Large Synoptic Survey Telescope

www.lsst.org

LSST Astroinformatics and Astrostatistics: Data-oriented Astronomical Research

Kirk D. Borne[1], K. Stassun[2], R. J. Brunner[3], S. G. Djorgovski[4], M. Graham[4], J. Hakkila[5], A. Mahabal[4], M. Paegert[2], M. Pesenson[4], A. Prša[6], A. Ptak[7], J. Scargle[7], and the LSST Informatics and Statistics Team

[1]George Mason University, [2]Vanderbilt University, [3]University of Illinois, [4]Caltech, [5]College of Charleston , [6]Villanova University , [7]NASA

The LSST Informatics and Statistics Science Collaboration (ISSC) focuses on research and scientific discovery challenges posed by the very large and complex data collection that LSST will generate. Application areas include astroinformatics, machine learning, data mining, astrostatistics, visualization, scientific data semantics, time series analysis, and advanced signal processing. Research problems to be addressed with these methodologies include transient event characterization and classification, rare class discovery, correlation mining, outlier/anomaly/surprise detection, improved estimators (e.g., for photometric redshift or early onset supernova classification), exploration of highly dimensional (multivariate) data catalogs, and more. The LSST ISSC team members present five sample science results from these data-oriented approaches to large-data astronomical research: the EB (Eclipsing Binary) Factory, transients characterization (for rapid follow-up and classification), classification of complex data sets, semantic e-Science, and dimensionality reduction through data compression (for visual analytics).

Background information:

- Astrostatistics conferences since 1991 (Statistical Challenges in Modern Astronomy SCMA, Babu & Feigelson).
- Astroinformatics developments since 2001 (= the establishment of Virtual Observatory projects worldwide).
- 2009: Astroinformatics position paper submitted to ASTRO2010 Decadal Survey Committee.
- Several additional ASTRO2010 Decadal Survey position papers and white papers were submitted related to astronomical sky surveys, astrostatistics, and data science.
- 2009: LSST Informatics and Statistics Research Collaboration Team proposed, approved, and formed.
- 2010: Astroinformatics and Astrostatistics special sessions at AAS-Washington DC.
- Websites: **http://astrostatistics.psu.edu/** and **http://www.practicalastroinformatics.org/**
- 2010: Astroinformatics 2010 International Conference: **http://www.astroinformatics2010.org/**
- **What is Astroinformatics?** – It is the set of data-oriented research methodologies for data-intensive astronomy:
  - Large-scale scientific data modeling, organization, and management
  - Scientific metadata, tagging, and annotations for astronomical information search and retrieval
  - Data mining, machine learning, and knowledge discovery from data
  - Astrostatistics
  - Information and data visualization, including data structures and data compression
  - Semantic e-Science
  - Data-intensive astro-computing and analysis
  - Reference: Borne, K. *"Astroinformatics: Data-Oriented Astronomy Research and Education,"* **Journal of Earth Science Informatics**, 3, 5-17 (2010).

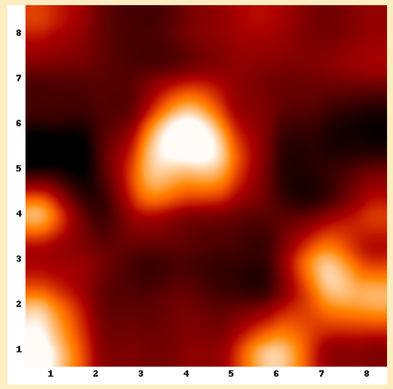
Current Topics of Interest for the LSST Informatics and Statistical Science Collaboration team:

- Advancing the field = Community-building:
  - Our team consists of Astronomers, Statisticians, and Computer Science data mining (machine learning) experts – we are developing a strong collaboration among these 3 research communities for the benefit of LSST and similar big data projects
  - Promoting Astroinformatics + Astrostatistics to astronomers (with several workshops sponsored since 2009)
  - Education, education, education! (including Citizen Science, undergraduate and graduate education, …)
- LSST Event Characterization vs. Classification
- Characterization of sparse, irregularly sampled time series and the LSST observing cadence
- Challenge Problems, such as the Photo-z challenge and the Supernova Photometric Classification challenge
- Testing algorithms on the LSST simulations: images/catalogs PLUS observing cadence – can we recover known classes of variability?
- Generating and/or accumulating training samples of numerous classes (especially variables and transients)
- Proposing a mini-survey during the science verification year (Science Commissioning):
  - *e.g.,* obtain high-density and evenly-spaced observations of extragalactic and Galactic test fields, to generate training sets for variability classification and assessment thereof
- Science Data Quality Assessment (SDQA): R&D efforts to support LSST Data Management QA activities

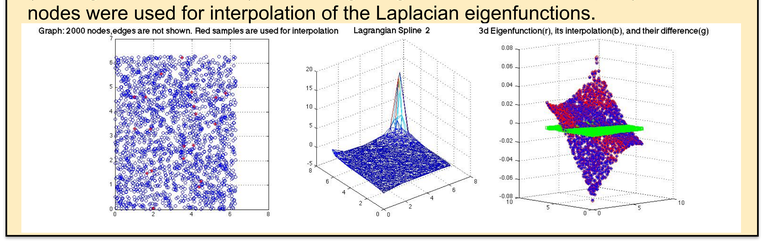
WHAT ARE THE PRIMARY SCIENTIFIC GOALS OF THE LSST ISSC TEAM?

- ***To address the LSST "Data to Knowledge" challenge***
- ***To help discover the unknown unknowns in LSST's petascale databases and archive***

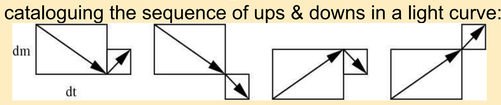
Semantic e-Science: Semantics adds intelligence to the handling of complex data by capturing domain knowledge and expressing it in a machine-processable fashion. At its simplest level, this allows disparate data sets to be commonly understood by both human and computer through mappings between terms, concepts and relationships used to describe and model the data. At a more sophisticated level, knowledge is incorporated directly into data mining algorithms, making them smarter and more powerful. The ***ontological self-organizing map*** (OSOM; **Graham *et al.* 2011**, in preparation) is an example of such a semantic data mining technique: it is a modified version of the traditional (dimension-reducing) SOM algorithm with the distance metric dependent upon the conceptual similarity of terms used to annotate data objects, *i.e.,* the degree to which they share information. The plot shows the trained toroidal network for a 21-dimensional data set comprising 100 annotated objects randomly drawn from the SIMBAD astronomical database. The axes are the transformed coordinates (reduced from 21 dimensions to 2) that mimic the "separation" between concepts. The clusters denote groups of objects which are conceptually similar: in this case, they are 'galaxies' (4,6), 'stars' (7,3), 'infrared sources' (6, 1), 'radio sources' (4,1) and 'variable stars' (1,1). In this case, similarity is a measure dependent on both the hierarchical structure of annotation terms (as expressed in an ontology) and the probabilistic frequency of terms in the SIMBAD corpus.



Data Compression through Dimensionality Reduction for Visual Analytics: Exploration of high-dimensional (multivariate) data sets. The figure shows an example of interpolation of eigenfunctions by Lagrangian splines on graphs with nontrivial topology. This illustrates a random graph with 2000 nodes (the edges are not shown); the average degree of the nodes is 35; only 2% of all nodes were used for interpolation of the Laplacian eigenfunctions.



Real-time Characterization of Transients for Follow-up Observation Support: The variety of astronomical phenomena and attendant transient events ensure that classification is a complex process. ***The path to classification is through characterization*** = cataloguing the sequence of ups & downs in a light curve:



…aided by statistical methods and astronomy knowledge, which in turn are based on heuristics from a select set of data embedded in huge multi-dimensional datasets. Light curves with a few points lead to early characterization:
(1) Simple heuristics and Gaussian processes (Fig. A),
(2) delta-mag delta-time surfaces (Fig. B), and
(3) Statistics of delta-mag over characteristic periods (one-day boxplots, Fig. C).

Fig. A

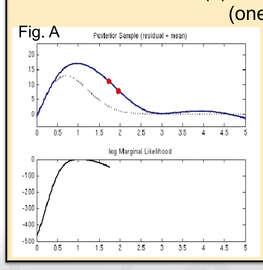


Fig. B

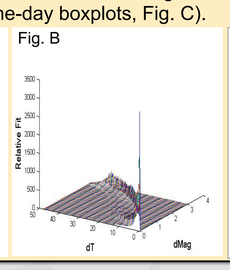
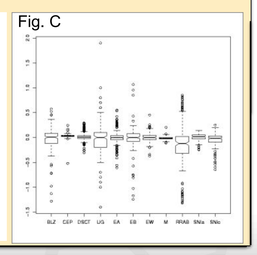
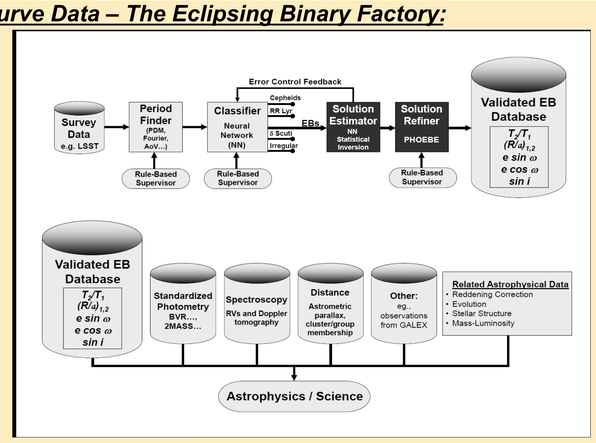


Fig. C



Data Mining and Knowledge Discovery from Light Curve Data – The Eclipsing Binary Factory:

The EB Factory is envisioned as a pipeline for processing large quantities of light curve data through a series of artificial intelligence-based filters, classifiers, and solution finders in order to produce a set of "validated EBs" whose physical properties are well constrained and are thus suitable for detailed follow-up. The Solution Refiner is fully developed (PHOEBE; **Prša & Zwitter 2007**), and the Solution Estimator has been piloted (**Prša *et al.* 2008, 2010**) with OGLE and Kepler light curves. The Period Finder and Classifier components are under development, thus completing the basic assembly line (top row of figure). The EB Factory will be used immediately to harvest EBs from LSST light-curves. We also seek to further enhance the EB Factory by adding capability to incorporate ancillary catalog data on the fly (bottom row of figure), such as radial velocities from the RAVE and SDSS MARVELS surveys, standardized photometry, and others.



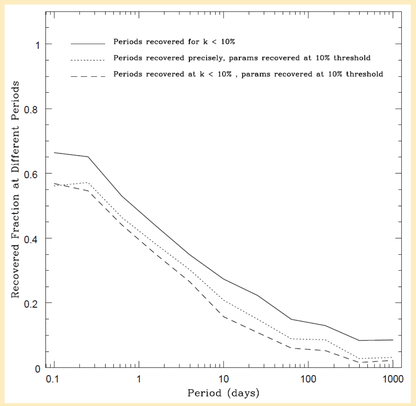
Pepper, J., Stassun, K., & Prsa, A. 2010, Bulletin of the American Astronomical Society, Vol. 42, pp. 218

Prša, A., Zwitter, T. 2005, IAU Symposium, Vol. 240, pp. 217-229

Prša, A., et al. 2008, ApJ, Vol. 687, pp. 542-565

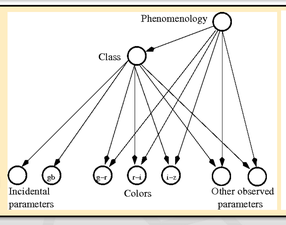
Prša, A., et al. 2010, ArXiv e-prints, arXiv:1006.2815

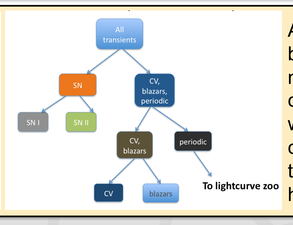
The figure to the right illustrates the expected yield of eclipsing binary stars with LSST. 10,000 eclipsing binary light curves are synthesized and sampled according to the LSST universal cadence and passed through the period finder. The phased data are then passed to a neural network-based estimator of principal eclipsing binary parameters. The solid line depicts the fraction of sources where the estimated period is within 10% of the actual value. The fractions of sources for which the principal physical parameters are recovered to a 10% accuracy, with exact periods (dotted line) and the recovered periods (dashed line), represent the ideal and realistic expected yields, respectively. Overall, we estimate that LSST will observe ~16 million eclipsing binaries down to r~22, for which S/N should be sufficient for analysis. Our yield calculations here suggest that ~1.6 million of these eclipsing binaries are likely to be successfully recovered and their physical parameters well estimated (Pepper et al. 2010; see also LSST Science Book, section 6.10).



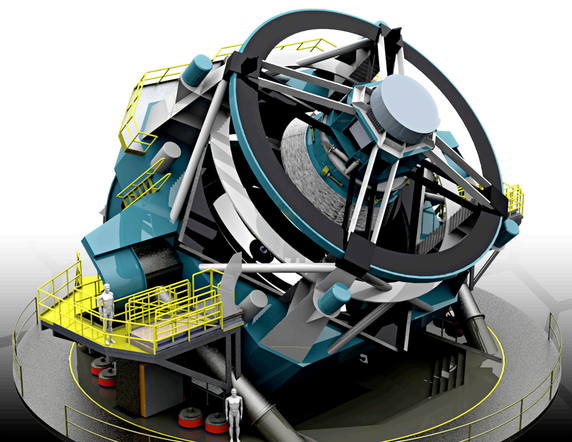
Classification of Complex Data Sets:

A series of Bayesian Techniques including Gaussian Process methodologies lead to classification from complex datasets. Priors based on basic observables like colors and other auxiliary information lead to broad early classification. Research is focused on making these algorithms scalable to large datasets.





A breadth-first tree based on small number of characteristics is a way towards a complex classification tree, as illustrated here.



January 2011

LSST is a public-private partnership. Design and development activity is supported in part by the National Science Foundation. Additional funding comes from private gifts, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.