Project Summary:

Name of project: **Geoweaver: a web-based 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 30, 2018 - Jan 31, 2019
Budget Requested: $7,000
Budget Summary: web wrapper of deep learning/cloud computing libraries - $1,200; reformed workflow graphical designer and instantiation module - $1,500; Landsat and CDL data-ingest pipeline - $1,100; web data visualization module - $500; OGC service digest pipeline - $700; connection bridge assembly between Geoweaver and computation facilities - $800; module integration joint kit - $600; ESIP winter 2019 attending cost - $600.

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 Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) 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 LSTM RNN, and realize highly across-institute collaboration and faster agricultural land use recognition. 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 create 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 simulating agricultural land use changes from a massive volume of satellite images.

This project is motivated by our ongoing research - using Landsat images and deep learning 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 are used in 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, 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 single Landsat 8 scene contains more than 34 million pixels). We eagerly need management software to sort out the configuration steps and processes, and provide us with an overview dashboard portal to operate and manage the underlying facilities via the Internet, and track issues and data provenance. In our case, we need it to serving functions to create an intuitive compound workflow for the LSTM-based classification from raw data to predicted CDL, run the workflow on GeoBrain Cloud with Deeplearning4j and Spark, track the provenance of each pixel to the raw Landsat scene and share the compared results to other scientists via Email or social media.

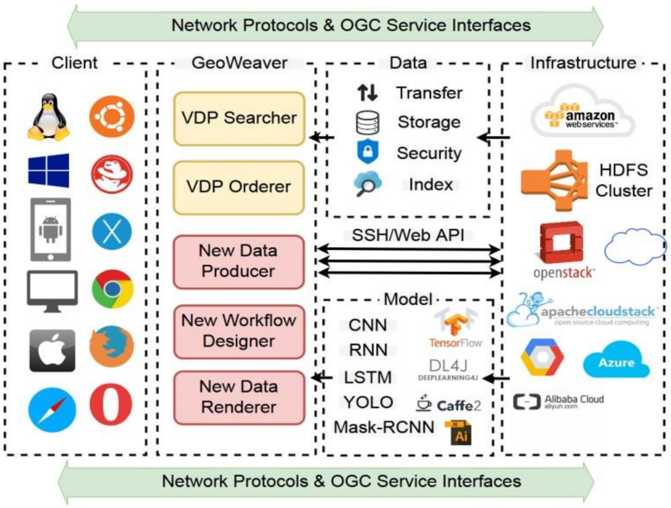

Figure 1. Architecutre design of Geweaver

The system 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:

1) The workflow designer will have a graphical panel for drag-n-drop of processes and linking them into an intuitive workflow, and an editor panel for inputting Shell scripts, Java/Python code snippets and system 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 between Geoweaver and the infrastructures will be conducted 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 results and compare them with the results from other methods or the original Landsat scenes. It will also provide buttons to easily generate reports and charts about the changes across map time series. The module also enables 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. We will continue to keep Geoweaver available online after the project is finished.

Project objectives, significance and impact:

The position of Geoweaver in cyberinfrastructure big picture is illustrated in Fig. 2. 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 together and make them feel like conducting modeling experiments on the same table while the execution actually 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 thriving agricultural land use researches in big data era. The old production scheme of land use products 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:

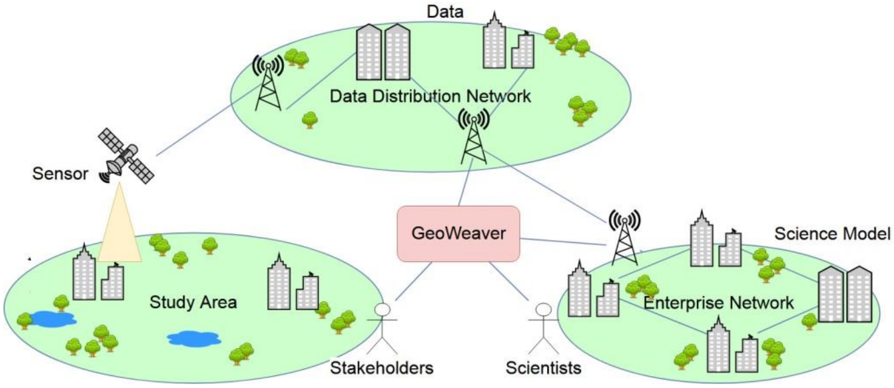

Figure 2. The prospective position of Geoweaver

1) turning large-scale distributed deep network into manageable modernized workflows;
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) proof the effectiveness of operationally using large-scale distributed LSTM network in classifying Landsat image time series.

Evidence has shown that open data and tools will directly lead to better science, especially a good tool could lead to an impressive scientific return per research dollar. Geoweaver is one of tool-developing efforts to assist in locating open remote sensing data and making it easily usable, for example, in generating predicted CDL and extracting new knowledge about agricultural land use changes. Publicly sharing workflows and products in Geoweaver increases citation rate, and encourages scientists to reuse the existing resources. Geoweaver also enables researchers to more rapidly answer complex queries using a streamlined interface rather than manually dealing with all sorts of low-level technical details.

Description of key project steps and timeline: 1) July 30 ~ July 31: kick off, set up the development environment, develop web wrapper on top of open-sourced deep learning/high performance computing library; 2) Aug 1 ~ Sep 30: develop workflow designer and data producer, complete bridge assembly 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, 5) Dec 1 ~ Jan 31: complete source code wrap-up, upload demonstration video, snapshot cloud instance, finish the GitHub final report and demonstrate it in ESIP winter 2019.

Outreach:

What groups/audiences will be engaged in the project? We are going to align Geoweaver efforts with the goals of the Agriculture and Climate cluster, Cloud Computing cluster, Energy and Climate cluster and Science Software cluster in ESIP. Discuss with the members in these clusters, help them use Geoweaver on their purposes, and make Geoweaver a useful tool to create a successful use case for them. We will communicate with community members from EarthCube, AGU and AAG geospatial CI 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 system and give us feedback. Any volunteer contribution to the project development is very welcomed.

How will you judge that project has had an impact? We will evaluate it by involving people from the mentioned ESIP clusters above, letting them review and try it and collecting in-situ feedbacks. An instant statement of the impacts of Geoweaver will be written to record those opinions, and eventually uploaded into GitHub repository.

How will you share the knowledge generated by the project? The knowledge about Geoweaver system will be shared via technical reports, publications and conference presentations. The knowledge stored in the workflow will be shared as workflow package which can be imported, reviewed and reused in any properly installed Geoweaver instance. Geoweaver will be open sourced on GitHub and take advantage of the power of open source communities to evolve and sustain. This team will also continue to contribute voluntarily after the project period is over.

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. Two students in GMU will join and contribute to the project together.

How will this project engage members of the ESIP community: We will provide online updates and regular bi-weekly telecon links to ESIP community to absorb advices and opinions. ESIP members can contribute as either advisors or test users. We will join in the four mentioned clusters in ESIP and try to engage people there. We will attend the ESIP winter meeting 2019 to demonstrate Geoweaver to ESIP members and collect their advices and try-on feedbacks.