Astronomical Data Reduction and Analysis for the Next Decade

Henry C. Ferguson (STScI)
Perry Greenfield (STScI)
Tim Axelrod (LSST)
Stefi Baum (RIT)
Alberto Conti (STScI)
Dennis Crabtree (Gemini)
Eric Feigelson (Penn State)
Mike Fitzpatrick (NOAO)
Wendy Freedman (Carnegie)
Kim Gillies (STScI)
Brian Glendenning (NRAO/ALMA)
Paul Goudfrooij (STScI)
Karl Gordon (STScI)
John Hibbard (NRAO/NAASC)
Paul Hirst (Gemini)
Jeffrey Kantor (LSST)
Anton Koekemoer (STScI)
Kathleen Labrie (Gemini)
J. Martin Laming (NRL; Chandra Users Committee)
Robert Lupton (Princeton)
Mario Mateo (U. Michigan; HST Users Committee)
Steve Myers (NRAO/EVLA)
Bryan Miller (Gemini)
Chris Miller (NOAO)
Chris Packham (UFL, US Gemini SAC)
Joe Pollizzi (STScI)
Rob Seaman (NOAO)
Chris Smith (NOAO)
Verne Smith (NOAO; US Gemini SAC)
Ingrid Stairs (UBC; NRAO Users Committee)
Massimo Stiavelli (STScI)
Elizabeth Stobie (NOAO)
Lisa Storrie-Lombardi (Caltech/SSC)
Michael Strauss (Princeton)
Doug Tody (NRAO; USVAO)
James Turner (Gemini)
Frank Valdes (NOAO)
Richard L. White (STScI)

Primary Contact: Henry Ferguson
Space Telescope Science Institute
3700 San Martin Drive
Baltimore, MD 21218
(410) 338-5098
ferguson@stsci.edu

Executive Summary

The astronomical community has become very sophisticated in setting requirements and figures of merit for the technical capabilities of new observatories. Sensitivity, field of view, spatial and energy resolution, observing efficiency and the lifetime of the facility are all set out to address scientific problems as efficiently as possible. The ultimate goal of these facilities, however, is not simply to gather data, but to create knowledge. It is thus important to consider the process of converting data to knowledge and ask whether there are ways to improve this for the coming generation.

Software for data reduction and analysis provides a key link in this chain. Modest investments in this area can have a very large impact on astronomy as a whole, if they are made wisely. Conversely, it is possible to waste significant amounts of money on software efforts that never fulfill their promise. We need to learn from the successes and failures of the past to try to maximize our productivity in astronomy as a whole. That means working more closely together across agencies, projects, institutions and disciplines to share in building and maintaining this essential infrastructure.

There is a strong need for a coherent cross-agency, cross-institution strategy for funding and developing the next-generation general-purpose data reduction and analysis software system(s) for astronomy.

Future Challenges for Data Reduction and Analysis

The main focus of this white paper is on software infrastructure that is used to support astronomical research, after receipt of the data. This includes the building blocks of calibration pipelines, as well as the suite of tools that astronomers use on a day-to-day basis to view, analyze and interpret data. Looking forward to the next decade, we see three major challenges: (1) data rates are growing rapidly, while individual CPU processing rates have leveled off; (2) industry trends in the computing hardware and software are likely to imply major changes to astronomical algorithms and software; and (3) computationally demanding analysis techniques are becoming more and more essential for astronomy.

Data Rates. Astronomical data rates have kept pace with advances in processing power, doubling roughly every two years. Today, data rates for large instruments range from a few (e.g., the Keck DEIMOS spectrograph) to about 100 GB per night (e.g., CFHT Legacy Survey). Projects such as PanSTARRS, LSST and JDEM will be pushing to much higher data rates with large sky surveys. The LSST data rate is expected to be ~20 TB per 24 hours. EVLA could reach 4 TB per 24 hours in some configurations; SKA data volumes are expected to be significantly higher. Studies of planetary transits will push toward high signal-to-noise ratios and rapid time series. We expect that cutting-edge astronomy will continue to require high-performance computing, ranging from supercomputers for the most demanding applications to state-of-the-art commodity computers for most projects.

Industry Trends. Until recently, advances in computing power have largely been brought about by advances in the effective speed of single CPUs. Most software developed for astronomical data analysis has worked well on typical desktop computers. Upgrading the hardware generally did not require radical changes in software. Astronomy has responded well to the appearance of computing clusters; many data reduction pipelines now run on such systems. Astronomical software has been slower to adapt to other industry trends. Standard desktop computers have had 64-bit CPUs for several years, offering the capability to address data arrays larger than 4 Gbytes in physical memory. Because this affects such fundamental issues as the length of an integer or a memory pointer, adapting to this change is much more complicated and has been only partially accomplished in the major software systems used for astronomy. CPUs with multiple cores have become the norm in desktop and even laptop computers. Dual cores are common today; the typical doubling time would predict ~32 cores in a typical astronomer's workstation by 2016. Astronomical software today takes advantage of multiple cores generally only when the astronomer runs two tasks in parallel (e.g., processing two separate images at the same time). The individual steps of the reduction have not been optimized to take advantage of the inherent parallelism of the CPU. This may need to change if astronomers are to realize the performance advantages of modern hardware. *According to some analysts, we are entering an era of the most disruptive advances in information technology since the computer was born.*[1]

An emerging trend is the use of Graphics Processing Units (GPUs) and field-programmable gate arrays (FPGAs) in high-performance computing. A high-end GPU in a gaming computer today has a speed of 4 TeraFLOPS (floating-point operations per second), more than an order of magnitude higher than a typical desktop CPU for comparable price. Tests of data processing steps for the Murchison Wide-Field Array, a low frequency radio interferometer under construction in Australia, have shown order of magnitude advantage for their GPU implementation relative to CPU in cost/performance, with an order of magnitude savings as well in power consumption.[2] While libraries are becoming available to facilitate coding for GPUs, it will require significant development effort to bring GPU processing into the mainstream for astronomical data reduction. A challenge of both multi-core CPUs and GPUs is that one needs to think in terms of "data parallelism" rather than "task parallelism," and that sometimes involves a radical change in algorithms.

It is important to keep up with software trends as well. In many respects, astronomy software was ahead of the industry when systems like IRAF and AIPS were developed. The user interfaces, device independence, networking capabilities, and inter-process communication (e.g. with image display software) were quite sophisticated for their time. However, industry infrastructure has evolved significantly over the past two decades and astronomy software has generally not kept pace. Compared to many software systems outside of astronomy, our systems can be difficult to install, the user interfaces often seem primitive, and the performance often seems slow. In moving forward, astronomy should take advantage of modern industry standards and should make a more concerted effort to draw on expertise from outside astronomy.

Analysis Techniques. Data processing and analysis techniques in astronomy have grown in sophistication over the past decade, and will no doubt continue to evolve, some driven by new computational algorithms, others by by advances in technology and instrumentation (e.g., wide-bandwidth correlators in the radio, high-dynamic-range coronographs in the optical). Examples include:

- PSF modeling and matching
- Iterative self calibration[3]
- Artificial source injection for error analysis
- Atmosphere, telescope and instrument modeling
- Resampling statistics[4]
- Clustering and classification analysis[5]
- Bayesian statistical analysis and Monte-Carlo Markov chain approaches[6]
- Wide bandwidth, wide field, full polarization interferometric image reconstruction
- Increasingly sophisticated spectral modeling relying on up-to-date laboratory data

New techniques often arrive on the scene from outside astronomy and percolate from a few early adopters to a wider community. Collaboration with statisticians and applied mathematicians is particularly useful for choosing the best algorithms, or developing new methods, for astronomical problems. Software begins with a few private efforts, but the techniques become useful for the field as a whole only when standard software packages are created and distributed with documentation and interfaces that reduce the effort needed to learn and apply the technique. Standard packages do not yet exist for many of the items in the list above.

Analysis techniques become increasingly complex as the dimensionality of our data increases. Three-dimensional data sets are becoming increasingly common and even higher dimensionality data can be expected (e.g., images as a function of wavelength and time). To date most visualization tools have focused on two-dimensional data. Because such data sets map well to image displays, visualization tools are comparatively straightforward, though not trivial. Coupled with the generally large size of 3-d and higher data sets, finding good ways to visualize and inspect such data sets interactively becomes a real challenge. This is a general software research issue, but may require tools specifically geared to the needs of astronomers.

It is also increasingly apparent that future data reduction and analysis will have to deal with both data and services distributed over the web. As much as possible, the software should be able to deal as transparently as possible with such distributed elements. An obvious component of this is being able to make use of Virtual Observatory protocols and data formats. Interaction with the VO will need to be integrated into the standard astronomical data reduction and analysis software packages.

The Current Landscape

Astronomical data reduction currently relies on a mixture of off-the-shelf operating systems and compilers, general mathematical and statistical libraries, and special-purpose

astronomical utilities. Many of the software systems in use today originated more than two decades ago, in an era where FORTRAN was the dominant language for scientific computing, VAX/VMS was the dominant operating system, and nine-track tapes were the dominant medium for transporting data. Data analysis systems in use today include AIPS MIRIAD, GIPSY and CASA, for radio astronomy; CIAO and HEASOFT for X-ray astronomy; and IRAF/STSDAS/Pyraf, MIDAS, XVISTA and MOPEX for optical and infrared astronomy. Together, these packages represent several million lines of code and hundreds of person-years of development effort. Astronomy departments and individual astronomers rely heavily on these systems to post-process data from the national observatories as well as process data from private facilities.

These complement general-purpose scientific computing software products such as IDL, Mathematica, and Matlab. These commercial systems have matured over the past two decades to provide a rich suite of mathematics, statistics, and signal-processing tools, as well as graphics and image display. While it is possible to develop astronomical libraries and data-reduction systems within such environments, several issues are generally seen as show-stoppers to using these as platforms for the national facilities. The issues include relatively high license fees, lack of access to source code, and the risks associated with locking into the proprietary system of a single company. Nevertheless, smaller teams have developed and distributed large code bases around such systems, and continued evolution and maintenance of such libraries is an important part of the general landscape of astronomical computing. Open-source scientific, mathematical, and statistical packages – such as R, scipy and GSL – have also evolved rapidly and are becoming part of this landscape.

A significant fraction of the total effort in astronomy goes into software development. The vast majority of this is code developed by individual scientists and research teams for their own research, often within one of the scientific computing environments mentioned above. As these software systems fall further behind the curve in support for new computing hardware or use of the latest scientific programming languages, the temptation will be for individual researchers and teams to develop their own software infrastructure. In moderation, this kind of progress can lead to innovation. However, as an overall strategy for the astronomical community it is an inefficient use of our collective resources.

The software systems in use for astronomy must evolve to meet the challenges of the next decade. This evolution will need to be more rapid and substantial than it has been over the past decade to meet the needs of new ambitious projects and respond to changes in the computing industry.

Structural Funding Challenges

Current funding mechanisms are failing to provide the needed support for this element.

Several factors contribute to the problem. The basic problem for individual astronomers and small groups is that whatever effort they devote to data analysis software is

necessarily fragmented. It must be provided by part-time efforts of individuals whose primary expertise is not in software, and whose efforts are necessarily directed toward solving their most immediate projects. This rarely results in software that is usable by others since extra effort must go into a good user interface, documentation, packaging and installation, user support, and ensuring that the software works on a variety of data sets. There is usually no time or incentive to provide that extra effort. Worse, the efforts at the infrastructure level developed by individual research teams are almost never incorporated into general community software libraries. Most reuse of such software is confined to the original research group. Frequently such software is modified on an ad-hoc level to meet the immediate needs and often reaches a fairly un-maintainable level after a few years.

Most of the major astronomical data-reduction packages were developed at the national facilities with the aim of supporting the data-reduction needs of their community. These systems have been installed and heavily used in universities and observatories worldwide, and often provide the framework within which small research groups develop their own software. The development efforts for many of these systems have tapered off over the past decade, with staff cuts putting systems such as IRAF and AIPS onto basic "life support." Progress on possible replacement systems such as AIPS++/CASA and Pyraf has been relatively slow, and priorities tend to be focused on project-specific software rather than general infrastructure. Like replacing plumbing or wiring in a house, some of the infrastructure work is not glorious or cutting-edge. Libraries need to be replaced with more efficient system calls. Algorithms need to be ported to modern computer languages or revised to make use of new CPU architectures. Documentation needs to be kept up to date.

For the national facilities, the difficulty stems from the project-oriented nature of the funding. In recent years, this funding has generally provided support for developing software specific to a new telescope or instrument, rather than support for developing and maintaining the general-purpose infrastructure. The goal of the project-specific software is usually to deliver calibrated data to the community, rather than provide tools for the community to analyze those data.

The last decadal survey set the establishment of the Virtual Observatory as a high priority. This drew attention to the importance of data archiving, data distribution and data mining. Cross-cutting efforts in the astronomical community and the funding agencies have resulted in significant progress toward the establishment of the VO. However, it is not the VO's purpose to provide tools needed for data reduction and analysis. While existing tools such as IRAF have served well, current development of reduction and analysis software is too slow and too fragmented to support the needs of the next decade.

Elements of a National Strategy

Our goal should be to ensure that the development of the next generation of astronomical data reduction and analysis software is carried out efficiently. The following are some of

the elements we view as important to a national strategy to improve our software infrastructure.

- *Roadmaps for major software systems*. These roadmaps should be developed for existing software systems, with leadership from the development teams and close consultation with the user community. We need to somehow avoid the Catch-22 of not being able to make realistic roadmaps without reasonable expectations for funding, and not being able to generate reliable funding without realistic roadmaps. Funding agencies can facilitate the long-term planning effort with very modest investments.
- *Evolution*. New systems will need to offer graceful entryways for astronomers who are heavily invested in existing systems. This involves providing mechanisms to translate code from old to new languages or coding standards; it involves developing and maintaining mechanisms for efficiently passing data between different analysis packages; and it involves substantial effort in documentation and training.
- *Effective coordination across wavelengths and between projects and institutions*. Development teams for the different software systems interact at ADASS and AAS meetings, but there are currently few examples of active coordination or collaboration. Serious coordination and collaboration requires effort and resources that typically cannot be found within specific project budgets. The development and adoption of common standards, protocols and libraries is essential for success. The AAS Working Group on Astronomical Software could take an expanded role in helping to coordinate national efforts. This will require focused attention and a long-term commitment.
- *More effective funding mechanisms for supporting the development and maintenance of the general-purpose astronomy data reduction and analysis infrastructure*. This work generally cannot be justified on the basis of a single project or single scientific result, and so is not suited to most funding opportunities in NSF and NASA. To a certain extent the goals fall under the NSF cyberinfrastructure category or the NSF SciDAC initiative. However, to date, none of the funded programs have involved astronomical data reduction and analysis (and no equivalent program exists within NASA). Funding mechanisms should provide long-term support for teams large enough to make significant annual progress, with mechanisms to promote collaboration among institutions. European astronomy today appears more organized in this area than the U.S., with funded efforts such as RADIONET[7] and OPTICON[8] both including a software component.

Summary

Astronomers are often ambivalent about software. On the one hand, it is essential for making scientific discoveries. On the other hand, software development competes for resources with pure research and instrument development. The path forward must recognize that much of the software development in astronomy takes place in small teams focused on specific research projects, and that this will continue to be the case. However, those development efforts rely heavily on a scientific computing infrastructure that includes both commercial software and more astronomy-specific tool suites provided by

the national astronomy facilities. We need to ensure that these tools evolve sensibly for the next decade, investing both in cutting-edge tools that push the envelope of high performance computing as well as the more mundane but equally essential infrastructure that ties the tools together and makes them accessible to a broad community.

References

[1] http://havemacwillblog.com/2008/06/26/8-disruptive-technology-changes/; http://elementaltechnologies.com/blog/?p=36; http://www.gartner.com/it/page.jsp?id=681107;
[2] Ord et al. 2009, http://arxiv.org/pdf/0902.0915v1; Dale et al. 2007 http://www.astrogpu.org
[3] e.g. Fixsen, Moseley & Arendt 2000, ApJS, 128, 651. Lonsdale et al. 2009, arXiv:0903.1828
[4] Babu & Feigelson 2006, ASPC, 351, 127; Clowe, Gonzalez & Markevitch 2004, ApJ, 604, 596. Kemball & Martinsek 2005, AJ, 129, 1760; Lee & Penn 2008, ApJL, 686, L1
[5] e.g. Richards, G. et al, 2009, AJ, 137, 3884
[6] Verde et al 2003, ApJS, 148, 195; Koen, C., 2009, MNRAS, 393, 1370; Corless, King & Clowe, 2009, MNRAS, 393, 1235; Carvalho, Rocha & Hobson 2009, MNRAS, 393, 581
[7] http://www.radionet-eu.org/
[8] http://www.astro-opticon.org/