Distributed and Peer-to-Peer Data Mining for Scalable Analysis of Data from Virtual Observatories

Hillol Kargupta

University of Maryland Baltimore County & Agnik

Kirk Borne

GMU

Chris Giannella

MITRE

Acknowledgement: Sugandha Aurora, Nikhil Kumar, Rajarshi Mallik, Sandipan Dey, Tushar Mahule, Kanishka Bhaduri, Kamalika Das, Haimonti Dutta,Wes Griffin, Codrina Lauth

Supported by **NASA NRA NNX07AV70G**

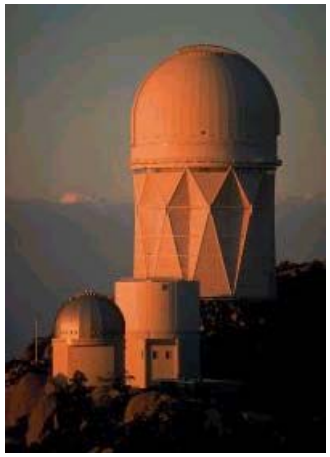
Road Map

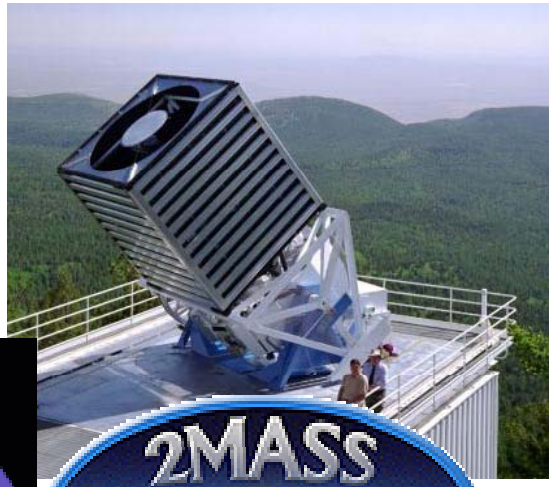
- Space/Earth Science and Data Mining
- Distributed Data Mining
  - P2P Mining of Virtual Observatory Data
  - Collaborative Tagging and P2P Text Classifier Learning
- Future Work

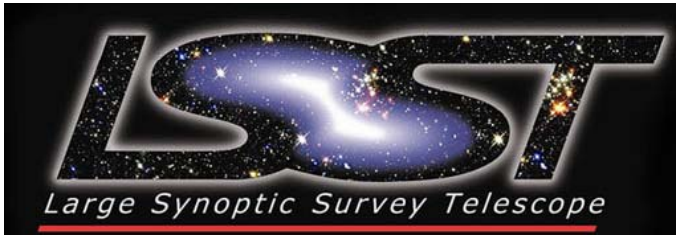
Future of Astronomy and Earth Science Data Processing Environments

- High throughput data streams
- Multiple data sources
- Heterogeneous distributed computing environment
- Increasing number of users; scientific communities forming peer-to-peer networks
- Increasing demand for faster response time

Multiple Data Sources

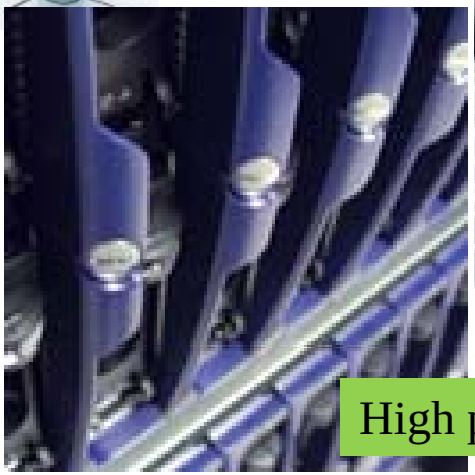






Distributed and Peer-to-Peer Computing Environment

Distributed computing environments (community of users)



High performance grid computing

GALAXY ZOO: How Community-based Science is Shaping up….

How Do We Analyze Data in Distributed Environment?

- Multiple data tables and streams
- Computing Environment:
  - High performance computing
  - Cluster of relatively low-end desktop machines
- Objective
  - Quickly sifting through distributed data and identify potential matches

Project Objectives

- Objectives: Develop distributed and P2P data mining algorithms and systems

- Enabling Technical Innovations:

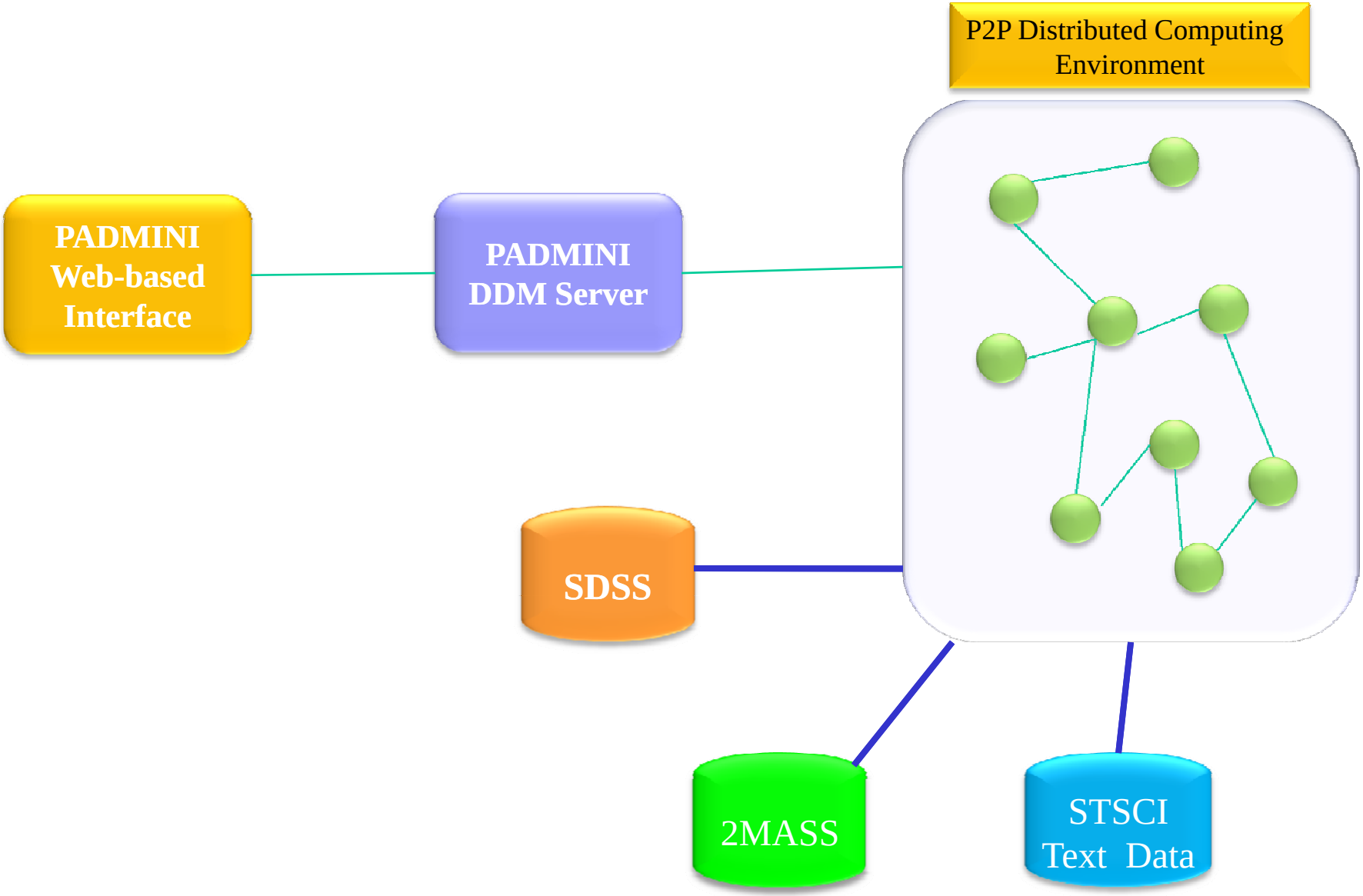
  Algorithmic Innovations:

  – Distributed classifier learning
  – Distributed eigenstate monitoring
  – Distributed outlier detection

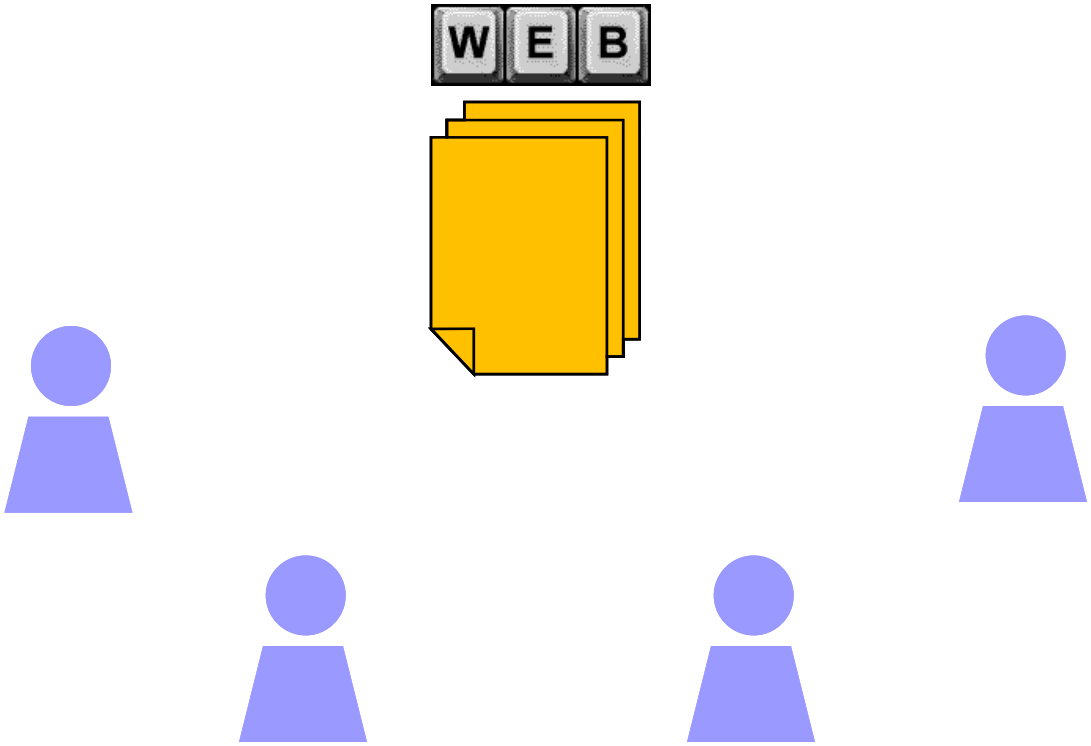
  Systems Innovations:

  – Google-Sky powered PADMINI System
  – Web-based user interface
  – Plug-n-play backend distributed data mining modules

Architecture of PADMINI



Peer-to-Peer Text Classification & Classifier Learning

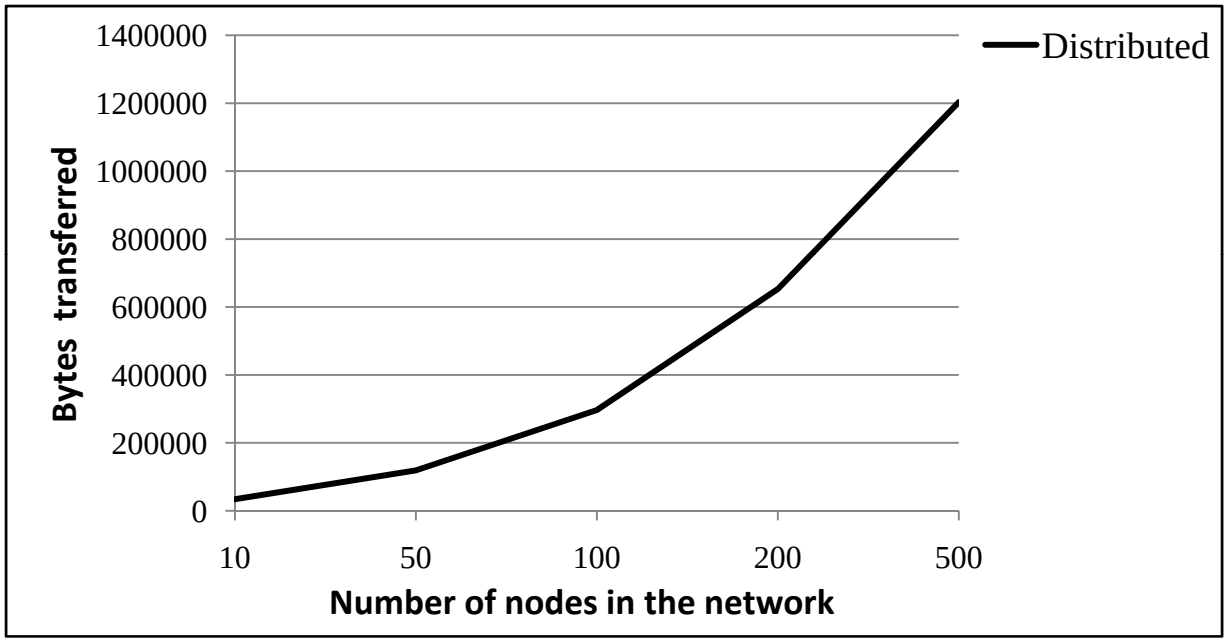


- Large Astronomy Text Repositories
- Collaborative text labeling
- P2P classifier learning from distributed labeled data

Algorithmic Approach

- Linear classifier construction
- Distributed data:
  - Each site has a collection of data tuples
- Can be posed as linear programming problem
  - Minimizing the error
- Distributed linear programming
- Distributed simplex algorithm

Communication Cost vs. Network Size



- Number of nodes in the network is varied from 10 to 500 nodes
- Number of variables in a constraint equation is kept constant at 35

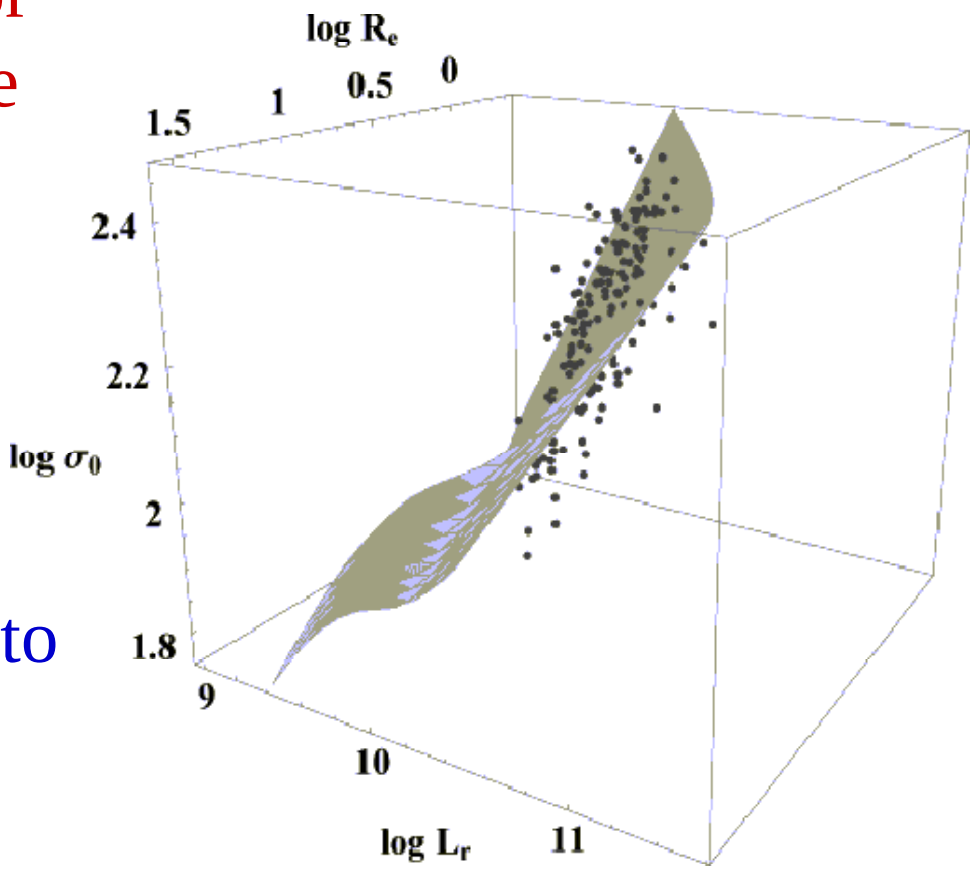
Peer-to-Peer Virtual Observatory Data Monitoring

- Detecting changes in streams of VO data using P2P data mining algorithms

- P2P Eigenstate monitoring algorithms

The Fundamental Plane of Elliptical Galaxies

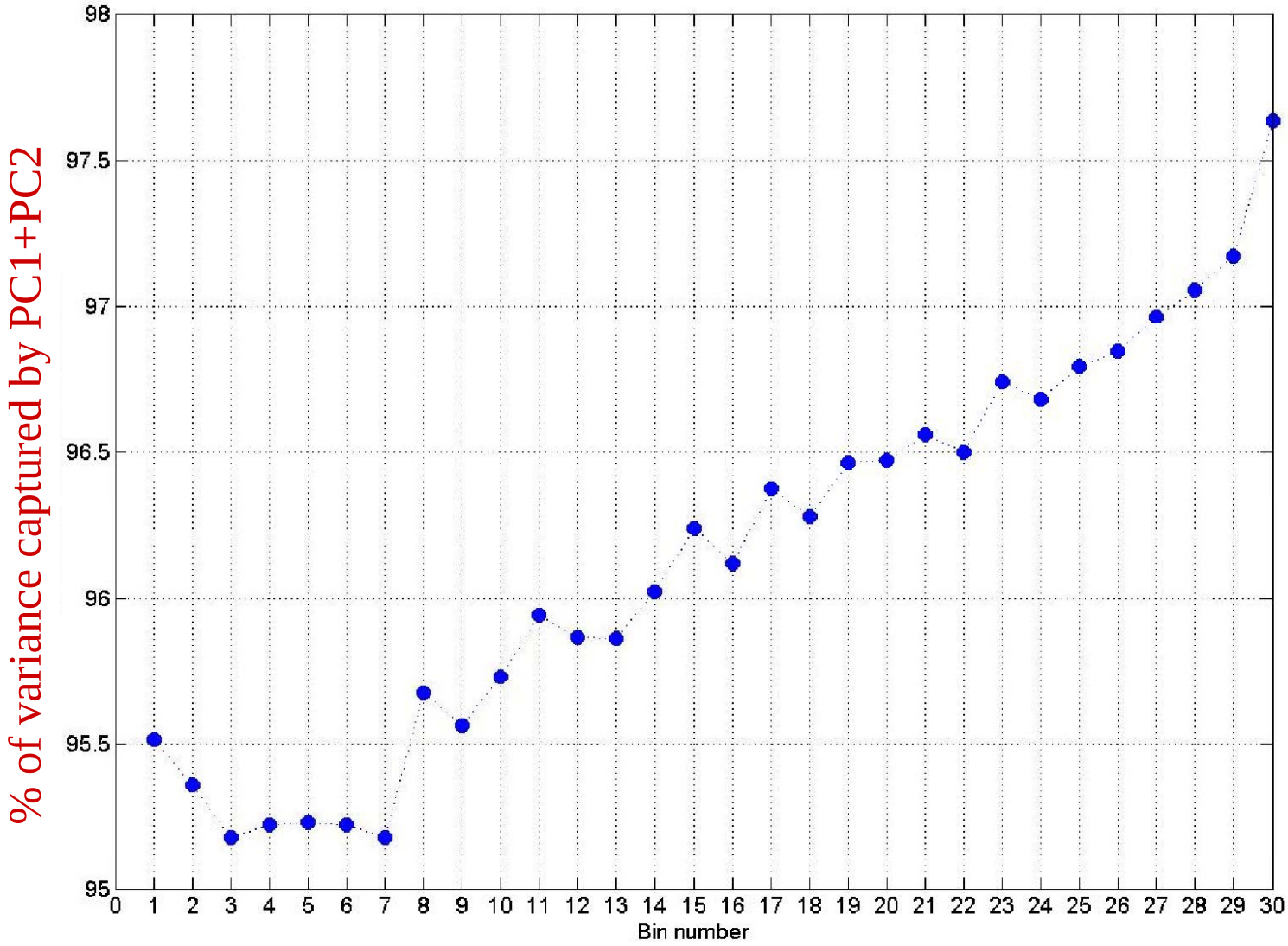
- The **fundamental plane** for elliptical galaxies tracks the correlation between the effective radius, average surface brightness, and central velocity dispersion.
- With this correlation, one can determine the distance to galaxies, which is a critical but difficult task in astronomy.



Our Prior Observation

- Produced a 156,000 cross-matched galaxy dataset with attributes from SDSS and 2MASS
  - SDSS: velocity dispersion, Pertrosian I band angular effective radius, redshift
  - 2MASS: K band mean surface brightness
- Partitioned into bins w.r.t. local galaxy density ρ
  - Estimated ρ using Delaunay tessellation methods
  - The local density around a selected galaxy is inversely proportional to the local volume that contains only that one galaxy – measured by the Delaunay tessellation
- Estimated the fundamental plane parameters for each bin (e.g., variance captured in the first two principal components, as $f(\rho)$)

Results



low (Local Galaxy Density) high

Future Work

- Finish building the research prototype of PADMINI
- Include more algorithmic support for P2P data mining
- Continue to interact with STSCI
- Extensive testing and evaluation.
- Transfer the technology

Sample References & Recent Accomplishments

- K. Das, K. Bhaduri, S. Arora, W. Griffin, K. Borne, C. Giannella, H. Kargupta. Scalable Distributed Change Detection from Astronomy Data Streams using Local, Asynchronous Eigen Monitoring Algorithms. SIAM International Conference on Data Mining, pp 245-256. Nevada. 2009.
- R. Wolff, K. Bhaduri, and H. Kargupta. A Generic Local Algorithm for Mining Data Streams in Large Distributed Systems. IEEE TKDE, 2008.
- S. Datta and H. Kargupta.(2008). A Communication Efficient Probabilistic Algorithm for Mining Frequent Itemset from Peer-to-Peer Network. Statistical Analysis and Data Mining Journal. Accepted for publication. (In press).
- K. Das, W. Griffin, H. Kargupta, C. Giannella and K. Borne. (2008) Scalable Multi-Source Astronomy Data Mining in Distributed, Peer-to-Peer Environments. Astronomical Data Analysis Software and Systems (ADASS) XVIII Conference. Quebec, Canada.
- K. Borne, Scientific Data Mining in Astronomy, Next Generation Data Mining (CRC Press: Taylor and Francis), pp. 91-114 (2009).
- K. Das, H. Kargupta, and K. Bhaduri. (2009). A Local Distributed Peer-to-Peer Algorithm Using Multi-Party Optimization Based Privacy Preservation for Data Mining Primitive Computation. Accepted for publication in the proceedings of the 9th International Conference on Peer-to-Peer Computing.

- 2008 IBM Innovation Award
- Paper selected for "Best of 2008 SIAM Data Mining Conference (SDM'08)" selection.