1.0 Project Summary

Name of Project: Developing Workflows for Assessing High-Resolution CubeSat Imagery to Infer Detailed Snow-Covered Areas for Studying Changes in Ecosystems and Water Supply

Project Lead and Contact Details: Nicoleta Cristea, Civil and Environmental Engineering and eScience Institute, University of Washington (cristn@uw.edu)
Project Partners and Contact Details: Amanda Tan, eScience Institute and IT Research Computing, University of Washington (amandach@uw.edu); Anthony Cannistra, Department of Biology, University of Washington (tony.cannistra@gmail.com); Kavya Pradhan, Department of Biology, University of Washington (kavyap2@uw.edu)

Proposed Start and End Date: January 2019 - July 2019

Budget Requested: $9000
Budget Summary: Refer to Section 3 for Deliverables

- Amazon Web Services cloud computing resource ($2000)
- Conference travel, presentations and reports to 2019 ESIP summer meeting and AGU ($3000)
- Travel and boarding for 2 students to attend Waterhackweek 2020 at the University of Washington, Seattle ($4000)

2.0 Project Outline

Project Description:

The ability to observe the Earth from space at relevant spatial and temporal scales is key to understanding changes and related responses of species and water systems to climate change. The recent perfusion of commercial earth imagery with high spatiotemporal resolution may be able to bridge the gap between ground-based instrumentation and coarsely-captured satellite data. In particular, observations of snow-covered area in montane areas at high resolution (meter scale) are of critical interest as snow drives much of the seasonal hydrological regimes and can have significant ecological impacts on phytocoenosis. Currently, remotely-sensed snow cover observations with adequate temporal resolution (daily) are either captured at a spatial scale far too large to be relevant to snow cover models and phenology studies (e.g. MODIS, 500m), or are appropriate in spatial scale (1-10 m) but have inadequate temporal resolution and are cost-prohibitive (e.g. LIDAR).

Planet Labs, Inc. (Planet) is a promising source of high-resolution imagery that can be used in environmental science studies, as it has both high spatial (0.7-3.0 m) and temporal (1-2 day) resolution. Planet imagery is acquired by a constellation of nanosatellites collecting data mostly in visible and by some at near-infrared bands with relatively narrow spectral bandwidth. However, its immediate utility with respect to inferring snow cover is limited due to the narrowness of the near infrared band which makes distinguishing snow from clouds difficult using a radiometric index (such as the Normalized Difference Snow Index, NDSI), and therefore require an alternative approach. The success of machine learning (ML) algorithmic approaches in geosciences led us to become interested in discovering whether a ML approach could be applied to the challenge of the spectrally-narrow Planet imagery for the purpose of snow cover classification. To this end we will employ a classification algorithm to discriminate between snow-containing pixels and snow-free pixels. A cloud-based approach is necessary as we will be using a large subset of Planet imagery and training and executing the ML model requires large computing power.

Our goals are to develop an open-source computer vision pipeline for orbital CubeSats Planet data to infer snow cover at meter-scale resolution using a machine learning approach. While we use mapping of snow-covered areas as an example, the workflow can be utilized by other Earth Sciences applications. Planet imagery has been used to identify a large range of changes in ground features (e.g greenness), to detect disaster affected areas (e.g. by floods, hurricanes and fires) or to assess changes in urban or agricultural patterns. Because the workflows are of interest for larger communities, we plan to develop tutorials and case studies to be used in training events such as hack weeks and incubator programs.

Project Objectives, Significance and Impact:

Objective 1. Develop code and workflows to process Planet data imagery to infer snow-covered area. To apply our ML approach, we will couple airborne-derived snow observations with Planet imagery in the Tuolumne River Watershed, California, USA. This watershed is important due to its role as the source of San Francisco's water supply. This site is also an ideal candidate for our analysis as we can train and validate our ML algorithm using airborne snow observations at high spatial resolution (3m) provided by the NASA JPL Airborne Snow Observatory (ASO). The ASO program (https://aso.jpl.nasa.gov/) used airborne LiDAR technology to collect snow depth data during the snowmelt seasons of 2013-2017 (including some of the lowest (2015) and highest (2017) snowpacks on record 6-11 times per year. We plan to use this extensive dataset to derive 3-m snow covered areas to train and validate the ML classifier (see Figure 1). The core algorithmic framework for this project is a Gaussian Process classifier coupled with a radial basis function (RBF, or squared exponential) kernel (Rasmussen and Williams, 2006). This approach is a general nonparametric machine learning algorithm capable of fitting to any function given by training data. Theoretical model performance will be assessed by examining quantitative cross validation-based metrics of classification accuracy (F score) using pixels extracted from training data as described above (Figure 1). The project will be designed for scalability and usability for spatial analysis, common features for a wide-range of Planet imagery-based applications.

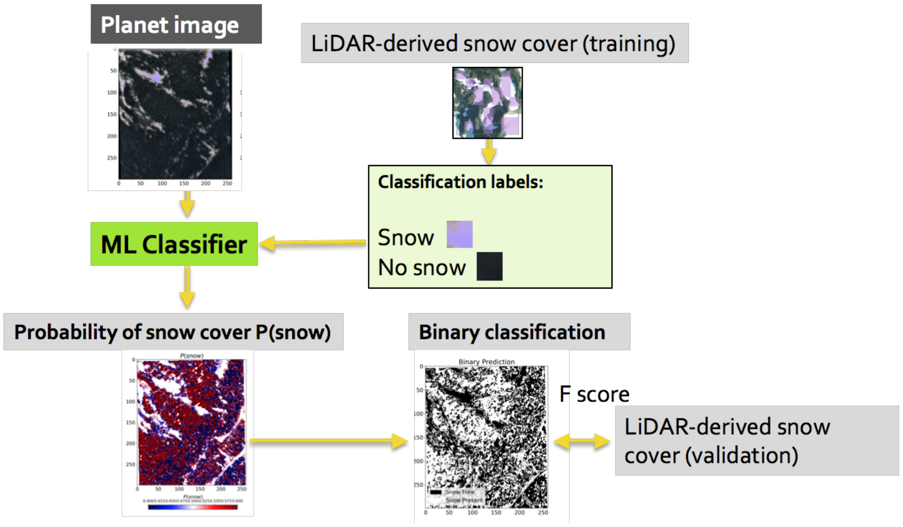

Figure 1. Workflow for a ML algorithm implementation to derive snow-covered areas from Planet imagery

Objective 2. Develop tutorials and case studies for Hackweeks and training events

We will develop tutorials based on our data processing and analytical pipeline. As the tools we intend to use are open-source with a focus on collaboration, reproducibility and sharing, the learning resources created based on this project will tie in well with tutorials, hackweeks and educational event developed by the UW eScience Institute and ESIP. The outcomes of this project will also be in line with educating the next generation of domain scientists in data science methodologies.

Key project steps and timeline:

1. **Cyberinfrastructure** (Feb. 2019 - Mar. 2019)

   We will build our pipeline on a public cloud platform i.e. Amazon Web Services (AWS) to utilize larger processing and storage capabilities. We will employ a test-driven development approach i.e. ensuring our processing pipeline and model is sufficiently tested before deploying massive computational workloads. This is to ensure that the cloud credits are utilized in a cost-effective manner. We will utilize the Spot market instances as necessary (not applicable during model training). We will also set up our own instance of Pangeo (http://pangeo.io) for processing the Planet dataset. Pangeo is a parallel computation platform (using Dask) for processing of big datasets (including arrayed datasets and imagery) with a Jupyter-based interface. Finally, all virtual machines that have been created will be saved as a machine image and all code/notebooks written will be shared through Github to ensure reproducibility.

2. **Data acquisition** (Mar. 2019 - May 2019)
   We will submit an academic data request to Planet.com for our Tuolumne River basin. Next, we will write a pipeline to access, download and clip Planet imagery to our area of interest (AOI). We will then extract related pixel values to derive the NDSI (Figure 1).

3. **Model setup and validation** (Mar. 2019 - Jun. 2019)

We will develop the algorithmic framework using a Gaussian Process classifier coupled with a radial basis function kernel. Thereafter, we will train and validate our algorithm using airborne snow observations. Finally, we assess model performance by cross-validating metrics of classification accuracy.

4. **Develop tutorials and case studies for hackweeks and training events** (Jun. 2019 - Aug. 2019)

3.0 Outreach

Our target audience includes researchers in academia, government, and industry engaged through hands-on teaching through simple but effective real-world examples using Planet imagery and cloud-based workflows. Our goal is to exemplify relevant methodologies to allow others to efficiently use the tools and instances to gain valuable insights about the domain and improve decision-making. Planet imagery has a world-wide coverage and we aim to bring this Planet-user community together by making the utilities available and inviting others to collaborate. Users will acquire experience on applying data science methods to improve their own understanding of the use of Planet data and methods, as our case study examples addresses broad questions and provides explicit details to ensure success of the proposed workflow.

The project's design end goal is scientific impact through peer-reviewed publication, engaging academic researchers and government bodies through sharing of open-source software and making all code and notebooks collaborative by publishing them on Github.

We will present the project and workflows at the ESIP 2020 summer meeting. Funding for this project will leverage existing ESIP lab support for the "Operational data provenance and cybersecurity for anticipatory disaster communication built on mesh networks" that is already funding two students to participate to Waterhackweek 2019. Continuation of ESIP funded participation in hackweeks creates a semi-trained student and early career community who can lead the open source research software development transition. This project outcome will be used at Waterhackweek and Geohackweek aimed at training researchers by using a hackweek educational model (Huppenkothen, 2018).

Integration with the ESIP community: ESIP lab will have the opportunity to participate in and sponsor Waterhackweek activities, which includes hands-on teaching and project work. The code developed through the proposed project, as well as code developed at the hackweeks on related projects will be shared with the ESIP community through ESIP communication channels and made available on Github. ESIP community members will benefit from the broad applicability and code reproducibility integrating cloud-based computing with geospatial analysis.

Project Partners (as applicable): Description of project partners (individuals and/or organizations) and their involvement: Dr. Nicoleta Cristea will oversee the cyberinfrastructure procedures, synthesis and workflow design. Dr. Cristea will coordinate yearly Waterhackweek events and collaboration with the ESIP Lab partners. Dr. Amanda Tan will build the cloud-based pipelines and will insure the scalability and flexibility of the methods. Graduate students Anthony Cannistra and Kavya Pradhan will implement the machine learning classifier procedures and algorithms and will manage the project GitHub account.