The Astronomical Information Sciences: A Keystone for 21st-Century Astronomy

Abstract

Twenty-first century astronomy faces unprecedented data analysis challenges. Many current and forthcoming data sets are so huge or so complex that astronomy does not possess, and cannot easily develop on its own, the means to effectively distill understanding from the data. In recent years astronomers have begun partnering with information scientists—statisticians, computer scientists, applied mathematicians, and engineers—to address a variety of these challenges in several isolated areas. Two new disciplines—*astrostatistics* and *astroinformatics*—are emerging from these astronomy/information science ("Astro/Info") collaborations. Astro/Info research has made important contributions to modern astronomy, but its impact and rate of growth have been seriously limited by several obstacles, largely arising from the novel, interdisciplinary nature of the work. Primary obstacles include: (1) The absence of community-level support mechanisms for Astro/Info research and education. Such mechanisms would foster training of new Astro/Info researchers, and communication (of both expertise and software) within the Astro/Info community, between that community and the wider astronomical community, and between astronomers and information scientists. Absent these mechanisms, astronomers ineffectively build on the accumulated expertise of information scientists and Astro/Info colleagues, increasingly often wasting resources "reinventing the wheel." (2) The absence of adequate research support. Astro/Info research is a poor fit for the majority of astronomy research programs, which are oriented around the traditional observer and theorist specialties. Also, *inter*disciplinary proposals fare poorly in the disciplinary review process followed by almost all programs. The few programs tailored to interdisciplinary astronomy research have minuscule resources in comparison to the need.

As a result of these obstacles, the ability of US astronomers to effectively use modern data is falling behind both the needs of our discipline, and the capabilities being developed in other data-intensive disciplines, at an accelerating pace. We recommend specific initiatives to address these issues. These initiatives would require a modest funding investment of a few million dollars per year, paying back disproportionately large dividends in data analysis capability.

Authorship

This Position Paper was prepared and endorsed by the following team of 84 US-based astronomers and information scientists (listed separately). The lead author is Thomas J. Loredo (Dept. of Astronomy, Cornell University, loredo@astro.cornell.edu). The team maintains a web site hosting information about the authors (including email addresses and links to web sites) and supporting information for this document, including BibTeX source for the references: http://inference.astro.cornell.edu/Astro2010/.

Astronomers

1. Alberto Accomazzi, Harvard-Smithsonian Center for Astrophysics
2. Joshua Bloom, University of California, Berkeley
3. Kirk Borne, George Mason University
4. Robert Brunner, University of Illinois at Urbana-Champaign
5. Douglas Burke, Harvard-Smithsonian Center for Astrophysics
6. Nathaniel Butler, University of California, Berkeley
7. David F. Chernoff, Cornell University
8. Brian Connolly, University of Pennsylvania
9. Andrew Connolly, University of Washington
10. Alanna Connors, Eureka Scientific
11. Curt Cutler, California Institute of Technology
12. Shantanu Desai, University of Illinois at Urbana-Champaign
13. George Djorgovski, California Institute of

Technology
14. Eric Feigelson, Penn State University
15. L. Samuel Finn, Penn State University
16. Eric Ford, University of Florida
17. Peter Freeman, Carnegie Mellon University
18. Matthew Graham, California Institute of Technology
19. Carlo Graziani, University of Chicago
20. Edward F. Guinan, Villanova University
21. Jon Hakkila, College of Charleston
22. William Jefferys, University of Vermont and University of Texas at Austin
23. Vinay Kashyap, Harvard-Smithsonian Center for Astrophysics
24. Brandon Kelly, Harvard-Smithsonian Center for Astrophysics
25. Kevin Knuth, University at Albany
26. Donald Q. Lamb, University of Chicago
27. Hyunsook Lee, Harvard-Smithsonian Center for Astrophysics
28. Thomas Loredo, Cornell University
29. Ashish Mahabal, California Institute of Technology
30. Mario Mateo, University of Michigan Ann Arbor
31. Bruce McCollum, California Institute of Technology
32. August Muench, Harvard College Observatory
33. Misha (Meyer) Pesenson, California Institute of Technology
34. Vahe Petrosian, Standford University
35. Frank Primini, Harvard-Smithsonian Center for Astrophysics
36. Pavlos Protopapas, Harvard University
37. Andy Ptak, Johns Hopkins University
38. Jean Quashnock, Carthage College & University of Chicago
39. Graca Rocha, Jet Propulsion Laboratory California Institute of Technology
40. Nicholas Ross, Penn State University
41. Lee Rottler, IPAC/California Institute of Technology
42. Jeffrey Scargle, NASA Ames Research Center
43. Aneta Siemiginowska, Harvard-Smithsonian Center for Astrophysics
44. Inseok Song, University of Georgia
45. Alex Szalay, Johns Hopkins University
46. J. Anthony Tyson, University of California, Davis
47. Tom Vestrand, Los Alamos National Laboratory
48. Ben Wandelt, University of Illinois at Urbana-Champaign
49. Ira M. Wasserman, Cornell University
50. Michael Way, NASA Ames Research Center
51. Martin Weinberg, University of Massachusetts Amherst
52. Andreas Zezas, Harvard-Smithsonian Center for Astrophysics

Information scientists

1. Ethan Anderes, University of California, Davis
2. Jogesh Babu, Penn State University
3. Jacek Becla, SLAC National Accelerator Laboratory
4. James Berger, Duke University and Statistical and Applied Mathematical Sciences Inst.
5. Peter J. Bickel, University of California, Berkeley
6. Merlise Clyde, Duke University
7. Ian Davidson, University of California, Davis
8. David van Dyk, University of California, Irvine
9. Bradley Efron, Stanford University
10. Chris Genovese, Carnegie Mellon University
11. Alexander Gray, Georgia Institute of Technology
12. Woncheol Jang, University of Georgia
13. Eric D. Kolaczyk, Boston University
14. Ji Meng Loh, Columbia University
15. Xiao-Li Meng, Harvard University
16. Andrew Moore, Carnegie Mellon University
17. Robin Morris, Universities Space Research Association
18. Taeyoung Park, University of Pittsburgh
19. John Rice, University of California, Berkeley
20. Joseph Richards, Carnegie Mellon University
21. David Ruppert, Cornell University
22. Naoki Saito, University of California, Davis
23. Chad Schafer, Carnegie Mellon University
24. Philip B. Stark, University of California, Berkeley
25. Michael Stein, University of Chicago
26. Jiayang Sun, Case Western Reserve University
27. Xiao Wang, University of Maryland, Baltimore County
28. Larry Wasserman, Carnegie Mellon University
29. Edward J. Wegman, George Mason University
30. Rebecca Willett, Duke University
31. Robert Wolpert, Duke University
32. Michael Woodroofe, University of Michigan

1 Introduction

Twenty-first century astronomy is increasingly an *applied information science*, with new discoveries relying more and more on use of advanced tools from statistics, computer science, machine learning, data mining, and other information sciences. Rapid developments in two complementary directions are driving this revolutionary change in our discipline. Most prominently, we are in the midst of a period of explosive growth in the sizes of survey data sets. Modern survey data sets do not merely strain conventional data handling and analysis techniques; they completely defeat them. Astronomers need new tools for data management, exploration, and analysis, that not only *use* cutting-edge information science technology, but that *push* the edge, requiring new information science research. The second direction of development is less dramatic and circumscribed, but no less important. It is the steady growth in complexity of modest and small data sets (some produced by complicated reduction of large data sets), and of the physical and phenomenological models developed to understand them. This complexity raises a host of problem-specific challenges for data analysis that increasingly demand fundamental interdisciplinary research in methodology and computational implementation, not just to maximize the science return from the data, but also to ensure basic accuracy of scientific conclusions in settings where new complexities can actually invalidate familiar methods.

Two disciplines have arisen to meet these needs. **Astrostatistics** weds the tools of statistics to the needs of astronomers. It focuses on probabilistic modeling of data and quantification of uncertainty. **Astroinformatics** addresses the wide variety of concerns arising in managing, exploring, and analyzing extremely large data sets—observational or from theory-based simulation—whose size thwarts straightforward application of standard methods (including optimal astrostatistical methods). It relies most heavily on developments in the emerging fields of informatics and knowledge discovery from databases (KDD). Together, astrostatistics and astroinformatics comprise the advanced research arm of an emerging new *data analyst* specialty in astronomy, complementing the traditional specialties of observer and theorist, and focusing on the interface between observation and theory. For brevity, we refer to them jointly as **Astro/Info**, denoting astronomical applications of information sciences.

The majority of the over 80 signers of this Position Paper (and a companion paper focusing on astroinformatics) are astronomers and astrophysicists working in these new disciplines. The authorship also includes over two dozen leading information scientists working in the disciplines of statistics, computer science, signal processing, and applied math, who have joined forces with astronomers to pioneer astroinformatics and astrostatistics. We argue here that *vigorous growth of these new disciplines is crucial to the health of twenty-first century astronomy, but that they are poorly served by existing support structures in astronomy and information sciences*. We offer a variety of recommendations for improving the situation.

In fact, the situation is dire. Just as the need for Astro/Info research is escalating, support is significantly declining, as we document below. This trend stands in stark contrast, not only to the growing need, but also to trends in other disciplines with similar needs (most notably the biological and geological sciences). These developments move in the opposite direction of what is required to achieve many of the science goals Astro2010 will be advocating for, and set up the US astronomical community for a dramatic mismatch between its capability to produce data, and its capability to understand that data and use it to guide theory. This constitutes a potential impending crisis for US astronomy that we urge the Astro2010 survey to address in strong, concrete terms.

2 The Key Role of Astro/Info in 21st Century Astronomy

Statistics emerged in the 18th-19th centuries in response to data analysis problems arising in astronomy and geodesy. But through most of the 20th century, biological, industrial, social scientific and medical problems became driving forces in the development of statistics. Through the 20th century, astronomers were for the most part consumers rather than developers of the

resulting statistical methods, mostly relying on basic methods developed in the first quarter of the century (e.g., χ^2 and maximum likelihood fitting, periodograms and spectral analysis). But by the last quarter of the century, a number of astronomers in different corners of our discipline began to see the need for improved methods. Some developed new methods on their own (e.g., Lynden-Bell's C^- method for adjusting luminosity functions for detection thresholds; the Lomb-Scargle periodogram for periodicity detection with unevenly sampled data). Others, sometimes in collaboration with statisticians, mined the statistics literature, adapting methods developed for other disciplines to meet astronomers' needs (e.g., survival analysis methods to account for upper limits in population analysis; linear regression with errors in X and Y; advanced spectral analysis; Bayesian analysis for Poisson processes and CMB data analysis).

By 1991 the community of astronomers and statisticians doing significant astrostatistical research was large enough to justify an *interdisciplinary* conference, the first Statistical Challenges in Modern Astronomy (SCMA) conference, organized by astronomer Feigelson and statistician Babu at Penn State. The community grew steadily through the 1990s, increasingly driven by the advent of large surveys. Tools from machine learning and data mining made their appearance, especially for automated discovery and classification in survey databases (e.g., naive Bayes, empirical Bayes, and neural network classifiers).

For the third SCMA conference in 2001, theoretical cosmologist Joe Silk was invited to observe the conference and provide a theorist's perspective on the activity at the conference's closing. He argued that the turn of the century marked the emergence of an important new specialty in astronomy, which he dubbed *data analyst*; but he predicted obstacles to the growth of this specialty:

> Astronomers divide into three types: observers, theorists and data analysts. . . . The data analysts are a relatively recent breed . . . who are having a difficult time . . . being neither fish nor fowl, not completely acceptable as either observer or theorist. [Silk03]

He urged data analysts to persevere in building the new specialty, noting that "now is an especially opportune time to explore more extensive collaborations," deeming the need for interdisciplinary research "urgent," and arguing that scalability would require computer scientists to become increasingly important in such collaborations. He closed by noting that "the challenges are immense, but so are the potential rewards."

Data analysts must share an observer's expertise in understanding the data and a theorist's expertise in physical modeling, but must also master tools in knowledge discovery and uncertainty quantification from the information sciences. As a result, despite significant recent growth of the Astro/Info community, its efforts still do not have a natural place within the organizational and funding structure of astronomy.

Despite these difficulties, impressive results have recently emerged from Astro/Info research. We highlight a few cases here, with additional information and numerous further examples presented on our web site.

Massive survey data sets: SDSS, Pan-STARRS, LSST. The most prominent driver of Astro/Info research is the rapid growth in size and dimension of data sets produced by large-scale surveys. Modern data sets are so enormous that there is no hope for exhaustive examination of a significant fraction of the data by humans. Discovery and analysis must increasingly rely on sophisticated automated methods. Moreover, many data sets are so large that even implementing straightforward techniques, such as nearest neighbor searching or kernel density estimation, is utterly hopeless without use of sophisticated algorithms that push the edge of computational statistics research.

The best current examples of these challenges have arisen in the management and analysis of SDSS data. SDSS Data Release 7 (DR7, released in 2008) contains over 15 TB of images and 3 TB of raw spectra, with catalog summaries containing photometry for 357 million objects, and reduced spectra for over 1.6 million objects. To manage this huge data set, SDSS astronomers

partnered with information scientists from Microsoft and academic computer scientists and particle physicists to build the Catalog Archive Server (CAS) and its various interfaces, including the SkyServer web interface. While many routine astronomical queries can be handled with CAS deployed on off-the-shelf systems, large-scale queries require specialized tools. To support such queries, Szalay et al. have developed the GrayWulf system [BHS09]. This systems takes a "storage brick" approach, optimizing each component of a cluster for high-speed data access. GrayWulf won the 2008 Storage Challenge by executing an SDSS query in 12 minutes that would take 13 days on a traditional database system.

For large-scale analysis of SDSS data, the *International Computational Astrostatistics* (InCA) group, hosted at Carnegie Mellon University and University of Pittsburgh, assembled an interdisciplinary team to develop efficient algorithms capable of performing key statistical algorithms on large data sets. A foundation of their approach is adaptation of modern proximity data structures, such as kd-trees and metric ball trees, for statistical algorithms such as nearest-neighbor searches and kernel density estimation. InCA algorithms have seen numerous applications to SDSS data. A recent milestone is the cataloging of a million SDSS quasars with photometric redshifts [Rich+09a]. The automated quasar classifiers and photo-z estimators used for this work relied on InCA algorithms; the same team has just produced a new, eight-color Bayesian quasar classifier that combines SDSS and *Spitzer* data to improve accuracy [Rich+09b].

During the coming decade, Pan-STARRS and LSST will produce 10 and 30 TB of data per night. The LSST image database will be ~ 150 PB and its multi-epoch source catalog will occupy ~ 50 TB. The temporal dimension of the datasets will enable spectacular breakthroughs—provided that methodology and algorithms commensurate to the tasks are developed. Early Astro/Info efforts motivated by this upcoming data deluge include fast identification of planetary transits [Prot+05], a classification broker for rapid transient classification [Born+08], and intelligent dimension reduction [Rich+09, Pese+09]. Our companion paper discusses LSST astroinformatics challenges in more detail; see also the LSST petascale data challenges web site.

Complex data and models: Imaging. Images are the lifeblood of astronomy, and exhibit a diverse range of complexity. Raw imaging data—counts in CCD pixels, interferometric visibilities, directions of photons in a gamma-ray tracker—may bear a complex relationship to the underlying true image (intensity vs. direction). The imaged system may itself be complex, with a mix of diffuse structures on many scales, and unresolved point sources. Astronomers were early innovators in image analysis, particularly image deconvolution. But in recent decades the the image processing community outside of astronomy has made great progress in deconvolution, feature extraction, and image modeling; much of this work has great potential for astronomy. Astro/Info researchers have just begun to tap into this expertise. We cite a few recent highlights; further examples are on our web site.

An important theme of recent image processing research is *multiscale methods*: methods capable of adapting to spatially-varying smoothness in images. A team of statisticians and astronomers in the *California-Harvard Astrostatistics Collaboration* (CHASC) recently developed a probabilistic multiscale image model, using it as the basis for a deconvolution algorithm for *Chandra* images [Esch+04]. Importantly, this algorithm provides not just a best image, but also error estimates. A team of Caltech astronomers at the *Spitzer* Science Center have adapted multiscale methods from the computer vision and image processing literature to *Spitzer* image analysis. Their approach is based on a mathematical correspondence between convolution and a linear diffusion equation (and its nonlinear generalization) that enables smoothing without smearing diffuse features or localized objects [Pese+08]. Scargle's influential "Bayes Blocks" algorithm for optimal multiscale segmentation of time series is being generalized for images and other multi-dimensional processes [Scar+03]. Statisticians and engineers are currently generalizing wavelet-based methods to the low-counts Poisson regime, expanding the wavelet basis to include components such as curvelets and platelets, and further generalizing to hyperspectral "image cubes." A monograph and recent papers introduce these methods to astronomers [Star+06, Will07].

Complex data and models: Precision cosmology. Precision cosmology is one of the most impressive success stories of modern astronomy. Tremendous advances in observational technology were the key enabler for this revolution. But analysis of the new data required significant advances in data analysis methodology that have played an important role in precision cosmology.

Cosmological parameter estimation per se is challenging, not because of data set size, but because of *complexity* in the data and models. The data directly analyzed in these calculations are modest-dimensional data products derived from the large CMB and LSS raw datasets (e.g., CMB multipole and large scale structure (LSS) correlation function estimates). But calculating predicted values for these quantities from physical models, and calculating the probabilities for the noisy, convolved observables from the predictions (the likelihood function), are both computationally expensive. Strong nonlinearities and significant noise and correlations preclude making simplifying Gaussian approximations. Finally, information from diverse types of data (e.g., CMB, LSS, weak lensing, and SN Ia data) must be combined correctly and optimally.

A Bayesian approach, implemented using Markov Chain Monte Carlo (MCMC) computational techniques, has become the standard tool for cosmological parameter estimation and model comparison. Such methods first became widespread for analysis of *COBE* CMB anisotropy data; today they are the mainstay of this field, with recent *WMAP* papers providing good examples of the approach in action [Dunk05, Dunk09]. It is noteworthy that early MCMC implementations, using simple algorithms, were inefficient and of limited accuracy. Significant improvements came when astronomers invested substantial effort to learn advanced methods in the statistics literature. In the US, the work of Wandelt's group pioneered this effort [Wand04, Chu+05]; UK astronomers are similarly exploring new MCMC directions; see the CosmoMC project site.

A team of statisticians and physicists at Los Alamos National Lab is pioneering an important new direction for this research: acceleration of cosmological model exploration via use of a carefully trained fast, nonparametric Gaussian process (GP) "emulator" for the cosmological model [Heit+09]. This emulator approach is already common for uncertainty management with complex climate and hydrological models. In a related development, a team of applied mathematicians and astronomers have developed clever approximation techniques enabling scientists to implement flexible, adaptive GP models for large data sets; their showcase application was a new photo-z algorithm for the SDSS galaxy catalog [Fost+09].

Complex data and models: Chandra X-ray spectroscopy. Detailed modeling of X-ray spectra demands accounting for diverse and complex instrumental effects (response functions, pulse pile-up), and modeling photon counting data in the non-Gaussian low-counts regime. Accurate methods must handle strong model nonlinearity, and quantify uncertainty not only in parameter estimates, but also in the number of model components underlying the data (e.g., the number of spectral lines). The CHASC collaboration was initially formed to address just these challenges; it was initially supported by the *Chandra* X-ray Center (CXC). CHASC scientists have developed algorithms using cutting-edge MCMC methods to model *Chandra* data accounting for all the effects just mentioned; they can flexibly handle both continuum spectra and both broad and narrow lines [vD+01, Park+08]. Their current research is exploring how to *quantitatively* account for systematic errors in X-ray spectral modeling and imaging.

Our web site provides a "Showcase" with capsule summaries and references for numerous other Astro/Info research efforts—from the recent past and in progress—including work on: flexible modeling of astronomical populations from survey data with selection effects and measurement errors; MCMC methods for analysis of exoplanet radial velocity data; nonparametric methods for estimating dark matter distributions in dwarf spheroidal galaxies; new inversion methods for helioseismology data with improved uncertainty estimates; adaptive experimental design methods for optimal scheduling of observations; optimal tradeoff between statistical power and computational cost in searching large feature spaces; and other diverse problems.

Astro/Info training efforts. Supplementing and supporting such research efforts, astronomers are also making increasing efforts to train young astronomers with information science skills tar-

geting our discipline's needs. A number of astronomers have recently authored books and monographs on statistics for astronomers; our web site contains a list of these with full bibliographic information. CHASC astrostatisticians have organized numerous astrostatistics sessions, at AAS meetings and at topical conferences, where new methods have been discussed in a tutorial manner. These sessions are always well-attended. The *Center for Astrostatistics* (CASt) at Penn State has hosted an annual Summer School in Statistics for Astronomers and Physicists since 2005, providing students a solid foundation in well-established statistical methods. (CASt has also organized two schools in India, and one in Brazil.) The school is regularly oversubscribed with only minimal advertising; there is clearly great interest in statistics among young astronomers. It teaches roughly 70 students annually, about a third from the US. This amounts to reaching about 10% of the the world's astronomy graduate students.

Astro/Info impact on statistics. Astro/Info also has a growing role within the information sciences, particularly in statistics. The last decade has seen a number of high-profile overtures from statisticians to the astronomy community. Some of the nation's most distinguished statisticians (including NAS members Berger, Bickel, Breiman, Donoho, Efron, and Johnstone) have turned their attention to astrostatistical research problems and have published in the astronomical literature. In the last year alone we note the following events: two special issues of prominent statistics journals were devoted to astrostatistics, with articles authored or co-authored by astronomers; the annual Joint Statistical Meetings (JSM, the counterpart to AAS meetings) hosted not just one but two invited sessions on astrostatistics; and the departing president of the International Society for Bayesian Analysis (ISBA), in his farewell letter, singled out astronomy as an area in which members should build collaborations. Two years ago the NSF-supported *Statistical and Applied Mathematical Sciences Institute* (SAMSI) in North Carolina ran a semester-long astrostatistics program that hosted an interdisciplinary workshop, week-long astrostatistics and astronomy schools, and three astrostatistics working groups. These are just the most recent examples of overtures from the statistics community (we provide a more extensive list on our web site). Clearly, information scientists see a breakout role for Astro/Info in the 21st century.

Astro/Info career skills. In his recent study of the production and employment of PhD astronomers in the US [Metc08], Metcalfe reported that though 87% of first-year astronomy graduate students plan on academic careers, the limited supply of permanent academic positions results in fewer than 50% ultimately obtaining academic positions. He describes this as a "persistent gap between expectations and reality," arguing that "graduate programs in astronomy should prepare their students for this reality" and urging faculty to "nurture appropriate skills in the next generation of astronomers." Information science skills represent a sound investment for young scientists; such skills can serve them in academic positions, in the astronomy research and support positions that employ an increasing fraction of astronomer PhDs, or in technical jobs in the emerging information economy, should some of these scientists leave astronomy. By providing these skills, astronomy departments serve not only the students and our discipline, but also the national need for an information sciences workforce; indeed, astronomy may act as a new channel attracting people to such careers. Few astronomy departments are in the position to provide these skills to their students. But growth in Astro/Info research and training support can improve this situation, simultaneously serving the scientific needs of our discipline, and our ethical duty to our students to prepare them for sustainable careers.

3 Current Support for Astro/Info Research and Education

Long-standing Astro/Info collaborations, such as InCA and CHASC, illustrate the potential for interdisciplinary research during the coming decade. These collaborations started in the late 1990s and each have involved 20–30 researchers and students from several fields. Funding has come from various NSF grants (usually initiated from the statistics program in the Division of Mathematical Sciences (DMS) with AST co-funding), the CXC, and NASA grants (ROSES Ap-

plied Information Systems Research (AISR) and Astrophysics Data Analysis (ADP) programs). For a brief period during 2004–06, the NSF had a Mathematical Sciences Priority Area with Astronomical Sciences (MSPA05), that funded astronomical research with a strong research component in any DMS-funded area. The NSF also provided a large grant to develop methodologies for the Virtual Observatory. The recent demise of the Long-Term Space Astrophysics (LTSA) program and the explicit orientation of ADP away from methodological research has essentially closed off two support channels for Astro/Info research. NASA had a Research in Intelligent Systems (RIS) program but it primarily targeted planetary and Earth science missions and was terminated around 2004; the burden of funding RIS-nurtured work fell to AISR, further straining its limited resources. The *Chandra*, *WMAP*, and *Spitzer* science centers have supported intramural astrostatistical research, and the LSST project has methodological work embedded in its data management and science working groups. But no satellite mission or telescope project provides extramural or competitive funding for Astro/Info research. The LSST system architects recognize the crucial role of extramural Astro/Info research and are building a data analysis framework that supports integration of algorithms developed by science collaborations not affiliated with LSST, with the expectation that the majority of science with LSST data will be done by external collaborations. But the machine learning Astro/Info research that must be undertaken by these collaborations will not be funded via the LSST construction project.

Other structural impediments to Astro/Info funding are consistently faced. Cross-disciplinary proposals rarely receive high ratings from disciplinary review panels; a proposal with a strong computer/statistical science research component must necessarily devote less space to its astronomy component, and vice versa. Only NASA's long-standing AISR and NSF's 2004–06 MSPA-AST programs have had strongly cross-disciplinary review panels. Poor communication to and within the Astro/Info community led to MSPA-AST being under-utilized (the vast majority of our authorship did not know it existed during its operation), and the program ended due to apparently limited interest. AISR has had visionary leadership and effective organization, but serves too many constituencies (including Earth, space, solar, and planetary sciences as well as astrophysics) with resources too minuscule to be effective.

The NSF has sponsored a number of large, foundation-wide interdisciplinary initiatives including Knowledge and Distributed Intelligence (KDI), Information Technology Research (ITR), Science and Engineering Information Integration and Informatics (SEIII), and the large and growing Cyber-Enabled Discover and Innovation (CDI). For reasons that are not clear, most Astro/Info attempts for support from these programs have failed although a few are active.

Altogether, most Astro/Info groups have found funding to be erratic and inadequate to the tasks confronting the field; it is currently decreasing just as need is escalating. Agencies which fund initial ideas are often not prepared to continue funding so that new methods are fully developed with tools promulgated to the wider community. No agency provides a consistently supported mechanism for delivering and maintaining codes. Funding for Summer School training is not provided in a continuous manner. It is difficult to keep collaborators on board: talented computer scientists and statisticians can readily find more lucrative and stable collaborations in other fields such as biostatistics and bioinformatics. Program administrators are often enthusiastic advocates of Astro/Info research, but their efforts are thwarted by resource and structural constraints beyond their control.

4 Models From Other Disciplines

Astronomy is hardly the only discipline to experience astonishing growth in the size and complexity of data sets in recent years. Two other areas experiencing similar revolutionary changes are the biological and medical sciences, and the geosciences. They have complementary lessons to offer astronomers. The typical biological scientist historically has not had the level of mathematical and computational training of a physical scientist; accordingly, the biosciences have a long tradi-

tion of turning to statisticians for help with their data analysis problems. In contrast, geoscientists have comparable mathematical and computational training to astronomers, and have similarly followed a "do-it-yourself-first" pattern throughout much of their history. But in the last decade, the geoscience data deluge has forced geoscience to forge strong ties with information sciences, and to develop new structures to support the growth of geostatistics and the emergence of geoinformatics. Thus the biosciences provide an example of the value of long-term investment in information sciences, while the geosciences offer lessons in how to quickly bootstrap such investments within a large, technologically savvy scientific community.

The biological and medical sciences have a long tradition of community investment in information science research, extending back perhaps a century. As a consequence, *biostatistics* has been a stand-alone discpline for decades, with multiple journals and entire academic departments devoted to it. *Bioinformatics*—focusing on large data set issues mostly arising in genomics and molecular biology, but increasingly arising in health sciences—is much newer but is probably the leading growth area in modern statistics. Biology research is funded by both the NSF and the National Institutes of Health (NIH), at levels of about $600M and $28.5B (2007 numbers). The NSF funding of biostatistics research is integrated within its biology and mathematical sciences programs; the level of investment has proved difficult to track. As an emerging discipline, bioinformatics investments are more visible. NSF currently funds bioinformatics investments via diverse programs in *four* directorates: Biology (multiple programs); Computer & Information Science & Engineering (two programs); Engineering; and Mathematical and Physical Sciences (our web site lists the programs). This constitutes a large and diverse portfolio of support. Many programs support multiple scales of investigation, spanning 3, 4, and 5 years. In addition, the NSF-wide Cyber-enabled Discovery Initiative (CDI) expects to support multiple computational biology and bioinformatics projects as it progresses.

Several of the 27 institutes comprising NIH have independent Biostatistics divisions supporting both intramural and grant-based biostatistics and bioinformatics research; in addition, many general research grants include funding for biostatistics consultation. In some institutes, bioinformatics has become so important that there are separate Biostatistics and Bioinformatics divisions. A representative example for which we have a fairly complete picture is the National Institute of Allergy and Infectious Diseases (NIAID). NIAID currently funds 30 biostatistics grants with a total investment of $8M in FY08, comprising 2% of total NIAID grant funding; it separately devoted $1.3M to biostatistics training. NIAID supports three types of biostatistics research grants: 2-year non-renewable Exploratory/Developmental Research Grants (typically $100k), 2-year renewable Small Research Grants ($50k), and 3-5 year renewable grants (up to $500k or above, with approval). The NIAID training grants are long-term investments (14–20 yr as of 2009) in specific academic institutions for training future biostatisticians, at both pre-doctoral and post-doctoral levels. These grants directly fund 19 biostatistics graduate students and 3 postdocs per year. This significant investment in training via targeted grant funding is echoed by other NIH institutes. Other institutes also support ambitious summer schools, such as the National Heart, Lung and Blood Institute (NHLBI) Summer Institute for Training in Biostatistics. Courses offered during this 6-week program are developed with grant support; the 2009 solicitation anticipates awarding 7 grants totaling $1.7M for courses to be offered in 2010–2012.

In the geosciences, the primary mechanism for supporting geostatistical and geoinformatics (geo/info) research is the NSF Collaboration in Mathematical Geosciences (CMG), a *partnership* between the four NSF geoscience divisions (Earth, ocean, atmosphere, and polar science) and both DMS and Computer & Information Science & Engineering (CISE). This program began operation in 2002 and is considered very healthy. The program makes 15 to 28 awards per year totaling $12M–$14M/yr. As an example, the Earth Sciences Division (EAR) CMG awards amount to 4-5% of the EAR research budget. CMG funds interdisciplinary research, interdisciplinary post-graduate summer training, and interdisciplinary post-doctoral research appointments.

Also of interest are community support mechanisms for information science research in these

areas. As noted, biostatistics is a mature field; biometrics (including both biostatistics and bioinformatics) is the largest Section of the American Statistical Association, and there are several more focused national and international societies for biostatisticians.

In the geosciences, significant community support has only just emerged in the last several years. The American Geophysical Union (AGU), supporting Earth, space and planetary scientists, formed its Earth and Space Sciences Informatics (ESSI) Focus Group in 2005 (focus groups support scientists whose work cuts across multiple AGU science sections). ESSI is "concerned with issues of data management and analysis, large-scale computational experimentation and modeling, and hardware and software infrastructure needs, which ultimately provide the capability to change data systems into knowledge systems." Its focus is predominantly geoinformatics; it also covers geostatistics, but geostatistics support is also strong within individual AGU sections. It has become the fastest growing AGU focus group, and the number of geo/info abstracts at the annual AGU meeting doubled from 2007 to 2008. The Geological Society of America (GSA) is an older agency, supporting the professional growth of earth scientists. Among its 17 divisions is a Geoinformatics division devoted specifically to supporting the geo/info community. Two new geoscience journals are substantially devoted to geostatistics and geoinformatics research: *Earth Science Informatics* (started in 2008 with an issue on Virtual Observatories in Geosciences) and *Geoscientific Model Development* (started in 2008).

The contrast between the relative size and scope of support for Astro/Info versus that for geo/info and bio/info is stark. The observational and theoretical astronomy specialties can boast a level of excellence rivaling or even exceeding that of their counterparts in these disciplines. For Astro/Info the converse is true: it is instead a weak link in the astronomical chain of expertise whose successes to date have been achieved despite support obstacles not present in other disciplines with similar needs.

5 Recommendations for Improved Support of Astro/Info

The above discussion demonstrates the growing scientific need for astrostatistics and astroinformatics research and training during the early 21st century, combined with inadequate and decreasing resources provided for the tasks faced. It does not make sense that US astronomy develops the finest telescopes and detectors and the most sophisticated astrophysical understanding while neglecting the modernization of the intermediate stage of data analysis. To accomplish this modernization, there is a critical need to develop and maintain active collaborations between astronomers and information scientists, and mechanisms to propagate their advances in data analysis and computational thinking into the broad astronomical community. Accomplishing this will be challenging, and the recommendations for action outlined here should not be considered a final answer. But the modest yet noteworthy achievements of Astro/Info to date, and the successes in other disciplines, provide useful models for future development of the field.

We believe the components comprising a viable solution to the Astro/Info crisis facing astronomy will have these key features: (1) Research funding must involve explicit partnerships between discipline-specific funding sources, all the way down to the level of review panels, which must be interdisciplinary. (2) Research funding must be sustained, and take an integrated, multi-faceted approach to supporting the variety of Astro/Info research activities. (3) There must be substantive support for Astro/Info training of young scientists in both astronomy *and* the information sciences. (4) Community support mechanisms must be created to foster communication and resource sharing among Astro/Info scientists, between the Astro/Info community and its partner disciplines, and between the Astro/Info community and funding agencies.

With these features in mind, we offer the following recommendations. We note that the total new cost for implementing our specific research and training recommendations is a few million dollars annually, a small fraction of annual spending in astronomy. This small investment will have a disproportionately large impact on Astro/Info and on astronomy as a whole.

1. Community support via AAS. The AAS should be urged to create a mechanism to officially support the growing Astro/Info community. Possible mechanisms include a Committee, Working Group, or Interest Group on Data Sciences. An AAS Astro/Info group would help coordinate communication within the Astro/Info community, and between this community and other astronomers and funding agencies. This group should work with funding agencies to implement the recommendations below.

2. NSF Astro/Info grant support. The NSF should be urged to quickly and permanently reinstitute an independent, cross-disciplinary solicitation along the lines of MSPA-AST, specifically targeting information science research in astronomy, and partnering AST with both DMS and CISE (that is, adding computer science as an additional partner). Based on the need, historical MSPA-AST funding, and the models of other disciplines, a funding level on the order of $2M/yr is an appropriate start.

3. NSF large project partnerships. NSF Astronomy should vigorously pursue partnerships in support of large projects that have a significant Astro/Info component. For example, LSST and VO should receive support from CISE and DMS in return for supporting researchers in the information sciences.

4. NASA's astrophysics data analysis programs. With community input, NASA should be urged to reorganize its support of data analysis and information science research, possibly replacing the current ADP and AISR programs with a more integrated portfolio of support for both routine and advanced data analysis serving space-based astrophysics. Key new features we hope to see in a revised portfolio include: (1) Support for archival data analysis where astronomy-driven information science research is encouraged to play an equal role to the astronomical science; (2) Support for extramural data analysis research for both operating and pre-flight missions. The current ADP and AISR budgets are $3.7M/yr and $2.5M per 18 months (this represents 1/2 of AISR's originally planned 2008–2009 funding). We estimate an annual budget of $6.5M as appropriate for an integrated data analysis portfolio targeting space-based astrophysics.

5. Multi-tiered grant funding. Both NASA and NSF Astro/Info research programs should implement explicitly multi-tiered support, with different categories of research of various duration and levels of funding. The NIH biostatistics example described above provides a model. Long-term funding must be included, especially targeting young researchers; the now-defunct NASA LTSA program offers a model.

6. Information infrastructure and science. Support for Astro/Info research targeting infrastructure (e.g., data management and computational resource management research, including development of astronomy-oriented parallel, grid, and cloud computing software environments) should be separated from support from focused, science-driven Astro/Info research, either via separate programs, or via explicitly identified proposal categories within a single program.

7. An Astro/Info career path. The community and funding agencies should work together to establish Astro/Info as a recognized career path for astronomers, so that within the decade departments and centers are making permanent data analyst appointments as routinely as they now do for observers/instrumentalists and theorists. This work should include development of information science courses for astronomers at the undergraduate and graduate levels, and broadened and sustained support of summer schools and cross-disciplinary workshops on advanced methods, both to train Astro/Info researchers and to integrate Astro/Info into mainstream astronomy.

8. Data Sciences Fellowships. A 3-year interdisciplinary fellowship in astronomical data sciences would encourage young scientists to pursue Astro/Info careers, and bring recognition to these scientists and to the discipline. It should support both astronomy and information science PhDs; in the latter case, it would encourage graduates who could easily find more lucrative positions elsewhere to pursue Astro/Info careers. We estimate that a rolling roster of a few fellows could be supported for $1M/yr; the funding source should be interdisciplinary.

9. Program administration. Interdisciplinary programs are especially challenging to administer. Any reorganization of funding to better support Astro/Info should include a fresh look at the

programmatic resources required to enable administrators to effectively run complex programs.

The task of implementing support sufficient to the needs and promise of Astro/Info research is complex; it may warrant further study by a dedicated panel. But the need for improved Astro/Info support is urgent in key areas of our discipline, and concrete action should not be delayed.

Bell, G., Hey, T., Szalay, A. (2009) "Beyond the Data Deluge," Science, 433, 1297

Borne, K. (2008) "A machine learning classification broker for the LSST transient database," Astronomische Nachrichten 329, 255

Chu, M. et al. (2005) "Cosmological parameter constraints as derived from the Wilkinson Microwave Anisotropy Probe data via Gibbs sampling and the Blackwell-Rao estimator," PhysRev D71, 103002

Dunkley, J. et al. (2009) "Five-Year Wilkinson Microwave Anisotropy Probe Observations: Likelihoods and Parameters from the WMAP Data," ApJS 180, 306

Dunkley, J. et al. (2005) "Fast and reliable Markov chain Monte Carlo technique for cosmological parameter estimation," MNRAS 356, 925

Esch, D. N. et al. (2004) "An Image Restoration Technique with Error Estimates," ApJ 610, 1213

Foster, L. et al. (2009) "Stable and efficient Gaussian process calculations," J. Mach. Learn. Rsch., in press

Heitmann, K. et al. (2009) "The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum," arXiv:0902.0429

Metcalfe, T. (2008) "The Production Rate and Employment of Ph.D. Astronomers," PASP 120, 229

Park, T. et al. (2008) "Searching for Narrow Emission Lines in X-ray Spectra: Computation and Methods," ApJ 688, 807

Pesenson, M. et al. (2008) "Multiscale Astronomical Image Processing Based on Nonlinear Partial Differential Equations," ApJ 683, 566

Pesenson, M. et al. (2009) "High-Dimensional Data Reduction, Image Inpainting and their Astronomical Applications," ADASS 2008, O 5.4., in press

Protopapas, P. et al. (2005) "Fast identification of transits from light-curves," MNRAS 362, 460

Richards, G. et al. (2009a) "Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ˜1,000,000 Quasars from Data Release 6," ApJS 180, 67

Richards, G. et al. (2009b) "Eight-Dimensional Mid-Infrared/Optical Bayesian Quasar Selection," AJ 137, 3884

Richards, J. W. et al. (2009) "Exploiting Low-Dimensional Structure in Astronomical Spectra," ApJ 691, 32

Scargle, J. (2003) "Adaptive Piecewise-constant Modeling of Signals in Multidimensional Spaces," in *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, ed. L. Lyons, SLAC eConf C030908, p. 157

Silk, J. (2003) "An astronomer's perspective on SCMA III," in *Statistical challenges in astronomy, III*, E. D. Feigelson & G. J. Babu (eds.), New York: Springer, 387

Stark, J.-L. & Murtagh, F. (2006) *Astronomical image and data analysis*. Berlin: Springer

Wandelt, B. et al. (2004) "Global, exact cosmic microwave background data analysis using Gibbs sampling," PhysRev D70, 083511

Willett, R. (2007) "Multiscale Analysis of Photon-Limited Astronomical Images," in *Statistical Challenges in Modern Astronomy IV*, ASP Conference Series, Vol. 371, p. 247

van Dyk, D. et al. (2001) "Analysis of Energy Spectra with Low Photon Counts via Bayesian Posterior Simulation," ApJ 548, 224