American Geophysical Union (AGU) Fall Meeting, 2008: Abstract

Kirk Borne (George Mason University)
Hillol Kargupta, Kamalika Das, Wes Griffin (University of Maryland, Baltimore County)
Chris Giannella (Loyola College)

Data-intensive science and knowledge discovery involving very large sky surveys are playing increasingly important roles in today's astronomy research. This Discovery Informatics scientific approach is evolving as a core research paradigm in all science disciplines. In particular, Astroinformatics is developing as the formalization of data-intensive astronomy for research and education. Nearly completed projects (such as the Sloan Digital Sky Survey SDSS, the 2-Micron All-Sky Survey 2MASS, and the GALEX All-Sky Survey) and future projects (such as the WISE All-Sky Survey, Pan-STARRS, and the Large Synoptic Survey Telescope LSST) are destined to produce enormous catalogs of astronomical sources. These collections are naturally distributed and heterogeneous, in addition to being terascale, petascale, and beyond.

It is this virtual collection of terabyte and (eventually) petabyte catalogs that will enable remarkable new scientific discoveries through the integration and cross-correlation of data across multiple survey dimensions (time, wavelength, and sky coverage). However, this will be difficult to achieve without a computational backbone that includes support for queries and data mining across distributed virtual tables of de-centralized, joined, and integrated sky survey catalogs. Moreover, use of local data management systems such as MyDB, MySpace in AstroGrid, and Grid Bricks for storing and managing user's local data is becoming increasingly popular. This is opening up the possibility of constructing Peer-to-Peer (P2P) networks for data sharing and mining. We will report on our research in these areas. We are exploring the possibility of using distributed and P2P data mining technology for exploratory astronomy from data integrated and cross-correlated across these multiple sky surveys. We will report on new scientific results, including new explorations of the classical fundamental plane problem, in which multiple dimensions of galaxy parameter space can be reduced to a hyperplane in lower dimensions. Since the attributes which define the fundamental plane span two data repositories (SDSS and 2MASS) instead of one, we focus on cross-matching them through the NVO, and we then apply distributed data mining algorithms to analyze these data distributed over a large number of compute nodes. Distributed data mining techniques will not require scientists to download massive chunks of data for scientific discovery and will thus enable them to use distributed database queries across distributed virtual tables of de-centralized, joined and integrated sky survey catalogs. This will make the existing client-server-based astronomy data services richer by providing the power of distributed and P2P data mining technology.