ORNL
Dinh, Thao	WPAF Base
Doucette, Peter	USGS
Doyle, Raffianne	Navy
Duke, Chris	ORNL
Eldawy, Ahmed	University of Cal. Riverside
Ferrier, Nicola	ANL
Frame, Mike	USGS
Gaikwad, Neil	MIT
Gleason, Shaun	ORNL
Gorman, Shaun	PixelI8.Earth LLC
Gupta, Umesh	ORNL
Hedrick, John	USAF
Hinkle, Jacob	ORNL
Hogan, Daniel	In-Q-Tel
Hogan, Ian	USAF
Howe, Jonathan	NVIDIA
Hughes, David	ORNL
Jacobs, Nathan	University of Kentucky
Jiang, Zhe	University of Alabama
Kannan, Ramakrishnan	ORNL
Kanu, Pamela	NGA
Kerekes, Ryan	ORNL
Kontgis, Caitlin	Descartes Labs
Korver, Mark	Amazon
Kumar, Vipin	University of Minnesota
Kurte, Kuldeep	ORNL
Layton, Christopher	ORNL
Lim, Seung-Hwan	ORNL
Liu, Frank	ORNL
Liu, Tao	ORNL
Liu, Yan	ORNL
Lunga, Dalton	ORNL
Makkar, Nikhil	ORNL
March, Don	ORNL
Marchetti, Amanda	NearMap
Maretto, Raian Vargas	NISR Brazil
Martinez Manso, Jesus	Planet Labs
McKee, Jacob	ORNL
Medeiros, Henry	University of Marquette
Menon, Sud	Esri
Miller, Corey	Ursa Space Systems
Munoz, Margell	USAF
Munsell, Mark	NGA
Myers, Todd	DoD
Newsam, Shawn	University of Cal. Merced
Nichols, Jeff	ORNL
Pacifici, Fabio	Maxar Technologies
Page, David	ORNL
Papamarkou, Theodore	ORNL
Parente, Leandro Leal	Federal University of Goiás
Pike, Bill	PNNL
Ramachandran, Rahul	NASA
Rapstine, Natalya	USGS
Reed, Tom	NVIDIA
Santos, Hector	ORNL
Scott, Kimberly	Astraea, Inc
Shankar, Arjun	ORNL
Shook, Eric	University of Minnesota
Simi, Christopher	DARPA
Sparks, Kevin	ORNL
Stewart, Robert	ORNL
Sukumar, Rangan	Cray Inc.
Thakur, Gautam	ORNL
Thapliyal, Himanshu	University of Kentucky
Theodore, Jay	Esri
Thompson, Mitch	Riverside Research
Thornton, Peter	ORNL
Tsaris, Aristeidis	ORNL
Tuttle, Mark	ORNL
Van Etten, Adam	In-Q-Tel
Vatsavai, Raju	North Carolina State Univerisy
Vaughan, Chris	DHS/FEMA
Voisin, Sophie	ORNL
Wang, Dali	ORNL
Weir, Nick	In-Q-Tel
Womble, David	ORNL
Yang, Lexie	ORNL
Yi, Zhuang-Fang	Development Seed
Yin, Junqi	ORNL

EXECUTIVE SUMMARY

Rapid innovations in satellite and airborne remote sensing capabilities holds the promise of collecting high-resolution imagery with daily, even hourly cadence across the entire planet. Availability of such earth observation data streams at varying spatial and temporal scales, coupled with astonishing progress in AI and transformational advances in high performance computing bring into view the possibility of mapping and interpreting the surface of our planet at unprecedented detail. The implications of such capability for science, technology, policy, and security are far reaching. Moreover, even at a modest 5m resolution, 100 trillion pixels will describe the surface of the Earth every day. Consequently, monitoring the pulse of our planet will entail collecting, refining, analyzing, and curating those 100 trillion pixels within a 24-hour period. That is a challenge with unknown solutions as of today's advances and therefore serves as a key motivation for this workshop.

The GEOINT and AI communities have a unique opportunity to enable detailed monitoring the pulse of our planet and obtain new insights into how humans occupy, interact with, and alter the surface of the Earth over time. From a science and technology standpoint, this is a modern-day moonshot that presents numerous challenges particularly related to scaling the analysis to every pixel covering the planet. The challenge of making sense of all these pixels in a timely fashion include image processing at scale, selective analysis of only the pixels that matter, generalizing AI models across heterogeneous locations, computing on hardware that is memory constrained, and deploying automated feature extraction on edge devices. The designing of capable GeoAI workflows and data pipelines will require interdisciplinary efforts and partnerships. The Trillion Pixel Challenge for GeoAI workshop was is promoted as a first community effort to discuss priority research and development directions and to establish the partnerships and collaborations needed to make progress on critical scientific and operational applications from local to regional to planet scale.

The two-day workshop focused on the progress and gaps in six topic areas. A summary of the gaps identified through speaker presentations include:

- *Global challenges:* Key applications and trends of broader adaptation of automated feature extraction methods by government agencies.
- *AI scalability and generalization:* Capability to learn from few labeled samples, the need for domain aware learning, lack of spatio-temporal generalization capabilities.
- *Geo-spatial data infrastructure:* AI for geospatial data integration utilizing machine learning to automate, streamline, or assist with geospatial data infrastructure.
- *Edge computing for GeoAI:* Systems to push analytics to data on edge platforms to minimize information loss at the points of collection and avoid costly data movement to the cloud or high-performance computing environments.
- *Hardware design and high-performance computing:* Systems capable of making sense of trillions of pixels streaming at a daily rate. Future systems will require tight integration of network, storage, and computing resources that span CPUs, GPUs, accelerators, and potentially domain-specific architectures such as GeoAI spatial processors.
- *Opportunities, collaboration and partnerships:* Focus on human talent to advance research and development in GeoAI. How can we attract more talent (e.g. students) from the data science community to work on geospatial problems?

Thirty two panelists from among ninety four invited participants, representing various sectors including United States Federal agencies, private industry, academia, international agencies, and the DOE national laboratories (Table 2), shared their critical perspectives on the above six themes.

Table 2. Workshop participation by sector

Sector	Number of participants
US Federal Agencies	17
Private Sector Organizations	22
DOE National Laboratories	42
International Organizations	3
Academic Institutions	10

Overall, the panelists and participants collectively concluded that future community-wide engagements should focus on multiple activity pathways:

1. Providing scientific community a shared space environment to enable user access to both data and AI capabilities.

2. Develop and disseminate uncertainty quantification (UQ) capability to allow users to assess uncertainty associated with both GeoAI systems and application data.

3. Seek new domain adaptation and transfer learning strategies as way to foster AI robustness on varying factors encompassing imagery acquisition conditions.

4. Experimenting with edge devices to compute on GeoAI problems that are aimed at limiting the movement of data.

5. Cultivate stronger partnership and collaboration between hardware design engineers and AI and domain scientists to bring-forth the physical demands of current trillion-pixel workloads as test cases for early prototyping of accelerators.

The convening of this workshop and similar gatherings in the future represent a promising step towards fostering a community-wide engagement on the development of GeoAI systems for processing trillion to quadrillion pixel data sets.

1. WORKSHOP SESSIONS

1.1 INTRODUCTION

The workshop format consisted of six interactive panel sessions delivered over one and half days. What follows are summary reports about the six sessions consisting of a description of the session theme, summaries of each of the presentations, and Q&A engagement with the audience.

1.2 SESSION 1: GEOAI IN GLOBAL CHALLENGES

Moderator: Budhu Bhaduri, Director, National Security Emerging Technologies Division, ORNL

With a growing number of global challenges, it is imperative that societal impact due to AI in the context of geographic knowledge discovery is fully explored and understood. Designing GeoAI systems that will scale with greatest challenges requires engaging the front lines and visionary societal perspectives for guidance. The global challenges session presented a forward framing of the GeoAI initiative as a bridge toward uncovering unlimited possibilities for impacting global sustainable development goals and challenges for society's benefit. The session engaged on the key role and highlighted important problems that GeoAI systems could resolve for various domain needs across the GIScience community.

Frank Cooper, Deputy Group Chief, Foundation GEOINT, National Geospatial Intelligence Agency (NGA)

The diversity of the Foundation GEOINT Group at the NGA presents operational benefits to the nation, in particular, intelligence gathering through the analysis of a wide variety of data-sets. Cooper's talk provided an overview of the Office of Geomatics. The office consists of scientists conducting geodetic surveys, astronomical surveys and developing gravitational and magnetic models for the earth. The offices of Geography, Maritime and Aeronautical form the backbone of the products at Foundation GEOINT group.

His talk emphasized the amount of data available to NGA, comprising of billions of features and millions of data points across all domains. Figure 2 summarizes some of the examples that were presented. Having access to large volumes of data is inadequate to meet customer demands. Adaptation by government agencies is a necessity. Through participation in the Trillion Pixels Challenge for GeoAI, government agencies can obtain a better understanding of current technological advances from the research community. NGA is embracing automation, augmentation, and artificial intelligence to adapt to the demands of industry and global challenges. Cooper pointed out that changes are currently being made within NGA recruitment strategies to hire persons with knowledge about the industry from the ground level.

The vast amount of data analysis effort within NGA is carried by manual extraction, taking 12 to 18 months to put forth product time lines that are not acceptable to meet decision making for critical missions. By embracing automaton through artificial intelligence, the key driver is a fast product cycle. Developing algorithms which are scalable and adaptable to changes in environment (e.g. changes in coastline infrastructure) it is necessary to develop products in a quicker and timely fashion.

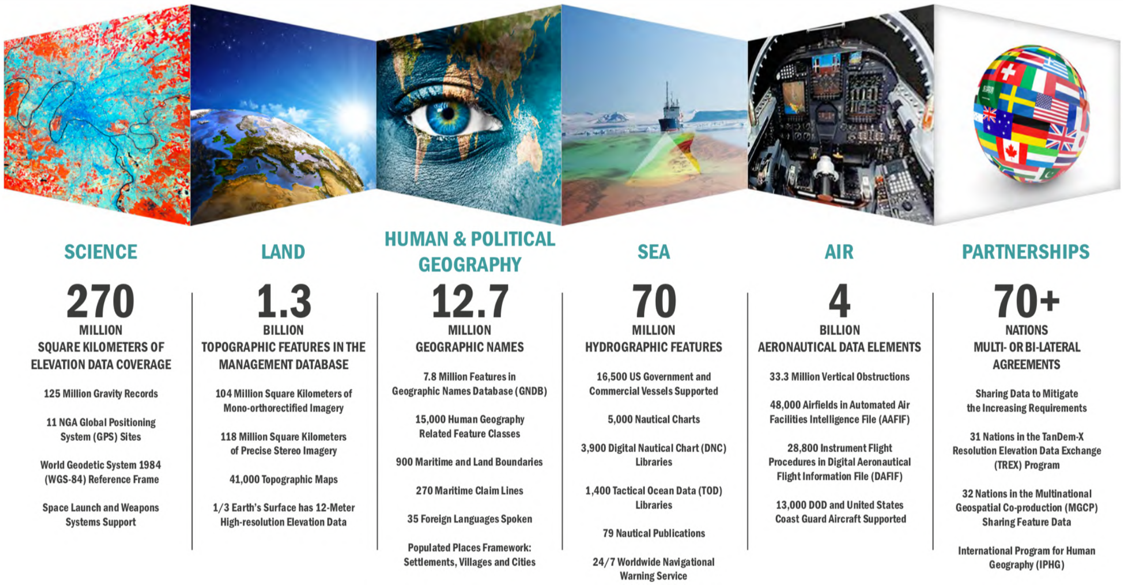

SCIENCE
270
MILLION
SQUARE KILOMETERS OF ELEVATION DATA COVERAGE
125 Million Gravity Records
11 NGA Global Positioning System (GPS) Sites
World Geodetic System 1984 (WGS-84) Reference Frame
Space Launch and Weapons Systems Support
LAND
1.3
BILLION
TOPOGRAPHIC FEATURES IN THE MANAGEMENT DATABASE
104 Million Square Kilometers of Mono-orthorectified Imagery
118 Million Square Kilometers of Precise Stereo Imagery
41,000 Topographic Maps
1/3 Earth's Surface has 12-Meter High-resolution Elevation Data
HUMAN & POLITICAL GEOGRAPHY
12.7
MILLION
GEOGRAPHIC NAMES
7.8 Million Features in Geographic Names Database (GNDB)
15,000 Human Geography Related Feature Classes
900 Maritime and Land Boundaries
270 Maritime Claim Lines
35 Foreign Languages Spoken
Populated Places Framework: Settlements, Villages and Cities
SEA
70
MILLION
HYDROGRAPHIC FEATURES
16,500 US Government and Commercial Vessels Supported
5,000 Nautical Charts
3,900 Digital Nautical Chart (DNC) Libraries
1,400 Tactical Ocean Data (TOD) Libraries
79 Nautical Publications
24/7 Worldwide Navigational Warning Service
AIR
4
BILLION
AERONAUTICAL DATA ELEMENTS
33.3 Million Vertical Obstructions
48,000 Airfields in Automated Air Facilities Intelligence File (AAFIF)
28,800 Instrument Flight Procedures in Digital Aeronautical Flight Information File (DAFIF)
13,000 DOD and United States Coast Guard Aircraft Supported
PARTNERSHIPS
70+
NATIONS
MULTI- OR BI-LATERAL AGREEMENTS
Sharing Data to Mitigate the Increasing Requirements
31 Nations in the TanDem-X Resolution Elevation Data Exchange (TREX) Program
32 Nations in the Multinational Geospatial Co-production (MGCP) Sharing Feature Data
International Program for Human Geography (IPHG)

Figure 2. NGA domains of interest

Pete Doucette, Deputy Program Coordinator, National Land Imaging Program (NLIP), U.S. Geological Survey (USGS)

Earth Map has been a grand challenge for decades. Doucette's talk focused on presenting Figure 3 to succinctly describe the Earth system challenge characterization from a temporal, spatial, and a data synthesis perspective. The compute power and the resurgence of neural nets in recent times has given us a basis to pursue this challenge in a practical way. Assimilation of heterogeneous data layers is necessary for predicting epidemics or natural hazards and its aftermath crisis. The number of land imaging satellites are increasing over time and help in gathering imaging data. For example, the Landsat program has amassed the longest collection effort and built the most comprehensive record of the Earth's land surface in existence. Starting in 2008, the free and open data policy for Landsat has allowed the cloud community to use the Landsat data archive and build products like Google Earth Engine.

There is evidence that global economic benefits from Landsat dramatically rose after the free and open data policy was adopted. This has inspired many other nations to follow suit with adopting the policy. Landsat offers analysis ready data, a new way to tile the data which eases the access to the data. Data is made accessible for stacking, facilitating time-series of analysis. Meta-data is provided to release the burden of the user to tasks such as cloud removal, handle cloud shadows, and other steps in data preparation. This helps the user to focus and pursue the business of analysis.

According to Doucette, the time-series change modelling for different observations in time of a particular location is a big challenge for the AI/ML community. Modelling seasonality or disrupting change and understanding long-term changes can be made possible by solving this challenge. Forecasting models at global scale have huge applications in national security.

His talk further highlighted the relevance for adapting cloud compute architectures and modern data stacks. Many agencies are slowly moving to cloud computing platforms, although limitations of infrastructure and its accessibility for big data remains a challenge. At a human capital development, cultural shifts in teaming would also give rise to cross agency interactions between teams. Focus and cooperation between inter-agency and inter-disciplinary teams will be necessary for solving the challenges of the future.

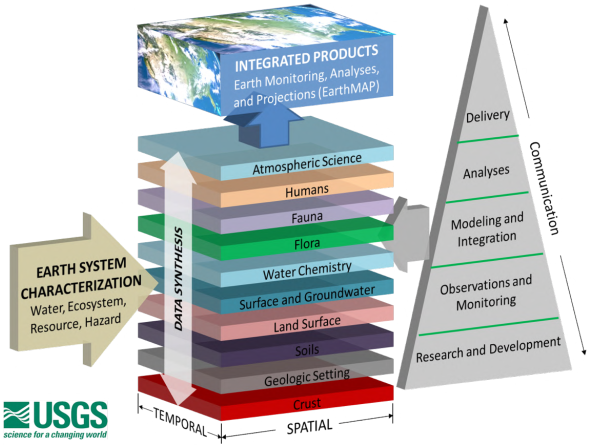


Figure 3. The Earth Map challenge

Raffianne Doyle, UNACORN Lead, Digital Warfare Office, U.S. Navy

Data is the key component sought by all agencies around the world. Doyle's talk informed the audience about the U.S. Navys desire to utilize all the data made available by the increase in computing and storage over time. The talk points that knowledge domain has grown as a complete domain of war fighting. It is being sought after not only by the United States but also by many other countries including adversaries of the U.S. Countries are investing huge money and manpower towards advancing new technologies. In 2019, the Chief Naval Officer made artificial intelligence and machine learning a top priority of the U.S. Navy. AI and ML efforts are to help in augmenting capabilities along with automating them across various levels of missions. Following the Chief Naval Officer's call, the U.S. Navy has put forth the formulation of an AI ecosystem for the Navy called the Universal AI Core Environment (UNACORN).

This ecosystem is tasked with the objective of enabling the war fighters with better and faster decisions for advanced predictive analytics. As a first stage, UNACORN is undertaking the task of building Naval training datasets for AI/ML development. The initiative is currently seeking partnerships with the research community towards efficient data labelling for U.S. Navy applications. An UNACORN AI pipeline has been established and is being used to build foundation data sets. Integration of the pipeline with open source capabilities to automate workflows to acquire, label, enhance, curate and do feature extractions on the data specifically needed for the war fighter.

Rahul Ramachandran, Manager, Inter-Agency Implementation and Advanced Concepts Team (IMPACT), Earth Science Program, National Aeronautics and Space Administration (NASA)

There is greater need for AI/ML for use in improving the quality of the data produced by the sources at the data production stage. Ramachandran's presentation highlighted the Earth Science program at NASA, where data obtained from different assets are managed with the help of a systematic engineered process. AI/ML can be applied to improve certain steps in this data life cycle. Research life cycle is supported by the data system life cycle and thus maximizing the returns on NASA datasets is the goal of the Earth Science program. As an overview of the Earth Science programs, Ramachandran highlights the mission's three-fold focus themes:

- Fostering innovation by improving the processes of data production increasing the effectiveness of the data.
- Building strategic partnerships with other agencies and private entities to leverage combined capabilities to maximize returns on our datasets.
- Enabling adoption of new promising technologies and incorporating them in the processes to be proactive about technology.

NASA does not lack data but lacks the effective utilization of the data. As such, ML for the Earth Science data system is necessary to exploit the large archives of Earth Science data collected from multiple instruments. Moving to cloud platforms for data storage and processing helps in removing the data-transfer bottleneck. ML in Earth Science also provides an opportunity to augment and improve the existing data systems operations and services. It also provides an opportunity for novel research and innovation.

Research on Earth data is lacking benchmark datasets which would help in effective measurement of different approaches on the data. Also, the available training datasets are limited. Most datasets are small and are not shared within the community. Building training data is much harder in earth science than in regular research. It requires investment of time along with field work and post processing. Providing information resources for such a process could be key to effective utilization of data in research. Data heterogeneity problem is rampant in Earth data and is key challenge to solve in this space.

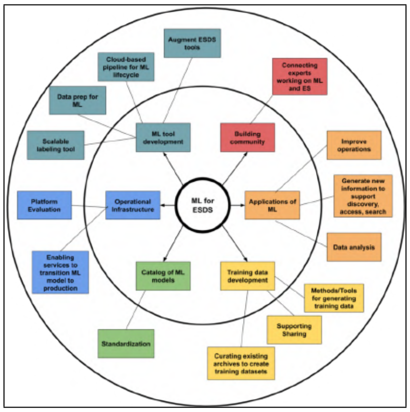

Augment ESDS tools
Cloud-based pipeline for ML lifecycle
Connecting experts working on ML and ES
Data prep for ML
Scalable labeling tool
ML tool development
Building community
Improve operations
Generate new information to support discovery, access, search
Platform Evaluation
Operational Infrastructure
ML for ESDS
Applications of ML
Data analysis
Enabling services to transition ML model to production
Catalog of ML models
Training data development
Methods/Tools for generating training data
Supporting Sharing
Standardization
Curating existing archives to create training datasets

Figure 4. Various ML challenges at the Earth Science program

Christopher Simi, Program Manager, Defense Advanced Research Projects Agency (DARPA) Strategic Technology Office (STO)

The improvement in satellite and computer technology has paved the way for the collection of large volumes of high resolution earth observation images. Simi presented a common viewpoint highlighting both how capable the current technology is at capturing huge amounts of quality data, yet lag in technologies for processing this data even with availability of computing power that has increased by orders of magnitude. Earlier it was not possible to capture both large area coverage and better resolution. This changed with the class of Worldview and Planetscope satellites. New commercial space industries can cover large portions of earth on a daily basis while simultaneously providing higher resolution imagery. The coverage of new commercial satellites is shown in Figure 5.

His talk emphasized a perspective that, processing such enormous amounts of data is a key challenge and automated and augmentation assisted techniques must be developed for analysis of such data in a timely manner.

DARPA STO, being a strategic technology oce, is trying to focus on future tactical environments which may unfold within very short timelines. From a strategist point of view, "It's not a trillion-pixel problem, but rather 'which part of the trillion pixels' are needed at a given time, is the problem," he said.

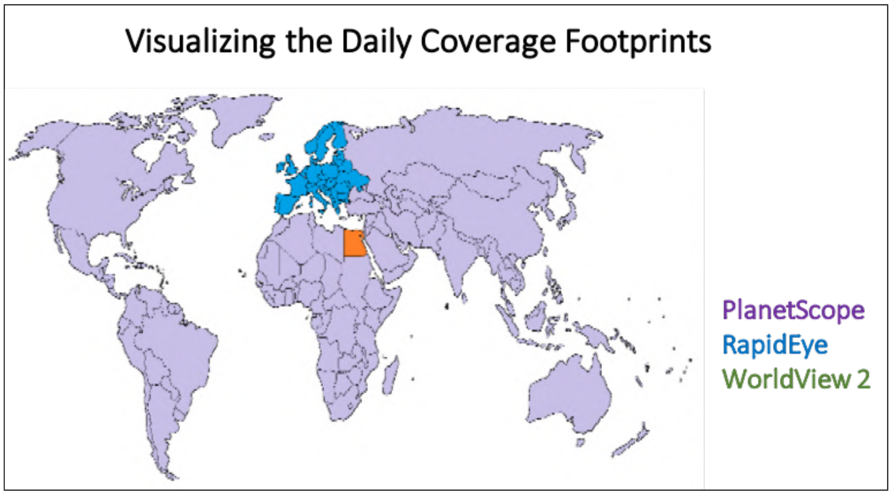


Figure 5. Coverage of new commercial satellites

Chris Vaughan, Geospatial Information Officer, Federal Emergency Management Agency (FEMA)

Modelling a disaster is super-complex. Occurrence of a disaster event is accompanied by disruption of normal data flows and supply chain. This provides a significant challenge in understanding the disaster event. Vaughan's talk armed that gathering data and labelling the data is a key step in disaster modelling. For this task, FEMA has partnered with Oak Ridge National Laboratory for mapping building footprints of the United States.

FEMA is also trying to add meta-data information about the buildings to the footprints. With this data, impact assessment estimation can be done, and remote sensing can be used for validation of the assessment. Crowd sourcing of the assessment can be used for understanding the ground level situation and impact of the disaster event.

For example, under current capabilities, FEMA is able to identify disruption in various markers like food, water, shelter, transportation, and communication after the disaster impact. This information comes from disparate sources. FEMA is now proposing a Prioritizing Operations Support Tool (POST). Information about this tool is presented in the Figure 6. This is a framework which uses information about the hazard, the estimated impact on community, the surveys about the population vulnerability and the exposure and impacts of the disaster around 48 to 24 hours prior to the impact to get an aggregate of any disaster type. All this information is being used to model 1km 1km prioritize areas of interest. Such a model of prioritized areas can help in estimating the resources requirement for such an area in case of a future disaster event. This is the grand challenge that the disaster modelling puts forth for the community to have an impact in emergency management.

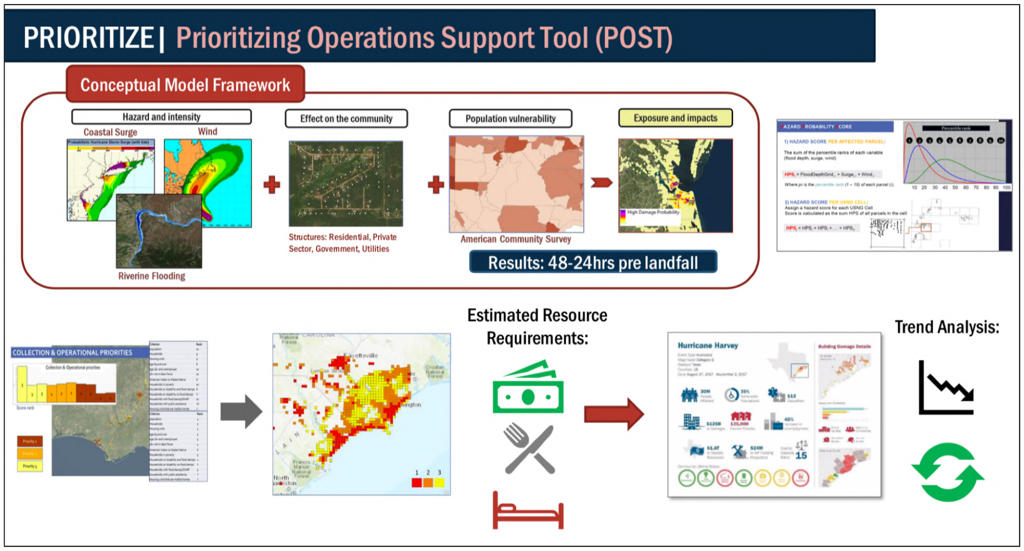

PRIORITIZE | Prioritizing Operations Support Tool (POST)
Conceptual Model Framework
Hazard and intensity
Coastal Surge
Wind
Riverine Flooding
Effect on the community
Structures: Residential, Private Sector, Government, Utilities
Population vulnerability
American Community Survey
Exposure and impacts
Results: 48-24hrs pre landfall
Estimated Resource Requirements:
Trend Analysis:
COLLECTION & OPERATIONAL PRIORITIES
Hurricane Harvey

Figure 6. Prioritizing operations support workflow

Q&A

Question 1: How do you keep updating or revising the model in real time, so that bad actors cannot modify their behaviours to beat the model? Raffianne Doyle: Our vision is to have digital twin of the model apart from the deployed model. This digital twin model can continue to train on the new data that is being streamed through so that we can detect the modifications to the behaviour and this information can be provided to supervising personnel to take a decision on what model to use.

Frank Cooper: We solve a similar situation by adapting the algorithm on the new data as it is being evaluated.

Question 2: Although all speakers mention that these data are available online for the public, it is not easily accessible according to my students. Could these data made more accessible or is there any ways we can help to do it more accessible.

Pete Doucette: Google Earth Engine is a great resource for the datasets. Also, as most speakers mentioned, organizations are moving to cloud platforms for data storage and which would help the accessibility issue to be driven out as more researchers adopt the cloud for the running algorithms.

Rahul Ramachandran: For NASA, data catalogue can be found at *https://science.nasa.gov/earth-science/earth-data*. Across agencies, *https://www.data.gov/* is repository where agencies are required to push their geospatial data. There is some learning curve required to understand science data. For this purpose, tutorials and notebooks are provided. There are various resources but they are distributed and making them accessible is key.

Question 3: My question is for stakeholder agencies. Deep learning techniques were developed by computer vision researchers for camera photos and are now applied to remote sensing. I want to know if you encountered any new challenges that you want the remote sensing community to address?

Pete Doucette: I think deep learning is projected as ultimate solver for many projects. But it really is not. Beyond Cats and Dogs, deep learning should be researched solutions on actual problems. Using deep learning to understand trends in temporal domain should be researched more thoroughly. I am proponent of this shift towards deep learning solutions but there huge discernment process that must undertaken by agencies towards the shift to deep learning.

Rahul Ramachandran: Our group applied deep learning for hurricane estimation. We ran into a problem while running deep learning models operationally with the shifts in the data and other things never encountered in a prototype. One other area we are looking at is Event detection, which has a much lower threshold of acceptance than science. Deep learning must be successfully applied to such problems with lower threshold of acceptance before moving to problems with high thresholds.

Chris Vaughan: From FEMA perspective, we deal with GeoAI at scale. We have publicly posted our damage assessment dataset and the USGS has provided the post-disaster imagery datasets. Using these two datasets, tell us what the right methodology is to integrate these trillion pixels in real-time.

Question 4: With all the talk about openness and sharing of the data being done, is there any protection being done in that we might be arming or preparing our adversaries with our plans and reactions to a disaster event.

Raffianne Doyle: That is on the forefront on the design and architecture that i am trying to set up for the data leaks. We are looking into techniques like mirrored datasets and camouflage to be able to provide subsets of the data that don't have the specific critical features in them. The US Navy is also keeping the data under government control and access is only given to vetted parties.

Frank Cooper: NGA is vets all the partners as well as the data to check for data corruption. We are also putting tools in place for vetting the data.

Peter Doucette: My answer is related to the openness and sharing the data part of the question. We provide data on the free and open policy, but it is often time a quid pro quo with other nations. We collaborate with number of international partners and we encourage them to adopt free and open policy for their data as well.

Question 5: This question is directed to official from FEMA. Is FEMA looking at some sort of Automated planning and scheduling tool for logistical planning and resource distribution in disaster event.

Chris Vaughan: Yes, such a logistical planning and quick distribution of necessary resources like grocery and medicines is done with the help from private sector as private sector is good in just-in-time delivery systems.

Question 6: As AI technology gets mature with the growing interest and the number players, is there anything from the panel that the community can do to accelerate GeoAI and not wait till the technology matures?

Raffianne Doyle: I would say the biggest challenge is the development of training dataset for validation. A human and machine teaming capabilities that is one of the issues we need to tackle and we need to have good methodology to be able to truly validate the recommendations provided by the ML models have a certain level of confidence and then we can take them into the calculus of decision making. This is area that where I believe that the research and academic community can aide us. This challenge of validating that the ML models are helping and not having a negative impact on the operation is critical challenge to undertake.

1.3 SESSION 2: AI SCALABILITY AND GENERALIZATION

Moderator: Shawn Newsam, Associate Professor and Founding Faculty, Electrical Engineering and Computer Science, University of California at Merced

The volume, velocity and variety of geospatial data are constantly growing at an unprecedented pace. To enable transforming and disruptive geo-knowledge extraction capability with such rich and diverse data, scalability and robust generalization aspects are critical to understand the design of new generation of AI systems. This session discussed AI scalability constraints and opportunities as motivated by various trillion pixel challenges toward extracting decision-critical and space-time relations from large scale geospatial data. Some of the key questions that where addressed during presentations included: key challenges in scaling GeoAI systems to the trillion pixel scenarios? How to ensure that future GeoAI systems are robust and reliable and why is this a core issue in their design? How to evaluate GeoAI systems at scale? GeoAI systems, particularly those based on deep learning, face the same challenges as in other domains: large collections of labeled training data, model architecture search, etc. Are there unique opportunities in the geo-domain to overcome them?

As an introduction to the session, Shawn Newsam, moderator, emphasizes that *scalability* and *generalization* carry different meaning depending on the context. Scalability is often understood in terms of hardware, compute, or algorithmic time complexity. Scalability and generalization in this session also extended to challenges in the realm of developing AI that is robust across heterogeneous data.

Bill Pike, Director, Computing & Analytics Division, Pacific Northwest National Laboratory

From a GeoAI perspective, there are five challenges that motivate the Computing & Analytics Division at Pacific Northwest National Laboratory (PNNL). These challenges include: 1) multi-scale inference, 2) spatiotemporal transferability, 3) geoinferencing, 4) generating insight from geographic data efficiently, and 5) trust and robustness of GeoAI systems. Pike's talk focused on techniques for addressing the challenge in generating insight from geographic data efficiently and approaches for improving the trust and robustness of GeoAI systems. Pike highlighted three ways to generate insight from geographic data efficiently. One technique is low-shot learning. This is the idea of being able to train a classifier from a small amount of labeled data ultimately enabling humans to ask new questions of the existing data. Domain aware learning is a second technique that can help improve our ability to gain insight from geographic data more efficiently. This idea recognizes that ML methods are good at interpolation but not so good at extrapolation. Furthermore, they do not allow us to express the phenomenology of the things that we care about nearly as clearly as we need in order to take action in the real world. PNNL is investing resources into the data model convergence problem, as shown in Figure 7, by working at the intersection of hardware and software for high performance computing (HPC), ML, modeling and simulation, and data analytics.

Important to this is the integration of domain-aware ML methods. Algorithms such as auto-encoders with predefined priors or developing ways in which physical phenomenology can be represented in ML algorithms are two approaches to domain-aware learning. A final example for improving our understanding of geographic data more efficiently is leveraging humans and machines together. The challenges of reliability and robustness was first highlighted by the general difficulty in model interpretability. One way to approach this challenge is by developing visual interfaces that allow operators to understand a model's performance and the features it leverages to make decisions. This allows the operator to assist in finding features that have the potential to work across more than one data modality and improve the model's robustness. Final thoughts focus on AI assurance and understanding how to mitigate adversarial influence in GeoAI. Examples of adversarial influence that GeoAI needs to be aware of include model reuse attack, data poisoning, model inversion, bit-flip, and adversarial attacks that manipulate a model's ability to perform the way it was intended.

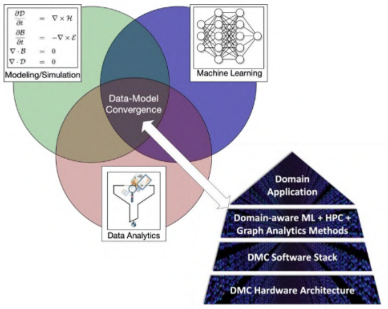

$\frac{\partial \mathcal{D}}{\partial t} = \nabla \times \mathcal{H}$
$\frac{\partial \mathcal{B}}{\partial t} = -\nabla \times \mathcal{E}$
$\nabla \cdot \mathcal{B} = 0$
$\nabla \cdot \mathcal{D} = 0$
Modeling/Simulation
Machine Learning
Data-Model Convergence
Data Analytics
Domain Application
Domain-aware ML + HPC + Graph Analytics Methods
DMC Software Stack
DMC Hardware Architecture

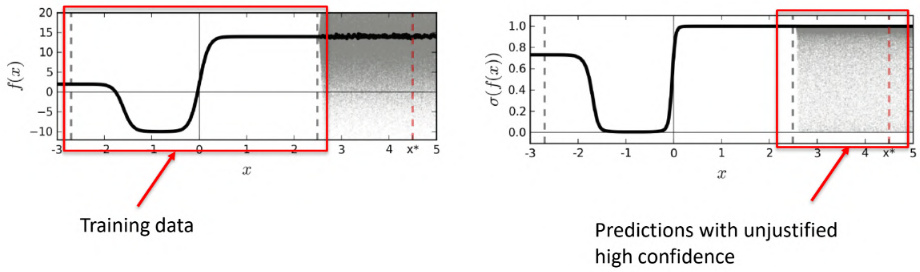
Figure 7. Data-model convergence diagram

Henry Medeiros, Assistant Professor, Electrical and Computer Engineering, Marquette University

Scalability and robustness for deep neural networks are important issues in the context of trillion pixel problems. The relationship between these two concepts is thoroughly discussed throughout Medeiros' presentation. At the present time, neural networks are not that robust. Neural networks don't know what they don't know. They often output predictions with unjustifiably high confidence on things that they have not seen before, see Figure 8. Progress towards systems that are uncertainty aware would not only improve robustness but also provide scalability. Scalable methods include self-supervised, semi-supervised, active, or reinforcement learning. All these learning methods require good measures of uncertainty.

Designing uncertainty aware models can come from leveraging Bayesian-like tools in deep learning models for uncertainty estimation. One strategy in particular is incorporating Monte-Carlo sampling for correcting high-confidence mistakes among different predictions. Monte-Carlo dropout is yet another strategy for estimating uncertainty. Another technique would be to train the network to explicitly learn the uncertainty. For example, one use case is in gaze estimation where incorporating uncertainty terms in the loss function gives more accurate error predictions, improved task performance, and reduces dataset biases. In summary, there are mechanisms to estimate the uncertainty in neural networks and by using these methods we can build systems that are robust and scalable.

There are still challenges that make it difficult to build these uncertainty aware models. These include choosing or designing the right network architecture for the problem. There are issues related to identifying effective prior distributions. There are challenges in bootstrapping the models and relying on transfer learning strategies. One other challenge is manual annotation of data and finding ways that we can minimize the effort. Evaluating the accuracy of the estimated uncertainties is also a challenge. For example, how do we leverage a large number of variances or uncertainty to develop useful model information.



Training data

Predictions with unjustified high confidence

Figure 8. Neural networks don't know what they don't know. Network predictions are not confidence estimates*

Kimberly Scott, Co-Founder & Vice President of Data Science, Astraea

As an illustrative example of challenges in generalization for GeoAI, Scott presented a case study focused on identifying utility-scale solar farms across the world with Sentinel-2. There were many instances of successfully identifying solar farms over regions across the United States and China, however, there were also failures. The use of low spatial resolution data resulted in spectral mixing in pixels, unclear spatial boundaries, and difficulty in excluding background noise. Interesting cases of false positives included tennis courts and agricultural fields. More broadly, challenges for generalization in GeoAI was broken down into three categories including data diversity, model inference, and rare object detection Figure 9.

There is wide diversity of data as a result of differences in geography, infrastructure, and architecture. These differences make it difficult to develop a single solution. Model inference remains a challenge due to the inability to train across all possible geographies. Extrapolation of models happens frequently in GeoAI and while it is sometimes necessary, it is risky. Detecting rare objects is difficult due to unbalanced samples and a broad area search. Suggested solutions to these challenges included human-in-the-loop strategies, active learning techniques, semi-supervised learning, and over/under sampling. Human-in-the-loop strategies emphasize the importance of humans being present in machine learning process via manual inspection to help verify and update workflows.

Challenges in scalability for GeoAI also extends to the usability of software and workflows. Building software that gives non-traditional users and non-experts the capability to generate insight from pixels requires a comprehensive look at data access, geospatial analytics, and procedures for developing actionable insights. Usability challenges outlined here include choosing the right hardware, parallelizing the embarrassingly parallel, and minimizing the number of applications by creating a streamlined replicable process. In conclusion, there is opportunity for scaling GeoAI to new users and new sectors by

*Gal, Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." International Conference on Machine Learning 2016 improving usability.

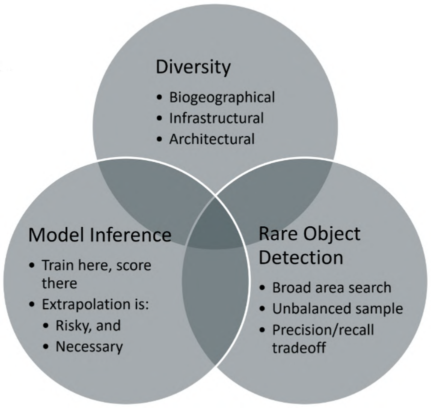


Figure 9. Challenges in generalization for GeoAI

Vipin Kumar, Regents Professor, Computer Science and Engineering Department, University of Minnesota

The trillion pixel GeoAI challenge is a result of the confluence of three different trending technologies. Kumar recognizes the technologies: 1) Geospatial Big Data, 2) AI/Machine Learning, 3) High-performance computing. The intersection of these technologies help answer societal relevant questions in a variety of sectors ranging from agriculture, water, forestry, and many others. The overarching challenge in generalization and scalability of ML and GeoAI is that remote sensing data can only provide an approximate and incomplete view of earth system processes that occur at multiple scales in space and time. Heterogeneity in space and time makes it difficult to transfer models built for a localized region and time. As an example, water bodies exhibit heterogeneity in space and time due to their varied appearance in different regions of the world and the fact that the same water body can look drastically different at different time-instances as shown in Figure 10. A second challenge is in data quality. Satellite data often suffer from missing data and labeling errors due to clouds, shadows, and atmospheric disturbances. These data quality issues cause problems in building robust models.

There is an opportunity to overcome these challenges through physics-guided ML methods which can use some of the underlying physical processes, laws, and process-level information to assist in the learning framework. Domain knowledge in this sense is key to building models that overcome generalization and scalability challenges. Several examples that leverage ML and process knowledge to overcome heterogeneity in space and time were presented, including mapping water bodies globally and then inferring their dynamics over space and time at a monthly scale. A second example is monitoring river width globally using deep learning. A final example is mapping of palm oil plantation dynamics in tropical forests between 2001 and 2014.

Dalton Lunga, R&D Scientist, ORNL

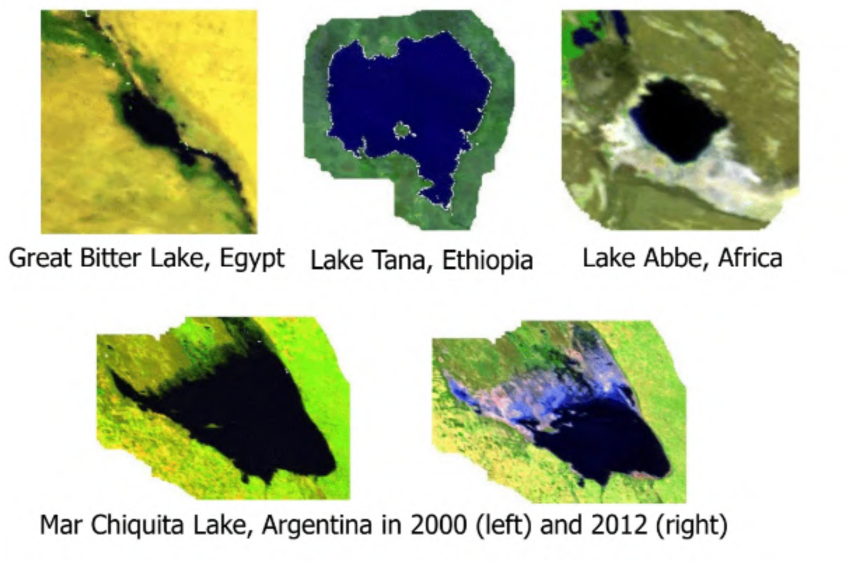

Great Bitter Lake, Egypt
Lake Tana, Ethiopia
Lake Abbe, Africa
Mar Chiquita Lake, Argentina in 2000 (left) and 2012 (right)

Figure 10. Heterogeneity in space and time of lakes

There are four broad Trillion Pixel Challenges that are often encountered in GeoAI: 1) generalization of GeoAI systems, 2) learning with limited labeled data, 3) scalability in computing, and 4) scalability in science. In this talk, Lunga highlights the primary problem with data driven approaches. Models often do not ’travel’ as far as deployment tasks requires. GeoAI for large-scale applications is often required to operate in heterogeneous environments where data characteristics can potentially be vastly different. To overcome these problems, we need to leverage methods that allow us to identify core features and capture the diversity of the data in terms of geography, ground features, and image characteristics. Often neglected, is defining the extent of object class type and their semantic definition across different geographies. For example, consider the visual difference in buildings or roads in different geographic regions. The structural/class definition of *what is a building or road* in Knoxville, United States of America differs to the physical appearance of building or roads in South Sudan? Designing models that can reconcile such structural inconsistencies is a relevant and hard problem that GeoAI systems should address.

A second challenge is learning with limited labeled data. It is expensive and laborious to collect labeled training data. Furthermore, most data annotation is carried out with a single task in mind. New opportunities in this direction include multi-task and semi-supervised methods. In this case, there are opportunities to share labeled data and learned features between tasks. An example of this is labeled data and the features learned for an object segmentation task can also be used in a regression task for counting in images.

The third challenge is in scalability in computing which recognizes that there is a significant performance bottleneck in efficiently training and iterating on a deep architecture for various applications. Variability in object size presents challenges in scalable computing (Figure 11). Opportunities are increasingly growing for algorithmic development and custom GeoAI computing hardware that is problem specific, leading to the scalability in science capabilities. Failure to handle multiscale aspects in a way that is suitable for the problem at hand will cause the most cutting-edge learning techniques and high performance computing platforms to struggle in finding an optimal model solution. An example of scalable computing limitations comes in terms of the memory required for processing images whose memory footprint when combined with that of given deep learning architecture does not fit within the capacity of current GPUs. The current workaround resorts to processing the entire image rather in small image patches or re-scaling the image size which distorts the original full resolution and amounts to degraded context information that maybe useful for estimating the effective receptive field for modeling larger size objects.

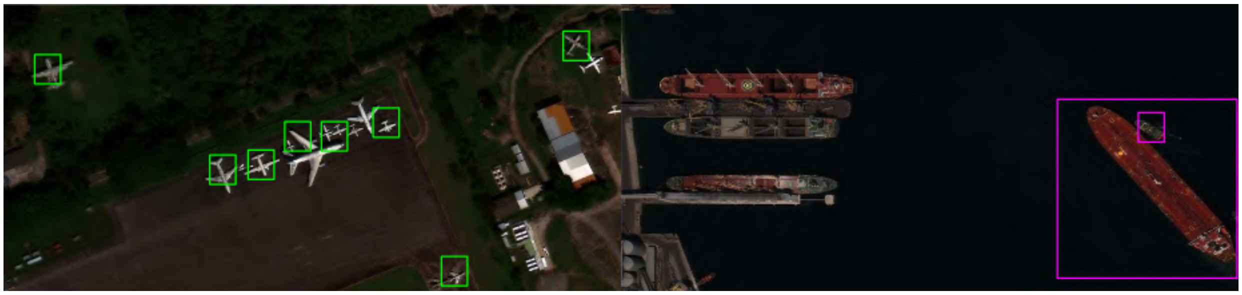


Figure 11. Objects of varying sizes*

Todd Myers, Senior Architect, National Geospatial Intelligence Agency (NGA)

The National Geospatial Intelligence Agency's (NGA) Data Analytic Architecture Service (NDAAS) is unifying efforts to leverage analytic capabilities within the agency, said Todd Myers. The way software developers are interacting with the environments needed a unifying process. Starting with the abstraction of orchestrated environments. However, there remain challenges in the way, including legacy systems e.g. pipelines, processes, controls are fully changed to meet a continuous integration and deployment (CICD) practice as in the private sector. In its efforts, NGA has developed a framework called Scale to Deploy on a distributed system with Docker for payloads that were originally written in C++, Matlab. These legacy systems required monolith systems to run. The ability to have one light process within the container was needed and Apache MESOS and NGA Scale to orchestrate the processing via containers. Orchestrating these resources whether on cloud or locally is a challenge and there is an understanding that current pipelines for NDAAS cloud native applications will need to evolve from the MESOS Gen 1 to Kubernetes Gen 2, which adheres to the Cloud Native Computing Foundation (CNCF) for future developments across the Enterprise of NGA.

Q&A

Question 1: What would be the one breakthrough in GeoAI that you are looking forward to?

Dalton Lunga: AI methods that are able to better leverage unlabeled data sets and achieve performance comparable to supervised methods within large scale applications.

Vipin Kumar: Being able to handle the heterogeneity of the large-scale problem. It is a breakthrough that will not occur in one fell swoop but instead, there will be small pieces that will fall into place along the

*image source: http://xviewdataset.org/

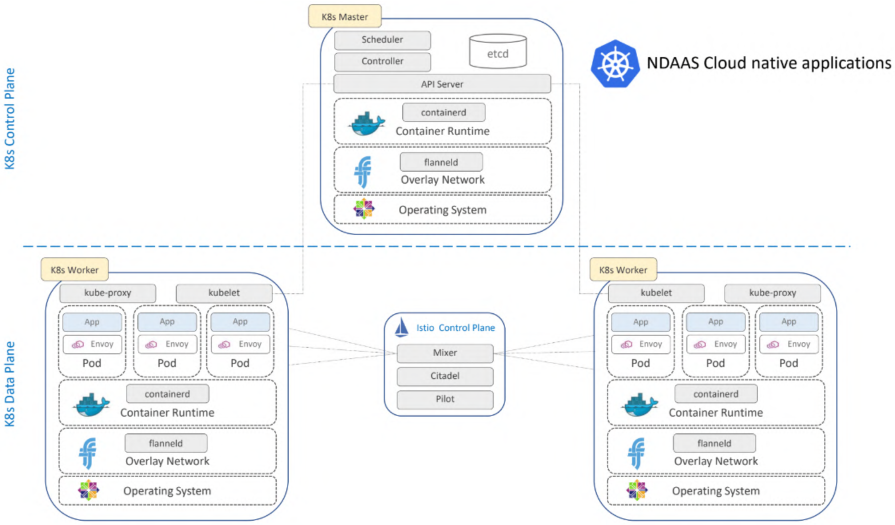


Figure 12. NGA's data analytic architecture service platform

way. To come back to the previous answer, many of the biggest advances and successes we have seen have come through using techniques with unlabeled data such as auto-encoders.

In a more general sense though, being able to incorporate our understanding of the physical processes into ML. That will be the key. We have to leverage existing science as opposed to replacing it. Learning how to incorporate this into ML would be a big breakthrough.

Todd Myers: I have two things I would like to see. The first one is an advancement in preprocessing on the actual collectors. I would like to see intelligent master-node execution in a unified fashion regardless of what you are executing.

Question 2: I want to understand more about the generalization problem like in the case of water bodies or roads in different geographic areas. What are your approaches to that problem? Do you take the model and feed it additional samples from the new area and create a model that can discriminate across all areas, or do you use some other approach? In general, is there any notion from a data infrastructure perspective actually managing taxonomies of models that are distributed geospatially for a particular problem so that when doing inference in a particular area, the appropriate model for that area is identified and used for that area.

Vipin Kumar: We found great success in incorporating process-level information into our models to accomplish global scale inference.

Dalton Lunga: How to address challenges of generalization. I'd like to go back and highlight the presentation given by Pete Doucette from USGS which gave a view of looking at the world in terms of grids. If there is a partitioning of the world into a grid and then defining a localized model for each cell perhaps we can begin developing models that work across different time steps. But then if things change within the grid over time, when or how do we know to re-adapt our models. Diversity within the grid cells will hamper the performance of the models.

1.4 SESSION 3: GEOSPATIAL DATA INFRASTRUCTURE

Moderators: Fabio Pacifici, Principal Scientist, Maxar Technologies and Yan Liu, R&D Scientist, ORNL

Massive geospatial datasets collected from survey, sensing, and social media provide rich data sources and geospatial context for GeoAI. Currently, these datasets have been largely unexplored by the machine learning community and AI research in geospatial data sciences community is still at an early stage of directly adopting existing machine learning frameworks. This session discussed major geospatial data challenges and opportunities for broad and specialized GeoAI RD, including difficulties in and impracticality/practicality of labeling massive geospatial datasets, the potential of leveraging existing rich data sources for GeoAI, HPC-based data-intensive AI computations, and, consequently, imminent and future needs for geospatial data infrastructure solutions that couple geospatial processes, data analytics, and AI as a scalable platform for empowering large-scale GeoAI applications. Planet scale GeoAI requires lots of ground referencing. Panel speakers for this session focused their presentations on data infrastructure challenges for GeoAI systems. Some of the key questions that where addressed included: How existing geospatial data infrastructure is being used to support GeoAI? What are key research and technological challenges that existing geospatial data infrastructure or new design must address in order to facilitate broader and large-scale GeoAI RD in our community? What are computational performance concerns when GeoAI is applied on large datasets?

Christopher Brown, Software Engineer, Google Earth Engine

Google earth engine gives the public free access to both data and computation resources. In terms of data catalog, Google earth engine provides surface temperature, terrain elevation, weather, and remote sensing imageries such as Sentinel, Landsat, MODIS, Terrain, Landcover, etc. Right now, those dataset covers more than forty years and is closer to 20 petabytes, and thousands of images are generated and added to Google Earth Engine's database every day. Regarding the computation, the user only needs to provide code using Python or JavaScript, and Google earth engine handles all the computation using its computing resources. Although, Google earth engine had already started to incorporate GeoAI into its platform, it is still relatively new to Google earth engine.

Brown's talk covered several examples including using GeoAI with Google earth engine to map human settlement using Landsat data. In the inference phase of this binary classification task, Landsat images were exported from Google earth engine and processed in Google AI Platform for prediction. To emphasize the efficiency, the platform only took 8 hours and 175 USD for an intern with a laptop to finish processing data of 40 gigapixels. Another user case was to create earth time lapse, which is a global, zoomable time-lapse video that allows users to view and explore changes to the Earth's surface from 1984 to 2018. To realize such a visualization application, Google processed 10 quadrillion pixels coming from Landsat and sentinel images collected each year from 1984 to 2018 to determine the existence of cloud for each pixel, and it only took four days using 20K CPUs. Currently, Google earth engine users employ Google AI platform for model training and inference, which is not free right now. However, In the future, Google earth engine not only plans to make its AI applications free, but also integrate more AI frustration into its platform like the AutoML.

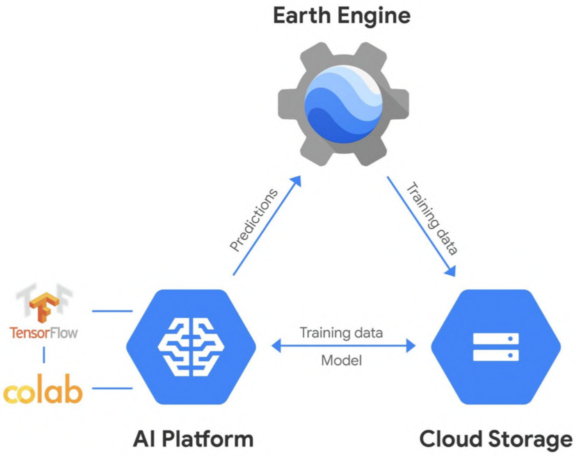


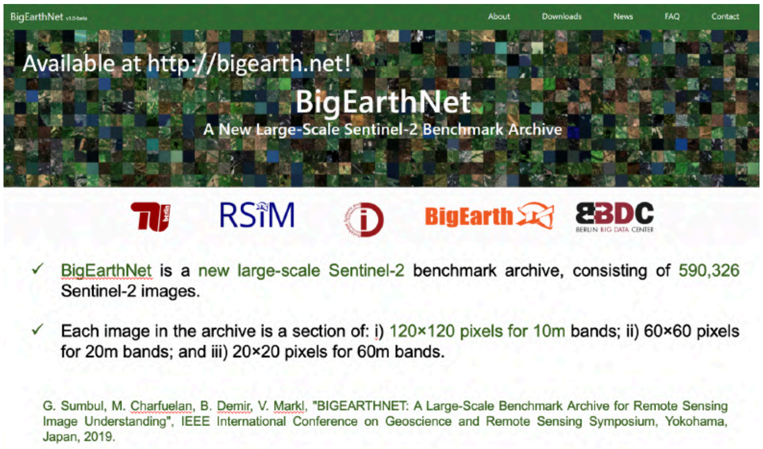
Figure 13. Google Earth Engine users utilize Google AI Platform and Cloud Storage for processing images

Begüm Demir, Professor and Head of the Remote Sensing Image Analysis Group, Technische Universität Berlin

Remote sensing images have been generated at an unprecedented speed due to the deployment of an increasing number of remote sensing sensors, resulting in a huge Earth Observation (EO) archive. Therefore, it becomes important to develop a scalable and accurate method to search and retrieve images from the EO archive in order to best manage and utilize the EO archive as suggested during Demir's talk. The BigEarth project which aims to develop tools to search images from the big EO archive in a scalable and fast way was presented. With this image search and retrieval tool, users can query the EO archive for images that contain a certain type of land cover feature, such as forest fire. In addition, it also allows the user to find temporal images that show the transition between different land cover features. Demir presented two tasks that are involved in such an image search and retrieval system. The first one being to learn a discriminative and robust EO data descriptor, and the second one is to map the data description output by the descriptor into a hashing code, such as a binary code that allows a fast search. Regarding the first task, Demir presented the limitations to remote sensing applications when using ImageNet pre-trained models. This is because satellite sensors have special characteristics in comparison with images used in the computer vision community.

Furthermore, images used by computer vision community usually has three bands, but remote sensing images often have more than three bands. Instead, they trained their model to learn descriptor using a remote sensing image archive called BigEarthNet. BigEarthNet is a new large-scale Sentinel-2 benchmark archive, consisting of 590,326 Sentinel-2 images. Each image in the archive is a section of: i) 120×120 pixels for 10m bands; ii) 60×60 pixels for 20m bands; and iii) 20×20 pixels for 60m bands. The number of labels associated with each image patch varies between 1 and 12, whereas 95% of patches have at most

5 multi-labels. Images acquired in different seasons are considered. Within 5 months after BigEarthNet was launched, it attracted 4000 users from around the world. Although use of Volunteered Geographic Information (VGI) might be one avenue to help conduct scalable applications, Demir shared some example limitations of VGI datasets - those often arise since that data can be noisy, incomplete, and redundant.



Available at http://bigearth.net!

BigEarthNet

A New Large-Scale Sentinel-2 Benchmark Archive

✓ BigEarthNet is a new large-scale Sentinel-2 benchmark archive, consisting of 590,326 Sentinel-2 images.

✓ Each image in the archive is a section of: i) 120×120 pixels for 10m bands; ii) 60×60 pixels for 20m bands; and iii) 20×20 pixels for 60m bands.

G. Sumbul, M. Charfuelan, B. Demir, V. Markl, "BIGEARTHNET: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding", IEEE International Conference on Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019.

Figure 14. Large volume of EO archive needs scalable data management

Lexie Yang, R&D Scientist, ORNL

Are we processing the Earth Observation (EO) fast enough?
To find the answer to this question, Yang first talked about her internal workflow at ORNL. To approach a large-scale project, the workflow first prototype a workable model in a small area. If the accuracy under monitoring is good enough for that small-scale dataset, then the algorithm is then tested on a larger scale using HPC or GPU clusters to evaluate if the model presents similar performance. Yang said, 10-15 years ago processing Landsat imagery covering a whole city could be termed as a large-scale project and it can be done fast and easy. However, as the sensors are becoming increasingly available and the meaning of "large scale" is transformed to country or even global level necessitating the design of efficient models. Yang acknowledges commonalities with workflows presented by other panelists, in particular, toward seeking a generalized image analytics workflow, which includes imagery acquisition, training sample collection, and model building and deployment. Several iterative steps may be required during the model building procedure to collect more samples for areas where models do not perform well. During the workflow model estimation, training is accelerated through distributed learning methods

To give an example of how fast it is for the current workflow, Yang used one project for Nigeria as an example. In this project, they processed 20,000 individual scenes, including 25 trillion pixels, covering 1,345,006 sq km, and it took only a few days to finish. In comparison with computation, Yang said, the training sample collection now became the bottleneck for accomplishing a large-scale project in a fast manner.

Yang’s concluded her talk with a suggestion to address the bottleneck incurred by training samples collection. She said, in the future exploiting the relationship between different sensors in order to reuse the training samples collected that are specific to a certain type of sensor should be the key.

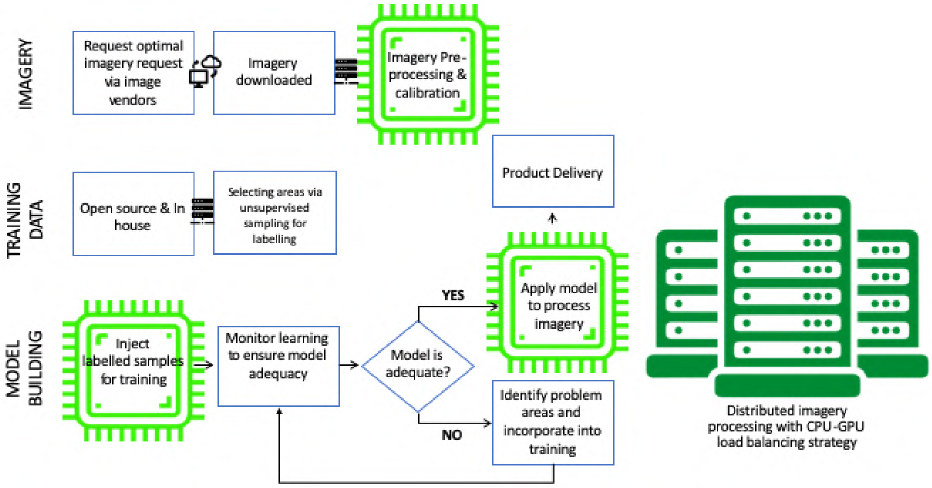

IMAGERY
Request optimal imagery request via image vendors
Imagery downloaded
Imagery Pre-processing & calibration
TRAINING DATA
Open source & In house
Selecting areas via unsupervised sampling for labelling
Product Delivery
MODEL BUILDING
Inject labelled samples for training
Monitor learning to ensure model adequacy
Model is adequate?
YES
NO
Apply model to process imagery
Identify problem areas and incorporate into training
Distributed imagery processing with CPU-GPU load balancing strategy

Figure 15. The generalized image processing workflow

Caitlin Kontigis, Applied Science Lead, Descartes Labs

Descartes lab is a Geospatial AI startup in Santa Fe, New Mexico. It was started by former employees from Los Alamos National Laboratory in 2014. They have grown from a startup of only 10 employees to a company that currently has 110 employees, with 25% of them having a PhD degree. In his presentation, Kontigis demonstrated company products to illustrate how Descartes Labs is dealing with EO dataset. Processing the EO data on a global level is challenging, since it requires the fast speed to process the large volume of data that has diverse characteristics. To address the challenge, Descartes Labs constructed a data refinery platform that allows data storage, data cleaning, training sample augmentation, model training, model testing, data analyzing, and result animation. Kontigis introduced a python package developed by her company. This python package allows the users to obtain the EO imagery given a region of interest and a period of time fast and easy.

Descartes Labs' business is broad with diverse customers and partners, among which are Cargill, DARPA, and the New Mexico government. Next, Kontigis demonstrated a series of use cases developed by her company. The first application is the boundary delineation cropland for corn and soybean using Rapid Eye Optical imagery in North Africa. The second application is land cover/land use mapping using high-resolution NAIP images, and the classes covers bare land, impervious, forest, shrub, grassland, and water in the downtown area. Kontigis discussed the subsidence mapping for Mexico City from 2016 to 2018, which enabled the estimation of the rate of subside. Finally, Kontigis showed the deforestation mapping, urban tree mapping, global nitrogen dioxide mapping, and U.S. methane concentrations mapping.

Sud Menon, Director, Software Product Development, Esri

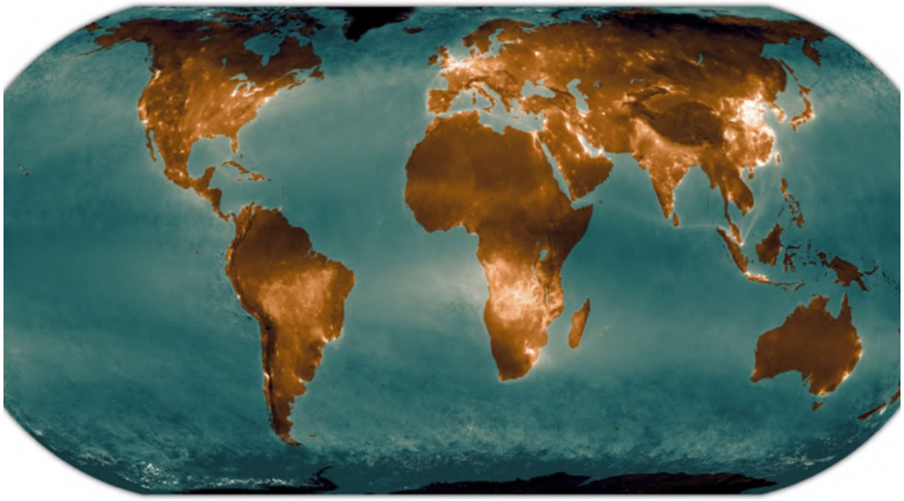


Figure 16. Mapping global nitrogen dioxide (NO2) using Sentinel-5P

The ArcGIS platform targets all the categories of geodata, including imagery, tabular, uninstructed, vector, 3D, and LiDAR. Menon's talk shared the overall goals of ESRI which include 1) making it easy for Geospatial Analysts new to DL as well as for experienced Data Scientists working with DL Engines to develop and train models against data in cloud based Geospatial Data Infrastructures, 2) making it easy for DL Model producers to share their trained geospatial DL models with those who need to use them, and 3) making it easy to deploy DL models for scaled out execution in cloud based Geospatial Compute and Data Infrastructures. Most of ESRI clients are interested in applying various computer vision models for the GIS and remote sensing tasks. For example, classification model can be used for damage assessment of building, object detection is useful for tree detection, instance segmentation is appropriate for building footprint extraction, etc. To satisfy those needs, ArcGIS created a flexible deep learning workflow that allows the user to collect training samples, export training data in standard DL formats, train DL model, and perform inference within or outside ArcGIS.

To make this workflow more convenient for the user, the ESRI integrates Python Notebooks into the Geospatial Data Infrastructure that allows the access to more than 275 data science libraries, all types of data, all the ArcGIS API for python and analytic servers. Menon highlighted the arcgis learn module in ArcGIS API for Python. This module makes good default choices regarding model architectures and conv net backbones for the different use cases. It leverages pre-trained models, transfer learning and automatic learning rate detection, allowing analysts with limited knowledge to easily adopt deep learning to their workflows. ESRI even provides more flexibilities for experienced data scientists. With Python Raster Function, the experienced data scientist can wrap the model developed by themselves in the form of json file and inject the model into the pixel block processing pipeline. The Python Raster Function now supports various DL frameworks such as tensorflow, CNTK, Pytorch, Keras. Next, Menon used oil well pad change detection and damaged structure detection as examples to show the capability of their workflow.

In addition to creating a convenient DL workflow for users, another goal of ESRI for building geospatial infrastructure is to develop backend geospatial infrastructure (as deployable Software, as SaaS) for image processing and raster analysis including management and storage of large image collections, and scalable computation for visualization and analytics. In this regard, Menon highlighted how they handle the processing and storage of mosaic datasets efficiently through a series of techniques.

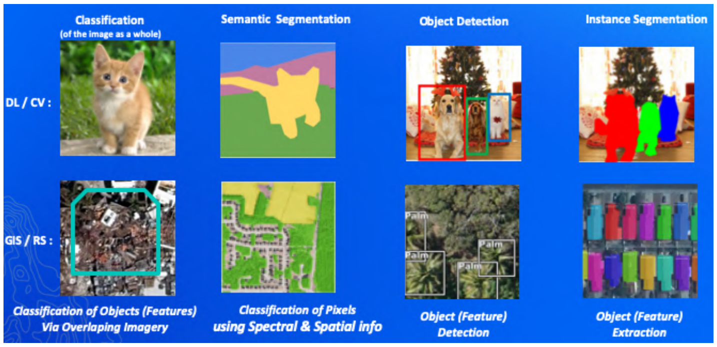


Figure 17. Applying computer vision models for GIS and remote sensing tasks

Q&A

Question 1: Given that deep learning model is data-hungry, it requires a large volume of labeled data for supervised model. One of satisfying the needs is to use OSM dataset. The question: what challenges did you encountered when you used OSM dataset and how did you handle those challenges?

Begüm Demir: When we used OSM dataset, we needed to be careful because the OSM may not be complete and label may not be accurate. In this case, we need to assign the confidence of accuracy for each OSM label and integrate that confidence value into the training procedure. The way you handle the uncertainty of OSM is also dependent on your applications. For example, if you try to classify forest and cropland, the uncertainty of OSM may not relevant.

Lexie Yang: The first challenge for us to use the OSM data is that the labelled data may not match the satellite imageries. This is because we usually only use the latest satellite images for mapping tasks, but OSM date was created using the relatively old images and not updated to the latest images. The other challenge is that for some regions especially in developing countries the OSM is not available. In that case, we have to rely on other resources.

Question 2: You mentioned different data types that you've been using in your research. One I have not heard of is the oblique images. I wanted to just hear a little bit about your experience with oblique aerial imagery, like have you applied oblique imagery and how is it improve your accuracy?
Caitlin Kontigis: The primary way we used oblique images is to calculate the height and volume.

Question 3: What about the feature extraction? Did you use the feature from nadir view and side views?

Caitlin Kontigis: We have not used that in our projects.

Question 4: This is the question for every in the panel. Can you talk about the efforts that you have done or the standards out there in the community that make it easy to share the data for different users?

Sud Menon: There are different levels of measurements to enable the interoperability. For example, all the geospatial data infrastructure is based on a web service interface. Anybody who makes HTTP request can access the data and the data is delivered to people in standard formats like jason format for a tabular.

Question 5: I have two questions. The first one is to how you decide your accuracy is good enough. The second one is to how you deal with situation where a fast response and a low latency is required.

Sud Menon: Before deep learning techniques, remote sensing people have accuracy assessment procedure to validate their model. You know all of that needs to be in place before you take action and what are your determining. I think in terms of the performance there's a whole range of different measurements people can take to improve the response speed. For example, people can process their data in parallel and train their model using GPU.

Question 6: Regarding the OSM data I want to add the comments here that Facebook can track back to which images were used to create the OSM with metadata. Then, you do not have mismatch problems. After training the model in a specific area and moving to another region, we augment the training samples and continue to train the model and the model got improved. Based on your experience, would model trained in this way have better performance than having individual models?

Lexie Yang: When we moved from one region to another, it is better to have pre-trained model. We are not sure if there exists a global model. That is an assumption that needs to be verified. Right now, we try to adopt a more intelligent method to collect samples when we moved from one region to another. We are trying to identify the most informative samples.

1.5 SESSION 4: EDGE COMPUTING FOR GEOAI

Moderators: Hector Santos-Villalobos, Research Group Leader, Multi-Modal Analytics and Architectures, ORNL and Sophie Voisin, R&D Scientist, ORNL

The cloud has become the ideal powerhouse for processing and storage of GIS information given that it outperforms the capabilities of at-the-edge devices. However, geospatial datasets are growing at an exponential rate, and there is an increased demand for real-time processing and storage that the bandwidth of current communication networks cannot match. Therefore, edge computing emerges as the solution for GIS processing needs in the not so distant future. Edge computing offers an evolving, energy-ecient, distributed processing and storage network that can facilitate real time, customizable information sharing and AI-based decision making.

The session focused on addressing the following key questions: why edge computing is key for GeoAI? What are and how to address edge security and reliability concerns? How to adapt existing devices and AI embedded technology for GIS use? Is 5G enough for GIS edge computing? How does the community envision edge-computing disrupting the GIS field? What are the most successful platforms to take deep learning to the edge?

Jay Theodore, Chief Technology Officer, Esri

Edge computing is a very vast topic and this talk addressed few things among those. Creating a digital twin that is closer to the physical world which can be extremely accurate and can give timely answers is the case of edge computing which Jay Theodore made in his talk. There are few challenges in seeking a consolidated edge based GeoAI system in conjunction with cloud computing platforms. First challenge is the Latency and it can be improved by off loading things from the cloud and put it on edge. Privacy, Governance, and Security are some other issues which can only be handled at the edge and not at the cloud. There is lot of data generated and not just asset data. It is also human movement data and machine data. There is a gap between the data being generated in couple of years and the capability of clouds to crunch the data. This gap is going to be filled by edge computing. Today a lot of this data is being discarded at the edge itself where it is being collected because it can't move up and find immediate use of that data.

Edge computing can mean many things to many people. Smarter devices are there, but the next thing is Edge servers. Edge servers and Edge portals are different in a sense based on who the consumer is. Edge servers are meant for machine to machine communications. These are fully automated systems providing services with no human interaction. Edge portals are basically representation of clouds contents like an emergency management system.

Theodore's talk contrasted with some examples on when to use edge computing and when to use the centralized cloud computing infrastructure. For massive analysis using historical data at cloud but specialized, focused and real-time analysis, GeoAI systems are mostly likely to benefit when computing with edge devices. Similarly, for large data analysis and inferencing GeoAI systems could benefit from the cloud but localized training and inferencing at edge because of localized data. Esri is using platforms like

Kubernetes and KubeEdge and hybrid as way of doing intelligence by utilizing cloud and edge with same or similar platform.

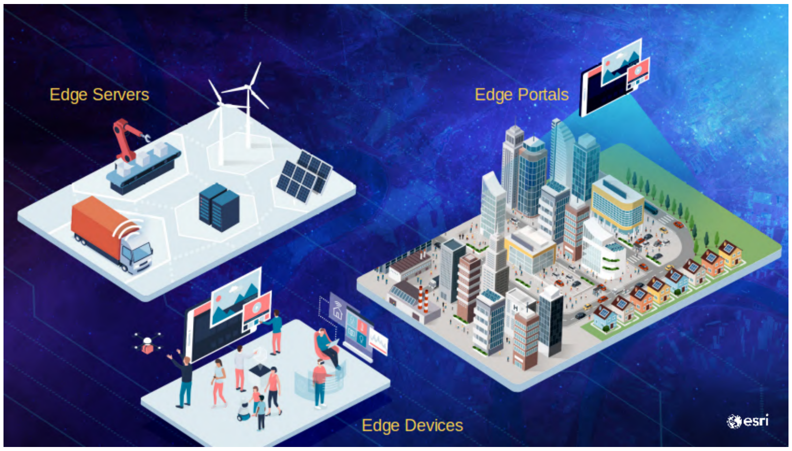


Figure 18. Edge computing at Esri

Himanshu Thapliyal, Assistant Professor, Electrical and Computer Engineering, University of Kentucky

IoT edge computing is affecting our lives in every possible way. We have IoT in many domains including retail, communications, connected vehicles, medical devices, industries. Adaptation of devices is growing very rapidly. There will be around 50 billion connected devices by the year 2020. The working frequency of these IoT devices change according to the applications. Furthermore IoT devices are battery operated, small in size and do transmission through wireless sensors. The major challenges involved with IoT edge devices are energy efficiency because of limited battery capacity and security as the system is on edge. One example of IoT security is cybersecurity of the vehicles. There is a need to develop hardware with better battery capabilities and which are resistant to security threats.

Other research activities being pursued by Thapliyal include energy recovery logic for building ultra-low power systems, solar cell and peizo based physical unclonable function which can provide security features to the edge devices and sensor based IoT edge security.

Jonathan Howe, Senior Solutions Architect, NVIDIA

Nvidia is among the key industry participants in deploying edge devices performing inference in imagery analytics. Some of the customers of Nvidia use Nvidia's edge devices to perform AI inference and parallelization of different types of algorithms such as for radio frequency comparison, DNA sequencing, aerospace and defense, smart cities, agriculture, retail, GeoAI, etc. Nvidia approaches these different applications with 'eating your own dog food' way to make sure all the underlying libraries, SDKs are as usable as possible and easy to use. For this Nvidia develops their own platforms and systems. One example being the

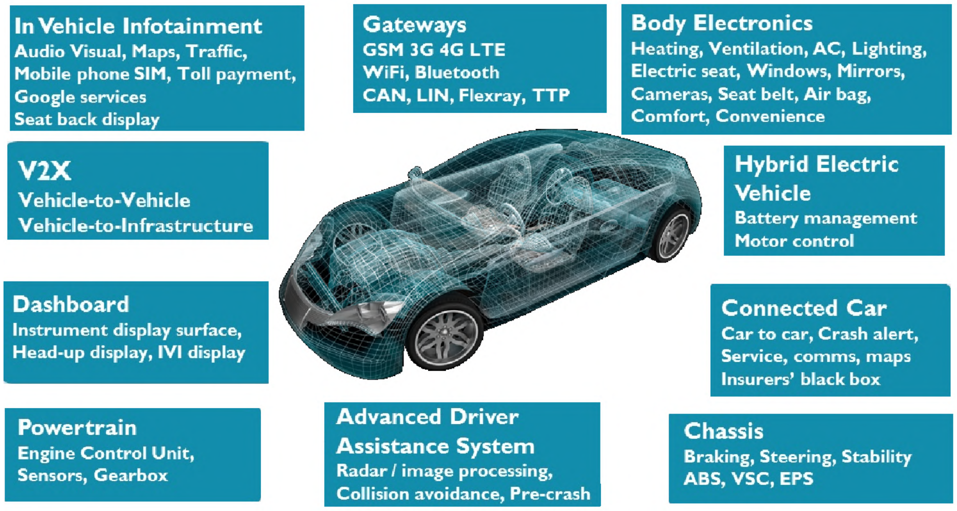

In Vehicle Infotainment
Audio Visual, Maps, Traffic, Mobile phone SIM, Toll payment, Google services
Seat back display
Gateways
GSM 3G 4G LTE
WiFi, Bluetooth
CAN, LIN, Flexray, TTP
Body Electronics
Heating, Ventilation, AC, Lighting, Electric seat, Windows, Mirrors, Cameras, Seat belt, Air bag, Comfort, Convenience
V2X
Vehicle-to-Vehicle
Vehicle-to-Infrastructure
Hybrid Electric Vehicle
Battery management
Motor control
Dashboard
Instrument display surface, Head-up display, IVI display
Connected Car
Car to car, Crash alert, Service, comms, maps
Insurers' black box
Powertrain
Engine Control Unit, Sensors, Gearbox
Advanced Driver Assistance System
Radar / image processing, Collision avoidance, Pre-crash
Chassis
Braking, Steering, Stability
ABS, VSC, EPS

Figure 19. Cybersecurity for vehicles

autonomous vehicles. Nvidia has a fleet of 30 to 50 autonomous vehicles across the world, driving every single day. The data collected is on the order of 1 PB per day significantly reaching beyond trillion pixels. For autonomous vehicles they have 16 different networks running at once and all feeding into the control system, and that control system is around 320 TeraOPS running in a single vehicle. Also, he gave examples of workflow of standard ecient video analytics running at the edge which takes around 30 TeraOPS and Drone Racing league, which is an autonomous drone racing so, lots of camera and other edge devices.

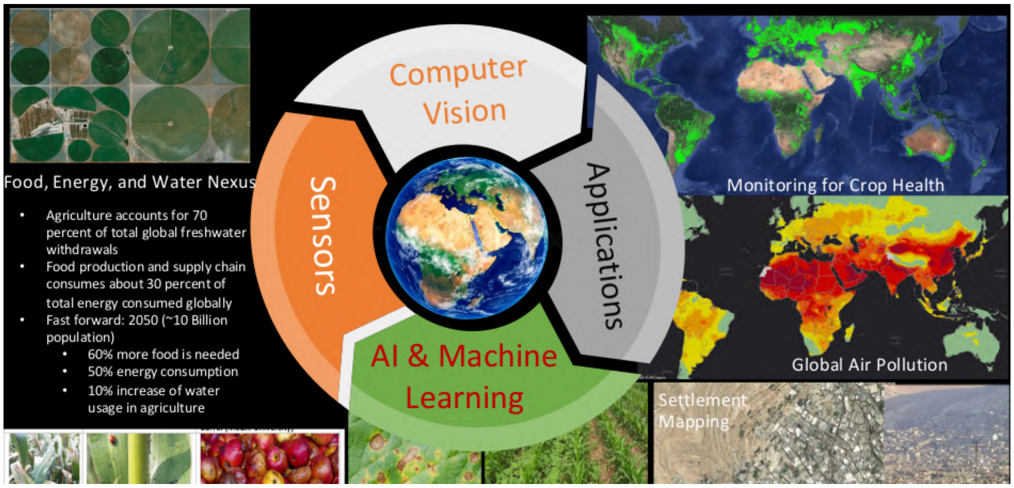
The presentation further highlighted Nvidia's JETSON family of edge devices like JETSON Nano, TX2 series and Xavier series, which have different processing capabilities and power usage. This Jetson family of devices comes with the underlying software stack which has various libraries like TensorRT, VisionWorks, CuBLAS and several multimedia and sensor drivers to import or stream the data directly into GPU memory which are required for several applications.

Figure 20. Drone Racing League

Raju Vatsavai, Associate Professor, Computer Science, NC State University

Spatio-temporal data computing is emerging and shaping the next generation of edge devices. The motivation being that data volume is increasingly growing and becoming too costly to move to the high-performance super-computing and cloud computing platforms. Informal estimates state that approximately 80 percent of all digital data is spatial and spatio-temporal. There is a need of technologies at the edge to deal with this volume of data. This data is not just the data through remote observations but also in-situ, for example in high tech agriculture equipment utilizing various sensors including cameras. With better edge computing we can do better and reduce the wastage of products like fertilizers. The third component is simulation, simulation model itself as an edge, each time step becomes specific amount of data and analytics can be done at that time step only. Edge computing can help many applications of national importance like Food, Energy and Water, Crop health, etc. Lots of sensors on field and UAVs are helpful in solving problems related to these applications on edge.

Vatsavai also made the case that edge computing is not going to replace traditional computing but there are certain applications like unsupervised data summarization and compression, where edge computing is very important. According to Gartner's report, today 91 percent of data is created and processed in decentralized data centers. By 2022 about 75 percent of all data will need analysis and action at edge.



Food, Energy, and Water Nexus

- Agriculture accounts for 70 percent of total global freshwater withdrawals
- Food production and supply chain consumes about 30 percent of total energy consumed globally
- Fast forward: 2050 (~10 Billion population)
    - 60% more food is needed
    - 50% energy consumption
    - 10% increase of water usage in agriculture

Figure 21. Applications of national importance

Ryan Kerekes, Group Leader, RF Communications and Intelligent Systems, Oak Ridge National Laboratory

ORNL is developing ground-based sensors and algorithms for object recognition at the edge. The goal is to do real-time data collection and data processing. The interesting problem is to do this on very low power budget so that the computer platform operates for years on battery or other power source like solar. The advantage here is that the sensor doesn't have to do recognition when there is no object and the object to be recognized is not going to be in field of view very often. One of the solutions Kerekes' group is looking at is 'Low power- Low duty cycle AI'. Data collection is carried out only when there is a need and stays asleep when there is no need. This can save a lot of power. Some of the challenges with 'Low power - Low duty cycle AI' are about implementation of sensor, time required to wake up the sensor, and amount of power consumption during sleep mode of sensor.

Kerekes's talk shared some of the edge platforms his group is using like NVIDIA Jetson, Google Coral Edge TPU, smartphones, and specifically into Intel Movidius Myriad. Intel Movidius Myriad is available as a USB stick and also embedded in cameras, for example Clear makes a camera called Firefly which streams data directly to Movidius which is a Vision Processing unit (VPU) which helps in applying algorithm directly to image screen from sensor. These are very low power systems and Movidius Myriad II claims to consume less than 500mW of power.

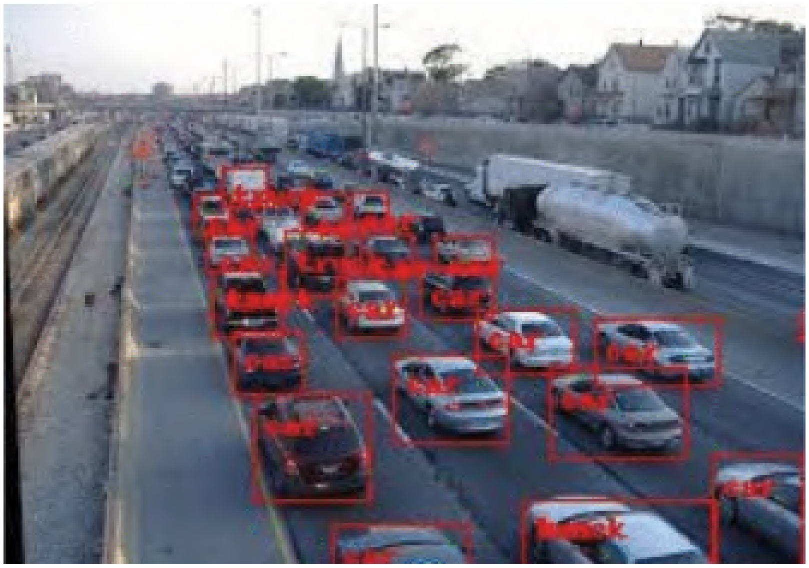


Figure 22. Real-time object detection

Nicola Ferrier, Senior Computer Scientist, Argonne National Laboratory

There is greater need for edge platforms that are modular, adaptable, and customizable because of the large quantity of data collection processing, which doesn't even require the 80 percent of the total data collected. 'Waggle' is an Argonne National Laboratory's edge computing platform which supports all current frameworks like tensorflow, pytorch, and caffe. It is powerful and capable of parallel computation at the edge for computer vision and deep learning frameworks. It also supports integrating sensors with plug-in architecture. Ferrier highlighted some of the projects where her team has utilized the Waggle. One of the projects mentioned is Array of things in the city of Chicago, which is a collaboration between University of Chicago, Argonne National Lab, and City of Chicago. In this project 'waggles' are deployed all over the city of Chicago for specific questions. Different groups in the community can request to address specific problems. Some of the problems are air quality, flooding, etc. By the end of this year 95 percent of the

Chicago residents will be in 2kms of a Waggle node and 75 of residents would be in 1km of a node. These nodes can be programmed to do different things such as pollution data. It also has two cameras on-board other sensors added by environmental scientists. This an open-ended system where the sensors and processors can be customized. This was deployed in field in different projects in drones, zoos, and trac flows.

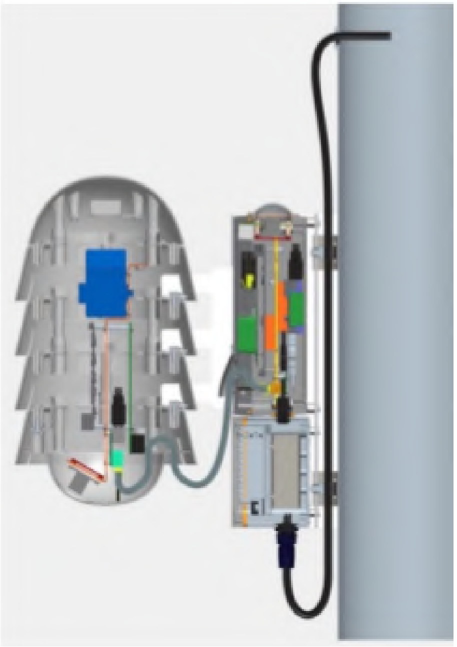


Figure 23. Waggle node: Argonne National Laboratory's edge computing platform

Q&A

Question 1: This is a question about the connectivity between edge and the centre to get contextual data which you need. Some of the data comes in from sensors which is used with models as part of that you need additional contextual data which you are not gathering and what if that data itself is periodically updated. The cadence of those is not at the edge, it might be at the centre. Are you thinking about, may be flows to get that data back to the edge.

Nicola Ferrier: We do have methods to update models running at the edge but at this point it is one of the big research questions we are asking. This is an open research questions where we are looking at how we go from edge to fog to cloud and other systems and push back.

Raju Vatsavai: It is application dependent. Right now 80-90 percent data is not even touched once. If an application require, not now, but may be in 10 years the entire dataset then ofcourse you have to store whole dataset but at the edge if you want to apply variational autoencoder , the encoding part can be done at the edge and reduce the resolution of data send to offline storage and regenerate full data. The issue here is privacy and security.

Kim Scott: This goes back to trust and generalization in models. Right now the state of the art in deep learning using earth observation just doesn't provide lot of trust in the results of the models. So the things you want to do is like ’Human in a loop’ or some other techniques to model where the uncertainty is. You need to keep that data so you can go back and see where you made that mistake.

Question 2: The question is for Jay. In your last slide you showed the real time map. How were you able to to something which has massive date and show that in real time? If you are using GPUs to offload any of these workload in the backend.

Jay Theodore: We used web sockets as a stream layer so a typical map itself can be made of many layers. One of them is to stream the data back to let’s just say to a browser to web sockets. SO it is a stream layer that can display on a map and then you just render that layer separately.

1.6 SESSION 5: HARDWARE DESIGN AND HIGH-PERFORMANCE COMPUTING FOR GEOAI

Moderator: Eric Shook, Assistant Professor, Department of Geography, Environment and Society, University of Minnesota

To address the growing number of global challenges using GeoAI will require a complex, high-performance computing (HPC) infrastructure to handle the processing, storage, and transfer of massive amounts of geospatial information. Whether in the cloud, in an HPC facility, or at-the-edge, these computing capabilities must adapt to meet the rapidly growing needs of GeoAI. This session explored the computational challenges facing GeoAI and discussed potential and holistic architectural solutions to overcome these challenges. These planet scale GeoAI architectures that are capable of making sense of trillions of pixels streaming daily will require tight integration of network, storage, and computing resources that span CPUs, GPUs, accelerators, and potentially domain-specific architectures such as a GeoAI spatial processors. The panel discussed promising architectural solutions and highlighted key challenges.

Several key questions were posed during this session. What are current architectural solutions, and will they scale to meet growing needs? Is there an "optimal architecture" that balances edge-computing, network bandwidth, storage, and high-performance and cloud computing? If so, what are its characteristics? How can accelerators and domain-specific architectures such as a GeoAI spatial processors address current and future challenges especially related to performance and energy eciency? What is the biggest challenge in hardware design and high-performance computing for GeoAI?

Tom Reed, Director, Solutions Architecture, NVIDIA

Given the growing need to seek solutions with trillion pixels, there is need for advanced hardware solutions in artificial intelligence (AI) research. Reed describes a "computational instrument" that would be to artificial intelligence research what the X-10 Graphite Reactor was to nuclear engineering. The research community intends to apply AI to some of the world's most significant challenges, such as climate, healthcare, and security. To overcome these challenges, new model architectures of increasing complexity have been introduced, and large quantities of data have been collected. Many iterations may be required to produce a useful model, so researchers need sucient computing resources to train models quickly.

The talk further shared factors leading to an increasing need for computational resources at the edge. In addition to small, low-powered devices, there is a role for higher-performance edge computing machines. He compared the proliferation of new machine learning architectures to the Cambrian explosion. Convolutional and recurrent neural networks have been joined by generative adversarial networks, reinforcement learning, and other kinds of architecture. Reed described three considerations that need to be addressed by new hardware. One is the exponential growth of available data, with tens of zettabytes already in existence and more on the way. The second is the rapid evolution of the AI field. Thousands of papers are submitted to NeurIPS and CVPR each year, so hardware needs sucient flexibility to handle whatever is the current state of the art. The third is the increasing complexity of deep learning models from a computational standpoint.

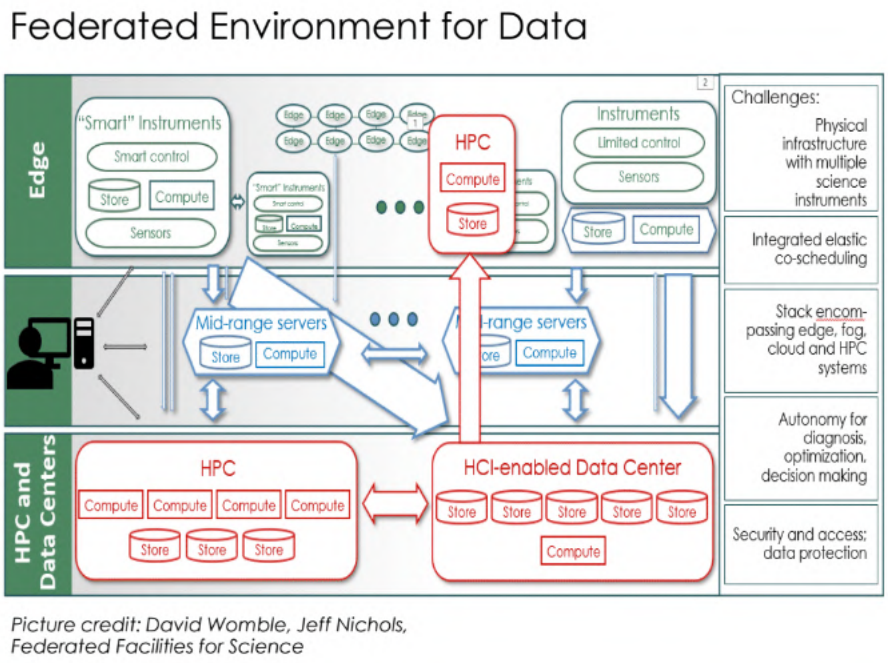
As an example of the sort of high-end edge device he is describing, he used the DGX SuperPOD. The DGX SuperPOD is an HPC architecture consisting of 64 DGX-2 Nodes joined with high-performance interconnects. DGX-2 nodes contain 16 Tesla V100 GPUs, so the system as a whole contains 1,024 GPUs. Mask R-CNN can be trained in 18.47 minutes using the whole system, or 1,024 instances of Mask R-CNN can be trained in 25.4 hours using each GPU separately.

Figure 24. NVIDIA's DGX superPOD

Mallikarjun Shankar, Group Leader, Advanced Data and Workflow, NCCS and CADES Director, ORNL

Shankar's talk provided a progression of architectures from Titan to Summit and Frontier. The talk described how many of the GPU-dependent applications that were written for Titan only made use of a node's 6 GB of GPU memory. This was because the memory bandwidth within a single GPU was much higher than the memory bandwidth across the node. Addressing this issue has been an important design consideration for subsequent machines. Summit features improved memory coherence across nodes, and this feature will carry over to Frontier as well. Near-node NVM storage is another feature included in Summit and continued in Frontier. He states that large amounts of memory, coherent memory access, and tiered data hierarchy are essential for processing data at scale. Some tension exists between general-purpose and domain-specific architectures. Abstractions and general-purpose systems are very popular in computer science, and it has seen success in many areas. Still, domain-specific architectures may have a role to play.

Shankar further discussed the structure of a typical federated data environment. There are three levels: the edge, the middleware or fog, and HPC or the cloud. These environments allow data to pass between edge devices and smart instruments, mid-range servers, and data centers or HPC systems. He notes that this sort of environment is very common in GeoAI and related fields, and that the edge will see an increasing number of smart instruments over time. There will also be debate about which edge systems should be kept lightweight, with processing ooaded to the fog or cloud. ORNL's Computing and Data Environment for Science (CADES) is an example of a federated data environment. He notes that Summit, based on ImageNet performance, could process a trillion pixels in around 30 seconds.



Federated Environment for Data

Challenges:

Physical infrastructure with multiple science instruments

Integrated elastic co-scheduling

Stack encompassing edge, fog, cloud and HPC systems

Autonomy for diagnosis, optimization, decision making

Security and access: data protection

Picture credit: David Womble, Jeff Nichols, Federated Facilities for Science

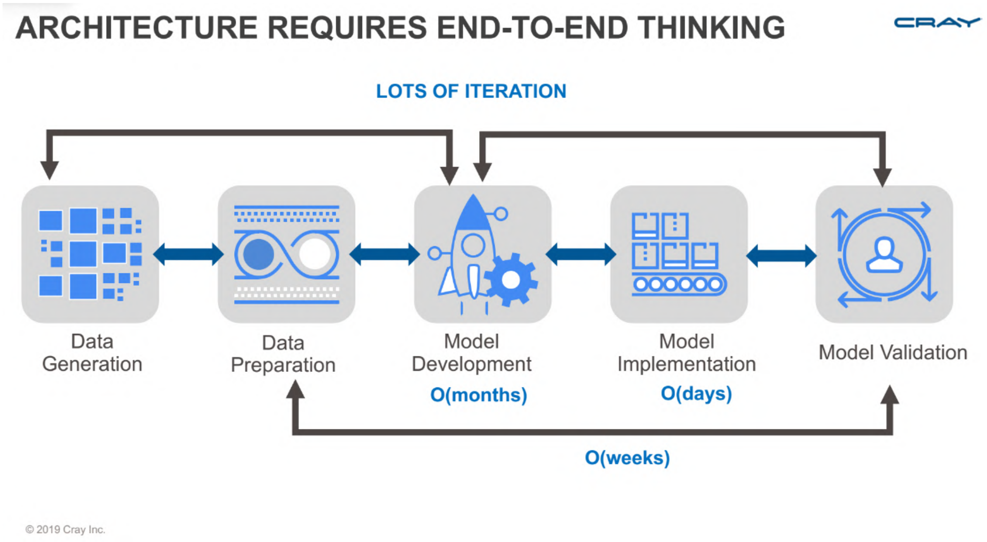
Figure 25. Federated environment for data

Rangan Sukumar, Senior Analytics Architect, CTO Office, Cray Inc.

Adapting a unique perspective, Sukumar describes the thinking behind Cray's current approach to architecture. The process of generating a production-ready machine learning model must be considered as an end-to-end workflow, from data generation and preparation to model development, implementation, and validation. Each step in this process may require iteration. The goal is to reduce the time needed to go from raw data to a working model. There is no one software stack or workflow that works for everyone, but there are commonalities between the life cycles of data sets. There is a process of maturation from raw data to an organized database. At each stage of this process, people need a means to organize, share, and analyze this data.

AI systems also have a cycle of maturity. Starting from the original problem formulation, the eventual goal is to have explainable intelligence. Different figures of merit are important at each stage of this process. Sukumar argues that the Shasta architecture can run a wide range of AI, modeling and simulation, and analytics workloads; support multiple processor architectures; and provide integrated storage. The software stack resembles a cloud environment. The architecture supports node and processor heterogeneity.

Past architectures were homogeneous. Workloads were primarily modeling and simulation, computing was performed on CPUs, storage consisted of hard disk drives (HDD), and high-performance computing interconnects were introduced. Today, many systems have complex, heterogeneous architectures. AI and advanced analytics workloads have become common. Systems often use both CPUs and GPUs for computing, and some include FPGAs. Primary storage incorporates burst buffers and flash storage, and standard Ethernet may be used alongside HPC interconnects.



ARCHITECTURE REQUIRES END-TO-END THINKING

LOTS OF ITERATION

Data Generation

Data Preparation

Model Development

O(months)

Model Implementation

O(days)

Model Validation

O(weeks)

© 2019 Cray Inc.

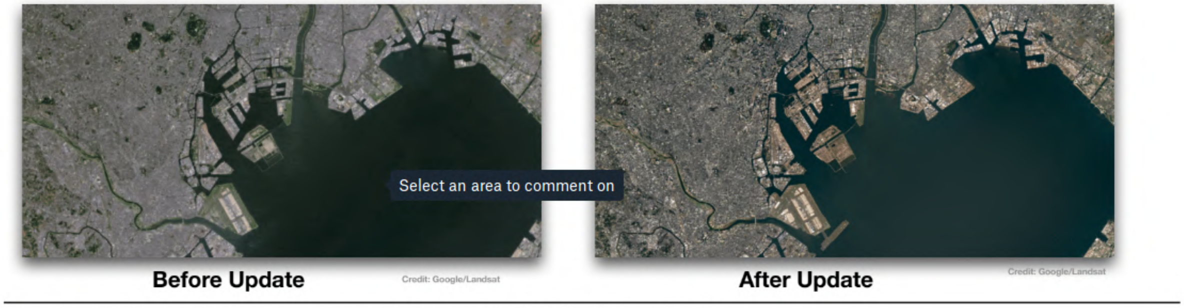
Figure 26. End-to-end architectural view

Frank Liu, Distinguished R&D Staff Member, ORNL

To provide an example of the challenges involved in working with trillion-pixel datasets, Frank's talk described the process Google used during its 2016 map update. This update incorporated data from Landsat 8. One petabyte of raw imagery, containing 700 trillion pixels, was processed into a final 1.5 terabyte product. A mosaicking algorithm was used to generate cloudless images. Six million CPU hours of computations were performed over a period of nearly one week using 43,000 GCP instances. Liu estimates that, using current prices, it would cost 355, 200 to repeat this data processing effort using on-demand virtual CPU instances, although this estimate does not include costs other than computing, such as networking and storage. Using NVIDIA Tesla T4 GPUs with an assumed 10Œ speedup, the price would be 570, 000.

Liu went to describe how increases in image resolution place demands on data storage. The surface area of Earth's dry land is around 150 106 km2. At one meter resolution, a raster dataset will have 150 1012 pixels and (assuming three 8-bit channels) require 450 terabytes (TB) of storage. This would require approximately 112 hard drives. Raster dataset size increases quadratic with resolution. Going to 0.1 meter resolution increases these figures by a factor of 100, and a 0.01 meter resolution dataset would have ten thousand times the number of pixels, and therefore file size, as a one meter resolution dataset.

Liu also gave a brief overview of computer memory hierarchy, from the fast, low-capacity processor registers to slow, high-capacity hard disk drives and tape storage. He notes that energy requirements for moving data from one level of this hierarchy to another can be substantially higher than the energy needed to perform computations on that data. He concludes with a final set of observations including that - geospatial computing is bound more by I/O concerns than by computing requirements. This means that factors such as data layout and indexing can be very influential on system performance. He recommends joint investigations at algorithmic, data, architecture, and system levels.



“Google’s satellite maps gets a 700-trillion-pixel makeover”, The Atlantic, June 27, 2016

Figure 27. Google’s 2016 map update

Ahmed Eldawy, Assistant Professor, Computer Science, University of California Riverside

Taking a different approach, Eldawy discussed the problem of trillion-pixel-scale GeoAI from a database perspective. Raster and vector data formats are both important to GeoAI applications, so methods to effectively combine these forms of data are vital. Borders are legal constructs, but their presence can impact the physical world. Differences in land use may be visible in satellite images, and differences in policy may affect deforestation or wildfire frequency. Information about national borders or administrative boundaries may serve as a useful supplement to training imagery for machine learning algorithms. Furthermore, the talk discussed some of the methods that may be used to combine raster and vector data. There are index-based algorithms that may prove useful. The data-transfer limitations of computer components must be taken into account. The talk concluded by highlighting libraries that could be useful to support storage-aware loading and processing, integration of vector and raster computation, and domain-specific computation patterns. Computing hardware might make better use of GPUs for demanding workloads, and FPGA-assistance may also prove beneficial. Opportunities in storage include near-disk filtering/aggregation; GPUDirect integration of GPUs and SSDs; and new file formats, such as cloud-oriented GeoTIFF.

Q&A

Question 1: What do you think about Cray’s HPC-AI roadmap?

Arjun Shankar: I’m more aligned with that than saying something should be fully domain specific. I like the idea that there’s an abstraction that can apply to multiple domain ideas, multiple domain needs. You want to find common abstractions as far up as you can take them and then repurpose them to specific other applications so that you can reuse them. So I’d be careful about thinking things should be domain specific and taking them all the way down, and so in those ways the effectiveness of GPUs across multiple domains is an example that supports this position. You can use a compute abstraction that applies to many different applications. So I’m for that, and we do a version of that in CADES.

Rangan Sukumar: What you want to look for in a system is not that one-size-fits-all, but you want to look for a system that allows you the flexibility to include whichever way you want to go afterwards. You want to be able to plan for growth, and tomorrow if there’s going to be a processor that’s specifically for GeoAI

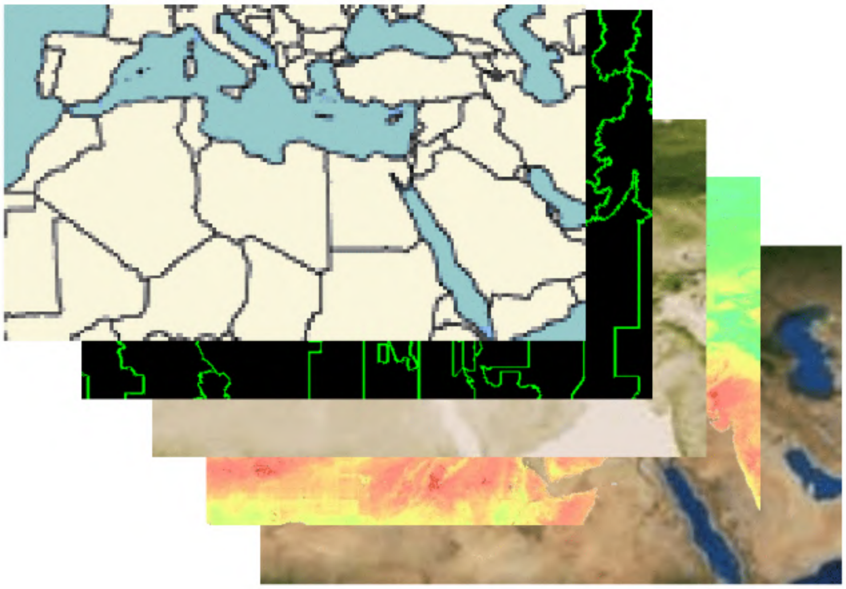


Figure 28. The effective combination of raster and vector data can provide useful insights

that needs to integrate storage, you don't need to move data, you just need to bring those processors to where the data is, and that can be one integrated system.

Question 2: Do you have any thoughts on Frontier doing the same thing?

Rangan Sukumar: I'll let Arjun comment on this one, but for Frontier, there are certain workloads that they prefer, and the architecture that you see in Frontier happened because there are certain workloads that Frontier was focused on. So in general, if you wanted a specific system for GeoAI workloads, you could architect a system that's uniquely the most performing system for that workload. So the way that I would answer that is that we think about it in four levels of hierarchy, and Frontier falls in the third or fourth bucket. The first one was "can we run HPC workloads and AI workloads on the same machine?" We were able to do that. "Can you run HPC and AI effectively?" Which means if you run an AI workload on an HPC system you're not being stupid, and if you run an HPC workload on an AI system you're not breaking the system. So we have to make sure that happens. And the third level you want to think about is "am I able to build a workflow that has both HPC and AI simultaneously in one system?" And I can build as many different workflows as possible. While I run multiple apps, my variation across apps is not as much. So with Frontier, the architecture that you see comes from that flavor. So I have fifty apps that people are going to run on this machine. Those fifty apps shouldn't vary across each other in terms of performance, and that needs to be bounded. The last level is where the specialized architectures, the domain specific processors, come in. Where you start thinking about "can I build one system that's going to be the most performant system?" So if you think about the Livermore announcement of El Capitan, that's one code that's going to be running at scale on that one machine. So that machine is architected for that one specific code. So if you want to think about how flexibility can be different knobs that you can tune one to the other and so forth.

Sud Menon: First of all, thanks for some really great presentations, really stimulating, I think Rangan you talked about the representation issue with respect to the input, like pixels are different than point clouds are different than meshes. I think that really resonated. You know we spend a lot of time talking about pixels, because that was the goal of the presentation. That whole thing with point clouds, photogrammetrically-generated point clouds as well as point clouds that are coming in from lidar, and people working with them. It's really a very important part. And in general, data science using there's the world of deep learning, but then there's all the other machine learning techniques, that can do things in an integrative way across all of these different types. That's really the larger challenge of GeoAI that we didn't get into. So I think thank you Ahmed for really calling that out. That's some context that we need to bring to this, and sort of zoomed in to this other aspect. So I think that really resonated, too. One sort of sense on that is yes, there's trying to do things at a global scale, where yes, you could use all of those things to guide what the machine needs to learn. Another perspective is, there are those kinds of global things, but there's a lot of people doing things locally, and the assumption is that that's being done where that contextual data is guiding the application of models locally. People engage in that kind of activity too. And that was sort of where the question yesterday about this global model, one global model to rule them all, versus local models that are indexed by space, was coming from.

The other thing I just wanted to say is that your last picture that just showed all of those layers in the stack and where things might happen. I would like to share my perspective on trying to do that, right, like in ArcGIS. So I'll just go back a long time, right, when people were doing things on the desktop. We had this system called ArcView, we had a scripting language called Avenue, and we tried to give people access to raster data, to vector data, and a set of functions that you could programmatically use to combine them together. And people did interesting things. We're still sort of trying to do the same thing now, except that the data is in the cloud. You know, you have point clouds, you have raster datasets. Of course, you have the tabular and vector datasets. Trying to build a Python based approach that has a somewhat uniform way of scripting, that resonated. People need to work with those things, they need to be well-indexed and well-accessible using space as the way to bring data together. I just wanted to share that. And I think, Frank, your comment that things are really I/O bound, I think that's really the reality of Geo. I think the thing I'd add there is that's why this co-location of compute and data is so important. And it's not just the image compute with the image data, it's the full geo compute with all the geo data that needs to be brought to bear on the problem. So, I just wanted to share those things.

Question 3: So today I can't fit my images, for training, in a GPU. You had a line that mentioned that as the area grows, to provide more context for the objects of interest, so does the targeted receptive field. What do I do? Do I get creative with the software side to bypass the footprint limitation that I'm having on the GPUs, or do I now look at you guys in the industry in terms of how you can come up with new hardware architectures or chips to accommodate such large images?

Frank Liu: I think the challenge is the amount of data required to make a decision. If you look at autonomous cars, if I want to do instance segmentation, I have one frame, I can work on it. I can make a decision right there, where are the pedestrians, where are the cars. No if you go back to the very simple example I gave you, to detect a vehicle, the receptive field is, let's say, 100 by 100. No matter how bi the area is, this is an embarrassingly parallel workload you can throw at a Spark cluster. It's well known. Now if you look the other way around at the other network example, that's much more challenging because you hypothetically can have a few hundred gigabytes of data you have to put in. Computationally, you need GPU or CPU or whatever it is in the future. I'm not sure everything can be done on the hardware side. To me, one natural way to look at it is how to index geospatial data in a very nice way. In database science, people worked for decades on how to do this, and companies go out to build special hardware, not necessarily chips, to help to do this indexing. So I don't know if the research has been conducted on how to do this effective indexing. I think yesterday there was this from Berlin, and she mentioned some hashing, which is one way to look at indexing. There are probably many other ways. So the question is you have this high-level task I want to ask. So you'd say how many cars are there, what's happening on the global scale at each grid? How do you translate that high-level task into a language that we can, from a hardware perspective, optimize on? I'm not sure it's there yet. I haven't seen this, I haven't done enough research on this. If you look at the HPC side, there are thirteen typical workload kernels. All of the tasks can break down into one of the thirteen, most of them at least. I don't know if it's something we should look at from the GeoAI perspective, whether we should do something similar to this. A special kernel, or maybe a few kernels. Not necessarily in computing, but maybe also include indexing as well. I think it probably makes sense to look at all these different things together.

Tom Reed: I'm not sure I know much more to add to that other than I do believe we've reached a point, and I'll just address the computer science because that's nearly beyond my scope of knowledge anyway. In that sense, we need to be way more precise with the utilization of the memory systems that we have. Think of it like a carbon tax that we pay for data. Both from data at rest but also data in motion and then data in the high-speed storage systems. If we did think about how many bits do I really need, and put that into our calculus of the questions we are asking, and only use those, or try to be conscious of what our use footprint is of memory, we'll do better. We will. I think it's kind of been proven out to do so. Now I know that throws a lot of burden on an already overburdened population of intellectuals and scientists, but I don't know that there's another really reasonable alternative towards the future. Systems will get bigger; memories will get larger, too. We are continuing to scale that up. But I think the scale of problem that you're on, we're going to have to find a way to compress that data down into a smaller footprint.

Rangan Sukumar: I'll try to give you a slightly cheeky answer. I know you're trying to find something to work for tomorrow, right? So if you have time for a year and a half, I think you'll have solutions that are coming out from all kinds of people. So far you've seen the cloud providers driving what is required, and services, and what's going into the toolkits and so forth. I think the contributions of what HPC could do to AI haven't come into production yet. For example, you're bringing up how "hey, my imagery doesn't fit into a GPU," we've been doing model parallelism in HPC for a long time. So if you can borrow some of those concepts to say "I want to have 4K images, I want to have 8K images, I want to have rasters that're going to be bigger than what memory can hold." If you think about ways in which, if you were to express that in Python, where you're saying my matrix itself can be distributed, and you're able to write your code that works on a HPC system out on the cloud in Python it's off and it's more performant. It's amazing. So now you're able to do what you can't with implicit model parallelism. But those kinds of concepts are still cooking. It's not production ready yet, but you're going to start seeing some of that concept coming up. So this is what I think is an opportunity for the domain-specific thing that you had discussions about. So for geo especially, the problem is you'll never be happy with the hardware you have. It'll always be high-resolution, it'll always be bigger than what you have in terms of hardware. So I think where you'll have to start thinking about it is, when I go from a problem, a point of resolution A to a mesh to a point cloud and so forth, what is my domain-specific requirement that's going to interface with these model. I think if we had that abstraction, we could implement some of those and say "we can can represent a point cloud in a distributed way in the following ways." We can build a parallel search system for point clouds. We can build a parallel search system for meshes. That has not happened yet. This where I think we'd talk more, if we had those kind of requirements. I'd be happy to take on the challenge from you guys.

Question 4: In matters of national security, early is on-time, on-time is late, and late is unacceptable. So we really have to be careful about computing in near real-time or in real-time systems. If you all had to add one slide that captures the real-time architecture side of things, what are the aspects that

you would like to add?

Tom Reed: I think real time is defined differently in many contexts. If you think of speech processing and telecom audio, you're talking about human response real-time requirements. That's on the order of tens of milliseconds. When you are talking about decisions at the edge, you might be talking about lower time frequencies. In those ways it's application specific. If I were to put one slide, it would be going from the timelyness requirements to defining the network requirements and the compute requirements.

Rangan Sukumar: I agree with the previous answer, and I just want to add another way to solve this problem. You can keep trying to think how to scale up the hardware and processing techniques to further speed this up, but you'll always hit a limit. There is a limit for how much data you can grab from the network, there is a limit for how much data you can put in memory, there are all these limits that you cannot go beyond. One other approach is to think about approximate computer processing. Instead of having the most accurate answer in maybe a minute (if you cannot wait for a minute and just have to have something in five seconds), maybe you can get away with an approximate answer. There is some work in approximate computer processing, even though for satellite data it didn't see much of this because the way it's stored you have to access the data. There are just simple ways of downsampling the data and getting an approximate computer processor, but I think there can be something that can be done better than this. One approach, for example, is GPUDirect or smart SSDs, where you can actually push some of the processes down to the SSD itself. So you can break the limit of SSDs, the rate that you can read from SSDs, by pushing some kinds of simple filters down to the SSD itself. You can do the processing near where the data is stored, maybe reduce the size a little bit, so that you can do more processing on the CPU or GPU at the higher levels. So approximate computer processing in this manner can help in getting something in five seconds rather than getting the best answer in a few minutes.

Frank Liu: I think a lot of existing technologies, HPMs, SSDs, flash networks, GPUs, so I won't add anything new. I would minimize the data required to make that decision. That can happen in a smart way to index data, so if you know you're interested in this particular grid, I can very quickly grab that data instead of going through the whole dataset. I think that includes not only the software, but hardware solutions to make this happen. It happened with database science in the past, and I think with some research we probably can do that too. That's just one chart.

Rangan Sukumar: It's hard, because everyone said the things I want to say. If I were to pick one thing, I would say I'd pick a library of models that you guys are going to be using on the edge. Because anything that's on the edge is a function of your model size, your sample size, and the architecture you can support. So if I had to add one slide to my deck, I would say given all these popular models that are being used in GeoAI, and given a sample size input that's going to come in, what is the latency I can expect from a single sample throughput on that model? What will happen if I have model multitendency? What will happen if I have model streaming? We heard yesterday that one model will not be enough; we need an ensemble of models if we are to make good decisions. I would have thought processes built around "how do I think that through on the edge?" So, do I have one big model that does it all, do I have multiple models that stream through fast? I'd have all of that analysis on my deck.

Tom Reed: I just want to really briefly say thank you for raising the idea that there is a fourth dimension here we really need to care about, and that is temporal. And so absolutely in every way I view this as a spatial-temporal issue and it should always be considered a temporal issue as well.

Question 5: What's your idea of a GeoAI spatial processor, and how do you see it benefiting or even

closing some of the gaps that we learned of today?

Eric Shook: Well, I sit on the domain-specific architecture side. I think that we have a lot of spatial data, and I fully agree, in terms of the generalizability of the hardware. I do think that there is a trade-off that was hinted at with the panel in terms of the data-intensive and I/O-intensive aspects of Geo. We tend to push though a lot of data and not have a lot of compute cycles applied to each pixel or polygon or point. But at the same time you have the AI pieces which are very compute heavy, so trying to find a balanced architecture is actually quite challenging, which is why we have the multiple architectures and why so much thought has to go into it. One of the benefits of GPU when it came out was that it allows you to have some of this balance, but getting the data into the GPU and these sorts of things is always a bit of a bottleneck, especially for geo-applications. So I sit on the domain-specific architecture side. I think there's a market for it, I think it's niche compared to regularized GPUs, for sure, but my suspicion is that if you built something like that, it would have a lot of domain applications in other areas.

Tom Reed: I fully support the idea. I do think that this is a software and hardware solution, a combo, and I don't think I know yet, maybe you do, where does that dividing line go. If you push it too far down too soon, you're going to find that innovation is going to be hampered, not helped, because the rate at which you can innovate at hardware is on a completely different time scale from software. I would argue let's do it, let's start by innovating the software down absolutely as far as we can possibly go, and then that gives us the motif and understanding to understand how to most appropriately build that hardware in cost-effective, generalizable form.

Arjun Shankar: I fully agree, as you saw when I spoke. I do want to just say also that what happens in geo-data is that there is a lot of sensor information that is directly relevant to our human experience. So the diversity of sensor deployments and the collection of data is very palpable for this community and for us to understand. But data across the science domains there are great similarities and it being an I/O problem, every data-driven discovery has an I/O bottleneck in it. So in that sense, you have a lot to learn and share with other communities. Bioscience, for instance, is huge in terms of data problems. They way they solve it is with some domain-specific chips, but they solve it a lot in software. With data structures, better representations of multi-modal data, how to tile the data better and to look for boundaries. You bring in other information. So I think it is a software and hardware problem. Software should move faster, obviously, because it can.

Frank Liu: If we look at the pixel level, from the workload perspective, at GeoAI versus traditional computer vision. At this moment I think they're very close. I think the question, to me, is if you want to make hardware, as Tom said, it's a very lengthy, expensive process. I would look at what's a real difference between the geo side, which has much bigger datasets, much bigger required memory footprint, than the traditional computer vision application. And maybe at the end of the day, something like a beefed-up GPU, maybe that's what we need. Or maybe we really have to look outside the box to something completely wacky, like people looking at coarsley reconfigurable blocks, for example. That's what I can think of.

Rangan Sukumar: As a technology geek and academic, love it, bring it on. As a person in the CTO office, I would say I haven't heard enough from folks in this room that say "I have this workload, I need the latency optimized." I asked this question specifically yesterday, and I had conversations afterwards. I haven't heard enough from the folks in the room that says "I have this limit, I need to make it 10X faster." Or 100X faster. So until you have that, you still have a niche market.

Ahmed Eldawy: I like the idea of having something domain-specific. I'm also seeing, from all the talks yesterday and today, in many applications we are bound by the capability of the hardware, whether it’s CPU, GPU, or whatever. We are reaching the limits of these, so it might be a good idea to have domain-specific hardware. However, unless we know exactly what we want to do, I wouldn’t really push for this. It’s a nice idea, but what exactly do we need? We need to further think of what exactly we need this hardware to do to go beyond the limits we have right now, and then after that it could be a good idea. So far I don’t really see what specific hardware that we need that will boost all this processing up.

1.7 SESSION 6: GEOAI OPPORTUNITIES - COLLABORATION AND PARTNERSHIPS

Moderator: Robert Stewart, Research Group Leader, Geographic Data Sciences, ORNL

Earth Observations (EO) provide invaluable big datasets over large spatio-temporal scales for monitoring the Earth and its changing environment. Throughout the last decade, along with the advancements in data-driven techniques, many AI-based algorithms have been developed for EO. However, these data are still underutilized and have great potential to impact global development. Several key questions were posed during this session including: What are the existing barriers to scale the scope of geospatial research problems? How can we attract more talent (e.g. students) from the data science community to work on geospatial problems? What are the opportunities for generating new data or models from existing EO? Are there high priority areas (applications or tools) that are under pursued? If so, how can we raise awareness on their importance across the community?

Matej Batic, Team Leader, Earth Observation Research, Sinergise

How big data is challenging in every domain and how the non-EO and EO experts are adapting to different approaches in handling data remains a topic of interest to many societal needs. Data with much higher resolution is now accessible and acquired at higher re-visits rates. Unlike in the past, EO big data doesn't sound that exciting, mostly because it creates a lot of chaos in the community. Non-research experts surely struggle with the complex nature of the data; however the experts are struggling too, and they find data preparation tasks tedious and costly at the same time. Batic then explained how the community can try to solve such problems, and not alone for the EO data, but for any general big data. Batic showed Sentinel Hub services that facilitate the uptake of the data, taking care of all the new and old issues, say projections and map corrections. Sentinel Playground, built on top of the Sentinel Hub, displays how to run the pixel-based JavaScript calculations on data to retrieve added value products like NDWI and others. Taking a step ahead, why not contribute your custom script to a repository to share with the community? With Sentinel-Hub it is now easy to get the data; the missing part is how to work with the data. EO-learn is an open source collection of modules that allow bridging the data into ML environment. Collection of tools, both for raster and vector data, static and temporal, on your own local computer or on the bigger infra on cloud or prem. EO-learn is also the backbone of Bluedot, providing global archive of surface area of waterbodies, retrieved from Sentinel-2. Such global monitoring, if done properly, can be very cheap.

Many institutions contribute to the EO-learn framework, both partners from European Commission and European Space Agency projects, as well as companies using EO-learn in their pipelines. Typical workflow retrieves satellite data directly from Sentinel Hub, while some publicly available (vector) ground-truth data are made available through Geopedia. When users are not able to find the ground-truth data, the crowd sourced classification app for labelling of satellite data can be helpful. The lessons learned working with satellite data and ML, remote sensing is not computer vision, so physical models should often be used instead of deep learning. Ground truth data is critical in Geo domain, and problems we are facing are not enough ground truth data, bad labeled data, and limited good data. We must understand that any ground truth data are usually just snapshots in time, and how to fuse these data is still largely unknown. Batic ended his talk by throwing an important question about the benchmarks to the audience: Is the user happy with the results?

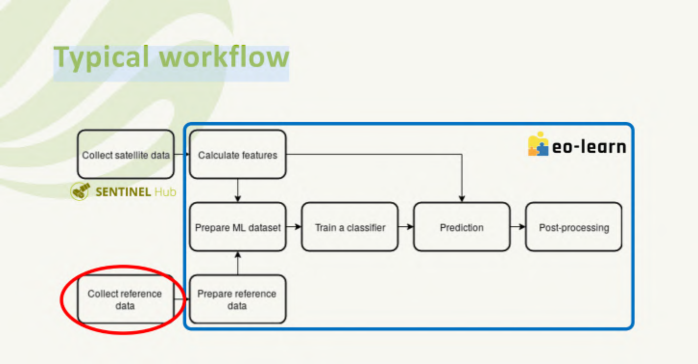


Figure 29. Typical EO-learn workflow with satellite data retrieved from Sentinel Hub services

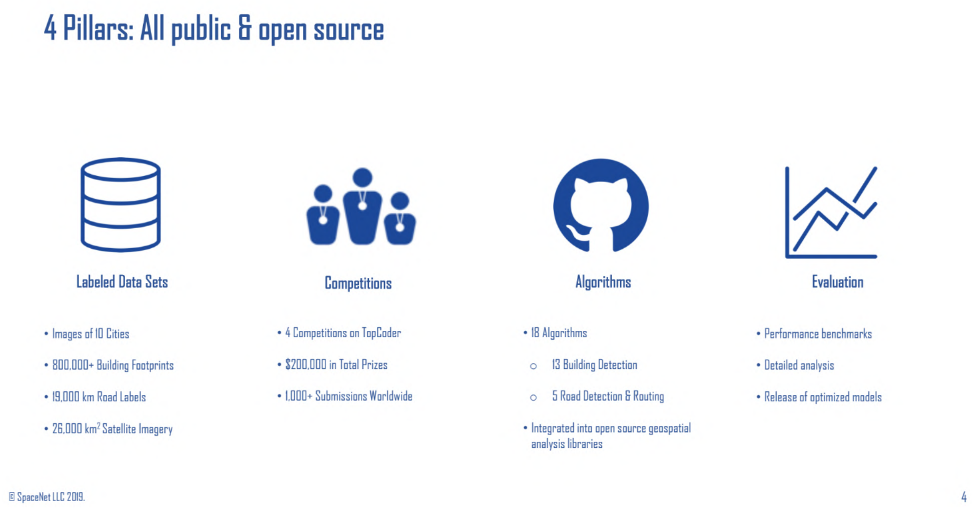
Nick Weir, Senior Data Scientist, In-Q-Tel

Introducing the SpaceNet Challenge, Weir is a Senior Data Scientist at CosmiQ Works, a geospatial focused lab under In-Q-Tel serving the US government and open source geospatial community. CosmiQ Works manages the SpaceNet LLC, a not-for-profit partnership aimed at advancing geospatial AI applications through its four pillars (shown in Figure 30): (1) releasing open source labeled datasets, (2) running public data science challenges against the datasets, (3) releasing challenge-winning algorithms to the public, and (4) providing robust evaluation metrics for geospatial AI models. SpaceNet provides a set of very high-resolution satellite imagery open datasets with high-quality labels.

Weir describes the history of SpaceNet and its activities that are advancing the computer vision applications for geospatial, including what lessons the team has learned with time. The talk highlights challenges with generating high-quality labels, and limitations to simply using OpenStreetMap data. Machine learning models need to have very high quality labels for training data, and he highlights others' work showing that if you reduce label quality to what OSM has, performance declines. Open source datasets enable sharing solutions from one organization to another. SpaceNet has a clear mission and through partnership is helping the remote sensing community to bring in different experts and aims to grow more.

Daniel Getman, Director, Solutions and Market Development, Maxar Technologies

Creating value from EO data is a process that includes data generation, algorithmic model development for machine learning, data product generation, human capital development to maintain data science capable talent, and a knowledgeable customer. Getman's talk presented perspectives on these elements while acknowledging shortfalls in these areas as a limiting factor in advancing GeoAI. For example, the lack of mechanisms to transform



4 Pillars: All public & open source

Labeled Data Sets

- Images of 10 Cities
- 800,000+ Building Footprints
- 19,000 km Road Labels
- 26,000 km^2 Satellite Imagery

Competitions

- 4 Competitions on TopCoder
- $200,000 in Total Prizes
- 1,000+ Submissions Worldwide

Algorithms

- 18 Algorithms
  - 13 Building Detection
  - 5 Road Detection & Routing
- Integrated into open source geospatial analysis libraries

Evaluation

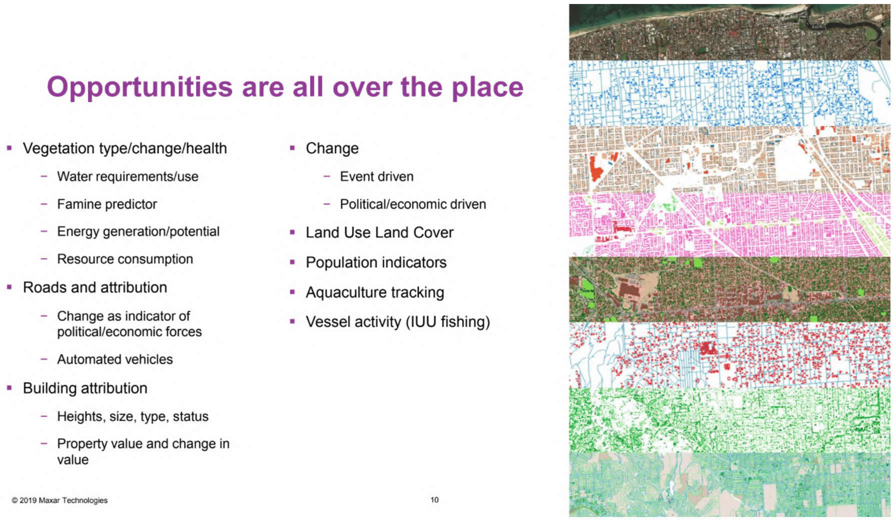
- Performance benchmarks
- Detailed analysis
- Release of optimized models

© SpaceNet LLC 2019.

4

Figure 30. SpaceNet activities

AI products into customer perceived value creates skepticism on the potential impact of AI to address business and societal challenges. Arguing that data and machine learning opportunities are now prevalent, as shown in Figure 31, Getman's talk highlights that the biggest gaps to bridge are between the spectrum of deep and creative technologists and information consumers who are adopting products to impact their markets. Posing questions for future pathways researchers and developers in GeoAI could focus on to address these concerns includes designing models that are less expensive to create and run, more accurate, and capable of creating information products that have clear and easily recognizable value to information consumers.



Opportunities are all over the place

- Vegetation type/change/health
  - Water requirements/use
  - Famine predictor
  - Energy generation/potential
  - Resource consumption
- Roads and attribution
  - Change as indicator of political/economic forces
  - Automated vehicles
- Building attribution
  - Heights, size, type, status
  - Property value and change in value
- Change
  - Event driven
  - Political/economic driven
- Land Use Land Cover
- Population indicators
- Aquaculture tracking
- Vessel activity (IUU fishing)

© 2019 Maxar Technologies

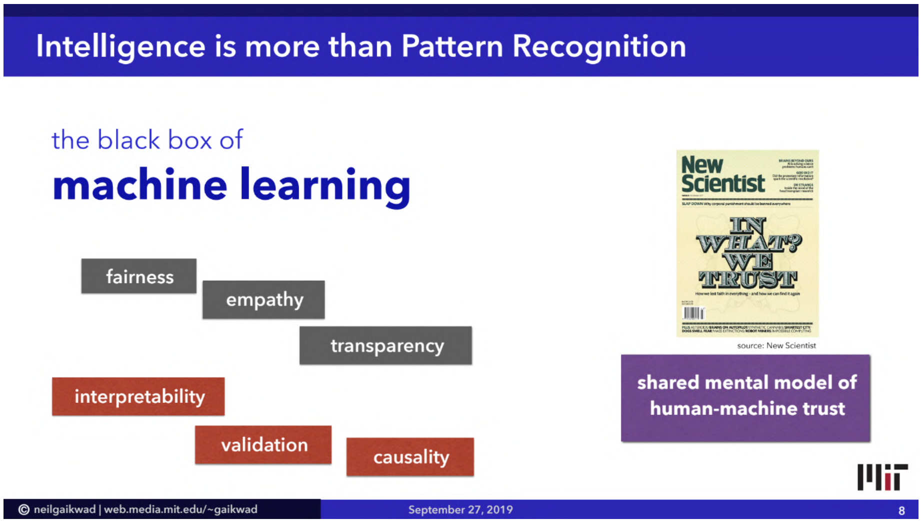
10

Figure 31. Applications opportunities for GeoAI

Neil Gaikwad, PhD Researcher, MIT Media Lab

Focusing on the human centered AI systems towards more socioeconomic development was a presentation by Neil Gaikwad. The talk presented a perspective on how the GeoAI community can include human values into machine learning. This is a challenging task that Gaikwad illustrated by posing a scene recognition test to the audience, from which the goal was to highlight the multiple perspective by humans on the same image scene. Even though the test is conducted with one image scene, it is indicative of what occurs with bigdata, where inference tends to be different than reality. From this illustration Neil's talk builds an argument to challenge the community to study physical processes together with observed remote sensing data. This can be achieved by incorporating social data that drives socioeconomic development into GeoAI systems and measuring the social impact thereof.

The source of complexity for many GeoAI may not solely be from the big data itself but the ill-posed problems, e.g. poverty and food security. Such problems tend to be unstructured and not presentable as binary problems that are common in machine learning and computer vision. What is often lagging is how to formalize these problems so that domain scientists and AI communities can leverage remote sensing data, human values, and machine learning as part of a solution. As a potential pathway forward, Gaikwad highlights the main gap that exists between applied AI research and the human social values: there needs to be metrics for measuring fairness, empathy to transparency in GeoAI systems. The adoption and success of machine learning should largely depend on the complexity of cultures, colonial histories, economic contexts in which they operate - a missing research component to current black box based GeoAI systems.



Intelligence is more than Pattern Recognition

the black box of

machine learning

fairness

empathy

transparency

interpretability

validation

causality

source: New Scientist

shared mental model of human-machine trust

© neilgaikwad | web.media.mit.edu/~gaikwad

September 27, 2019

8

New Scientist
IN WHAT? WE TRUST
MIT

Figure 32. Black box

Mark Korver, Geospatial Lead, Amazon Web Services

Data infrastructure for the future is at the center of Amazon Web Services. AWS has a growing footprint in handling big data. For example, supporting NASA SAR Mission to process around 85TB of data each day to the EOSDIS archive. In highlighting how AWS has enabled such a capability, Mark Korver presented the Cumulus Data Framework which is available through the Distributed Active Archive Center. The talk gave compelling examples while contrasting the benefits of a cloud storage vs on premise storage facility. One of the major takeaways was how easy it could be for individual scientists to leverage loosely coupled data architectures for machine learning as it is easier to manage compared to on premise databases.

Korver further summarized with viewpoints of how to advance GeoAI - the main closing argument being that enablement of GeoAI impact begins with sharing of data. Furthermore, there is greater need for data processing toolkits. AWS provides different levels of capabilities that are part of cloud infrastructure, for example, staging of data for in-situ rather than moving of data, using cloud optimized GeoTiff that supports http and just not file systems, STAC, the open source spatial temporal asset catalog which is a standard way to index data in deep object stores, all allows indexing for machine learning applications.

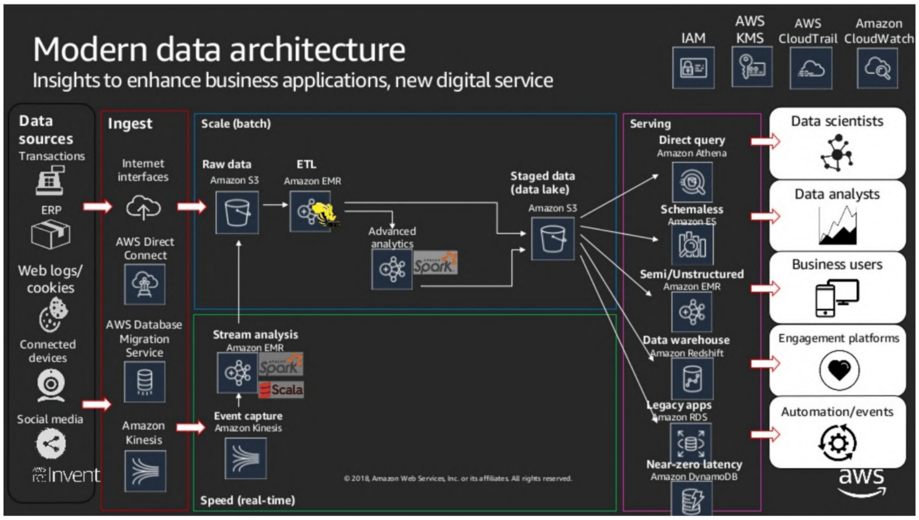


Figure 33. Modern data architecture

Q&A

Question1: For Mark, all the techniques we have today is fundamentally indexing the pixel, how all these technologies can be unified and what are the constraints to come to it.

Mark Korver: My general perspective on it is that probably unification is very difficult, right? It is more important to focus on proper curated data and metadata and making it generally available. We see all types of data on cloud today, genomics, geospatial, that doesnt matter, the point is content changes rapidly, methods changes rapidly, and really the focus should be on proper methods to share that data and making indexes available so that machine can understand and make use of that data. If you are using the different model tomorrow as it constantly changes, so again it is important to support the standard organizations like OGC, WFS, WMTS, that direction the support needs to go. Under the hood there will be a constant change so the system needs to be flexible. Its more important to have authorated data thats well curated and maintained that we can build on top of and refer to over time.

Question 2: Many agencies including the federals want to make the data free and open, it can be on

prem or cloud or on AWS. How to budget that how many people are actually going to use the particular data and for how long? Many civil agencies or special data sources want to budget the data for their user. Wat does panel thinks about that.

Mark Korver: It actually looks a simple problem for me or already many of them in the space are taking care of. Considering the loosely coupled data architecture, the data objects helps storing and access easy. Traditionally the system used catalog to index the data and capture the users requirements in order to budget the requirements. Unlike before when even to do the FTP a local copy was needed which we don't need any more. He further supported his comments giving an example of how Sinerize-AWS work together where interestingly the data is downloaded in the bucket to be used directly. Its a paradigm shift he says, how the system today operates the data at the storage level.

Matej Batic: I can think of a similar approach for this problem at Sinergise. Converging the access of data just behind the API makes it much simpler to handle the user data relationship. It doesn't really matter where the data lies, whether on cloud or in prem, what matter is how cheap it is to store, maintain and access that data. There are open standards like WNS and WFTS to make it more simple.

Question 3: Should the GeoAI community have their sort of grand challenge effort that should be applied across organizations, across the disciplines and have a set of representation on the table of how the US government is viewing science and computing today and what would that problem be?

Nick Weir: From my perspective I dont know if there would be one grand challenge unified under one platform because there are too many different tasks and problems and that is just my perspective on that.Because the tasks are different in nature, the algorithms are different too, and we just cant unify all of it on a single platform. Just to give an example of how different BigEarthNet and SpaceNet are and what specific space they focus in.

Neil Gaikwad: So we already have ImageNet and OSM for doing CV that enables researchers to access the data instead of buying the expensive data. The downside is that it pushes the researchers to that side and marry them to the specific data set which can be totally unstable and biased which might have propagated from the history. But the real world challenges requires to look at wide spectrum of different data set, like data set coming from remote sensing. Worlds problem is very different than the way it looks, it changes according to location and geography, the cultural plays huge role in it and the way data is collected matters.

Daniel Getman: I think it is really a challenging question for a domain where almost everybody starts with something that somebody else has given him for free. The retention of IP in AI is miraculous. You have got samples nobody has, you have got models, nobody has. People don't want to share that in the commercial space, there are exceptions obviously. The idea that as a community we can identify the foundational things that would benefit everyone that aren't necessarily the big advantage that takes lot of time. We can identify so that the organizations who do challenges, the organizations who fund research can collectively focus on those things so that everyone else moves faster and I think that is a fantastic idea. As a community I think there is definitely room for it and to discuss about that further.

Mark Korver: At amazon we have the amazon culture of innovation, like a briefing session, one of the key points of that is that you want to support lots of experimentation. You want to make it inexpensive to fail because that is what leads the innovation. I think how you facilitate much smaller things because they have now access to more democratic platforms, access to code, we now have issues around of IP. But cam we make someone who is say at the other side of the planet to participate if they couldnt participate before or didn't have the budget etc. One of the failures that I saw in our organizations was there something called the GDAL born raising that occurred last year, open source, that was trying to raise 140k dollars to make improvements in GDAL specifics to its user SRS, and we were not able to put money on the table to help out, and there were other organizations and they were typically smaller ones, could of larger ones, Esri contributed to that. So if you look at what drives lot of these innovations are these open source projects that are typically one or two people. So there are real interesting innovations happening and they are not big budget and big organizations, they are much smaller. So from my perspective I can give them to the AWS credits, but its a small shop, so they might need dollars on top of it, and there where we fail, we need to put money on the table for very small projects and there are lot of them and that leads to innovation.

1.8 KEYNOTES

ORNL's 10-Year Vision for Computing and Data

Nichols delivered a talk in which he used the previous decade of supercomputer development to extrapolate what might be expected from ORNL's 2030 computing resources.

The High Stakes History of ORNL

David Keim, ORNL's Director of Communications, gave a talk on the lab's history. Starting with James Chadwick's discovery of the neutron in 1932, David described the events that led to the Manhattan Project.