LSST AND BIG DATA

LSST Corporation, 933 N. Cherry Avenue, Tucson, AZ 85721 520 881-2626 www.lsst.org

The Future Has Arrived

Advances in technology and the ever-growing role of digital sensors and computers have led to an exponential growth in the amount and complexity of data that we collect in industry, engineering, and science. We are at the threshold of an era in which hypothesis-driven exploration is being complemented with data-driven discovery. This alternative way to pursue research affects all fields, from genomics in biology, to astrophysics and many domains in social sciences. The data collected are complex in size, dimension, and heterogeneity — all three generating what is generically referred to as "Big Data." These data provide unprecedented opportunities for new discoveries; they also come with challenges that must be addressed. As a recent (Feb 12, 2013) MIT Technology Review reports, the challenge "is not processing or storing this amount of data — Moore's law should take care of all that. Instead, the difficulty is uniquely human. How do humans access and make sense of the exascale data sets?"[1] We must develop new methods for extracting knowledge and understanding from Big Data.

In 2010, Eric Schmidt, Google CEO, noted that "every two days, we create as much information as we did from the dawn of civilization up until 2003." Whereas we had to struggle with Megabytes (10^6 bytes) of data in the 1990s, many datasets in 2013 include many Petabytes (10^{15} bytes) of data, and this number is expected to grow to the Exabyte (10^{18}) and even Yottabyte (10^{24}) scales before 2020.

LSST

Many regard LSST as the lighthouse project for looming Big Data challenges that are faced by most areas of science and engineering.[2] The complexity and high dimensionality of the data from LSST are analogous to those faced by industry and widening areas of science. Open source solutions developed for LSST will advance the field of Big Data analytics and the knowledge-from-data challenge generally. Discovering the unexpected in Petabytes of data is an exciting challenge with potential for significant spin-off. LSST will generate a hundred Petabytes of data. Many of the scientific questions that will be addressed by LSST require extraction of knowledge and understanding via statistical analysis of relationships buried in the time-space data. Algorithms for automated discovery which will be developed for this mission will have useful application in any area.

Over one million alerts per night will be issued worldwide within one minute for objects that change in position or brightness. Mining these data quickly and efficiently for the known knowns and the unknown unknowns presents unprecedented opportunities as well as object classification algorithm challenges. LSST scientists will develop new algorithms for this and for uncovering hidden trends. This requires novel database architectures supporting new scalable and affordable algorithms for analyzing big data spatial and temporal data sets. LSST scientists and engineers have already been recognized for advances in extremely large database technology.[3]

[1] http://www.technologyreview.com/vi-exascale-computing/

[2] http://www.economist.com/node/15557443

[3] http://go.usa.gov/Gsi