CS486C – Senior Capstone Design in Computer Science
Project Description

Project Title: Web Visualization of High Dimensionality Spatial Data	
Sponsor Information:	Jason Laura, Research Scientist / Geographer USGS Astrogeology jlaura@usgs.gov 928.864.7366 Trent Hare, Geographer USGS Astrogeology thare@usgs.gov

Project Overview:

As a the "big data" revolution continues to accelerate, computer-based analysis and powerful visualizations of large data sets are becoming indispensible tools for extracting information from very large data sets. As an example, USGS is working on data sets produced by the Japanese Kaguya Spectral Profiler instrument. This instrument captures hyperspectral spot data of the lunar surface. This ~1.4TB data set is just becoming available to planetary U.S. scientists and is an essential tool for analysis of surface composition. The data supports answering questions like: What exactly is the moon made of? What does that mean for how the moon was formed? Longer term, the developed proof of concept serve as a basis to support not only visualization (the focus of this work) but also semantic data search.

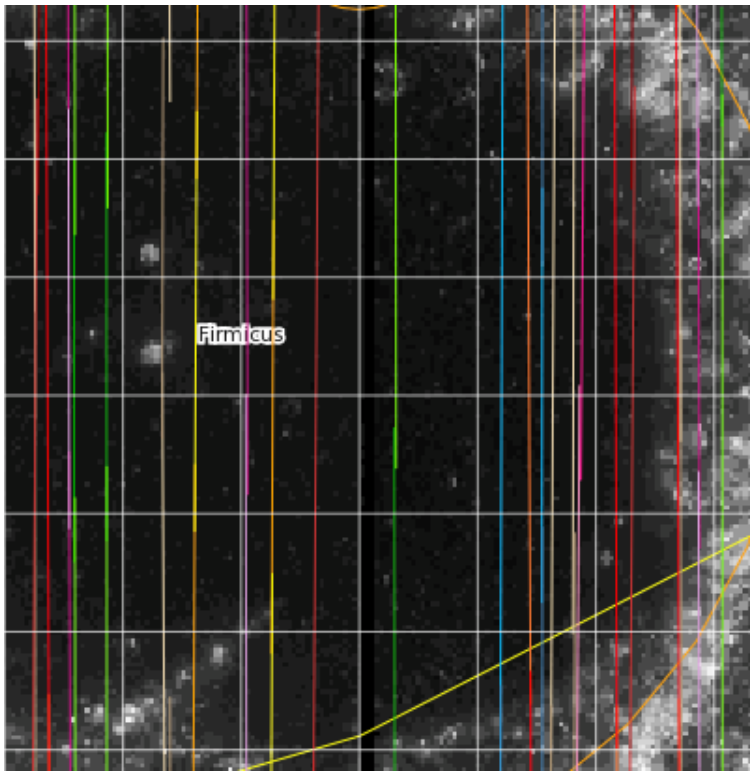

Figure 1: Sample data with spot observations captures in sequence colored and represented as a line. Each line is composed of ~96 points.

While the potential scientific benefit of these large new datasets is enormous, the visualization and analysis of increasingly large spatio-temporal data sets is an open problem with no clear single solution. This data set is composed of approximately 68 million spatially enabled point observations captured from the surface of the Moon; each observation is associated with a hyperspectral (~300 element vector)

observation, and associated observation condition metadata. Obviously, the sheer size of the data set makes efficient and timely analysis and visualization of such data sets a significant challenge.

The overall aim of this project is to develop a web application to support interactive visualization and Exploratory Spatial Data Analysis (ESDA) of a high-dimensionality spatial point database. The problem is that current visualization and data transfer techniques do not support a naive, bulk push of a ~68 million element GeoJSON (or other protocol) data blob while still supporting user interactivity. Visualization of such a high number of data elements also suffers from overdrawing, saturation, and binning issues. As a result, the ability for our users to interact with the data is severely limited.

One promising approach to tackling this problem is use of data shading techniques for small scale (global) view coupled with data streaming techniques for larger scale views. The primary use case for the proposed web application focuses on the scientist user, engaged in a multi-step process: First, a contextual basemap is loaded to support navigation of the scene. Next, the user will select a specific region of interest, based on additional meta-information about the dataset that must be provided by the application. Finally, the user will select a small subset of the total data volume (at most ~10 observations) and visualize the full ~300 element vector(s) as a line plot(s).

The hyperspectral data set describe above is recently acquired and is largely un-mined. A web application exposing basic ESDA functionality is an important first step in unlocking this data set for exploitation by planetary scientists. This proof-of-concept will be the first time that this data has been leveraged in this way by any U.S. scientists. Very exciting!

The goal of this project is to develop a prototype web application that:

- identify one of more visualization techniques for the entirety of the 68 million points and determine how to efficiently transmit that data to the user;
- assuming the above is not a bulk push of all ~68 million observations, determine at what spatial scales a different representation is appropriate and how best to transmit (stream) said data;
- supports visualization of the ~300 element hyperspectral data (vectors);
- allows users interactivity as the data loads or streams – the application should not freeze or hang during loading or navigation;

Of course, our team here at USGS can provide a lot of background information and strong leads for the Capstone team to pursue in designing and implementing a solution. We expect these overall specifications to become more precise as part of the early design and requirements process.

Knowledge, skills, and expertise required for this project

- Some familiarity with spatially enabled data, data visualization, and standard database access techniques will be invaluable in allowing the team to begin focusing on the visualization and data transfer problem immediately.
- Presentation and manipulation of high-resolution graphics within a webapp context.
- General web programming expertise, including familiarity with cloud architectures, and appropriate front and back end frameworks.

Equipment Requirements:

- We will provide a fully functioning single table database with a subset of the above ~68 million points. Sufficient drive space to store the database is all that is required. If this is problematic, we can determine if remote read-only access to the locally stored database is feasible.

Software and other Deliverables:

Basic deliverables include:

- Functioning web application to:
  - Visualization the entirety of the test data set;
  - Alter (if necessary) the visualization technique as map navigation occurs;
  - Visualize individual ~300 element data vectors for selected observations;
- Complete user manual for non-technical users;
- A strong as-built report detailing the design and implementation of the product in a complete, clear and professional manner. This document should provide a strong basis for future development of the product.
- If not included in the As-built report, documentation (via a wiki perhaps?) of design rationale detailing the exploratory design that happened, e.g. why was technique X selected over technique Y and what are the pros/cons?;
- Professionally documented (and tested) source code, in an online repository (Github, Bitbucket) that is viewable by our team.