1.0 Project Summary

Name of Project: Developing Tutorials for Satellite Imagery Analysis using AWS Open Data and Cloud-Native Workflows

Project Lead and Contact Details: Amanda Tan Lehr, eScience Institute, University of Washington (amandach@uw.edu)

Budget Requested: $10000
Budget Summary: Refer to Section 3.0 for Deliverables

- Develop new and expand existing tutorials to showcase value of AWS Open Data for Satellite Imagery Analysis ($3000)
- Work with AWS to ensure content is public-facing and technically accurate ($1500)
- Develop pedagogical tools for outreach and encourage adoption of NASA and AWS Open Datasets ($2500)
- Workshop to be held at University of Washington ($1500)
- Travel to and from ESIP meetings to hold workshops ($1500)

Proposed Start and End Dates: March 2020 - September 2020

2.0 Project Outline

Project Outline: The National Aeronautics and Space Administration (NASA) launched the Earthdata Cloud in July 2019 in anticipation of its growing data storage needs of upwards of 250 Petabytes (PB) by 2025 (https://earthdata.nasa.gov/eosdis/cloud-evolution). NASA is leveraging commercial cloud services (e.g. Amazon Web Services(AWS)) to create new opportunities for scientific discovery and revolutionize the way scientists interact with these datasets, from machine learning and artificial intelligence to sequencing data.

Many open-source analytical applications currently exist to work with increasing amounts of data. However, many in the Earth and Space Science (ESS) research communities have yet to adopt these tools. Many users lack experience working with collaborative tools like Jupyter Notebooks, utilizing Python-based software and integrating their work with version control software for reproducibility. Researchers are also stymied due to a lack of cohesive set of tutorials that are applicable to their own workflows. To fully draw on the power of working with data in cloud requires a shift in traditional methods of data processing and analytics.

In this project, we will develop and refine demonstration-ready Jupyter notebooks to exhibit the value of AWS Open Data (here we will focus on Landsat-8 and Sentinel-2 data), articulate the connection between cloud computing and computation at scale, and work to enhance the ability

of users to map a satellite image analysis workflow to their own use case. Some of the modules that will be developed include (but not limited to):

- Introduction to Basic Python Tools for Collaborative, Reproducible and Open Science
- Introduction to Cloud Computing
    - Systems administration (user roles, tagging, cost management), cloud best practices
    - Data storage, getting data in and out of the cloud
    - Cloud Computing for Scalable Science
- Introduction to AWS Open Datasets
    - How do we utilize and work with AWS Open Datasets?
- Introduction to xarray
    - Small computations using xarray on your personal computer
- Introduction to Dask
    - Dask for scalable analysis
- Introduction to the Pangeo ecosystem
    - Scalable computing on the cloud using Pangeo, core packages/infrastructure of Pangeo
    - Installation guides for specific packages and environments
- Examples of domain specific end-to-end scientific workflows

Project Significance and Impact: The overarching aim of this project is to develop a suite of executable notebooks to demonstrate the utility of AWS Open Data and to expand the ability of researchers and students to work with NASA satellite imagery. It is anticipated that the modules developed in this project will be **reusable** in short courses taught at ESIP meetings and other scientific workshops. All modules and content will also be publically available on Github and formatted in such a way that learners will be able to utilize the content at their own pace. We will set up environment files so that users are able to execute the notebooks on binder (http://mybinder.org) or Google Colab (colab.research.google.com). Users will be able to request access to a Pangeo compute platform to execute Pangeo-specific learning modules. The Project Lead will also integrate the developed materials within the Pangeo community's education and outreach efforts.

4.0 Key Project Steps and Timeline

1. March 2020 - June 2020: Develop new and curate existing tutorials to showcase value of AWS Open Data for Satellite Image Analysis
2. May 2020 - July 2020: Work with AWS to ensure content is public-facing and technically accurate

3. July 2020 - August 2020: Develop pedagogical tools for outreach and encourage adoption of NASA and AWS Open Datasets
4. September 2020: Convene a workshop at the University of Washington to test and solicit feedback on content development

5.0 Outreach

The content developed in this project will be hosted on Github in the manner of tutorials developed by Software Carpentry (https://software-carpentry.org/). In this way, the content can evolve through community input and collaboration. Further, we will use materials from this project to teach at ESIP meeting workshops, hold workshops at the eScience Institute at the University of Washington and use the content developed as a testbed to integrate with future NASA outreach programs.

6.0 Integration with the ESIP community

We will work closely with existing ESIP collaborators to convene a workshop at the ESIP summer meeting.

7.0 Project Partners (if applicable)

The Project Lead will work closely with the Pangeo community to solicit feedback on the content developed and to hold a joint workshop.