Project Summary:

Name of project: **Geoweaver: a web-based prototype system for managing compound geospatial workflows of large-scale distributed deep networks**

Project lead and contact details: Ziheng Sun, zsun@gmu.edu, 703-993-6124

Project partners and contact details: Liping Di (LAITS/CSISS), ldi@gmu.edu, 703-993-6114

Proposed start and end date: July 1, 2018 - Jan 1, 2019

Budget Requested: $7,000

Budget Summary: complete web wrapper on top of open-sourced deep learning/high performance computing library - $1,500; complete web-based workflow designer layout - $1,000; complete workflow instantiation module - $1,000; complete data visualization module - $1,000; complete connection bridge (SSH/Web API) between Geoweaver and data/function resources - $1,000; complete module integration and bug fix - $1,000; complete source code wrap-up, snapshot cloud instance and the Github final report - $500.

Project Outline:

Project description: Deep neural networks often run on distributed high-performance platforms to condense the long-lasting duration of training or testing. However, in spatial data related application, it is a daunting challenge to manage disparate spatial data storages and computational power, and dock the pre- or post- processing steps with the neural network. This project aims to prototype a web system, called Geoweaver, to allow users to easily compose and execute full-stack deep learning workflows in web browsers by taking advantage of the online spatial data facilities, high-performance computation platforms, and open-source deep learning libraries. The system will enable easier and more efficient integration of distributed resources, decrease the cost of building and managing deep networks, and realize highly across-institute collaboration and faster information extraction. In Geoweaver the data storages and software commands are represented as data entities and functional processes, which are chainable into workflows. The atomic processes in Geoweaver-created workflows could be web services (OGC web services), scripts or any other executables, which grants flexibility to Geoweaver users to reuse the function of the existing software or libraries. This project will develop a workflow designer prototype using D3 Javascript library and a workflow runner prototype based on tensorflow, deeplearning4j, Apache Spark, HDFS cluster and IaaS (Infrastructure as a Service) cloud. Geoweaver is a decentralized system and could be duplicated and installed on any instance VM to create and manage deep learning workflows in various hardware situations according to user requirements. We will showcase the concept by using the prototype in using Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to simulate land use changes from a massive volume of satellite images.

This project is motivated by our ongoing research - using Landsat images and deep learning classification algorithms to study land use changes and corresponding socio-economic influences. We are using the Cropland Data Layer (CDL) from USDA (United States Department of Agriculture) NASS (National Agricultural Statistics Service) as reference data to predict the unknown areas and periods. The classified maps will help yield estimation and agricultural drought monitoring. LSTM RNN is utilized in this research. We try to leverage cloud computing platform (GeoBrain Cloud) and parallel processing software (Apache Spark) to meet the challenge of tremendous number of pixels in remote sensing images (a single Landsat 8 scene contains more than 34 million pixels), but the entire experiment poses too many management issues for scientists to handle. We constantly run into disorganized confusion, hardware communication constraints and annoying configuration problems. A management software is eagerly needed to sort out the steps and processes, provide us with an overview dashboard to operate and manage the underlying facilities via the Internet, and track down issues and provenance. In our case, we need it to serving functions to create an intuitive compound workflow for the LSTM-based classification from raw images to land use maps, run the workflow on GeoBrain Cloud with Deeplearning4j and Spark, track the provenance of each map and share the results with other scientists via Email or social media.

The prototype architecture is proposed as shown in Fig. 1. The internal structure of Geoweaver is composed of five modules: VDP searcher, VDP orderer, workflow designer, data producer and data renderer. VDP represents **V**irtual **D**ata **P**roduct which corresponds to a workflow on demand generating real data product hereafter. Every newly created workflow in Geoweaver will be enrolled as a new VDP. VDP searcher and VDP orderer will directly reuse the modules of CyberConnector and the other three modules will be implemented in this project:

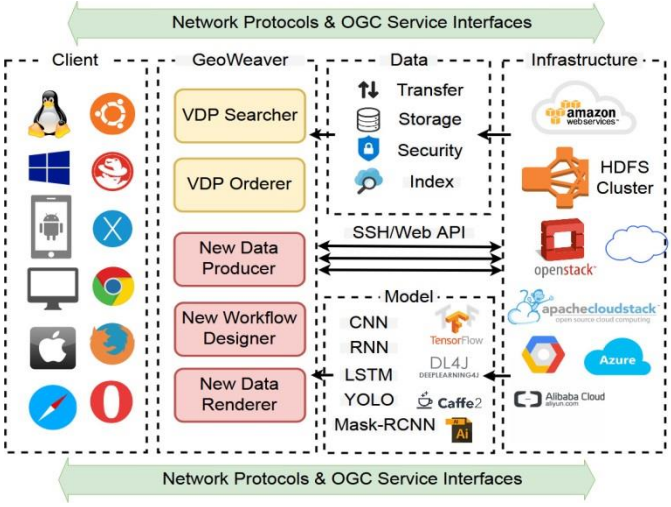

Figure 1. Architecutre design of Geweaver

1) The workflow designer will have a graphical panel for drag-n-drop processes and linking them into a workflow, and an editor panel for inputting Shell scripts, Java/Python code snippets and software commands as new processes. 2) The data producer will provide a dialog to configure the underlying infrastructures, such as cloud VMs, HDFS clusters, Spark clusters, and storage controller if possible. The communication will be conveyed via WebAPI (e.g., AWS-API, CloudStack API, OpenStack API, etc.) or SSH (Secure Shell). 3) The data renderer will provide an OpenLayers-based map page for scientists to review the classified land cover/land use maps and compare them with other methods or the original

Landsat scenes. It will also provide shortcut buttons to generate reports and charts of the land use changes in time series. The module also supports the downloading and sharing of the rendered maps, reports and charts.

All the modules will be developed completely by Web 2.0 techniques and available on all the mainstream operating systems and browsers, with no need for extra installation of plugins or clients.

Project objectives, significance and impact:

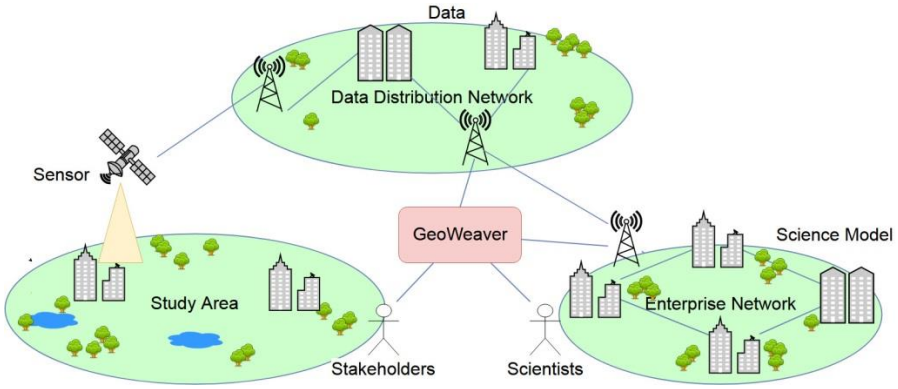

Figure 2. The prospective position of Geoweaver

The position of Geoweaver in cyberinfrastructure big picture is illustrated in Fig. 1. Data is obtained by sensors and transmitted to facilities where data distribution network will preprocess raw data into different levels of products and make them available via web services. Earth scientists have built all kinds of models which consume those products and export valuable information about Earth incidents, such as weather, hurricane, earthquake, drought, and wildfire. Geoweaver aims to help scientists pulling data and models (LSTM RNN in our case) together and make them feel like conducting modeling experiments on the same table while the execution happens on distributed hardware. Geoweaver also enables stakeholders to review the real-time extracted information from scientific workflow via the share button in data renderer module.

The presence of Geoweaver has significant meaning for cyberinfrastructure management in this big data era. The old production scheme is ambitious and confusing, and always requires too many involvements of scientists to deal with technical details to download data, set up models, streamline the processes, manage the outputs and track the provenance manually. Geoweaver brings a change to this landscape by realizing the separation of scientists from underlying infrastructures and manipulating resources in the streamlined workflow designer. Several instant beneficial impacts are expected from the adoption of Geoweaver: 1) turning large-scale distributed deep network into manageable modernized workflows composite of disparate atomic processes; 2) boosting higher utilization ratio of the existing cyberinfrastructures by separating scientists from tedious technical details; 3) enhancing the frequency and accuracy of classified land cover land use maps for agricultural purposes; 4) enabling the tracking of

provenance by recording the execution logs in structured tables to evaluate the quality of the result maps; 5) proofing the effectiveness of operationally using large-scale distributed LSTM network in classifying Landsat image time series.

Description of key project steps and timeline: 1) July 1 ~ July 31: kick off, set up the development environment, develop web wrapper of open-sourced deep learning/high performance computing library; 2) Aug 1 ~ Sep 30: develop workflow designer and data producer, complete connection bridge (SSH/Web API) between Geoweaver and data/function resources; 3) Oct 1 ~ Oct 31: complete data visualization module; 4) Nov 1 ~ Nov 30: complete module integration, create and conduct LSTM experiment and bug fix, 5) Dec 1 ~ Jan 1: complete source code wrap-up, upload demonstration video, snapshot cloud instance and finish the Github final report.

Outreach:

What groups/audiences will be engaged in the project? We will communicate with community members from ESIP, EarthCube CyberWay, AGU and AAG geospatial cyberinfrastructure groups on system interface and functionality design for best user experiences and usability. We will invite volunteer agricultural scientists with deep learning background from USDA/NASA to try the prototype and give us feedback. Any volunteer contribution to the project development is very welcomed.

How will you judge that project has had an impact? The impact will be judged by 1) tracking academic citations of our prototype system and publications; 2) tracking the website traffic and download times of Geoweaver website, installation package and cloud VM snapshot; 3) tracking the download and usage of the land use maps generated by Geoweaver; 4) creating a conference session about Geoweaver and survey the audience on site.

How will you share the knowledge generated by the project? The knowledge about Geoweaver prototype will be shared via technical reports, publications and conference presentations. The knowledge in the created deep learning workflow will be shared as a workflow package which can be imported, reviewed and reused in any properly installed Geoweaver instance.

Project Partners (as applicable):

Description of project partners (individuals and/or organizations) and their involvement: Professor Di will give guidance on the interoperability through standardized service interfaces and the training improvement of deep neural networks on behalf of agricultural informatics.

How will this project engage members of the ESIP community: LAITS, OGC Testbed, and EarthCube CyberWay members will be more or less engaged in this project as in-kind contributors. Other members can contribute as either advisors or test users. We will provide online updates and regular monthly telecon links to ESIP community to listen to advices and opinions.