Large Synoptic Survey Telescope

Data Mining Research with the LSST

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

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).*

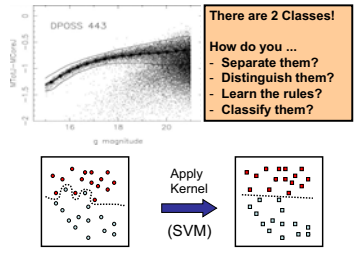
Data ➔ Information ➔ Knowledge ➔ Understanding / Wisdom!

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

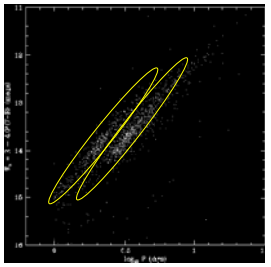
Some data mining methods explained:

Clustering (Segmentation): *Group data items according to tight relationships or greatest similarity, and separate the items that are most different.*
Principal Component Analysis: *Reduce the dimension of the input vectors (multi-attribute data records) by eliminating redundant components – captures the directions of greatest variance in the data.*
Independent Component Analysis (ICA): *Identify the mutual statistically independent components in multi-attribute data records.*
Outlier (Anomaly, Glitch, Deviation) Detection: *Find data items that fall outside the bounds of known or statistically robust clusters.*
Classification: *Assign data items to predetermined groups (classes, clusters).*
Bayesian Analysis: *Assess the probability of a hypothesis being correct (for example, whether a classification is valid) by incorporating the prior probability of the hypothesis and the experimental data supporting the hypothesis.*
Support Vector Machines (SVM): *Map input vectors to a higher dimensional space where the classes are divided by a maximal separating hyperplane.*
Nearest Neighbor Method: *Classify a data item according to its nearest neighbors (i.e., records that are most similar).*
Association Mining (Market Basket Analysis): *Associate data with higher than expected co-occurrence of attribute-value combinations.*
Link Analysis: *Associate data in a graph network that are connected through shared attribute values or semantic relationships.*
Artificial Neural Networks (ANN): *Non-linear predictive models that learn through training and resemble biological neural networks in structure.*
Rule induction: *Extract useful if-then rules from data based upon statistical significance and information gain.*
Decision Trees: *Hierarchical sets of decisions, based upon rules, for rapid classification of a data collection.*
Genetic Algorithms: *Rapid optimization techniques that are based on the concepts of natural evolution.*
Data visualization: *The illustration and visual interpretation of complex relationships in multidimensional data using graphics tools.*
Self-Organizing Map (SOM): *Graphically organizes (in a 2-dimensional map) the information stored within a database based upon similarities and links between concepts. It can be used to find hidden relationships and patterns in more complex data collections.*

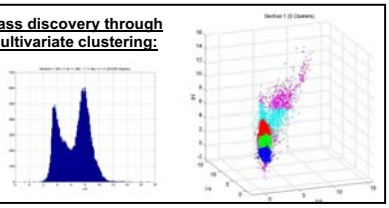
SVM for Classification:



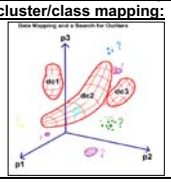
ICA for Classification:



Class discovery through multivariate clustering:



Outlier detection through cluster/class mapping:



Sample Machine Learning Applications for LSST: **(credit: http://www.thinkingtelescopes.lanl.gov/)**
***Automated Feature Extraction*:** Real-time identification of artifacts and transients in direct and difference images.
***Classifiers*:** Automated classification of celestial objects based on temporal and spectral properties.
***Anomaly Detection*:** Real-time recognition of important deviations from normal behavior for persistent sources.

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 database
- 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, unexpected 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 = 675M rows**
- **Source catalog = 260B rows**
- **Object catalog = 22B rows, with 200+ attributes**
- **Moving Object catalog**
- **Variable Object catalog**
- **Alerts catalog**
- **Calibration metadata**
- **Configuration metadata**
- **Processing metadata**
- **Provenance metadata**

Relationship discovery via Data Visualization:

Image Credit: http://dls.bell-labs.com/mirage/mirage.html

SOM for discovery of semantic relationships and for visual exploration of classes and concepts:

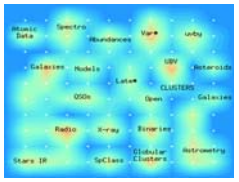
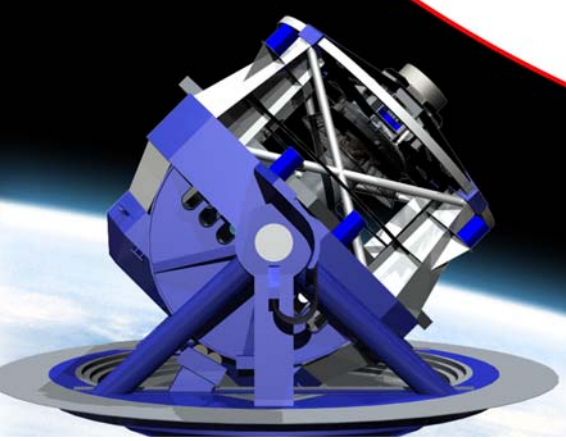


Image Credit: http://simbad.u-strasbg.fr/A+A/map.pl

Data Mining Research Challenge Areas: **(opportunities for LSST Data Management & Science Teams)**
- **scalability (at petabytes scales) of existing machine learning and data mining algorithms**
- **development of grid-enabled parallel data mining algorithms**
- **designing a robust system for brokering classifications from the LSST event pipeline**
- **multi-resolution methods for exploration of petascale databases**
- **visual data mining algorithms for visual exploration of the massive databases**
- **indexing of multi-attribute multi-dimensional astronomical databases (beyond RA-Dec spatial indexing)**
- **rapid querying of petabyte databases**



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

LSST is a public-private partnership. Design and development activity is supported by in part the National Science Foundation under Scientific Program Order No. 9 (AST-0551161) and Scientific Program Order No. 1 (AST-0244680) through Cooperative Agreement AST-0132798. Portions of this work are supported by the Department of Energy under contract DE-AC02-76SF00515 with the Stanford Linear Accelerator Center, contract DE-AC02-98CH10886 with Brookhaven National Laboratory, and contract W-7405-ENG-48 with Lawrence Livermore National Laboratory. Additional funding comes from private donations, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.

Brookhaven National Laboratory, California Institute of Technology, Columbia University, Google, Inc., Harvard-Smithsonian Center for Astrophysics, Johns Hopkins University, Kavli Institute for Particle Astrophysics and Cosmology - Stanford University, Las Cumbres Observatory Inc., Lawrence Livermore National Laboratory, National Optical Astronomy Observatory, Princeton University, Purdue University, Research Corporation, Stanford Linear Accelerator Center, The Pennsylvania State University, The University of Arizona, University of California at Davis, University of California at Irvine, University of Illinois at Urbana-Champaign, University of Pennsylvania, University of Pittsburgh, University of Washington