Large Synoptic Survey Telescope

www.lsst.org

Astroinformatics in the Age of LSST: Analyzing the Summer 2012 Data Release

Kirk Borne[1], N. De Lee[2], K. Stassun[2], Martin Paegert[2], P. Cargile[2], D. Burger[2], J. Bloom[3], J. Richards[3]

[1]George Mason University, [2]Vanderbilt University, [3] University of California Berkeley

The Large Synoptic Survey Telescope (LSST) will image the visible southern sky every three nights. This multi-band, multi-epoch survey will produce a torrent of data, which traditional methods of object-by-object data analysis will not be able to accommodate. Thus the need for new astroinformatics tools to visualize, simulate, mine, and analyze this quantity of data. The Berkeley Center for Time-Domain Informatics (CTDI) is building the informatics infrastructure for generic light curve classification, including the innovation of new algorithms for feature generation and machine learning. The CTDI portal (http://dotastro.org) contains one of the largest collections of public light curves, with visualization and exploration tools. The group has also published the first calibrated probabilistic classification catalog of 50k variable stars along with a data exploration portal called http://bigmacc.info. Twice a year, the LSST collaboration releases simulated LSST data, in order to aid software development. This poster also showcases a suite of new tools from the Vanderbilt Initiative in Data-instensive Astrophysics (VIDA), designed to take advantage of these large data sets. VIDA's Filtergraph interactive web tool allows one to instantly create an interactive data portal for fast, real-time visualization of large data sets. Filtergraph enables quick selection of interesting objects by easily filtering on many different columns, 2-D and 3-D representations, and on-the-fly arithmetic calculations on the data. It also makes sharing the data and the tool with collaborators very easy. The EB/RRL Factory is a neural-network based variable star classifier, which is designed to quickly identify variable stars in a variety of classes from LSST light curve data (currently tuned to Eclipsing Binaries and RR Lyrae stars), and to provide likelihood-based orbital elements or stellar parameters as appropriate. Finally the LCsimulator software allows one to create simulated light curves of multiple types of variable stars based on an LSST cadence. More information about the research activities of the team can be found at the Astrostatistics and Astroinformatics Portal (ASAIP): http://asaip.psu.edu .

LSST Informatics and Statistics Science Collaboration (ISSC) research team:

Addressing the LSST petascale data-to-knowledge challenges.

Enabling discovery through novel Data Science methods for KDD (Knowledge Discovery from Data):

- *Astrostatistics : inference and learning from complex high-dimensional data*
- *Astroinformatics : Big Data analytics for astronomy (data mining / machine learning ; visualization / visual analytics ; data structures and indexing for rapid search, discovery, and retrieval)*
- *Semantic e-Science : searchable tags/metadata (ontologies, taxonomies, Citizen Science tagging)*

Developing algorithms and techniques for discovering the unknown unknowns in LSST's massive data set:

- *100-200 Petabyte final image archive (after 10-year survey)*
- *20+ Terabytes of new imaging data every night*
- *20+ Petabyte final database, including massive Object catalog (50 billion) & Source catalog (20 trillion)*
- *1-2 million alerts every night, signaling new astronomical transient events (astrometric & photometric)*

Topics of Research Interest to ISSC team:

- Provide rapid descriptive characterizations and probabilistic classifications for millions of events each night
- Find new multivariate correlations and associations in high-dimension astronomical attribute parameter space (#p~100's)
- Discover voids in these high-dimensional parameter spaces (e.g., period gaps), perhaps signaling new astrophysical processes
- Discover new and improved rules to classify known classes of objects
- Discover new and exotic classes and subclasses of objects and astrophysical processes, along with new properties of known classes
- Serendipity – discover rare one-in-a-billion(trillion) types of sources through outlier detection ("Surprise Discovery") algorithms
- Identify novel, unexpected behavior in time series data
- Hypothesis testing – verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions (or billions) of training samples
- Quality Assurance – identify data errors through deviation detection

The Vanderbilt Initiative in Data Intensive Astrophysics (VIDA)

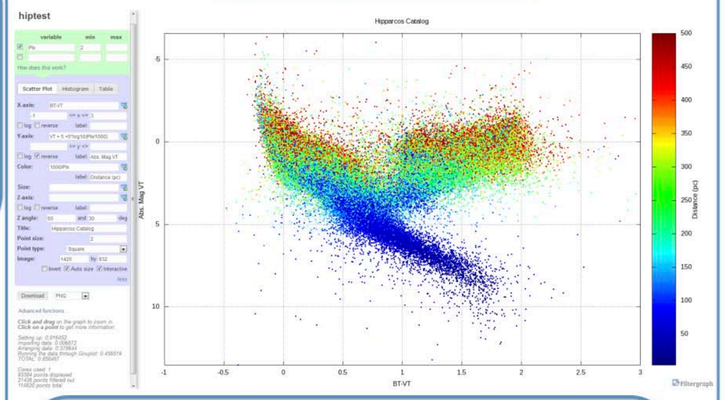
http://www.vanderbilt.edu/AnS/physics/vida/)

The VIDA Website contains a series a easy to use web-based visualization and analysis tools.

Filtergraph – Easily plot, filter, and share your data
EB Factory – Neural Network Lightcurve Classifier
Lightcurve (LC) Suite of tools:

LC Animator – Create Public Relations Animations
LC Chopper – Analyzing semi-periodic signals
LC Simulator – Simulate a wide variety of lightcurves

Filtergraph (http://filtergraph.vanderbilt.edu/)

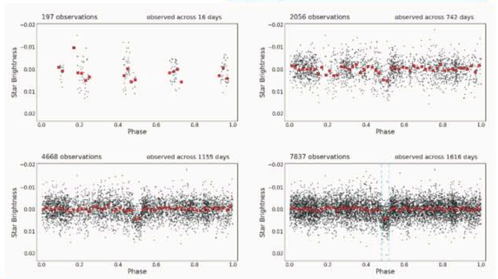


Upload an ASCII table, and within minute:
•Interactive histograms and up to 5-d scatter plots
•Easily share your data with collaborators
•Optimized for speed (can plot millions of points in a few seconds)
•Make many different cuts based on columns
•Allows you to use formulas on different columns
•Has many ways to download (PNG, PDF, JPG, PS, and GIF)

Try it yourself! Go to http://filtergraph.vanderbilt.edu/hiptest
Or upload your own data at http://filtergraph.vanderbilt.edu/

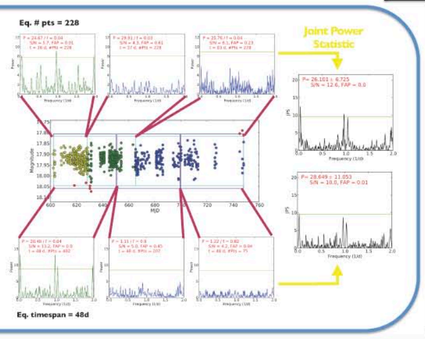
LC Animator

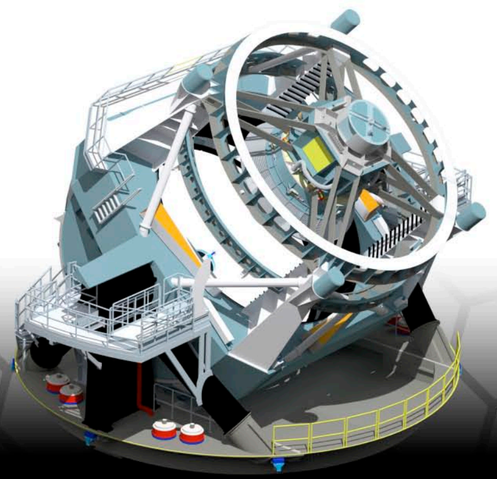
LC Animator: Screenshots from the KELT-2Ab animation . The slides are advancing in time (and number of data points) from left to right and top to bottom. White points are observations , red points are binned. YouTube video: http://youtu.be/3MG0KA7hRqI



LC Chopper

LC Chopper is an analysis tool designed to pick out periodic signals from semi-periodic data. The plots to the right show the periodogram analysis of the PTF lightcurve of a K dwarf star. On the left of the plot, Top: light curve analysis using "static" chopping (based on equal number of points), Bottom: using "dynamic" chopping (based on equal time span). The horizontal line is the power for a false-alarm-probability = 0.05%. These periodograms were then combined to create the joint power statistic periodograms (right column).





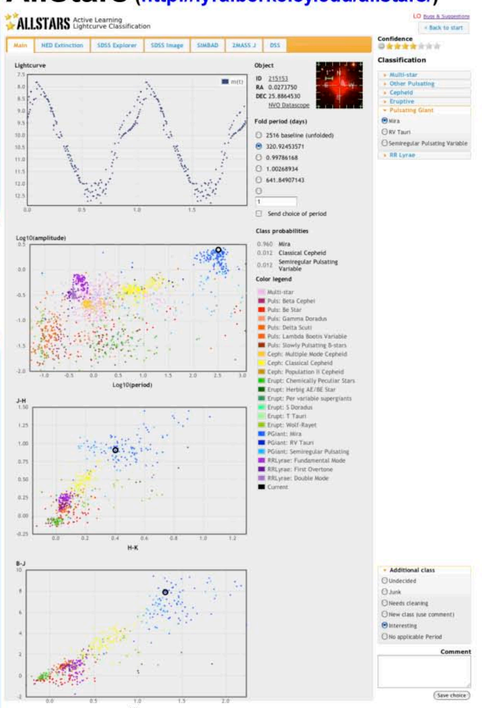
UC Berkeley Center for Time-Domain Informatics (http://cftd.info/)

Active Learning Lightcurve classification Service (ALLSTARS) is a crowdsourcing user interface for light curve classification

Designed to query users for classifications ONLY for the sources that would most improve the performance of a machine-learned classifier

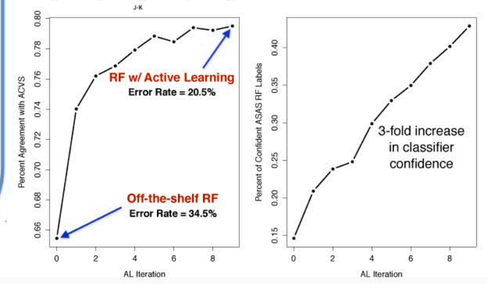
Uses Active Learning methodology, which has demonstrated success in varstar classification

AllStars (http://lyra.berkeley.edu/allstars/)



Machine-learned ASAS Classification Catalog (MACC) is a calibrated probabilistic classification catalog of 50k sources in the All-Sky Automated Survey into 28 variable star classes

- Easy-to-use interface for querying objects
- Customizable queries based on classification, anomaly score, or class probabilities
- Visualization powered by Google Fusion Tables
- Other survey catalogs coming soon…



January 2013

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.