Cosmology data analysis challenges and opportunities in the LSST sky survey

J Anthony Tyson

Department of Physics
University of California, Davis, USA

Abstract. Fueled by advances in software, microelectronics, and large optics fabrication, a new type of sky survey will soon begin. In a relentless campaign of 15 second exposures with a 3 gigapixel camera, the Large Synoptic Survey Telescope will cover the sky deeply every week for ten years. LSST will chart billions of remote galaxies, providing multiple probes of the mysterious Dark Matter and Dark Energy. Multiple probes of the effects of dark energy over an unprecedented volume of the universe will allow us to measure how dark energy behaves over time to high precision. Hundreds of petabytes of high dimensional complex data will be mined and compared with Exascale simulations. After reviewing the LSST project, I will describe some of the computational challenges and opportunities.

1. Introduction

It is remarkable that fully 96% of our universe remains a mystery, evidence for new physics. Moreover there are two “dark” components. Dark matter is responsible for the gravitational assembly of galaxies, and while we can produce images of dark matter utilizing the gravitational mirage it produces on the background galaxies, we have no idea of its physical nature. Similarly we as yet have no understanding of the physics of the late-time acceleration of the expansion of the universe, “dark energy.” Dynamically, these two dark components oppose each other: dark matter causes galaxies to cluster, but dark energy expands space-time at recent times.

To explore the nature of this, we need to measure the way the expansion of our universe and the dark matter mass structure changes with cosmic time. Since the expansion of the universe has been accelerating, the development of mass structures via ordinary gravitational in-fall will be impeded. Measuring how dark matter structures and ratios of distances grow with cosmic time – by observing the cosmic mirage of background galaxies produced by foreground mass – will provide clues to the nature of dark energy. A key strength of the next generation surveys is the ability to survey huge volumes of the universe. Such a probe will be a natural part of the all-sky imaging survey: billions of distant galaxies will have their shapes and colors measured. Sufficient color data will be obtained for an estimate of the distance to each galaxy by using redshifts derived from the multi-band photometry. By combining with dynamical data from spectroscopy, we can test whether the “dark energy” is due to a breakdown of General Relativity on large scales. These diverse goals require the construction of a new astronomical facility capable of surveying billions of galaxies – measuring their positions, shapes, and distances – over the full visible sky and back in cosmic time.
1.1 Exploration in astronomy

The various dimensions of exploration in astronomy (wavelength, angular resolution, depth, sky coverage, and time resolution) normally are impossible to achieve simultaneously with the same facility. The rate at which we can survey the sky to a given faintness is proportional to the product of telescope light collecting area and the size of the focal plane imaging array. This is also the data rate. The technologies of microelectronics, computation and software, and large optics fabrication drive the new sky. Aided by rapid progress in these areas, CCD camera sky surveys are changing the way we view and study the universe. Historically there has been complementary progress in detectors (imaging arrays) and computation.

Figure 1 charts this trend in optical sky surveys over 40 years. The effect of technology is clear. While the total light collecting area of telescopes (blue curve) has remained comparatively constant, the information content of sky surveys (using the number of galaxies measured per year to a given signal-to-noise ratio as a proxy) has risen exponentially. This is mainly due to high efficiency imaging arrays growing to fill the available telescope focal plane, and to the increase in processing power to handle the resulting data volume. Development of software to analyze these digital data in new ways has resulted in a corresponding increase in science output. Photographic surveys had the advantage of large focal plane area early on, but have been eclipsed by more sensitive CCD surveys, driven by the exponential rise in pixel count and computer processing capability – both enabled by the microelectronics “Moore’s Law”. Plotted versus time is the cumulative sum of all CCD pixels in sky survey cameras, as well as the number of transistors in a typical CPU. Processing capability keeps up with the data rate. Also plotted is one result of CCD surveys – the number of galaxies measured per unit time – ranging from a survey using a single 164 Kpixel CCD on a 4m telescope to a 3.2 Gpixel camera on an 8.4m telescope: the Large Synoptic Survey Telescope (LSST) sky survey.

![Figure 1](image)

**Figure 1.** Data trends in optical surveys of the sky. Data rate is proportional to the product of telescope light collecting area and the size of the focal plane CCD array. Information content (galaxies surveyed per unit time) in digital sky surveys roughly follows Moore’s law, and processing capability has kept up with pixel count. The LSST wide-fast-deep sky survey 2022-2032 will open new windows on the universe.
2. LSST

LSST combines a large light collecting capability with a giant wide-field digital camera: an 8.4 meter diameter mirror and a 3.2 gigapixel camera covering ten square degrees per exposure (an area on the sky equivalent to 40 full moons). The camera focal plane is tiled with 189 4Kx4K CCDs of special design which enables good spectral response from 300nm to 1100nm. Each CCD is segmented into 16 sections, each with its own output amplifier. Due to this massively parallel architecture, the entire focal plane is read out synchronously within 2 seconds. The LSST has the capability of surveying the deep universe in a novel way. This system will have unprecedented optical throughput and can image about 10,000 square degrees of sky in three clear nights using 30 second “visits” per each sky patch 2-3 times per night. Beginning in 2022 the LSST will begin tiling the entire visible southern sky with 1000 exposures per sky patch. This will continue every night for ten years, creating a catalog from its 30 trillion measurements of 37 billion objects. Figure 2 shows LSST at its location on Cerro Pachon in northern Chile. By October 2018 the facility was nearly complete. A live web cam at Cerro Pachon on the gallery page at http://lsst.org shows the current status.

The sky survey capability of LSST is unprecedented. Because each 30 second exposure goes so faint, rapid tiling of the sky enables a new kind of exploration of the time domain in astronomy. Unexplored regions of luminosity-timescale for explosive events will be explored. The huge volume of high-quality data enables a wide range of science opportunities, opening a new era of precision cosmology with unprecedented statistical power [1].

![Figure 2. The LSST Observatory: artist's rendering of the dome enclosure with the attached summit support building on Cerro Pachon in northern Chile. [LSST Project/NSF/AURA]](image-url)
2.1 The New Sky
Since LSST will cover the entire visible sky deeply and rapidly, unexplored regimes become accessible. The unprecedented data from the LSST sky survey will open new windows of understanding on cosmology, the assembly of our Milky Way galaxy, and near-Earth asteroids. New types of objects will likely be found, whose luminosity changes rapidly. In cosmology the combination of a survey of a deep volume of the universe (a wide range of cosmic time) with a wide sky area will bring novel probes of the physics of dark matter and dark energy. Surveying a wide range of cosmic time (wide redshift range) is important both for optimizing the gravitational lens mirage and for detection of the evolution of cosmic dark matter structure.

2.1.1. Multiple probes of cosmology. The mean density of dark matter in the universe is only one piece of the puzzle. More information is needed to decipher the physics of dark matter. The late-time cosmic acceleration [dark energy] is a profound challenge to our understanding of the universe. So far investigations have been focused on two classes of models: a new mass-energy having effectively negative pressure and modified gravity. There are also models that have no new mass-energy components or physics but attribute the acceleration to a breakdown of the cosmological principle or an oversimplification of General Relativity effects in the real universe. These models are less popular, but they will be tested with the multiple probes together with measurements of the isotropy of dark energy.

The data from the LSST sky survey may be used to probe the physics of dark energy and dark matter in multiple complementary ways. Measurements of galaxy clustering, dark matter distribution vs cosmic time, the evolution of geometry, and any anisotropy all give clues to the nature of these mysteries. For a recent review of LSST cosmological probes, see [2].

3. LSST Data
During operations the LSST facility will produce 20 TB of on-sky imaging data per night for ten years, in addition to daily calibration images. These data will result in a relational database including 20 billion galaxies and a similar number of stars, and will serve the majority of the primary science programs. While this data volume is not challenging either for storage or processing, the high dimensionality of the data is challenging for analysis. For example, each of the approximately 20 billion galaxies imaged and cataloged by LSST have hundreds of effective dimensions. Data collected by the LSST telescope and camera will be processed to data products -- catalogs, alerts, and reduced images -- by the LSST Data Management system. The science community will work directly with these data products, without having to work directly with the raw pixels.

The data processing from images to catalogs requires 1.8 PFLOPS continuous. This includes pixel processing, calibration, transient object detection, co-addition of all images of the same part of the sky, and sky catalog generation. Already by 2018 a million lines of code have been written. All the LSST code has been designed and implemented following software engineering best practices, including modularity, clear definition of interfaces, continuous integration, unit testing, and a single set of documentation and coding standards [3]. The primary implementation language is Python and, where necessary for performance reasons, C++. Most of the LSST software is licensed under the terms of the GNU General Public License (GPL), Version 3, and can be found at https://github.com/lsst . The documentation for the LSST Science Pipelines is available at https://pipelines.lsst.io .

The data management system will span four key facilities on three continents: the Summit Facility on Cerro Pachon in Chile; the Base Facility in La Serena, Chile (which will serve as a retransmission for data uploads to North America, as well as the Data Access Center for the Chilean community); the Data Processing and Archiving Facility at the National Center for Supercomputing Applications (NCSA) in Champaign-Urbana, IL and the Satellite Processing Facility at CC-IN2P3 in Lyon, France. The data will be transported between the centers over redundant existing and new high-speed optical fiber links from South America to the U.S.
3.1 Data Products

After the pixel processing, the data processing branches into a prompt alert processing pipeline and the “static sky” deep survey pipeline. In each case catalogs are created. The large data volume, the fast detection and alert aspect, and the complexity of processing involved makes it impractical to rely on the end users for the data reduction. Instead, the data collected by the LSST system will be automatically reduced to scientifically useful catalogs and images. Over the ten years of LSST operations and 11 data releases, this processing will result in cumulative processed data of about 500 PB for imaging, and over 50 PB for the catalog databases. The final data release catalog database alone is expected to be approximately 20 PB in size.

The data from the LSST camera will be automatically processed to data products: catalogs, alerts, and reduced images. These products enable the majority of LSST science cases, without the need to work directly with the raw pixels. Details of the LSST data products may be found in the LSST Data Products Definition Document [https://ls.st/dpdd], which is periodically updated. These data will be served via a relational database. The scale of the LSST data release catalogs -- together with the rapid needed response -- present some engineering challenges. The LSST project has been developing Qserv, a shared-nothing massively parallel processing database system, to meet these needs. Catalog data within Qserv is spatially partitioned, and hosted on servers running on dedicated hardware within the LSST Data Facility. The servers locally leverage conventional relational database management system technologies, running behind custom front-end codes which handle query analysis, rewrite, distribution, and result aggregation. The Qserv servers also provide cross-user synchronization of full-table scans in order to provide predictable query response times when serving many users concurrently. More details about Qserv can be found in the LSST document LDM-135 [4].

Two major categories of data products will be produced and delivered by the LSST: Prompt and Data Release. Prompt products are time domain data products which support the discovery, characterization, and rapid follow-up of time-dependent phenomena. Data Release products are the deep universe catalogs obtained from co-added multi-band images and are most relevant to cosmology: they are designed to enable systematics and flux-limited science, and will be made available in annual Data Releases. The LSST data products will be made available to the U.S. and Chilean scientific communities and to international partners with no proprietary period. The ultimate goal is to make LSST data products including the relational database available to the public and scientists around the world.

3.2 Data Enabled Discovery

Jim Gray commented that with the new generation of experiments we are entering a 4th paradigm: data driven discovery. As an example of this new paradigm, LSST is a kind of astronomy genome project. Rather than undertake separate new sky surveys to test new idea as we have done in the past, a single comprehensive sky survey will gather sufficient data to serve the entire astrophysics community. In effect the sky will be transformed onto spinning disks with high-dimensional catalogs of features for 37 billion astronomical objects. This Exascale data enables many separate “experiments”. It also enables automated searches for the unexpected.

The LSST science community has organized into several science collaborations to carry out specific experiments centered around supernovae, active galaxies, stars, etc. These organized teams are working together to optimize science analyses and assess the importance of systematic uncertainties on the derived results. For example, the LSST Dark Energy Science Collaboration (DESC) has members with interests in the study of dark energy and related topics in fundamental physics with LSST data. The LSST science community will perform its analyses using LSST computational facilities at NCSA, and the system has been sized accordingly.
4. Curse of Dimensionality: Analysis Computing

The LSST sky survey will produce large uniform data sets with high quality for multiple cosmological probes. These probes are affected by sample selection and various systematic errors and are sensitive to cosmology in different ways. Joint analyses of multiple probes will improve parameter constraints, and more importantly, also enable cross-calibrations of known systematics via breaking single-probe degeneracies, and finally detection and characterization of unknown systematics. It is important to have a consistent treatment of shared physics (there is only one universe with one dark matter distribution) and systematics across all probes.

This enables robust tests on cosmological models and exploration of the fundamental physics of the universe beyond current understanding. Automated data quality assessment is key; there is too much data to be looked at by humans. The challenges of dimensionality become apparent when we consider the computational effort for the LSST 3-probe investigation of dark energy. The data vector will have 6500 elements, the covariance matrix will have 43 million elements, and the cosmology likelihood analysis is expected to take 1M CPU-hours per run. There is also the challenge of optimization and sampling in high-dimensional parameter spaces. As described below however, there is a far larger effort required in forward simulating the observations themselves.

4.1 Cosmology simulations

We must undertake simulations of the universe for many different cosmological models in order to solve the inverse problem of determining cosmological parameters, investigating systematics in the survey observations, as well as astrophysical systematics which can mimic physics effects of the dark sector. These cosmological simulations are also needed for calibration of errors in the process of cosmological parameter likelihood estimation via precision covariance estimates, and for testing, optimizing, and validating observational strategies with synthetic catalogs. Data richness (high dimensionality) is an issue, requiring suites of very expensive cosmological simulations. These simulations must be sufficiently large in volume to cover the observed universe and of sufficient mass and force resolution to resolve the smallest relevant objects. The simulations must also span the model parameter space, and reach the precision enabled by the LSST data, and be thoroughly validated. Typically each of these simulations of cosmic structure formation in large volumes take weeks on supercomputers [5].

These n-body simulated catalogs of dark matter structure formation must also supplemented with realistic distributions of astrophysical objects, including galaxies which form gravitationally at the peaks of the dark matter. Their number, size, brightness, colors, and shape distributions must match what is known from deeper and complementary surveys. These galaxies are assigned self-consistently to corresponding peaks in the n-body simulated dark matter 3D field. It is critical to validate the consistency of these catalogs of the simulated astrophysical sky [6].

4.2 Simulations of observations

Ultimately observations must be compared with simulated observations based on the astrophysical sky. The cosmological simulated catalogs discussed above are used as input to simulate mock observed images and mock observational catalogs. It is important to include the selection effects due to observation parameters as well as the effects of image analysis and catalog generation software. Because the photometric pipeline processing from images to catalogs is nonlinear and complex, we must first simulate mock observed images, and then process those in exactly the same way that the LSST images are processed. There are two reasons to simulate the observations: (1) this permits unit testing of analysis algorithms, including cosmological parameter estimation (see below), and (2) production of realistic covariance matrices. These simulations of LSST images take months on supercomputers. It is useful to benchmark the computation required ultimately for the simulation of the survey. Using an LSST image simulator which follows each photon from the cosmos through the detector, and then extrapolating to the full 18,000 sq.deg implies we would need 1.3 billion CPU-hours! Better use of GPUs will help, but
there are also possible simulation algorithm efficiency improvements now being developed. Finally, there is the question of whether we need to simulate the full survey images, rather than a representative subset. Obviously the forward simulation of the imaging data is expensive. The computational burden of simulating these representative images is far larger than any other task, including all the data management required for processing the images to catalogs.

4.3 Cosmological parameter estimation
We wish to compare the theoretical prediction to the observed measurements. In a Bayesian analysis, a likelihood function is developed to estimate the probability of the observational data given a proposed theory (and its parameters). Modular methods have been developed which make it efficient to incorporate the observational selection and systematics in a self-consistent way for any combination of correlation functions for the observational probes [7]. The goal is to determine the posterior distribution of cosmological parameters given the simulation output and the observational data. In parallel with the observations, simulations of many candidate cosmologies are carried out, and converted to catalogs with the addition of galaxies (and their properties) and known astrophysical systematics. This parallel chain is shown in figure 3. Large simulations of the full observations yield a covariance matrix. After incorporating the known observational sample selection and systematics, the model is compared with observations via a likelihood analysis incorporating errors from the full covariance matrix. This must be repeated for each candidate cosmology model.

Figure 3. The Bayesian process of cosmological parameter estimation begins with two parallel campaigns: astronomical observations and simulated universes. After incorporating the known observational sample selection and systematics, the model is compared with observations via a likelihood analysis incorporating errors from a full covariance matrix. [S. Habib]
The level of sophistication of this whole process has developed over the years, tracking computation technology. This increase in capability has led to several generations of frontiers in complexity-volume space as applied to cosmology. **Terascale**: first generation surveys and single cosmological probe simulations. **Petascale**: second generation surveys, two-probe simulations, intermediate accuracy cosmological parameter estimation. **Exascale**: next generation surveys (like LSST), end-to-end multi-probe full survey scale simulations, and multiple cross-calibrated probes. Although the cosmic inference MCMC is not parallelizable leading to weeks of supercomputer time, by far the largest computation is the simulation of realistic observed images.

### 4.4 Opportunity for discovery of the unexpected

While the normal approach of hypothesis testing, i.e. comparing the data with various cosmological models will certainly be done, the richness of the data enables more adventurous explorations; we can in principle let the data speak for itself. Correlations over the high-dimensional database may discover unexpected cosmological trends. Non-parametric machine learning approaches – trained on the set simulations for of candidate cosmologies – may reveal outliers in feature space. There has been some initial progress, with ML applied to galaxy distance estimation.

The LSST holds great promise for breakthrough discoveries in cosmology, and it comes with challenges. While the LSST project completes the systems engineering and construction work, scientists have much to prepare for. Since the galaxy sample size is so large, the statistical noise is sub-dominant: the LSST data will be limited by systematic errors arising from instruments, observations, data reduction, analysis methods, and even deficiencies in theories. The good news is that the LSST survey will bring unprecedented statistical precision enabling knowledge beyond our current horizon, requiring extraordinary effort and novel use of the data. Efforts on detecting and controlling systematics, especially those not yet known, are crucial for the LSST to reach for its full potential. An imperative is the development of tools for new modes of data exploration.

**References**

[1] Abell P A *et al.* (The LSST Collaboration) 2009 LSST Science Book  arXiv:0912.0201
[2] Zhan H and Tyson J A 2018 *Rep. Prog. Phys.* 81 066901 (arXiv:1707.06948)
[3] Jenness T (LSST Data Management Team) 2017 *ASPC* 512 297 (arXiv:1511.06790)
[4] Becla J *et al.* 2017 https://ls.st/LDM-135
[5] Habib S *et al.* 2016 *New Astronomy* 42 49-65
[6] Mao Y *et al.* 2018 *ApJS* 234 36
[7] Zuntz J, Paterno M, Jennings E, Rudd E, Manzotti A, Dodelson S, Bridle S, Sehrish S and Kowalkowski J 2015 *Astronomy & Computing* 12 45 (arXiv:1409.3409)