

## MANAGING AND MINING THE LSST DATA SETS

Astronomy is undergoing an exciting revolution -- a revolution in the way we probe the universe and the way we answer fundamental questions. New technology enables this: novel detectors are opening new windows on the universe, creating unprecedented volumes of high quality data, and computing technology is keeping up with this explosion. In turn, this is driving a shift in the way science is produced in astronomy and astrophysics: huge surveys of the sky over wide wavelengths can be analyzed statistically for low-level correlations and inverse problems may be solved by statistical inversion, producing new understanding of the underlying physics.

This parallels progress in high energy physics. Decades ago, a handful of photographs of events sufficed for ground-breaking discoveries. This gave way to experiments in which the systematic measuring (scanning) of many bubble chamber pictures allowed the measurement of statistical properties, such as lifetimes. Current experiments extend the technique by recording all events electronically and subjecting Petabyte data sets to rigorous statistical analysis.

A key ingredient in mining our astronomical science from such huge databases is efficient algorithms for statistical analysis, but these have been under-emphasized in the rush to utilize new technology and get the data products out to the science community. Past data sets in astronomy (and indeed in most areas of science) have been small enough that one individual could visualize the data and discover unanticipated correlations. This is often how major discoveries have been made. Data sets are now growing so large that even the requirements for basic processing are becoming challenging, while analyses that examine the entire data set can be overwhelmingly expensive. In the near future, analysis of Petabyte databases will require the solution of this problem.

### ***New Horizons***

This change in the scale of astronomical datasets is exemplified by several recent projects. A giant departure from the tradition of one astronomer and one modest data set per project has been the *Sloan Digital Sky Survey*: a 15TB imaging data set covering multiple wavelengths and up to 10,000 square degrees of the sky (<http://www.sdss.org/>). Nearly 100 Co-Is will mine these data in prescribed ways. Current plans do not include mining the 15TB directly. Rather, 1TB of catalogs of detected objects and another 2TB of their "cutout" pictures will be produced and mined. Nevertheless, this will surely result in new understanding of our universe. Imagine what might be discovered if the full 15TB could be explored efficiently! Another refreshing and very successful departure from tradition is the 2MASS infrared survey of the sky

(<http://irsa.ipac.caltech.edu>). This group has poured major effort into usability of the data products and efficient remote searching.

On an even larger scale, we are proposing a new survey, to begin gathering data around 2010, that will produce several petabytes of data per year (a petabyte is  $10^{15}$  bytes, a million gigabytes). The optical instrument, an 8.4-meter aperture 3-mirror telescope with 7 square degree field of view and uniformly excellent image quality will provide an unprecedented figure of merit for deep surveys. Emphasis will be placed on the survey products, their unique science capability and distribution to the community, the data pipeline, the camera and data system, and the telescope. In the ranked projects in the recent NRC AASC Decadal Survey for Astronomy, this facility was named the Large-aperture Synoptic Survey Telescope (*LSST*) to emphasize its multiple missions. Advances in three areas of technology (large aspherical optics fabrication and metrology, microelectronics and terascale computation) have come together in the design of the *LSST* system. The *LSST* fills a nearly unexplored region of parameter space and enables programs that would take many lifetimes on current facilities, ranging from a nearly complete assay of near-Earth objects down to 300m in size, to unique probes of cosmic dark mass-energy.

The data reduction and analysis for the *LSST* will be done differently from most current observing programs. The data rate, combined with the need for real-time analysis and post-reduction data mining, requires a fresh approach making use of the best technology while developing innovative software for efficient data management. Much headway can be made in efficient algorithms and associated software, but there will also be CPU and disk hardware challenges. It will be particularly effective to have data analysis innovations in place when the telescope and camera system is in the verification phase, as early as 2009. Achieving this requires parallel efforts on optics, electronics, and software. The collaboration will include astronomers involved in current and upcoming ultra-large surveys, experts on statistics and algorithm development, computer science, and data mining and visualization.

### ***The Hardware Challenge***

The telescope's optical design is unique (it will have over fifty times the optical throughput of the current 4m wide-field cameras!), and it puts unprecedented demands on the computation and storage facilities to handle the data it will produce. The hardware associated with data processing and database mining must be studied now. The data rate from the 2.3 gigapixel camera will be of order 10 TB per night, compressed. Pipeline image processing of this data stream will be possible using parallel processors five years from now. More interesting challenges are presented by the archiving and, particularly, mining tasks. Storage technology is rapidly evolving, so that keeping all the data online (of order 20PB over 10 years) using commodity disks will almost certainly be

possible. However, accessing the entire dataset will be prohibitively expensive. Searching for correlations naively is an  $N^2$  procedure; when  $N$  is of order petabytes  $N^2$  is infeasibly large. It is therefore critical to discover ways to search for correlations in the resulting massive database efficiently. It is from studying correlations in the data that we expect to extract unanticipated science. Echoing the lessons of the Sloan Survey, while the required data hardware and software for the key (prescribed) science programs present challenges, assuring opportunity for unanticipated science using such huge databases presents an even greater challenge. Designing optimal data handling and search routines will be an exciting aspect of this project. Some science programs may need access to the full imaging data archive. Examples are: searches for low surface brightness objects not detected as individual objects, and searches for patterns of transient objects in time and space.

There are several hardware issues. Most of the computation will necessarily be done locally to the data disks, most likely in a hierarchical nesting of processors and a corresponding hierarchical nesting of slow and fast disks. The computation requirements, while large, appear to be within reach in the next five years (CPU Moore's law scaling is not expected to saturate until 2015, and there is the likelihood of disruptive innovation avoiding the quantum and charge confinement crisis even then). Capacity per disk is still increasing in a fashion comparable to, even exceeding, Moore's law for processors. Today, for low-bandwidth storage, a \$200 IDE disk holds about 100 GB; we can expect by 2009 that a comparably priced disk will hold several terabytes.

The bandwidth to disks is also growing, but not at nearly the same rate. This means that as the current magnetic technology advances, the total time required to read all the data from a disk actually goes up, not down. In the last five years, disk bandwidth has improved a factor of six, so we can estimate at least another factor of six for the next five years.

At a pace of one 2.3 Gpixel image per 20 seconds (15 sec exposure, 5 sec read/repoint), each night the *LSST* will generate 13 TB uncompressed processed data in 8 hours -- that's about 450 MB/s. This suggests that we can assemble the required pipeline processing fast disks by 2009. We predict from scaling that disks by then will handle over 300MB/s, but that is sustained transfer rate under ideal conditions. Several disks in parallel will provide more headroom to get the data written to disk in real time. It is also likely that new technologies such as holographic storage will break this scaling in the next five years, at least for slow storage.

Crafting the software pipeline and developing efficient database management tools and the algorithms for data mining will present more of a challenge than the pre-processing computational capacity. The demands of this post-processing will be hardware and software intensive. The effort invested in software, data system design, tools for visualizing and analyzing data, and the science data analysis,

may be comparable to that spent on the telescope optics.

### ***The Software Challenge***

The enormity of these data sets creates exciting statistical challenges beyond the computational. Many statistical techniques used by astronomers today have been optimized to deal with the small size of their existing data sets. At the same time the algorithms scale as a power of the number of objects, prohibitive for the new data sets with billions of objects, where even  $\log N$  is 30! Algorithms which scale much worse than linearly will be unacceptable computationally. At the same time, the main source of errors will be various systematic effects, and we also have to deal with the fact that we can only observe a single realization of the universe as a random process. There is also “cosmic variance,” the expected statistical variations in space of the power spectrum.

One can overcome the cosmic variance by surveying larger and larger volumes of the universe, but then the data sizes become even larger. We should think about approximate statistical techniques, where the approximation is within this variance, but the algorithm has a non-polynomial scaling. A recent example of such an attempt is Szapudi et al. 2000. For example, the new sky surveys are sufficiently large to allow accurate estimation of third and fourth order moment structure of galaxy locations. This immediately raises the issue of what functionals of higher order moment structures one should consider and how to relate these to theoretical models for the evolution of the universe. Deep surveys, which necessarily look far into the past, allow one to consider models for evolving large-scale structure, raising questions of how to estimate this evolution.

Furthermore, no matter what the size of the data set, there will always be features and scales for which the estimation error will be important, so the need for developing statistically efficient estimation schemes and methods for assessing estimation error will always remain. Combining statistical efficiency with computational efficiency will be a constant challenge, since the more statistically accurate estimation methods will often be the most computationally intensive.

With the ability to do fast correlation analysis on Petabytes of data, we could revolutionize how we detect faint moving objects or probe the underlying dark mass-energy of our universe. Weak gravitational lensing, the deflection of light by intervening clumps of dark matter, causes distortions in the observed shapes of galaxies. These data may then be inverted to yield a mass map of the intervening universe. Closer to home, potentially devastating near-Earth objects go undetected. New techniques of extracting relevant image parameters can be used on the Petascale imaging data to automatically find such objects. Similar image-mining techniques can be very relevant in other areas of science as well

(satellite observations, biology, oceanography, etc). Generally, inverse problems (and the regularization of their solution) are computationally more intensive than simple n-point correlations. With Petabyte databases, new algorithms for statistical regularization must be developed.

Finally, data visualization will present a formidable challenge. Efficient methods for statistical visualization and sampling of large databases are required. User-reconfigurable trees of image feature catalogs driving multi-dimensional displays could help, but the opportunities here are largely unexplored.

### ***A New Collaboration***

We see this research program attracting a broad range of mathematical, computer and physical scientists. In addition to the obvious connections to astronomy, statistics and large-scale computation, this program would also include probability, data visualization and data management. We would also seek to include representatives from the high-energy physics community, who have faced somewhat different problems involving massive data sets and immense data streams for many years now. Some representation from theoretical cosmologists who simulate universes would add to the mix and allow the question of comparing simulated universes to the actual universe to be more profitably addressed.

It will be particularly useful to study the characteristics of spatial processes, since it nicely combines the central computational and statistical challenges. Very little work has been done to date in this area, although a recent paper by Moore et al. (2001) recognizes the importance of this problem and describes an algorithm for computing estimates of higher order correlation functions that, for sufficiently large data sets, is much more efficient than the obvious approach.

We need not simply a theoretical study of how massive astronomical data sets should be analyzed, but major efforts to analyze the most recently available data sets. Data from the Sloan Digital Sky Survey should be publicly available by 2003. It will be useful to work with this database in new ways, searching for low-level correlations. Deeper imaging surveys, such as the Deep Lens Survey, are producing imaging data and catalogs nearly to the depth that LSST will reach, but over a very small area of sky by comparison to a decade of LSST operations. Such surveys are precursors to LSST and their data products will prove to valuable sand boxes for development of new algorithms.

A common technique in modern high-energy physics experiments is the “mock data challenge.” The data stream, from detector, through data acquisition and processing, to final science analysis, is simulated at the appropriate level of detail. This allows a final acceptance testing of all data systems to be completed

along with the hardware, so that full-up science operations can begin on a much better schedule, with good diagnostics in place. For the science, these studies are just as important. Analysis teams combing for subtle effects can, in the end, compare their result (and error estimate) with the “true” values of parameters that were in the simulation. Often, a sample of “real” data is used to get the background distribution of events correct. Using catalogs from the *SDSS* and the Deep Lens Survey as a basis for the mock data challenge for the *LSST* will make it more effective.

There is now a clear recognition in several communities of the benefits of collaboration. *LSST* and its data challenge provides a focus of effort involving astronomers, statisticians, computer scientists, physicists and others -- creating new multidisciplinary collaborations that will have major impacts far into the future.

*February 5, 2001, Revised December 30, 2002*

Rob Pike, Google  
Michael Stein, University of Chicago  
Alex Szalay, Johns Hopkins University  
Tony Tyson, Bell Labs, Lucent Technologies  
and *The LSST Collaboration*

### ***References***

Moore, A. et al. (2001). *Fast Algorithms and Efficient Statistics: N-point Correlation Functions*. astro-ph/0012333.

Szapudi, I. et al. (2000) *Fast CMB Analyses via Correlation Functions*. astro-ph/0010256