The LSST catalog database will exceed 10 petabytes, comprising several hundred attributes for 5 billion galaxies, 10 billion stars, and over 1 billion variable sources (optical variables, transients, or moving objects), extracted from over 20,000 square degrees of deep imaging in 5 passbands with thorough time domain coverage: 1000 visits over the 10-year LSST survey lifetime. The opportunities are enormous for novel scientific discoveries within this rich time-domain ultra-deep multi-band survey database. Data Mining, Machine Learning, and Knowledge Discovery research opportunities with the LSST are described. Specific applications to LSST include scientific data mining, object classification, outlier identification, anomaly detection, image quality assurance, and survey science validation.

The enormous LSST data archive and object database enables a diverse multidisciplinary research program:
- Astronomy & astrophysics
- Machine learning (data mining)
- Exploratory data analysis
- XLDB (extremely large databases)
- Scientific visualization
- Computational science & distributed computing
- Inquiry-based science education – using data in the classroom

Definitions of data mining:
1. An information extraction activity whose goal is to discover previously hidden facts (patterns, relationships, links, correlations, trends) contained in large databases.
2. The transformation of knowledge from a data format representation into a rule format representation.
3. Knowledge Discovery in Databases (KDD).

Data Mining Research with the LSST

K. Borne (George Mason University), M. Strauss (Princeton), J. A. Tyson (UC Davis)

Astronomers are trained as data miners because we:
- Characterize the known (clustering)
- Assign the new (classification)
- Discover the unknown (outlier detection)

Class discovery through multivariate clustering:

Outlier detection through cluster/class mapping:

Potential Science Use Cases: (opportunities for LSST Science Collaboration Team input)
- Provide rapid probabilistic classifications for all 10,000 LSST events each night
- Find new "fundamental planes" of parameters (e.g., the fundamental plane of elliptical galaxies)
- Find new correlations, associations, relationships of all kinds from 100+ attributes in the science databases
- Compute N-point correlation functions over a variety of spatial and astrophysical parameters
- Discover voids or zones of avoidance in multi-dimensional parameter spaces (e.g., period gaps)
- Discover new and exotic classes of astronomical objects. Discover new properties of known classes
- Discover new and improved rules for classifying known classes of objects (e.g., photometric redshifts)
- Identify novel, unique data behavior in the time domain from time series data of all known variable objects
- Hypothesis testing: verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions of training samples
- Serendipity: discover the rare one-in-a-billion type of objects through outlier detection
- Quality assurance: identify glitches, anomalies, image processing errors through deviation detection

LSST Database Contents:
- >100 database tables
- Image metadata = 670M rows
- Source catalog = 2608M rows
- Object catalog = 22B rows, with >200 attributes
- Moving Object catalog
- Variable Object catalog
- Alert catalog
- Calibration metadata
- Configuration metadata
- Processing metadata
- Provenance metadata

Relationships Discovery via Data Visualization:

SOM for Discovery of Semantic Relationships and for Visual Exploration of Classes and Concepts:

Are you interested in contributing or participating? Please contact Kirk Borne at kborne@gmu.edu