National Aeronautics and Space Administration

Report from the

NASA Machine Learning Workshop

April 17-19, 2018 • Boulder, Colorado

Sponsored by:
NASA Advanced Information Systems Technology (AIST) Program
Earth Science Information Partners (ESIP)

Table of Contents

1.0 Introduction

The workshop was held from April 17-19, 2018 at the University of Colorado Earth Lab in Boulder Colorado. Approximately 60 people were invited and 75 attended. Individual organizations are enumerated in Table 1. The workshop had a diverse mix of attendees from across academic, industry and government sectors as well as across domain from air quality, cryosphere and hydrology and finally, there was a mix of career stages with early and late career attendees.

Amazon Web Services (AWS)	Booz Allen Hamilton
Caltech	City of Los Angeles
DigitalGlobe	ESIP
Esri	GMU
Jet Propulsion Laboratory, California Institute of Technology	Lingua Logica LLC
Middle Path EcoSolutions	MIT
NASA Ames Research Center. SGT Inc at NASA Ames, USRA, NASA Ames, ARC, BAER Institute/ NASA Ames	NASA Earth Science Technology Office
NASA GES DISC, UMD/ESSIC & NASA/GSFC	NASA Goddard
NASA Headquarters	NASA Langley Research Center, LaRC ASDC
OpenAQ	SSAI
Stanford University	The HDF Group
UC Irvine	UCAR
University of Alabama	University of Colorado Boulder, Univ. of Colorado - LASP
University of Houston	University of Maryland, UMBC
USACE-CRREL	

Table 1 Institutional Attendees at the ESIP Machine Learning Workshop

1.1 Need for Expanded Conversation - Analytic Center Framework

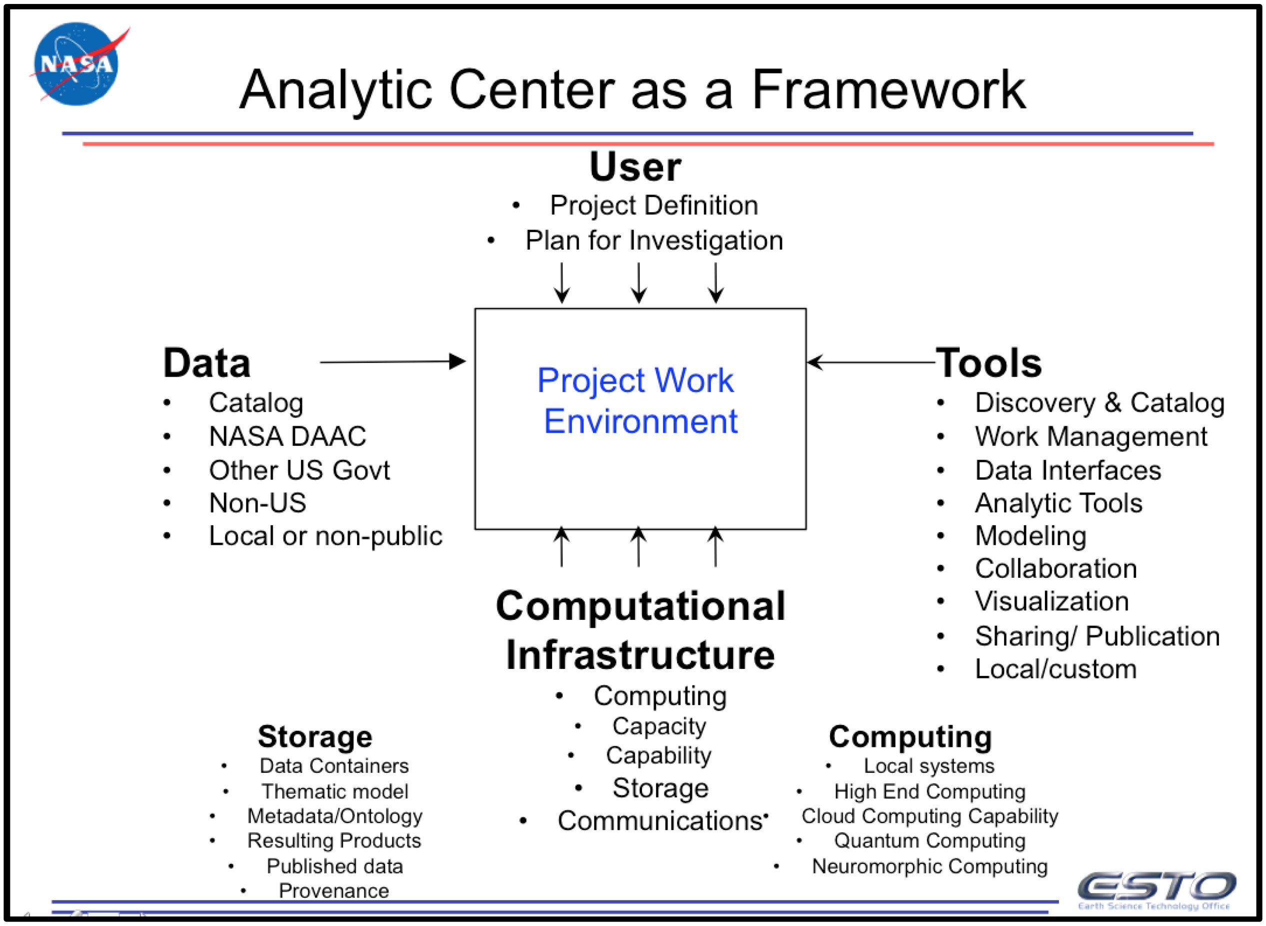
Over the past five years, considerable discussion has been held in the ESIP community as to how emerging information technologies could be made easier to use by the Earth science domains in order to improve the speed and effectiveness of scientific investigation. These discussions have been held in the Summer and Winter Meetings as well as in some of the clusters. Clearly, more information was needed and more direct interaction between the various communities.

One concept that has emerged is that of an Analytic Center, a framework that harmonizes the tools, data and computational resources to support the needs of the Principal Investigator and their team. The fundamental concept can be seen in Figure 1. Re-framing the discussion to focus on the investigation and the Principal Investigator, the Analytic Center is a complement to the data-centric approaches and integrates all the various data sources, regardless of origin. This framework is expected to vary in instantiation to support the way the PI wants to conduct their investigation. As a framework, it provides a common way to interface the tools to the storage system so that the tools do not require significant re-configuration to work with the storage and computing cyberinfrastructure. Some Analytic Centers which involve multiple users and a series of investigations may become persistent and take on the aspects of infrastructure where others may be instantiated, configured for a specific study or experiment, documented and disposed. A number of examples of Analytic Centers have evolved over the past few years, including the NASA Earth Exchange (NEX) at NASA Ames Research Center, supporting Land Change/Land Use studies and the Oceanworks system at the Physical Oceanography DAAC at the Jet Propulsion Laboratory (JPL) supporting the Physical Oceanography community.

In order for this evolution to move forward, further concentrated conversation among the science user community and the information technology communities (machine learning and data science) seemed appropriate. The information technology communities needed a better understanding of the needs of the science communities:

- How they differ by science domain, along with shared needs,
- What processes they use for conducting their investigations,
- What obstacles they meet in trying to leverage the technology, and
- What perceptions and sociological resistance to the use of the technology exist.

Many of the Earth Science domains have started to experiment or even to operationalize the use of Machine Learning in their research. However, many researchers indicated frustration in understanding how the tools worked and when to use which ones. They also wondered if they were using the best approach and felt that some of the more advanced capabilities could be useful to them but lacked time and collaborators in trying to apply them. They also expressed concerns about how to balance the need to understand how to use the tools correctly and to trust them against the time consuming process of developing robust, validated software themselves.



Analytic Center as a Framework

User

- Project Definition
- Plan for Investigation

Project Work Environment

Data

- Catalog
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

Tools

- Discovery & Catalog
- Work Management
- Data Interfaces
- Analytic Tools
- Modeling
- Collaboration
- Visualization
- Sharing/ Publication
- Local/custom

Computational Infrastructure

- Computing
- Capacity
- Capability
- Storage
- Communications

Storage

- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

Computing

- Local systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing

Figure 1 The Analytic Center is a Conceptual Framework for Harmonizing Tools, Data and Cyberinfrastructure to meet the Needs of a Scientific Investigation

1.2 Workshop Objectives

As an element of ESIP's objectives to build communities in Earth Science, the overall objective of this Workshop was to bridge the cross-cutting machine learning expert community with technically literate domain scientist to apply machine learning techniques in new areas.

This bridging discussion enables the group to define needs and constraints for machine learning tool development to support science investigations. It would also provide each of the separate communities an instructive overview of the other group's needs and capabilities. Specifically, it would explain the machine learning technologies to domain scientists and provide an understanding of domain science challenges to machine learning experts. Three science domains were selected based on interest and an assessment of the degree to which information technology could be useful. These were hydrology, cryosphere and tropospheric composition. Through consultation with program managers and program scientists, the focus in each was narrowed to Western States Water, sea ice classification from space and air quality in urban environments. Idea exchange, networking, collaboration and professional development were all intended activities of this interactive workshop. The attendees were all asked to identify their own objective for attending the workshop and are described in Appendix D. Common themes were learning about machine learning and understanding how it could be applied to remotely sensed data in specific domain problems. The workshop attendees tended to skew toward air quality and, as a result, there are more air quality specific objectives listed.

1.3 Workshop Logistics and Format

The workshop was designed to establish a baseline between machine learning and domain experts and then provide an opportunity for them to work together on real problems defined by the users. The workshop started by describing the current state of machine learning ahead of the workshop with read-ahead material and pre-recorded webinars going in depth on technical topics. Most attendees knew less than half of the other workshop attendees, so day 1 was designed to begin with networking activities to introduce the room, provide additional stage-setting from Mike Little and a Machine Learning panel moderated by Dan Crichton. The end of day 1 was the beginning of the idea collection phase, where attendees all shared one ‘Big Idea’ listed in Appendix D and the ideas were anonymously passed through the group and scored five times. This exercise gave an initial prioritization of interest in the room. Day 2 focused on bridging the divide between machine learning technologist without clear applications and domain users with challenges that might be solved with machine learning applications. The morning of Day 2 was devoted to technologist interviewing users in small groups and the groups reporting back. The second half of Day 2 the attendees formed mixed teams and sketched out proposals for these challenges. Day 3 was primarily devoted to proposal presentations with Q&A from the room and a very brief wrap-up. The detailed agenda can be found in Appendix B.

2.0 Current State of Machine Learning in Earth Science - Pre-workshop material & Tech Panel Take-aways

Machine learning has achieved significant success in the financial and security industries. The Federal security agencies have made significant investments in maturing the theories advanced during the 1990’s into a robust, fast and trusted capability.

In Earth Science, adoption has been irregular. While some researchers have continued to use more conventional techniques for analysis, others,including commercial interests have adopted tools and adjusted their approach to analysis to include these techniques, often using open source code for algorithms. Some commercial interests have organized around the effective use of these new tools and techniques and profited from their willingness to take chances.

In some industries, the users have moved on to more sophisticated questions, based on brief but effective experiences with the application of machine learning to analysis of large volumes of disparate data.

3.0 User Challenges and Proposals

User defined applications and the Current State
The attendees divided themselves among the nine tables for two rounds of user interviews. Users were asked questions like:

- Describe a day in the life? (Workflow – listen for bottlenecks)
- Where do you spend 80% of your time?
- Tell me a story about your work?
- Tell us about a problem you haven't solved?
- If you had a magic wand, what would you fix?

Nine user applications were identified from the attendees. The topics covered were:

Hydrology - The users indicated they want to combine machine learning and physical models to predict where water will be 10 years from now.

Cryosphere investigations envisioned:

- Autonomous rover on Greenland and eventually Europa
- Surface of sea ice (categorization from space)
- Bedrock topography under land ice to predict sea level rise
- Acceleration and changes in ice sheets and changes in ice dynamics

Air Quality (AQ) domain experts identified the following investigations:

- Improved AQ forecast at high resolution in both spatial and temporal dimensions,
- Characterizing AQ in cities
- Aerosol Optical Depth (AOD) -> PM2.5 (atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers) correlation over 11 Western States for 2008-2014.
- Heat island – how to connect policy goals with implementation and evaluation

The users identified a number of common challenges:

- Way too much time is spent on data wrangling –
  - Preprocessing archive data files
  - Finding data that is available and a lack of awareness of all the sources
  - Lack of quantified uncertainty/quality characteristics for remotely sensed data
  - Missing data from data sets

- The observation scale is sufficiently large to be daunting.
  - Data volume is often very large, both current and expected.
  - The decision to do analysis computation. The full range of choices include a workstation, data center cluster, commercial cloud computing or HPC?
  - Combining physical models with machine learning techniques to yield a better understanding of the phenomenon. In particular, how to constrain the models?

    - Training data issues. They reported that manual construction of training data sets is not scalable and that there are no/limited classified observations available.

The Groups Developed Proposals to solve Domain Science Questions using Machine Learning
Nine groups formed around the topics identified and began to think about solutions both on a near term scale and on a longer-full proposal time scale. One observation from the group was that there is a need to industrialize computing and machine learning capabilities to make it easier to focus on interesting science problems

4.0 Conclusions

4.1 Lessons Learned Regarding the Workshop

- The facilitators should calibrate expectations on funding up front.
- It would be helpful to announce a more structured agenda ahead of time and clearly articulating path for attendees at start of each section (this is where we have been, where we are now and where we are going).
- Trust issues were identified among potential new collaborators. There was expressed a general concern in sharing ideas for fear that a collaborator might “steal my idea and then drop me as a collaborator.”
- Remote access was less than satisfactory to permit true participation. It would be useful to run some exercises to figure out what this would take. Microphones in the room failed to pick up some participants. It was difficult for remote participants to insert themselves into the conversation.

4.2 Technology Gaps

- Cutting edge machine learning algorithms and techniques need to be available, packaged in some way and well understood so as to be usable.
- Techniques need to be developed for working with sparse environmental training data.
- Techniques and tools are needed for combining process (physical) models with machine learning models in a meaningful way.
- Computer security implementations are outdated and uncooperative with science investigations. Research in making computational resources secure and yet easily usable would be valuable.
- Expert system support and online help providing assistance to users is ineffective, unresponsive and lacks needed content.
- More options are needed in selecting and using Machine Learning applications.
- Regression and machine learning techniques generally fail to convey the physicality of the processes being models/forecasted and lack acceptance by some science communities. These shortcomings make it difficult to understand the boundary conditions. More work needs to be done on techniques that provide insight to the physical processes.

4.3 Obstacles to Adoption

A number of obstacles were identified in the discussion.

- Data wrangling is one of the biggest obstacles to using large volumes data from, potentially, different sources. This is labor intensive and requires substantial manipulation of the original source data.
- Data availability is sometimes an obstacle when the data has never have been published or made accessible. A roadmap pointing towards all Earth science measurements and access mechanisms to them would be useful.
- Reduce the learning curve for new comers with a catalog or guide of what Machine Learning tools are available and to which category of problem they can be applied.
- Specific machine learning methods need help to tackle specific problems.
- Clarity is lacking in understanding which techniques are likely to succeed, have heritage, community acceptance, etc.
- Existing data handling and processing infrastructure is inadequate to support capacity demands (including NASA, NSF and institutions),
- Expert Assistance without humans in the loop should assist in selecting Machine Learning algorithms would be invaluable. Sharing experience-based advantages and limitations of each.

Some suggestions were flagged in the discussion to help make the problem definition clearer. One thought was to create collaborative data hubs where users can contribute to data preparation (tagging, filtering, cleaning). Another involved tools which perform data hygiene on demand with certain specifications.

- Expert advice would be useful on the application of algorithm domains, including pros and cons.

There are so many different technology gaps that need to be worked through, some of which were common across groups:

- technology-end user interface:
- optimizing real time sensors,
- opportunities for augmented reality tools.

Other suggestions focus more on the community-science interface:

- transitioning the knowledge-worker workforce as their daily activities become more automated,
- understanding connections between science, technology and government policy,
- how to influence public thinking regarding science while balancing ethics.

4.4 Evaluation of Workshop Format

Attendees were surveyed for the week following the workshop. 47 of 75 responded completely and the overall sentiment was the meeting had been a worthwhile, productive experience, but the bar could have been higher for engagement. The facilities at CU EarthLab were great. The room was at capacity and the workshop could have benefited from a slightly larger room.

Attendees felt that their objectives had been met or exceeded, but that quantifiable next steps needed to be taken and that the workshop only scratched the surface of gathering requirements from the domain scientist. The interactive format was well-received, but a bit shocking to attendees who were used to a more traditional workshop of one directional information flow. In future workshops a more clearly scoped agenda should be given out ahead of time to allow the participants to prepare for more contributions.

With 75 participants who are generally new to each other, networking and introductions are clearly needed. A suggestion to have more shorter, lightning talks as another way to provide context is a good idea for future meetings. The activity with the most satisfaction was the user/technologist interviews on the morning of Day 2. It would be helpful to brainstorm a few more questions ahead of time that users could consider ahead of the meeting.

Constructive feedback around the idea generation activities on Day 1 included the need for more professional facilitators supporting the groups. One suggestion from participants addressed the idea generation component: "The first idea generation activity was prone to first reaction and bias. A series of quick (2 minute) idea generation, followed by 2 min q&a, done 3 times THEN take a few minutes to generate idea on post it's. These can then be binned in some way. The "votes" were not grounded in any one on day one."

Day 2 proposal activity also received useful feedback:

- The proposal process was a little chaotic and random. It could be streamlined by having the POC of each team to give a 5 minute overview to allow the rest of the participants to make a more intelligent choice as to which group to join.
- Might be a more effective way for CS people to select Users/Projects that they could contribute to. Process felt a bit random to me.
- it might have been smart to ask proposal presenters to say what the weakest or highest-risk portions of their proposal was/were, and where they (as researchers) felt the *real* interest is. I.e., "Parts A, B, and D are standard, but nobody has ever done C and it has broad applicability".
- Rather than the 90 day proposal using ML, perhaps frame the task ask a concept study to demonstrate ML, or demonstrate how a particular science domain would benefit in a significant way... ID what cannot be achieved right now with existing tools.

Complete survey results are available at:
https://data.surveygizmo.com/r/163668_5ae0d9dc0028f3.58406937

5.0 Outcomes and Next Steps

The desired outcomes of this workshop were: (1) identifying gaps in current machine learning technologies; (2) fostering new cross-cutting teams of domain and machine learning experts and (3) to socialize and refine the concept of the Analytic Center. These outcomes were met with the above gaps listed. According to the attendees they did form quantifiable new colleagues and of the 48 respondents, 43 anticipate collaborations from people met at the workshop.

The primary area of specific analytic center feedback was around the user input. From Day 1 of the workshop, the group felt that data, tools and computational infrastructure had all been covered, but the user input was where the framework was lacking. Day 2 and 3 of the workshop began to address that gap. The momentum gained at the workshop will be carried forward through three mechanisms within ESIP. ESIP Lab Incubator awards, Summer Meeting Sessions and travel and an ESIP Machine Learning cluster.

Incubator Call for Proposals -
The Earth Science Information Partners (ESIP) Lab is happy to announce our spring 2018 request for Incubator-project proposals. For this round of funding, we have identified the following topics as emergent areas of need in the Earth science community, and for this RFP, project proposals that address these areas will be given priority.

- Proof-of-concept for emerging technologies slated for operational use.
- Modernization of Earth science workflows using open source, machine learning and/or cloud computing.
- Data provenance to advance data-driven decision making.

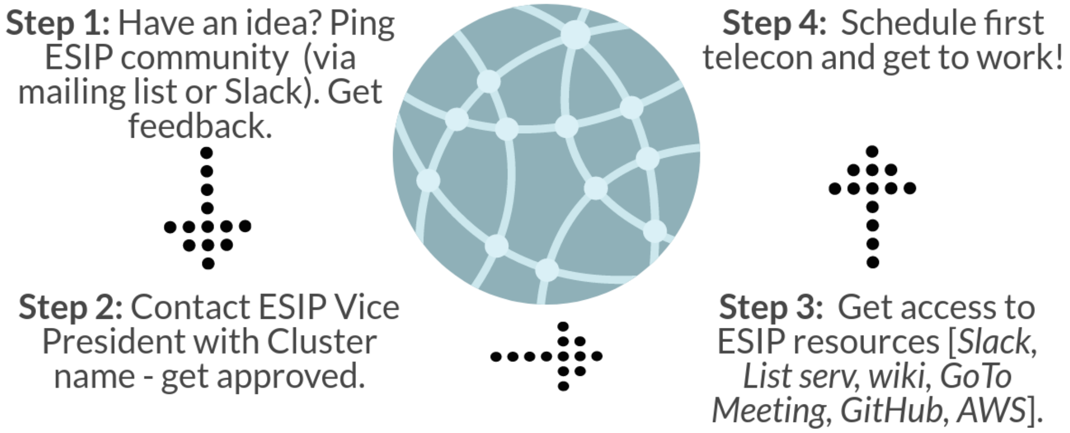
Projects have a 6-8-month duration. A typical budget for Lab projects is $7,000, however, larger budgets will be considered with the firm limit that costs may not exceed $10,000. Deadline for submission is May 30, 2018. You can read the full solicitation here: http://www.esipfed.org/wp-content/uploads/2018/04/May-2018-Request-for-Proposals.pdf

Summer Meeting Sessions Proposed:
The ESIP Summer Meeting will be July 17-20, 2018 in Tucson, Arizona. Details on the meeting are here.

ML Workshop Report	This session is a series of talks reporting the initial MLWS along with work performed and progress during the 90 day follow-up period.
ML Working Session	Machine Learning engagement activities to increase the connectivity among data providers, Earth scientists, machine learning practitioners and computer service providers

Work through ESIP Clusters

ESIP Clusters are communities of practice around specific technologies or application areas. Cluster take advantage of the backbone infrastructure provided by ESIP. The outcomes of this workshop may be well-suited for a machine learning cluster or to revive the Air Quality Cluster.



Step 1: Have an idea? Ping ESIP community (via mailing list or Slack). Get feedback.

Step 2: Contact ESIP Vice President with Cluster name - get approved.

Step 3: Get access to ESIP resources [*Slack, List serv, wiki, GoTo Meeting, GitHub, AWS*].

Step 4: Schedule first telecon and get to work!

Clusters

Appendix A - **Attendees**

First Name	**Last Name**	**Company**
Oleg	Alexandrov	SGT Inc at NASA Ames
Jason	Barnett	LaRC ASDC
William	Baugh	DigitalGlobe
Gerald	Bawden	NASA Headquarters
Sabrina	Bornstein	City of Los Angeles
Brian	Bue	Jet Propulsion Laboratory, California Institute of Technology
Megan	Cattau	Earth Lab, University of Colorado Boulder
Chris	Checco	Amazon Web Services (AWS)
Gao	Chen	NASA Langley Research Center
Ved	Chirayath	NASA Ames Research Center
Yunsoo	Choi	University of Houston
Marge	Cole	NASA / SGT, Inc.
Daniel	Crichton	Jet Propulsion Laboratory, California Institute of Technology
Tom	Cwik	Jet Propulsion Laboratory, California Institute of Technology
Kamalika	Das	USRA, NASA Ames

Jeremy	Diaz	Earth Lab, University of Colorado Boulder
Daniel	Duffy	NASA Goddard
Bart	Forman	University of Maryland
Steve	Fowler	NASA ESTO
Dejian	Fu	Jet Propulsion Laboratory, California Institute of Technology
Sangram	Ganguly	BAER Institute/ NASA Ames
Ted	Habermann	The HDF Group
Colene	Haffke	NASA Headquarters
Christa	Hasenkopf	OpenAQ
Daven	Henze	University of Colorado at Boulder
Ute	Herzfeld	ECEE and CIRES, University of Colorado at Boulder
Kimberly	Hines	NASA ESTO
Jeanne	Holm	City of Los Angeles
Hook	Hua	Jet Propulsion Laboratory, California Institute of Technology
Thomas	Huang	Jet Propulsion Laboratory, California Institute of Technology
Beth	Huffer	Lingua Logica LLC
Balaji	Iyer	Amazon Web Services (AWS)

Chris	Jenkins	University of Colorado
Brian	Johnson	Earth Lab, University of Colorado Boulder
Max	Joseph	Earth Lab, University of Colorado Boulder
Thomas	Kurosu	Jet Propulsion Laboratory, California Institute of Technology
Sari	Ladin-Sienne	City of Los Angeles
Barry	Lefer	NASA Headquarters
Alan	Li	NASA ARC
Michael	Little	NASA ESTO
Melissa May	Maestas	Earth Lab, University of Colorado Boulder
Ashish	Mahabal	Caltech
Joe	McGlinchy	Earth Lab, University of Colorado Boulder
Scott	McMichael	NASA ARC
Piyush	Mehrotra	NASA ARC
Nathan	Mietkiewicz	Earth Lab, University of Colorado Boulder
Mathieu	Morlighem	UC Irvine
Chelsea	Nagy	Earth Lab, University of Colorado at Boulder
Kumar	Navulur	DigitalGlobe

Grey	Nearing	University of Alabama
Nikunj	Oza	NASA ARC
Victor	Pankratius	MIT
Craig	Pelissier	SSAI
Chris	Polashenski	USACE-CRREL
Amy	Povak	DigitalGlobe
Erin	Robinson	ESIP
Gian Franco	Sacco	Jet Propulsion Laboratory, California Institute of Technology
Lynne	Schreiber	UCAR
Dustin	Schroeder	Stanford University
Michael	Seablom	NASA Headquarters
James	Sill	Esri
Jennifer	Sleeman	University of Maryland, Baltimore County
Ben	Smith	Jet Propulsion Laboratory, California Institute of Technology
Florence	Tan	NASA HQ
Brian	Tisdale	Booz Allen Hamilton
Michael	Turmon	Jet Propulsion Laboratory, California Institute of Technology
Arika	Virapongse	Middle Path EcoSolutions

Leah	Wasser	Earth Lab, University of Colorado Boulder
Jennifer	Wei	NASA GES DISC
Christine	White	Esri
Daniel	Wilson	Esri
Anne	Wilson	University of Colorado - LASP
Chaowei	Yang	George Mason University
Soni	Yatheendradas	University of Maryland/ESSIC & NASA GSFC

Appendix B - Agenda

Day 1 - Recording from GoToMeeting

1:00 Welcome & Stage setting - 5 min

1:05 Welcome and Introduction to Earthlab - 10 min (Brian Johnson, EarthLab)

1:15 Overview of Workshop, ESIP intro & Group intros - Erin (30 mins)

- Introductions from group - In no more than 30 seconds
    - i. Name
    - ii. Organization and project
    - iii. What do you hope to get out of the workshop
    - iv. Dots representing workshop interest and science problem domain

Active Listening: Audience using three colors of post-its to capture Roses or strengths, Thorns or problems, or Buds or opportunities. Think of Bold ideas too

2:00 Analytic Center Intro ()

2:30 - 4:00 Panel - Dan Crichton, Moderator

- b. Esri - Christine White
- c. AWS - Chris Checco
- d. Digital Globe - Bill Baugh
- e. MIT - Victor Pankratius
- f. NASA Ames - Kamalika Das

4-5 Each participant contributes a bold idea if needed based on conversations. Use the 25/10 exercise to mix and rank ideas. Identify top 11 ideas

5-5:15 Wrap up (shuttle back to hotel + drivers)

Day 2 - Recording from GoToMeeting

8:30 Recap day 1 - room check in

9:00 - User Interviews - Users self-identified out of the attendees and formed eight topical groups. The machine learning experts developed a few questions to guide the interviews. The group came back together and each topic had a table. Machine learners interviewed domain experts. Groups did two rounds of interviews - one tech person + user stayed at the tables for continuity

Interview Questions:

1. Day in the life? (Workflow – listen for bottlenecks)
2. Where do you spend 80% of your time?
3. Tell me a story about your work?
4. Tell us about a problem you haven’t solved?
5. If you had a magic wand, what would you do?

10:15-10:30 AM break

10:30 - 11 Wrap-up of interview results
11-12 - Users reported back to room on their challenges

12-1 lunch and break - can come back to board and try again if you don't like it and you can join other teams

1-4:30 - Break into teams and develop proposals based on the challenges described

Day 3 - Recording from GoToMeeting
8:30-9 Recap from day 2
9-11:30 Pitch final prototype proposals (10 mins + 5 mins for questions from the group)
11:30 - 12:30 Lunch & Brief discussion of next steps

Appendix C - Raw Material Collected from Workshop

Meeting Hopes - Day 1 Introduction

I hope to learn how to use machine learning with space images

Ongoing + new science opportunities to apply geospatial technologies (and understand those that already exists)

How to integrate ML into NASA's Water & Energy Cycle Focus area

New Ideas for making LA more resilient esp related to climate risks

Potential future collaborative talks in ML for remote sensing

A broad understanding of how others are using machine learning; an understanding if my gaps in knowledge are consensus gaps in knowledge

Ideas,; observation needs

learning latest & greatest ML for Earth science; meeting new collaborators and making new partners

I want to meet ML colleagues. Air pollution forecasting modeler

What advances in ML are taking place and what technology gaps still exist

Opportunities to advance science through machine learning

1. Identify interesting scientific applications that can benefit from use of ML; 2. special focus: physics infusion into ML/data modeling; 3. non-conventional data sources that help the discovery process

science ideas; what ML scientist are interested in

New ideas

Knowledge of open source ML availability

Applications for state & local government

How machine learning help in AQ science

I would like to see advancements in DL algorithms for Earth sciences and specifically scaling algorithm in hybrid cloud

learn about how data is used + documented

What cryo problems could benefit most from ML; What hesitation cryo science have to learning/using techniques

Brainstorm ways the openaq community can apply ML (or access ML community to our giant AQ data set)

Ideas for ML applications and Techniques in Air Quality

Learn new ML approaches; design cool pilot project with others

Improved understanding of how to ML w/high value to the public

Pilot to improve air quality in LA driven by Data, models + changed behavior and outcomes using ML

New ML applications

Better grasp of ML application; ideas/areas how to use it; Hands on example

A clear direction on how earth science data/machine learning can help w/ long-term planning for LA

How ML can help understand Air Quality

Explore new areas for machine learning applicability w/ new data types

Obstacles and gaps

Resources for learning more about machine learning

Learning of problems that can be solved by using existing ML

Better understanding of utility of DL/ML in applied remote sensing research

Learn about active and potential ML techniques/projects for cryosphere study

Requirements of the HW/SW infrastructure required to support ML technologist and scientist

Identify what ML approach can be used to better understand basal properties under the ice sheet

what are other disciplines doing that I can borrow? How can I make what I am doing relevant to more people

I would like to gain better understanding of where ML has been and needs to be applied to Earth science

Ideas

Collaboration w/ people more capable at handling massive datasets (I can bring science needs & algorithms)

Interest in user needs (how can high resolution data help); networking

understand ML applications; new connections; experiment w/ facilitation techniques to bridge gaps between science and tech communities

Identify what, if any problems in radio glaciology are optimally suited for machine learning

Learn about current SOA in Machine Learning; learn about current applications and research endeavors

I hope to learn more about the science looking for problems to work on

Ideas to incorporate ML into Autonomy applications; potential paths of development ML

new relationships; science user requirements

An understanding of what the leadership wants to accomplish with this initiative

I'd like to gain a better understanding of the skills we should be teaching to support this type of science

Find a solution to mitigate climate change by combining data

Interesting use cases for scientists in geospatial machine learning

Greater understanding of ML techniques and science applications

An understanding of gaps that machine learning might fill in climate science

Landscape of AI for Earth science

I need more info of available tools (e.g. NEX) that can be leveraged for deep learning and machine learning applications in hydrology

1. learning more about use cases, challenges to use ML, deep learning; 2. how ans ml services can help solve problems 3. Actionable next steps

An understanding of the array of use cases which Earth science deems impactful

Connect to potential collaborators; input from science & policy community about problems they would like tech to try to solve

Strengths, Opportunities & Problems with Machine Learning - Active Listening from Day 1 Session:

Identify problems with analytic center	Strengths
open source use cases + associated data, code, algorithms for easier replication	Strengths
workflows to capture end-to-end earth science problem solving, splicing in new methods, new datasets, etc	Strengths
Creating transparency for peer-review trust	Strengths
How to improve partial physics as a soft constraint on ML-learning/predictability	Strengths

Standardize data (metadata)	Strengths
Analytic center framework makes work more effective	Strengths
Cross-product integration	Strengths
Analysis with data from multiple sources	Strengths
Data organization - Pandas, x-array	Strengths
Apply existing Techniques for many sets they will work	Strengths
Need way for scientists to understand how the hardware affects data interpretation	Strengths
Build this into data analytics center, say flowchart but can launch tools directly	Strengths
Easy to use decision framework for users to choose model, viz, and understand risks/limitations	Strengths
ID pros and cons of tools	Strengths
chance to seamlessly move between learning via data versus learning via physics	Strengths
organizing data in cloud for performance	Strengths
prediction w/out understanding	Strengths
metadata	Strengths
research in new machine learning techniques	Strengths
tech advance opportunities for science	Strengths
build of community of liaisons between data works + data consumers (agriculture, health & energy)	Strengths

synergy of data, methods and tools	Strengths
Basic computational framework -> power of learning from data and physics	Strengths
new ML algorithms depending on science applications	Strengths
focus on the big gulf between data publishing + data adoption + use operationally	Strengths
understanding end user's data needs when designing the hardware, science measurements and data collection	Strengths
Explore physical chemical consisting of the data observations	Strengths
Data assimilation & osses for prospective remote sensing missions w/ ML	Strengths
Cross-pollination of expertise across fields	Strengths
improving accountability in use of data/methods through collaborations	Strengths
Abstract data sets for transfer learning from other domains	Strengths
A unified NASA AOD product	Strengths
collecting and sharing tools for accessing and manipulating data	Strengths
Improved data delivery & access (permission)	Strengths
Are we building infrastructure or tools that run within infrastructure	Problems
glaciers	Problems
Complexity	Problems
non-conforming data, units, gsd, projection	Problems
prediction w/out understanding the physics	Problems

more beer; facilitate conversations + collaborations	Problems
breakdown in stakeholder communication along value chain	Problems
Scientist change their minds	Problems
proposals are one sided science only or tech only	Problems
what tools do scientist need?	Problems
Barriers to entry	Problems
It's hard for policy makers to get data that they can use	Problems
gaps in skill and knowledge in big data technologies	Problems
Never going to get all data in perfect format	Problems
Problem accessing the data	Problems
How do you envision projects with narrower scope to achieve the analytics center "compatibility"	Problems
lack of investment in new computational techniques in machine learning	Problems
weak general understanding of what machine learning is and is not	Problems
Accessing and understanding interpreting satellite data	Problems
how do you employ ml for use without code? There are so many different options, this makes this a challenge	Problems
no framework for combining the strengths of data-based prediction w/ the strengths of physics based prediction/explanation. PDEs are bad at learning from data. Not constrained by ANNs or physics	Problems
Data wrangling and processing for use in model frameworks	Problems

Ensuring a consistent and working compute environment for multiple platforms	Problems
Need enough data to train the model	Problems
Need for more training data	Problems
Look for physical understanding from ML/DL	Problems
People making tools not talking to end users	Problems
Machine learning for data discovery	Problems
metadata	Problems
Documentation of Machine Learning algorithms + results	Problems
data migration to cloud ie ETL?	Problems
Scale	Problems
Identify problems with analytic center	Problems
ML + Earth science communities don't mix	Problems
Too many answers, not enough questions	Problems
Scientist may make assumptions about data that are incorrect - e.g. data have to be interpreted	Problems
How to look from observations to the events	Problems
ML, DL for detecting the events	opportunities
ML for finding rules & knowledge	opportunities
A suite of tools available + gaps	opportunities

How to know when we have events of interest. AQ problems, Heavy rainfall	Problems
reproducible work flows	opportunities
Benchmarking suite of ML models	opportunities
Lots of labeled data	opportunities
People hungry for ML	opportunities
Volume of data	opportunities
More objective analysis	opportunities
more accurate analysis	opportunities
Experimental analytic center in AQ	opportunities
Taking advantage of information from a dense set of data, time and space	opportunities
Leverage military investments	opportunities
New questions we have never considered	opportunities
"ingest" big datasets	opportunities
New entrants who think "outside of the box"	opportunities
ML is super-good at hydrological forecasting	opportunities
ML + cloud+commercial accomplish more off the shelf	opportunities
Jobs for grad students	opportunities
NEX demonstrates potential of analytic center for ML of earth science data	opportunities

Big ideas with Scores:

NASA library of training data and ML/DL models	21
Community-based library for data-merging, format, organization, management and processing with interchangeable pipelines -. Make it off the shelf	21
Identifying abrupt change in ice sheet configurations from combined radar altimetry and multi-spectral imagery	21
Exemplar based CBIR for searching large remote sensing archive (CBIR - content based image retrieval) choose examples to search for similar content in the dataset	21
How to impose physical constraints on ML models	21
Expand Victor's approach into a framework for change management which is adapted to the specific objectives of the particular project	20
More ML on spacecraft, construct ML analytics pipeline from the ground station; increase opportunities to share training datasets; teach NASA course on ML for science	20
Use machine learning to develop 3 year predictive models for water availability in the western u.s. using historic; current, ground based satellite data. Can the past help forecast the future trends	20
The topography under big ice sheets remains poorly known. It is only measured by radar along flight lines. ML could help us predict the topography between these flight lines based on other predictions(surface features, ice speed, …) for which we have full coverage - map sub ice sheet topography	19
Understand how urban built environment affects heat and quality of life using satellite data + smart city sensors + interventions	19
Use ML and satellite/ground data to improve air quality forecasting - ozone, pm2.5	19
Metadata describes how knowns were accounted for in complex results	18
ML based metadata data catalog for top five use cases; ML based data quality that feeds to data lake	18
Big idea: fleshing out a fuller understanding of sub-daily AQ patterns of cities of the world to: 1. find cities facing similar AQ temporal patterns, pollutant - combo issues; (2) find suitable cities to launch epi studies or intervention work studies on AQ + health (3)	18

foster international collaboration	
ML to fuse hyper or multispectral imagery with other land images to predict wildfire probability over various regions	18
Establishment of a hybrid model combining DNN/Data assimilation/AQ 3DCTM model to forecast better tomorrow air quality and provide a…	18
Apply machine learning to attribute the contribution of each source that drives the tropospheric ozone trend	18
Machine learning to aid constellations of satellites helping creating sensor-webs in space	17
Utilize machine learning techniques to refine and filter surface observation sites (solar or meteorologist) time series data in order to see their relationships to model/gridded datasets	17
Support for automated discovery of linkages between observed variables, for example, as links in a Bayesian network	17
Model Assessment, measurement consistency between different variables	17
Create an OPEN Big data repo with metadata and documentation and let ML methods be trained on them	16
Use Machine learning to help solve problems to reduce increasing temperatures + urban heat island effect. Specifically, at the interaction between temp and the built environment - both how trends will exacerbate projection, heat and mitigate	16
Support for automated adaption of ML hyper-parameters within an analytic center	16
Clearly understand the deep benefits deep learning offers for problems such as assimilation, feature engineering, etc	16
Facilitate/support interdisciplinary teams for ML applications in order to enhance the discovery/communication of science results to engage and reassure new science communities of the value of ML technologies	16
Genetic programming to compare M.L. mode performance to existing scientific models. Use scientific models to simulate training data to see if ML models can generalize/explore results	16
Global data/pattern analysis combining Air quality and greenhouse gas - remote sensing	16

Use of deep learning/GAN in the absence of data - send rover to greenland/antartica/artic to study how ice retreats temporally (1 year - 10 years) & spatially. Rover must be able to rover autonomous & detect/react to obstacles - active error correct/diagnosis	15
Quantify sea ice thickness from altimetry data	15
Set aside funding for a centennial challenge to solicit & award best ideas; use funding for computing infrastructure and data system support	15
A framework for evaluation of machine learning algorithm performance in the context of scientific discovery (not necessarily ml metrics such as squared error, AUC, etc.)	15
New ways to leverage ML for city of LA. Also identify NASA datasets for use	15
Finally time has come to create the satellite labelled training, database open to public - call it "SATNET" just like we have "ImageNet" for camera images - 5 mil image samples	14
ML: Troll science abstracts in a domain research need - discovery; use ML to ID high value science needed in a domain Hydro, cryo, AQ	14
End goal: automate the pre-processing of data that is necessary to make it analysis-ready; Idea: use ML + semantics + reasoning to extract metadata that can support such automation	14
Idea for prototype - machine learning to estimate object drag coefficient	14
Improve algorithms to better estimate surface air pollution concentrations from satellite data	14
Improve urban air quality by applying ML to longitudinal data, urban policies and interventions and health outcomes	14
Is it effective to fuse traditional physics based analysis with machine learning. Is there a possibility of a better and unique result than either endmembers alone?	13
Machine learning -> transparency in formulation -> adherence to first-order physics? -> right answer for the right reason?	13
Tell me an example (any science field) of using ML that is successful or well-received in the community (knowingly or unknowingly)	12
Define benchmark problem that ML can solve that is validated. Structure problem with layers of fidelity	12

Infrastructure creation: The unified earth data tool create an interface to all global raster layers that can be queried, subsetted in space + time and have user supplied code applied to generate additional layers, automatically publishing + archiving output to database. E.g. users may identify all sources global forests, agriculture fields	12
Automatically capture all scientists workflows. Machine learning over those workflows can be used to learn parts of workflows that contribute to success or failure	11
Leverage ML techniques to look for anomalies in Terra Fusion dataset - arcgis/AWS	11
Symbolic regression applied to multi-resolution, multi-system data fusion	11
Develop hardware to allow community to participate in earth science data collection. Use data to gain knowledge for decision-making/planning; Data air quality, water quality	10
Use the language processing services to "hear" what a software user scientist wants to do and then propose tools, models, info to help them	10
Automatically detect interesting features from proper data feeds, sensor, archive, internet etc	10
Standardization of remotely sensed data - metadata; opensource	9
I'd like to see deep learning architecture that can naturally (intentionally, meaningfully) incorporate spatial point information	8
To create an ML-ready data framework that cuts across disciplines, types, sources (internal + external) to help bolster discoveries and reduce the time to science	6
Synthetic hyperspectral imaging using MSI + pan to synthesize HIS	

Appendix D - Raw Notes

Day 1 - Tech Panel

Day 2 - User Report Back

1. 1:41 Greenland – autonomous to eventually send to another planet; Keep itself alive and adaptively sample + both unknown and refine models. Challenges – reluctance to accept AI system; utility afterward to justify pattern (weird pattern that robot sampled or did robot get fixated); Models and science are in discovery how do you separate anomaly detection vs instrument breaking; Captured vs new obstacles; Robot optimize sampling while keeping safe + diagnostic

a. Issue – data volume; downlink data is limited and precious; ML onboard that is large locally and smaller that sent down.
b. Project so future oriented that don't know
c. Current missions are automated not autonomous. Sending commands every morning; 80% of time spent planning every automated step
d. Q – how do deal with training on Europa – orbiter before rover goes. Mars have good training data.
e. Q – observed? Mass balance, water storage,
f. Robot run local models of hypothesis Challenge – build in models enough you want to improve models and opportunities to change
g. Certain can be solve today and lots of future learning.

2. 1:52 Surface of Sea Ice – dependent on albedo, high res with lots of data already have ML algorithm for classification and have a training dataset with 50,000 segments. System seems to be working reasonably well but LOTS more data. Want help with data volume 20 TB up to a PB – hitting scale issues and plan to grow exponentially. * How to do data processing to get to modeling ?
Data augmentation – training is manual process of experts classifying? How do we scale to better predict ?
How to scale models on HPC or on cloud? Right now there is a workflow. Have a preprocessing routine. How to modularize and create containers use cluster scalers to spin in parallel same process
End goal: Not just sea ice, but meter scale land cover of entire globe
What model? Rand 4 classifier – feature based classifier ??? imagery features calculated half are internal to segment and half are around
How are you going to do augmentation with feature based calculation? Different model

3. Univ of Houston ~ 1:56 AQ forecasts
Coverage 70-85% for air pollutant coverage??
Current problem – Atmospheric scientist not tech; Data quality problem – remote sensing data has uncertainty; training set needs to know quality of data also issue of **missing data**; Big data issues – 400 TB challenges. Ozone, PM2.5 forecasts; magic wand – funding issues; high

precision modeling system – combine data fusion with model and observation for better forecasts; Tech people to design better system. Questions – what was problem? Missing data and don't know how to deal with missing data? Focus was Air Quality

4. AQ – EarthLab 2:02
AQ, Roadcover, surface feeding into ensemble machine learning to get surface pm2.5 out at zip code level. Using EPA PM2.5 for training data; Useful advice – long and short term memory (AOD for day of T-1 as helpful predictor) MERRA 2 and data from Earth on AWS;
Comment: Interesting to have ensemble of models
How much data do you have to use? Span of years 2008-2014 by day and 11 western states (lat/lon of every pm2.5 during the time) and lat-lon of center of zip code.
Why 11 western states? Scaling up model originally done from California for the southern; Fire situation is different between east and west

5. Grey 2:11 – Hydrology model to estimate floods and droughts - estimate water on earth with 100s of variables, 1-10 km spatial. Trivial machine learning is beating the models and means there is extra info not being captured in science models. Machine learning regression - Wants to combine machine learning + predict where water will be 10 years from now – **regression models are not reliable for long-term. Allow not have to put in parameters that I don't know and let physical parts of model and constraints that they do know**
Solution: Bayesian network with different nodes and each edge is a neural network that can replace with machine learning or parameters
Combine hydrologist things you do know with power from learning from data
Q – Ideas on how to solve? Couple of ideas but don't fully understand one strategy – build mass conservation into constraint
Hook – teach machine learning to understand other kind of classes; codify elements of physical model into machine learning
Q – validation – never be able to evaluate 10 years from now Train on 80's for 90's; 90's on 00's

6. Chief resilience officer – City of LA Heat 2:14
Reduce urban/rural heat differential; released resilience strategy Prepare and protect those vulnerable to extreme heat – looking at models and built environment exacerbation. City has policy goals – veg, tree canopy, soil, roof, cool pavement pilots etc. How to prioritize rolling out policy and evaluating the strategy. Do we impact the things we care about? Q – How do we measure heat? What are datasets available and how do we get to relationships between heat and built environment?

- A lot of data out there. How do we access, understand datasets that exist
- Missing data (user not aware of data)
- Need block by block scale

Comment – temperature datasets derived from … could help (phil Yang data)
Sangram – High level tree top canopy; Data Awareness; Data available but not unified and not clear what sources are best to answer the questions
Block by block – need the scale or by category and then derive by neighborhood? Small intervention are at block scale

7. City of LA AQ – Global Health 2:20
Challenge – at a city level challenge of decision making global – forecasting and ML – classification
Goal – improve quality of air to save 4.6
Unleash power of ml and use nasa data. Get index but don't know what to do to fix it.
LA 1 in 10 kids have asthma

Lots of data available – remote, surface, - dense data rich space near LA; Correlation between coarse scale remote with surface obs; Esri tools – GIS application and can use AWS. Which chemical transport models would be useful?
ML opps in forecast and classification – how can it be translated to less data rich places (chemical and longitudinal studies) to identify sister cities
Anomaly pattern event detection
****AOD -> PM2.5 correlation ** major issue across all three AQ groups.**
Change policy to reduce pollution

Q – Asthma – pollution or healthcare. There is correlation between aq and asthma

8. Grounded ice not floating ice ~2:26- How much sea level over next century? What do ice sheet do over next century and need models.
Know enough about inputs, but don't know about bed topography? Need to have features of bedrock right. Can measure from radar along flight lines, but misses in between lines. Need full 3d model. Need high res (200m horizontal resolution) that model can ingest. Bumps in bed should be on surface. Have surface DEM and speed (also good indicator of bed) -> bed topography maps

ML prediction and what has been measured
Q – what kind of assimilation? Not an expert (good feedback)
Q – **Conservation of mass and add some geostatistical characteristic of the bed. Is there a way that ML could honor both of those constraints. This works well along coast, but interior is less accurate so need other methods. Add to cost function.** Weighting terms differently

9. Ute – Cryo ~2:30
Problem - Acceleration and changes in ice sheets and changes in ice dynamics is largest uncertainty in sea level rise IPCC. Changes are not linear. Ice dynamics in remote sensing have image, point cloud, altimeter etc .
Analytic side – data formats/tools and round two table – deep learning – action items:

1. Combine physical process with machine learning can take over (**which parts of scientific problem could be taken over by machine learning and which part is physical model – bring two together**) efficiency of ML Transferability from one system to another

2. Supervised vs. unsupervised classification
3. Test different types of ml algorithm; differences in complexity
4. Different methods on different components
5. Role of data fusion
6. Validation – data coming in to validate new systems

Day 3 - Proposals

Proposals described above were summarized and the obstacles and technology gaps discovered were identified.