Journal of Physics: Conference Series **125** (2008) 012091 doi:10.1088/1742-6596/125/1/012091

Sunfall: a collaborative visual analytics system for astrophysics

Cecilia R Aragon, Stephen J Bailey, Sarah Poon, Karl Runge and Rollin C. Thomas
Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

E-mail: CRAragon@lbl.gov

Abstract. Computational and experimental sciences produce and collect ever-larger and complex datasets, often in large-scale, multi-institution projects. The inability to gain insight into complex scientific phenomena using current software tools is a bottleneck facing virtually all endeavors of science. In this paper, we introduce Sunfall, a collaborative visual analytics system developed for the Nearby Supernova Factory, an international astrophysics experiment and the largest data volume supernova search currently in operation. Sunfall utilizes novel interactive visualization and analysis techniques to facilitate deeper scientific insight into complex, noisy, high-dimensional, high-volume, time-critical data. The system combines novel image processing algorithms, statistical analysis, and machine learning with highly interactive visual interfaces to enable collaborative, user-driven scientific exploration of supernova image and spectral data. Sunfall is currently in operation at the Nearby Supernova Factory; it is the first visual analytics system in production use at a major astrophysics project.

1. Introduction

Many of today's important scientific breakthroughs are being made by large, interdisciplinary collaborations of scientists working in geographically widely distributed locations, producing and collecting vast and complex datasets. These large-scale science projects require software tools that support not only insight into complex data but collaborative science discovery. The technique of "visual analytics," defined by Thomas in *Illuminating the Path* [1] as "analytical reasoning facilitated by interactive visual interfaces," combines statistical algorithms and advanced analysis techniques with highly interactive visual interfaces that support collaborative work. Such approaches offer scientists the opportunity for in-depth understanding of massive, noisy, and high-dimensional data. Astrophysics in particular lends itself to a visual analytics approach due to the inherently visual nature of much astronomical data (including images and spectra).

One of the grand challenges in astrophysics today is the effort to comprehend the mysterious "dark energy," which accounts for three-quarters of the matter/energy budget of the universe. The existence of dark energy may well require the development of new theories of physics and cosmology. Dark energy acts to accelerate the expansion of the universe (as opposed to gravity, which acts to decelerate the expansion). Our current understanding of dark energy comes primarily from the study of supernovae.

The Nearby Supernova Factory (SNfactory) [2] is an international astrophysics experiment designed to discover and measure Type Ia supernovae in greater number and detail than has ever been done before. These supernovae are stellar explosions that have a consistent maximum brightness, allowing them to be used as "standard candles" to measure distances to other galaxies and to trace the rate of expansion of the universe and how dark energy affects the structure of the cosmos. The SNfactory receives 50–80 GB of image data per night, which must be processed and examined by teams of domain experts within 12–24 hours to obtain maximum scientific benefit from the study of these rare and short-lived stellar events.

To facilitate the supernova search and data analysis process and enable scientific discovery for project astrophysicists, we developed Sunfall (SuperNova Factory AssembLy Line), a collaborative visual analytics system for the Nearby Supernova Factory that has been in production use for over 18 months. Sunfall incorporates sophisticated astrophysics image processing algorithms, machine learning capabilities including boosted trees and support vector machines, and astronomical data analysis with a usable, highly interactive visual interface designed to facilitate collaborative decision making. An interdisciplinary group of physicists, astronomers, and computer scientists (with specialties in machine learning, visualization, and user interface design) were involved in all aspects of Sunfall design and implementation.

The major contribution of this paper is the description of a novel visual analytics framework for supernova search and followup that is currently in production use and has significantly improved all aspects of the scientific data flow. Additional contributions include (1) an evaluation of multiple machine learning algorithms and their effectiveness in supernova search, (2) intelligent and highly interactive visual interfaces for facilitating scientific insight into supernova data, and (3) novel image-processing algorithms for supernova search, including Fourier contour analysis for image processing. Lessons learned from the development of this framework will have value for the design of future automated transient search projects such as the SuperNova Acceleration Probe (SNAP) [3] or Large Synoptic Survey Telescope (LSST) [4].

This paper is organized as follows. Section 2 describes the astrophysics background for supernova detection and spectral analysis, including the SNfactory project data flow. Section 3 contains information on previous systems built for other supernova experiments. Section 4 discusses the Sunfall design approach, and Section 5 describes the Sunfall architecture, including its four major components: Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse. Section 6 discusses our conclusions and lessons learned, and Section 7 presents ideas and plans for future work.

2. Science background

The discovery of dark energy is primarily due to observations of Type Ia supernovae at high redshift (up to z = 1, or a lookback time of about 8 billion years) [5,6]. This supernova cosmology technique hinges on the ability to reliably compare luminosities of high-redshift events to those at low redshift, necessitating detailed study of low-redshift events. Large-scale digital sky surveys are now being planned in order to constrain the properties of dark energy. These studies typically involve high-volume, time-constrained processing of wide-field charge-coupled device (CCD) images, in order to detect Type Ia supernovae and capture detailed images and spectra on multiple nights over their lifespans.

A stellar explosion resulting in a Type Ia supernova will occur on average only once or twice per millennium in a typical galaxy of 400 billion stars and, then, will remain visible for only a few weeks to a few months. Additionally, the maximum scientific benefit is obtained if the supernova is discovered in the first two weeks before it attains peak brightness. To further add to the challenge, the imaging data from which supernovae must be detected are extremely noisy, corrupted with spurious objects such as cosmic rays, satellite tracks, asteroids, and CCD artifacts such as diffraction spikes, saturated pixels, and ghosts.

The Nearby Supernova Factory is an international project designed to tackle these ambitious goals and to accumulate the largest homogeneously calibrated spectrophotometric (containing both images and spectra) dataset of Type Ia supernovae in the "nearby" redshift range ($0.03 < z < 0.08$, or about 0.4 to 1.1 billion light years distant) ever studied [2]. On a typical night, 50–80 GB of data (approximately 30,000 images containing 600,000 potential supernovae) are received by SNfactory image-processing software. These data are processed overnight; the best candidates are selected each morning by humans for further follow-up measurements. Likely candidate supernovae are sent to a dedicated custom-built spectrograph for follow-up imaging and spectrography. The resulting spectra typically each contain 2000–3000 data points of flux as a function of wavelength. The SNfactory database is expected to be released to astronomers worldwide and become a definitive resource for measurements of dark energy. This program is the largest data volume supernova search currently in operation.

The SNfactory obtains wide-field imaging data from the Near Earth Asteroid Tracking program (NEAT) [7] using the 112 CCD QUEST-II camera of the Mt. Palomar Oschin 1.2 m telescope [8], covering 8.5 square degrees per exposure. (For comparison, the full moon covers 0.2 square degrees.)

Images are transferred from Mt. Palomar via the High Performance Wireless Research and Education Network (HPWREN) to the High Performance Storage System (HPSS) at the National Energy Research Scientific Computing Center (NERSC) in Oakland, California. Each morning, SNfactory search software running on NERSC's 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF), matches images of the same area of the sky, processes them to remove noise and CCD artifacts, then performs an image subtraction from previously observed reference images on each set of matched images (figure 1) [9-11].

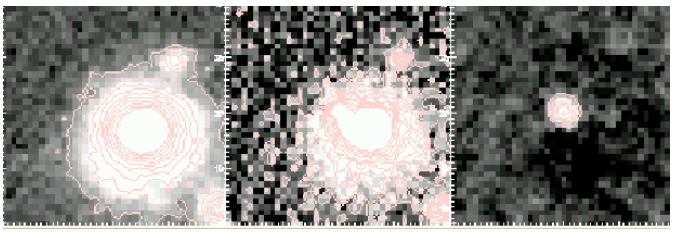

Figure 1. Example supernova images, from left to right: reference image of galaxy, new image of galaxy plus new bright spot, subtracted image reveals new supernova.

Supernova candidates are identified from among the over 600,000 objects processed per night by a set of image features including object shape (roundness and contour irregularity computed from Fourier contour descriptors) [11,12], position, distance from nearest object, and motion. These features are used as input to machine learning algorithms that select candidates to be sent to humans for scanning and vetting [9,13]. Promising supernova candidates that pass the human scanning and vetting procedure are sent for confirmation and spectrophotometric follow-up by SNIFS (the SuperNova Integral Field Spectrograph) [2] on the University of Hawaii 2.2 m telescope on Mauna Kea.

Candidates are imaged through a 15 x 15 microlens array on SNIFS and the spectral data is saved to a database in France. Currently the SNfactory has collected spectrophotometric data on over 150 Type Ia supernovae, and more data is being gathered each night.

Spectra are the best key to understanding the actual physics of exploding stars. Understanding Type Ia supernovae through their spectra is important because it places constraints on their presupernova evolution and the immediate stellar environment of the event and provides clues that can be turned into correlations between peak brightness and the physics of the explosion. With a better understanding of the physics of Type Ia supernovae through their spectra, it is believed that their utility as standardized candles for cosmological distance measurements can be refined for their use in future precision cosmology experiments such as SNAP, DESTINY, or ADEPT [3, 14, 15].

3. Related work

Since the discovery of dark energy via studies of Type Ia supernovae nearly a decade ago [5,6], astrophysicists have developed several large-scale supernova searches to collect as much data as possible on these rare events. These searches typically rely on custom image subtraction software containing highly complex, hand-tuned heuristics and "cuts" to extract supernova candidates. They are often plagued by large numbers of false positives, which must then be screened out by humans in a labor-intensive process. For example, the 2005 Sloan Digital Sky Survey II (SDSS-II) supernova program generated approximately 4,000 objects per night, which needed to be visually checked by humans for verification [16]. The ESSENCE and SNLS supernova searches both resulted in 100–200 objects to scan per night, with a significantly smaller input data load than the SNfactory [17]. Techniques used in these surveys will not scale to future, much larger datasets.

After supernova data has been collected, they are typically stored in a database accessible via a web-based interface. These web sites often present information in tabular form, listing the supernova coordinates and providing links to images. They are designed to provide all the necessary information for astronomers to observe supernovae with their own telescopes. Examples are the Supernova Legacy Survey (SNLS) and Sky Survey (SDSS) websites [18,19]. In [16], Sako et al. describe the SDSS software tools for supernova search and follow-up. They include sets of scripts that perform functions similar to parts of Sunfall, but require human intervention at numerous stages during the supernova search and follow-up process. None of these groups has built a complete, automated visual analytics system that encompasses the entire data search, analysis, and scientific discovery process.

4. Sunfall design process

To design an effective collaborative visual analytics system for the SNfactory, we first conducted, in mid-2005, an extensive, two-month evaluation of the existing SNfactory procedures and environment. Data sources used for evaluation included individual interviews, observation of team members performing typical project tasks, review of existing source code, literature reviews, examination of other supernova search projects, and consultation with physicists and computer scientists with relevant experience building similar scientific software systems.

We conducted over 100 hours of interviews with LBNL scientists, postdoctoral researchers, and students and with SNfactory collaboration members outside LBNL. This included 19 current and former team and collaboration members, and 12 scientists with relevant experience outside the SNfactory collaboration. We also performed a detailed software review of over 150,000 lines of existing SNfactory legacy code in C++, IDL, Perl, and shell scripts.

We established the following requirements for the Sunfall software framework: It must encompass the entire scientific data capture, processing, storage, and analysis process, enable collaborative, time-critical scientific discovery, and incorporate SNfactory legacy code (custom astronomical image-processing algorithms). It must reduce the number of false positives sent to humans and improve the quality of the subtraction image processing. It must automate repetitive data transfers and other manual tasks to leverage domain experts' unique processing and image recognition skills to the maximum extent.

The Sunfall user interface was designed and implemented using participatory and iterative design techniques; for example, interactive prototypes were used to evaluate areas where existing interfaces did not support scientists' workflow. Scientists' feedback sometimes led to major redesigns of the interface. The system was implemented from the beginning with this possibility in mind, so that changes were easily made and accepted as an appropriate part of the development process [16]. Sunfall has been in operation at the SNfactory since 2006.

5. Sunfall architecture and components

Sunfall consists of four major components: Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse (figure 2). This section describes each component's structure in detail.

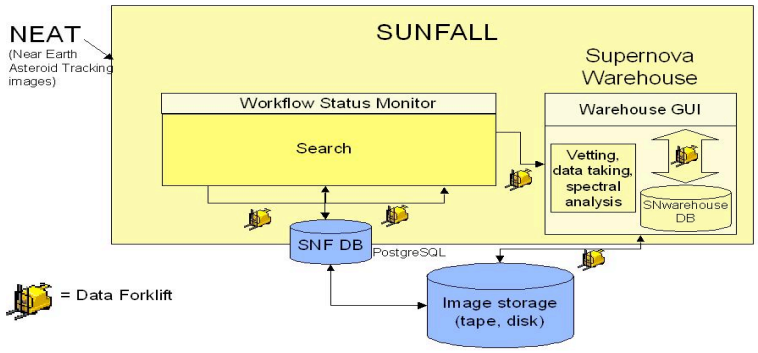

Figure 2. Sunfall architecture diagram, depicting the four components (Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse) and data flow between the components.

Search handles image processing and subtraction and includes machine learning algorithms and novel Fourier contour descriptor algorithms to reduce the number of false positive SN candidates. The Workflow Status Monitor is a web-based monitor that facilitates collaboration and improves project scientists' situational awareness of the data flow by displaying all relevant workflow (search pipeline) status data on a single site.

The Data Forklift is a middleware mechanism consisting of a coordinator and a suite of services to automate astronomical data transfers in a secure, reliable, extensible, and fault-tolerant manner. The Data Forklift also provides the middleware for the other three components, transferring data between heterogeneous systems, databases and formats securely and reliably.

The Supernova Warehouse (SNwarehouse) is a comprehensive supernova data management, workflow visualization, and collaborative scientific analysis tool. The SNwarehouse contains a PostgreSQL database, Forklift middleware, and a graphical user interface implemented in Java.

5.1. Search

The supernova search component of Sunfall is an example of *analytic discourse*, defined in *Illuminating the Path* as "the interactive, computer-mediated process of applying human judgment to assess an issue," [1] applied to the realm of astrophysics. Sunfall Search has increased efficiency of supernova detection and enabled more effective human intervention by reducing the number of false positive candidates by nearly an order of magnitude [6,13] and, through the use of intelligent interfaces, by increasing human efficiency in evaluating candidates by a factor of four.

The legacy search software computed approximately 19 photometric and geometric features on each of over 600,000 supernova candidate subimages per night and applied threshold "cuts" to each of these features. Subimages that satisfied all thresholds (1,092 on a typical night in April 2006, just before the new software was put into production) were sent to human scanners who selected potential supernova candidates for follow-up study. The majority of subimages that passed the thresholds were false positives that humans had to manually reject. These manually tuned thresholds were brittle and often resulted in both missed SNe and too many false positives for human scanners to evaluate daily. The process of scanning on the order of 1,000 subimages took 6-8 people several hours per day.

Machine learning algorithms, including support vector machines and boosted trees, were incorporated into the search software. By replacing simple threshold "cuts" on the image features with classifiers of the high-dimensional data in the feature space, we reduced the number of false positives by an order of magnitude (from 1,094 to 113 on typical nights in 2005 and 2006). This resulted in labor savings of nearly 90%, where the process of scanning now takes one person a couple of hours each day. Additionally, system tests determined that the new algorithms had a true positive ratio equal to or better than the original algorithm (in other words, they did not miss any more actual supernovae). The data is depicted in figure 3.

Currently, of the approximately 100 visually scanned subimages per night, about 10-30 are flagged as containing potential supernova candidates. These candidates (about .001% -.01% of the objects originally imaged) are sent to the SNIFS spectrograph [2] for spectrophotometric screening and follow-up. Typically from two to five of the objects examined by SNIFS will finally be determined to be supernovae, of which (on average) about one per night is the sought-after Type Ia supernova. Thus of the objects originally imaged, somewhere near .0002% prove to be Type Ia supernovae. This process is described in more detail in [9,13].

For confirmation and selection of promising supernova candidates, the scanners use a visual interface to view and analyze the candidates. This visual scanning interface was redesigned to increase interactivity and allow humans to focus exclusively on the imagery itself. Previously, operators of the interface spent much of their time cutting and pasting long filenames, pausing the scanning process to look up data on several different external web sites, and waiting for image data to be retrieved from tape storage. Previously, this process took two minutes per candidate on average; with Sunfall, a decision could be made within 30 seconds. By applying intelligent algorithms to the search data itself, we were able to leverage the unique human abilities that enable the final detection of true supernova candidates and not only minimize the number of false positives but also increase the efficiency of the spectroscopic follow-up.

5.2. Workflow Status Monitor

The Workflow Status Monitor is a web-based dashboard application, designed to improve project scientists' situational awareness of the data flow by synthesizing diverse information flows from various systems and displaying critical workflow status data on a single site (figure 4).

The supernova search software is highly parallelized as 30,000 images are queued for processing in a multistage pipeline that runs on NERSC's 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF). Nodes frequently go down or jobs fail, and failures must be detected promptly and jobs resubmitted quickly due to the time-critical nature of the search. Detection of such failures was challenging and time-consuming before the deployment of the Workflow Status Monitor. The monitor displays graphs and visual displays of job completion times on PDSF, job queues, PDSF node uptimes, and disk vault loads. The interface was prototyped, evaluated for efficacy by project scientists, and implemented via PHP on a web site hosted on an SNfactory server running SuSE Linux.

The dashboard also displays date and time information at project locations in several formats relevant to astronomers. One author's previous experience with a status monitor for the Mars Exploration Rovers [21] demonstrated that apparently minor features, such as a clock that displayed the time on Mars and at various project locations on Earth, yielded tremendous improvements in scientists' situational awareness and may very well have prevented critical errors. Additional information displayed includes telescope scheduling, supernova candidates found, and current operational status of telescopes, spectrographs, and cameras used in processing SNfactory data.

5.3. Data Forklift

Each night, over 50 GB of heterogeneous, distributed supernova data arrives at the SNfactory and needs to be processed, managed, analyzed, queried, and displayed – all within a time period of 24 hours or less (ideally 12 hours). The data is distributed over wide geographic distances; some of it is hosted on systems outside the SNfactory's control, and some on unreliable systems with frequent downtime.

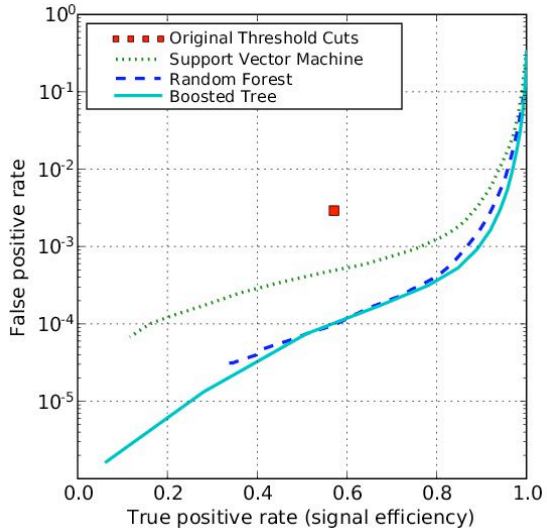

Figure 3. Comparison of Boosted Trees (solid line), Random Forest (dashed line), and Support Vector Machine (dotted line) for false positive identification fraction vs. true positive identification fraction. The lower right corner of the plot represents ideal performance.(from S. Bailey et al. [13]).

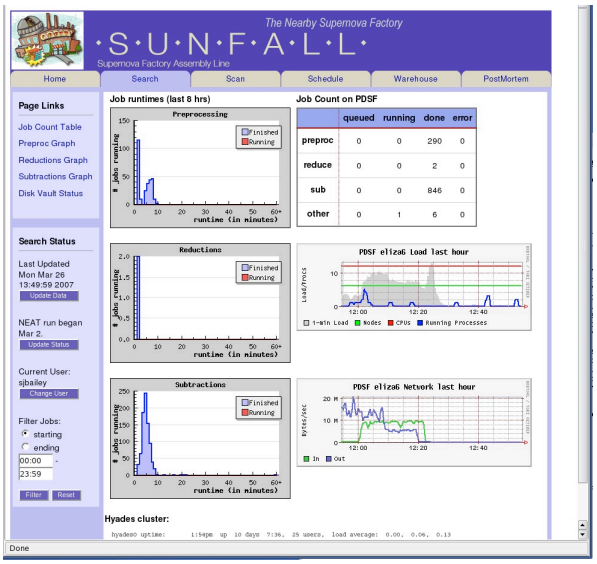

Figure 4. Sunfall Workflow Status Monitor. This web-based dashboard presents a single-page overview of many relevant information flows from multiple systems and, according to project scientists, increases operator efficiency significantly.

To solve these data management problems, we designed and implemented the Sunfall Data Forklift, a suite of services for automated data retrieval, storage, movement, querying, and staging for display. The Data Forklift consists of a coordinator and a set of individual services, each customized for its particular task, but having key attributes in common.

The Data Forklift supports "different place, asynchronous" collaborative scientific work by facilitating data and information transfer amongst a geographically separated team. Because of the time-critical nature of the data collection (telescopes must be operated at night and are located in different time zones), tasks must take place at distinct, specified times.

All Data Forklift services share the following properties:

- The mechanism resides at a central location (an SNfactory server), but processes data remotely from widely separate geographic locations, including making connections to external machines not under the SNfactory's control.
- All data transfers are secure, utilizing encrypted channels and authentication.
- The Forklift services and coordinator are reliable and self-restarting. The Forklift coordinator restarts itself if its process dies or if the server is rebooted.

- The mechanism is fault-tolerant; services detect unreliable connections and recover from errors in data transfer. Partially transferred files are detected and retransferred, and the process restarts the transfer at the point of failure.
- The mechanism is extensible (new tasks can easily be added). The Forklift coordinator was designed so that new services can easily be added to a central table.
- The Data Forklift enables data retrieval from external sources. Many astronomical databases are web-based, and designed for interactive rather than automated retrieval. Forklift services handle such interactions in a fault-tolerant manner. SNfactory spectral data is stored at a central processing database in France and must be retrieved according to external requirements.
- User actions can trigger Data Forklift requests. For example, whenever a supernova candidate is saved to the Supernova Warehouse (SNwarehouse) database, several Forklift services are automatically started, retrieving related image data from tape storage, accessing external web-based asteroid, galaxy, and other astronomical databases, converting image files into display format, and staging data for display.
- Forklift services act as middleware for the Supernova Warehouse, retrieving data from internal and external databases, performing program logic, and staging data for display by the SNwarehouse GUI.
- The Data Forklift is written in Perl and runs under SuSE Linux 9.3 on an SNfactory server.

5.4. Supernova Warehouse

The Supernova Warehouse (SNwarehouse) is a comprehensive supernova data management, workflow visualization, and collaborative scientific analysis application. It consists of a PostgreSQL database hosted on a dedicated SNfactory database server running SuSE Linux 9.3, Data Forklift services as middleware, and a graphical user interface (GUI) written in Java. SNwarehouse supports collaborative remote asynchronous work in several different ways.

Collaboration members can access the GUI from any networked computer worldwide via a Forklift remote deployment mechanism. Security is provided via password authentication and encrypted communication channels. SNwarehouse furnishes project scientists with a shared workspace that enables easy distribution, analysis, and access of data. Collaboration members can view, modify, and annotate supernova data, add comments, change a candidate's state, and schedule follow-up observations from work, home, while observing at the telescope, or when attending conferences. This access is critical due to the 24/7 nature of SNfactory operations. All transactions are recorded in the SNwarehouse database, and the change history and provenance of the data re permanently stored (records cannot be deleted in order to maintain the change history) and continuously visible to all authenticated users.

SNwarehouse centralizes data from multiple sources and supports task-oriented workflow. Project members perform well-defined tasks, such as vetting, scheduling, and analyzing targets, which collectively accomplish the goal of finding and following Type Ia supernovae. Typically, an individual or small group performs a given task, and the results of the task provide inputs for the next task in the workflow, often performed by another set of group members. Thus, the inputs and outputs of any task must be well defined and easily recognizable.

5.4.1. SNwarehouse overview. SNwarehouse's interaction design takes the approach of overview, filter, and drill down to details. The main overview page (figure 5) displays two tightly coupled representations of the list of targets registered in the database. The top visualization plots the targets in the sky; below is a sortable, tabular representation of the same data. From here one can drill down as needed based on the desired task.

5.4.2. Supernova candidate vetting. "Vetting" involves a process of scientific analytic discourse where the scientist must quickly decide, based on limited data, how to allocate scarce and expensive telescope time to the night's supernova candidates. At this point, before spectra are taken, it is often unknown whether a target candidate is a supernova, so the scientist must gather as much relevant data as rapidly as possible to make an informed prediction. This task involves retrieving any available

images of the target's location in the sky prior to discovery and querying several external astronomical databases.

Once a target is saved as a candidate supernova, the target name will appear in the SNwarehouse overview table color-coded in red, indicating that the target is newly discovered. On a rolling basis, a background process checks if previous data products exist on disk. If found, these data products are registered with the database, and a "!" appears next to the target name, a flag signaling newly found images. The combination of these two visual clues allows the scientist to quickly see which targets need evaluation. Once this determination is made, the vetter will then open the Details View for the target (figure 6).

At this stage, the vetter is trying to determine whether this target is potentially a Type Ia supernova, based on how consistently the magnitude is rising in comparison to standard Ia's. The Details View helps the vetter make this determination in two ways. First, a visualization of the magnitudes of all target observations against a standard Ia lightcurve (figure 6) allows the vetter to quickly determine whether the magnitudes are rising like those of a typical Ia. In addition, the vetter can compare the discovery image with prior images taken at this particular set of coordinates in the sky (figure 6).

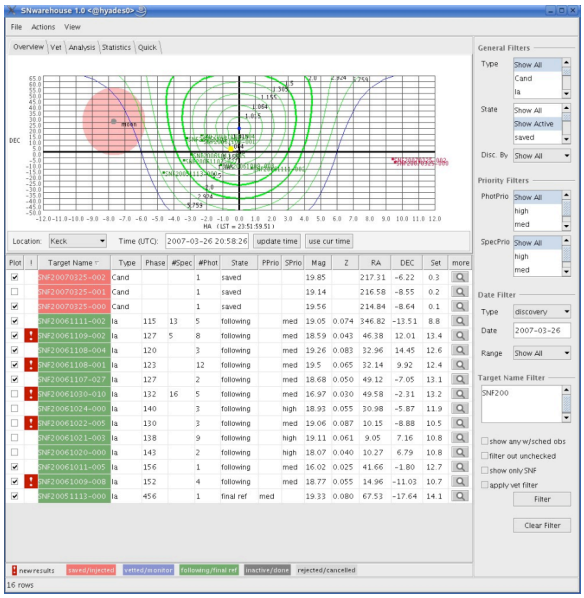

Figure 5. SNwarehouse overview tab. The Sky visualization at the top plots declination vs. hour angle (astronomical sky coordinates) in a view preferred by domain experts. The green lines represent airmass (the thickness of the atmosphere) at particular coordinates at the specified time.

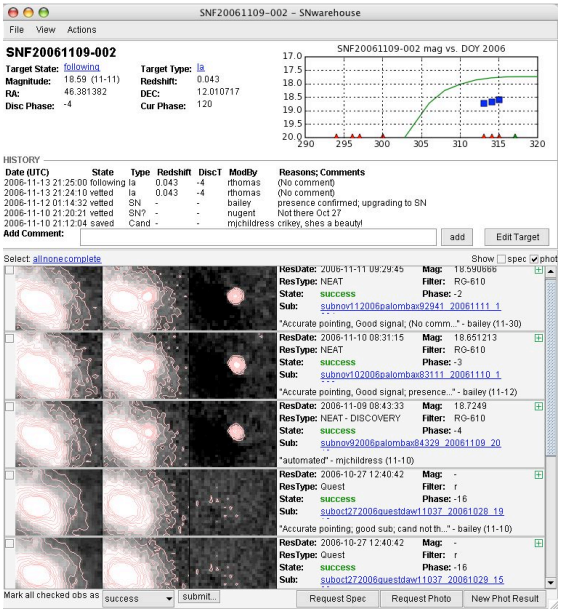

Figure 6. SNwarehouse details view. Comparison to standard lightcurve is in upper right-hand corner. The five rows at the bottom were imaged on different dates; note the target supernova is visible at the time of discovery but not on prior dates.

5.4.3. Observation scheduling. Prior to followup observation at the telescope, a schedule is made based on the vetter's assessments to order and allocate telescope time. In SNwarehouse, the scheduler starts by filtering for only those targets with observation requests. With this list, the scheduler must order and assign exposure times for each target using a variety of techniques. Exposure times are automatically determined using a lookup table based on the phase and redshift of the target. Considering these exposure times, the scheduler must also order the target list according to when the target will be visible in the sky by the telescope. A custom visualization offers insight for this task.

The Sky visualization, described in [22] and also shown at the top of figure 5, depicts the positions of targets in the sky at a given time and ground location. The green lines represent airmass (atmosphere thickness) for target coordinates at the specified time. The blue line is the horizon, and the red halo around the moon is the "lunar exclusion zone," where light cast by the moon makes it difficult to view faint targets. The yellow circle depicts the sun. Major telescope names and corresponding latitudes and longitudes are displayed on a drop-down menu so the visualization can be used world-wide. The time can be changed so the viewer can plan observations for the remainder of the night. If

the target appears within the blue horizon lines, it is visible to the telescope, assuming good weather. The scheduler must note the rise and set times of each target when creating the schedule for the night.

5.4.4. Data taking. During each night of spectral observations with SNIFS at the University of Hawaii 2.2 m telescope, an observer needs to point the telescope at each scheduled target in order. In the event of weather or mechanical problems, a rescheduling decision must be made quickly and efficiently. A visual data-taking interface facilitates this process. Much of the observation procedure is automated, allowing the observer to concentrate on troubleshooting, rescheduling, and determining the success of each target observation.

5.4.5. Analysis and post-mortem. Once observation at the telescope is complete, the data products (images and spectra) from the telescope are registered with the database and marked with a "!" in the SNwarehouse overview page. The next task is spectral processing and analysis, known as the "post-mortem" process. The scientist performing post-mortem filters for targets with new data products that are being followed (highlighted in green and blue, depending on the type of followup) and opens up the details view to evaluate the telescope results (figure 7).

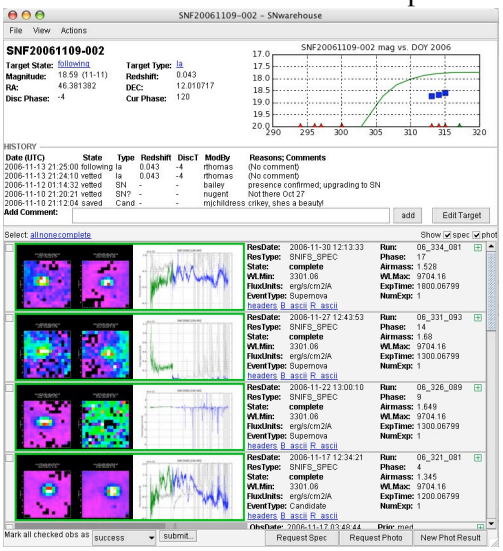

Figure 7. SNwarehouse details view depicting spectral data observations. The small images on the left-hand side of each observation subwindow indicate the accuracy and signal strength of the received data.

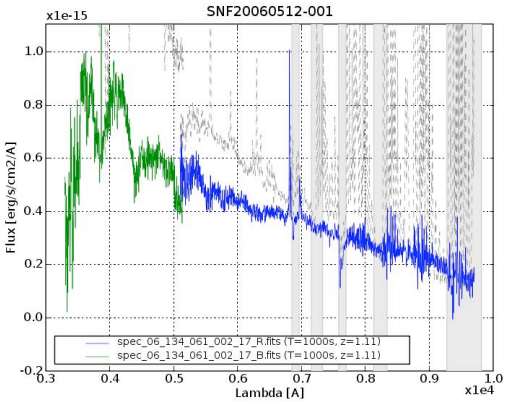

Figure 8. A spectrum from a very early Type Ia supernova, 15 days before peak brightness. The green and blue plots indicate the spectral data. The spiky grey lines in the background depict the background sky spectrum. This supernova is just starting to show the features that define a Type Ia.

The raw data from the SNIFS spectrograph is complex and requires a significant amount of processing in order to yield meaning to the scientist. A visual depiction of the accuracy of the pointing and signal strength provides much more information more quickly than tables of numeric data. The two small images on the left-hand side of each observation subwindow are a custom visualization designed to indicate whether the telescope accurately pointed at the target and if a good signal was received by the spectrograph (figure 7). Color-coding and position indicate the accuracy and signal strength of the received data. If there is a bright dot in the center of both squares, that is a clear indicator of a "good pointing."

An early Type Ia supernova (captured well before peak brightness) will display a certain pattern in its spectral plot (figure 8). A later supernova (imaged a few days past peak brightness) will show other characteristic spectral lines and features. Spectral data are plotted in green and blue. The spiky grey lines depict the spectrum of the background sky. The broad grey bands represent areas of atmospheric absorption. Because of the complexity of the data (thousands of points of flux vs. wavelength for each observation) and the necessity to make rapid, accurate decisions in order to maximize the use of limited, expensive telescope time, visualization provides the most efficient solution to the problem. The scientist can also access the raw numeric data by clicking on any of the images or spectra.

6. Conclusion

Sunfall, a collaborative visual analytics system in operation at a large-scale astrophysics project, has demonstrated that such systems that facilitate scientific analytic discourse and computer-supported collaborative work can have a positive impact on data-intensive science. In the process of design and implementation, we learned that an interdisciplinary team incorporating specialists from several fields, including scientific domain experts, will be most effective in designing an effective visual analytics system for science. Additionally, the framework architecture, efficient visual interfaces, and algorithms developed for Sunfall will be transferable to future large-scale automated transient alert pipelines such as SNAP or LSST.

7. Future Work

Current plans for improvements to Sunfall include the development of a multidimensional visual analytics tool for supernova spectra classification. We are currently applying several different clustering algorithms to the normalized spectral data. Tools are under development to apply various transformations to the calibrated and normalized spectra. The end goal is feature detection and similarity detection across supernova spectra. The optimal presentation of these clusters is still an open problem. We hope these visual analytic tools will facilitate the search for new correlations between classes of supernovae.

Acknowledgments

We thank the scientists of the SNfactory collaboration for their time and detailed feedback. This work was supported by the Director, Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET), and by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by a grant from the Gordon & Betty Moore Foundation. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] Cook Thomas J and Cook K 2005 *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (National Visualization and Analytics Center, IEEE)
[2] Aldering G, *et al.* 2002 Overview of the Nearby Supernova Factory *Proc. SPIE* **4836**:61.
[3] SNAP 2007 Supernova Acceleration Probe, http://snap.lbl.gov
[4] Izevic Z, et al. (LSST) 2008 arXiv:0805.2366v1, http://www.lsst.org/overview/overview_v1.0.pdf
[5] Perlmutter S, *et al.*1999 Measurements of Omega and Lambda from 42 High-Redshift Supernovae *Astrophys. J.* **519:**565–586
[6] Riess A, *et al.* 1998 Observational evidence from supernovae for an accelerating universe and a cosmological constant *Astron. J.* **116**:1009–1038
[7] NEAT 2007 Near Earth Asteroid Tracking, http://neat.jpl.nasa.gov
[8] QUEST 2008 Palomar-QUEST Survey 2002 http://hepwww.physics.yale.edu/quest/palomar.html
[9] Romano R, Aragon C, and Ding C 2006 Supernova Recognition Using Support Vector Machines *Proc. 5th Intl .Conf. of Machine Learning Applications.*
[10] Wood-Vasey W 2004 Rates and progenitors of Type Ia Supernovae (Ph.D. dissertation, University of California, Berkeley)
[11] Aragon C and Aragon D 2007 A Fast Contour Descriptor Algorithm for Supernova Image Classification *Proc. SPIE Symposium on Electronic Imaging: Real-Time Image Processing.*
[12] Zahn C and Roskies R 1972 Fourier descriptors for plane closed curves *IEEE Trans. Computers*, **21**:269–281

[13] Bailey S, Aragon C, Romano R, Thomas R, Weaver B and Wong D 2007 How to find more supernovae with less work: object classification techniques for difference imaging *Astrophys. J.* **665**:1246B
[14] Benford D and Lauer T 2006 Destiny: a candidate architecture for the Joint Dark Energy Mission *Space Telescopes and Instrumentation I: Optical, Infrared, and Millimeter, Proc. SPIE*
[15] ADEPT, Advanced Dark Energy Physics Telescope, 2006 http://www.physorg.com/news73758591.html
[16] Sako M, *et al*. 2008 The Sloan Digital Sky Survey-II supernova survey: search algorithm and follow-up observations *Astronomical J.* **135**:348-373
[17] Astier P, *et al*. (SNLS) 2006 *A&A* **447**:31-48
[18] SNLS, SuperNova Legacy Survey 2007 http://www.cfht.hawaii.edu/SNLS/
[19] SDSS, Sloan Digital Sky Survey, 2007 http://www.sdss.org
[20] Aragon C and Poon S 2007 The impact of usability on supernova discovery *Workshop on Increasing the Impact of Usability Work in Software Development, CHI 2007: ACM Conference on Human Factors in Computing Systems*
[21] Mak R, Walton J, Keely L, Heher D, and Chan L 2005 Reliable Service-Oriented Architecture for NASA's Mars Exploration Rover Mission *IEEE Conf. Aerospace*
[22] Aragon C, Poon S, Aldering G, Thomas R, and Quimby R 2008 Using Visual Analytics to Maintain Situational Awareness in Astrophysics *Proc. IEEE Visual Analytics Science Tech.*