


CS486C – Senior Capstone Design in Computer Science

Project Description

Project Title: A GUI interface for large data stream analysis for all-sky astronomical measurements	
Sponsor Information: 	David Trilling, Professor Astronomy and Planetary Science David.trilling@nau.edu Mike Gowanlock, Assistant Professor School of Informatics, Computing, and Cyber Systems Michael.gowanlock@nau.edu Northern Arizona University

Project Overview:

Understanding the physical properties of small bodies (asteroids, comets, etc.) helps us understand the formation and evolution of the Solar System and how similar we might be to other planetary systems. Additionally, creating a catalog of asteroid properties will help us prepare for the eventuality of an asteroid impacting the Earth. The primary limitation to acquiring this knowledge is simply the number of asteroids that can be observed, which in turn depends on the size of telescope and amount of data that can be collected. In other words, this science investigation is fundamentally a data science question.



Now, the big data revolution is coming to astronomy. We are entering the era of all-sky surveys, where the most or all of the entire sky is imaged every night. This is a dramatic change from old school astronomy, where each individual object (star, asteroid, galaxy) was observed one at a time; work in astronomy is becoming more and more a data science – that is, using big data tools to understand the nature of the universe.

As our data collection changes to a massive big data scale, it has become harder for astronomers to keep up with the volume of data being produced using traditional “manual” methods of analysis. In order to make progress in this big data context, astronomers must create new, automated computational analysis tools that are able to somehow pre-process this massive data stream to identify “interesting” elements, and then archive, organize, and index the results for easy access by scientists.

A great example of this big data revolution is the Vera C. Rubin Observatory’s **Legacy Survey of Space and Time (LSST, see pic below)**. The Rubin Observatory is being built in Chile, and first light (that is, the first data from the telescope) is expected in 2021. Full science operations will commence in 2023 or 2024. LSST will produce *20 terabytes of data every night*, for 10 years. At this scale, no person could ever look at an entire LSST image in any detail, and all discovery will need to be carried out with computational approaches.

There is probably no part of the astronomy research community that is ready for the massive volumes of data that will be produced by LSST at this time, and we urgently need to develop big data management tools and concepts so that we will be ready when it happens. Fortunately, a smaller (but still large-scale) prototype program called the **Zwicky Transient Facility (ZTF)**, located in San Diego County, California, is operating in an LSST-like mode,

producing essentially a one-tenth scale model of the LSST data stream. This makes it a perfect testbed for developing big data management tools that explore ways to manage the large data volumes that LSST will produce.

At NAU we are ingesting the part of the ZTF data stream that is broadcasting information related to asteroids, i.e., rocky bodies in the Solar System that are tracers of the formation and evolution of the Solar System since its formation 4.5 billion years ago. This information comes from ZTF in real-time during the night and passes through a “broker” in Tucson; we then ingest this data stream here at NAU.



To manage this large incoming data stream, we have created SNAPS: The Solar System Notification Alert Processing System. SNAPS has several facets, including this real-time data ingest as well as analysis tools for profiling the ingested data. This toolset is relatively mature from a computational standpoint; most of the analysis functions are developed and thoroughly tested. However, access to the SNAPS toolset is limited by its lack of a highly usable end-user interface. The **goal of this capstone project** is to address this shortcoming, by creating a powerful

graphical interface for visually indexing and profiling our ZTF datasets, and providing a web-based GUI to allow easy access to SNAPS analysis tools. Some key features of the envisioned product include the following.

Phase 0: Minimum Viable Product. Bare proof of concept

- Secure, modern Web2.0 web application with user authentication and role-based permissions.
- Basic GUI interface to browse our SNAPS database of measurements and derived properties.
- GUI to apply at least two analysis tools. Such analysis tools include interactive plots showing historical information of individual objects, including derived lightcurves, phase, and amplitude. Other interactive plots showing properties of the population of objects, such as 18/8 pixel aperture information and spatial distribution of objects in high-dimensional feature spaces, as projected into lower dimensions (2-D or 3-D).
- Interactive plots for data analysis: Plots will contain millions of data points. A challenge will be determining how to best render the plots so that the application remains responsive/useful to the user.

Phase 1: A nicely appointed application that is useful to actual scientists. Builds on Phase 0 features to provide:

- Advanced custom searching/filtering, ranking of objects based on outlier metric, and so on, to allow users to quickly zero in on data/objects of interest.
- An integrated GUI that includes all of the following SNAPS functions: rank list of targets of interest, allows custom filters that employ SQL scripting to allow users to customize how they access the database and receive alerts, and other scientifically-motivated sorting functions.
- Easy export of SNAPS analyses, e.g., graphs or data tables in a variety of forms: csv, pdf, png, FITS.
- Speed/bandwidth testing of the product with datasets of various sizes, specifically to show potential performance with larger scale projects like LSST.
- Automatic access other existing data sources, such as ANTARES, JPL Horizons, etc., to pull in additional information so that information regarding interesting objects is nicely aggregated in one place, avoiding the need to laboriously visit other webpages.

Phase 2: Stretch goals. Features that go beyond current functionality to explore future innovation.

- Mechanisms for scientists to save datasets, search results, and analyses in a sort of personal “lab notebook”. Could support multiple notebooks per user, e.g., for different analytic efforts. Allows users to quickly find and “restore” interesting datasets and analyses for continued analysis.
- Mechanism for scientists to share their analyses with colleagues, e.g., by making a particular lab notebook selectively available to other users who could then attach review/commentary.

- An extensible framework for supporting various future SNAPS modules. That is, a way to add a new SNAPS processing function to the GUI, e.g., by specifying the legal input parameters and ranges in some way, and the system uses this to automatically generate a GUI interface to the module.

When complete, this product could have a significant impact on our science community, and could become the cornerstone of a central data clearinghouse for all astronomers in this field of study. More broadly, the informatics concepts proved in this product could be modified and applied to other aspects of the ZTF data stream (not just asteroid data), and could serve as a model for data access management for large projects like LSST.

Knowledge, skills, and expertise required for this project:

- Some relational database knowledge and familiarity with SQL. This project will not involve building or maintaining a database but executing queries on our database.
- Interest in/experience with web design (both aesthetics and function). We can provide examples to emulate.
- Interest in/experience with API design.
- Working with (human) clients – both internally and external testers of our beta version.

Equipment Requirements:

- No special equipment outside of regular computer access and commonly available open-source tools and IDEs. Our database lives on NAU's HPC cluster, monsoon.

Software and other Deliverables:

- 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.
- Complete professionally-documented codebase, delivered both as a repository in GitHub, BitBucket, or some other version control repository; and as a physical archive on a USB drive.
- A web interface and API, as described above
- All documentation, including speed/bandwidth test outcomes