Griding the Sky: An Astronomical Perspective

Andrew Connolly
Department of Physics and Astronomy
University of Pittsburgh

New Data New Analyses

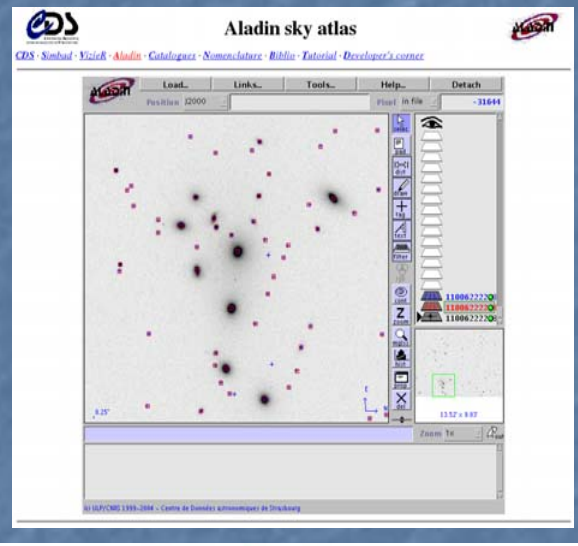
- Data Federation
  - All-sky surveys
  - Multifrequency federated data sources
- Growth of structure
  - Point processes
- Classification and Anomalies
  - High dimensional classification
  - Time domain analyses
- Moving sources
  - Several thousand fold dynamic range in motions

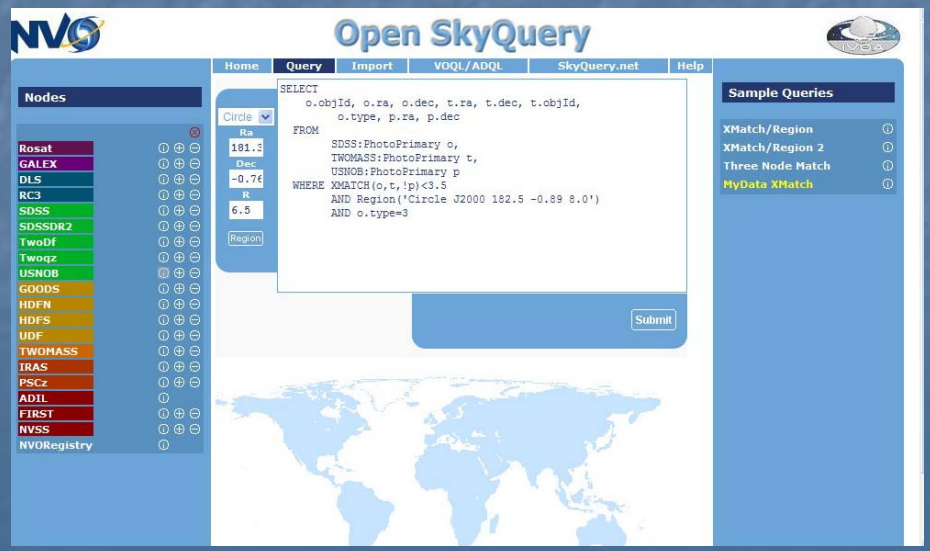
Current Surveys

- SDSS
  - 7000 sq degrees
  - 5 filters, UV to near-infrared
  - 1.5TB, 1.87×10^8 sources (DR4 2004)
  - 40TB raw data (over 8 years)
- 2MASS
  - 40,000 sq degrees
  - 3 filters, near-infrared
  - 5×10^8 sources, 100 TB (over 4 years, 2 telescopes)

The Virtual Observatory

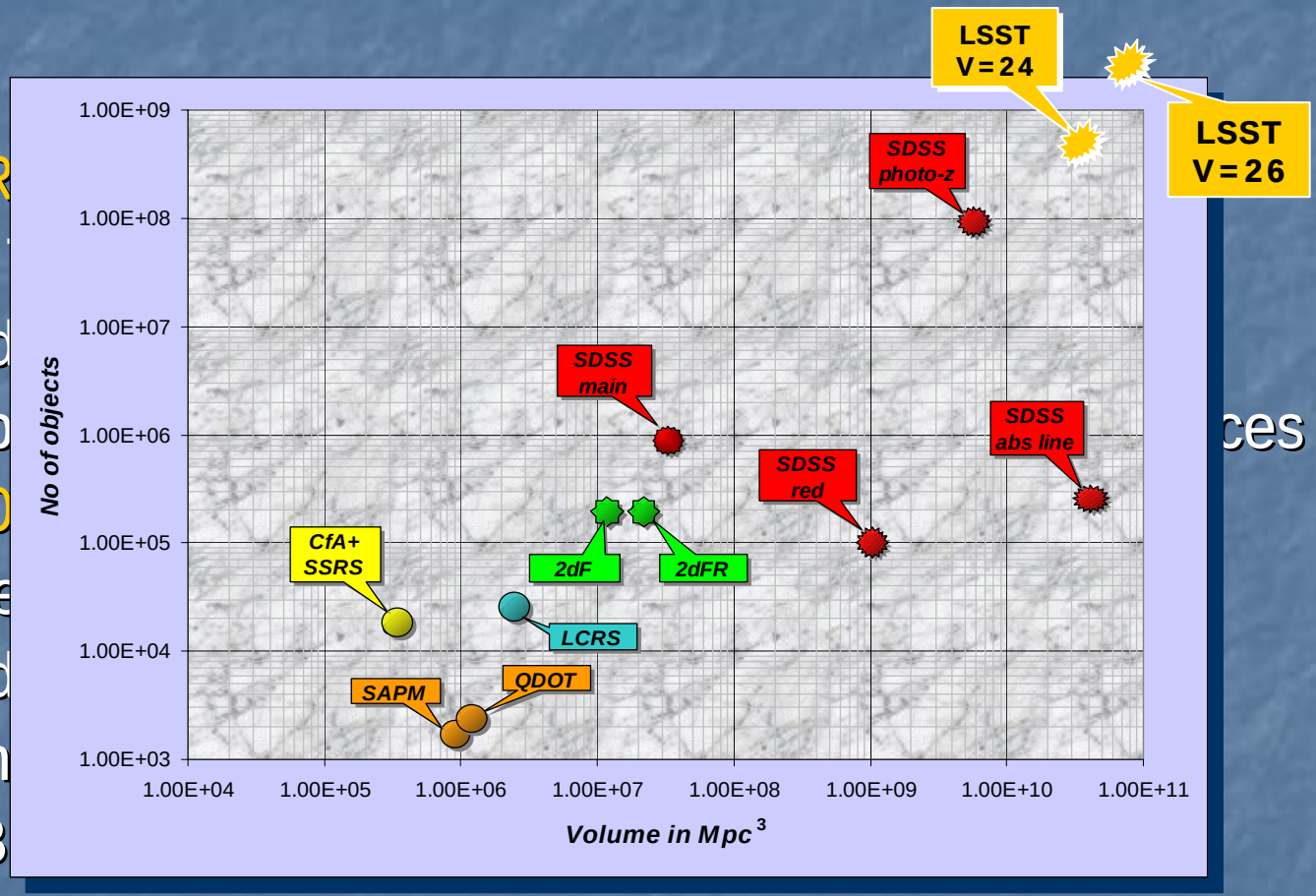
- Federation of Databases
  - Openskyquery.net
    - Cross matching 19 astronomical databases
    - Multiple database cross matches
    - Upload external data
    - Accessible through webservices
- Defined wire formats
  - XML, VOTables
- Extensible webservices
- Accessible gridservices



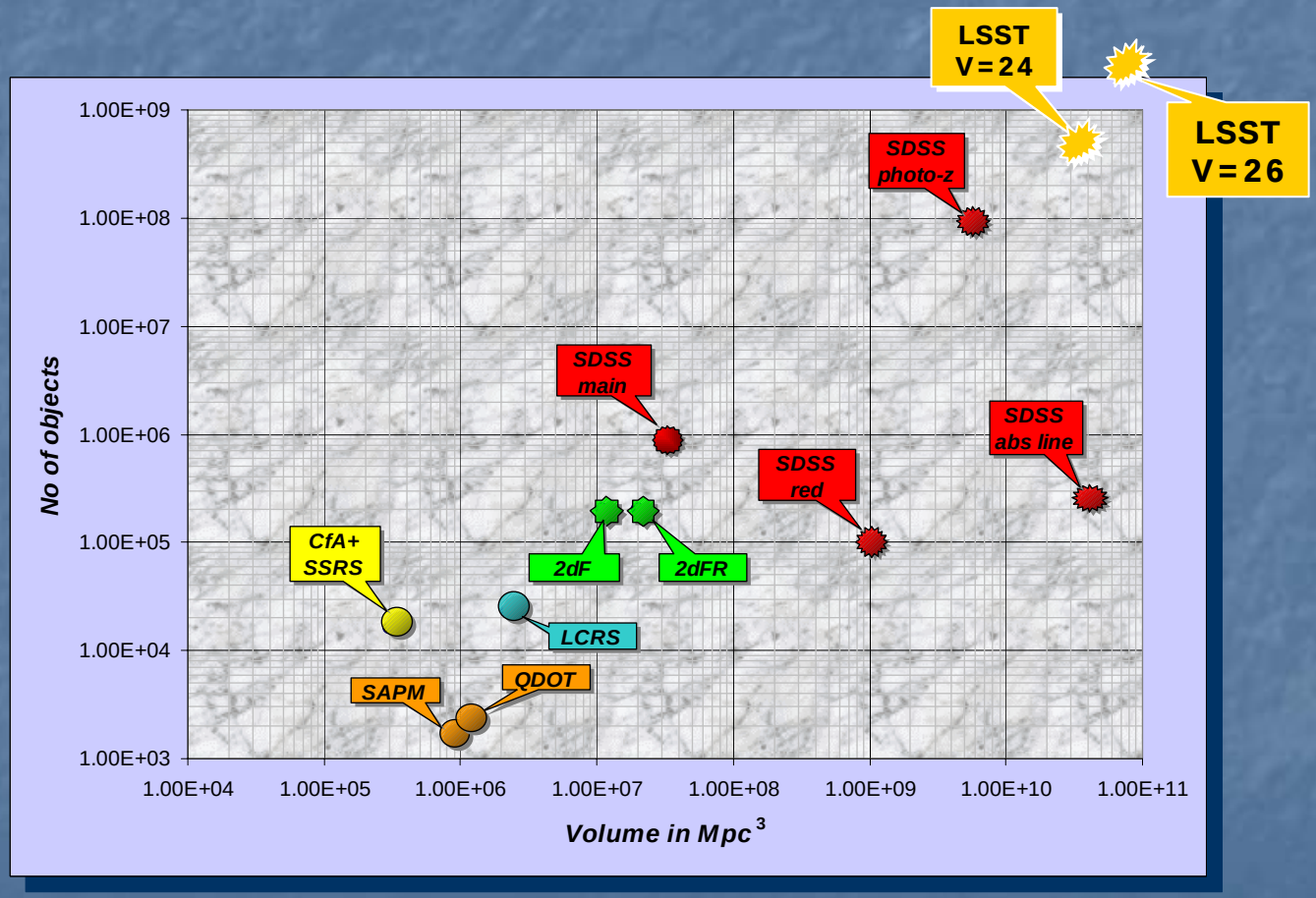


The Next Generation

- PanSTAR
  - 2.5m
  - 7 sq d
  - Multip
- LSST (20
  - 8m Te
  - 7 sq d
  - 10s in
  - 16 TB
  - >1 PB catalog and image database

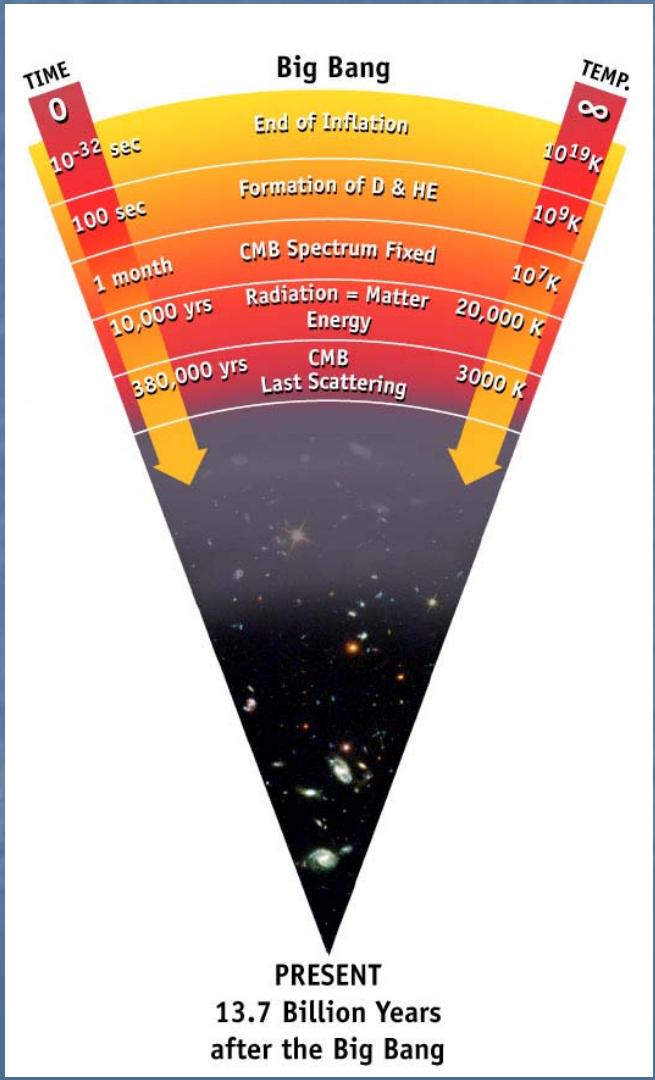


The Data Flow in 2010

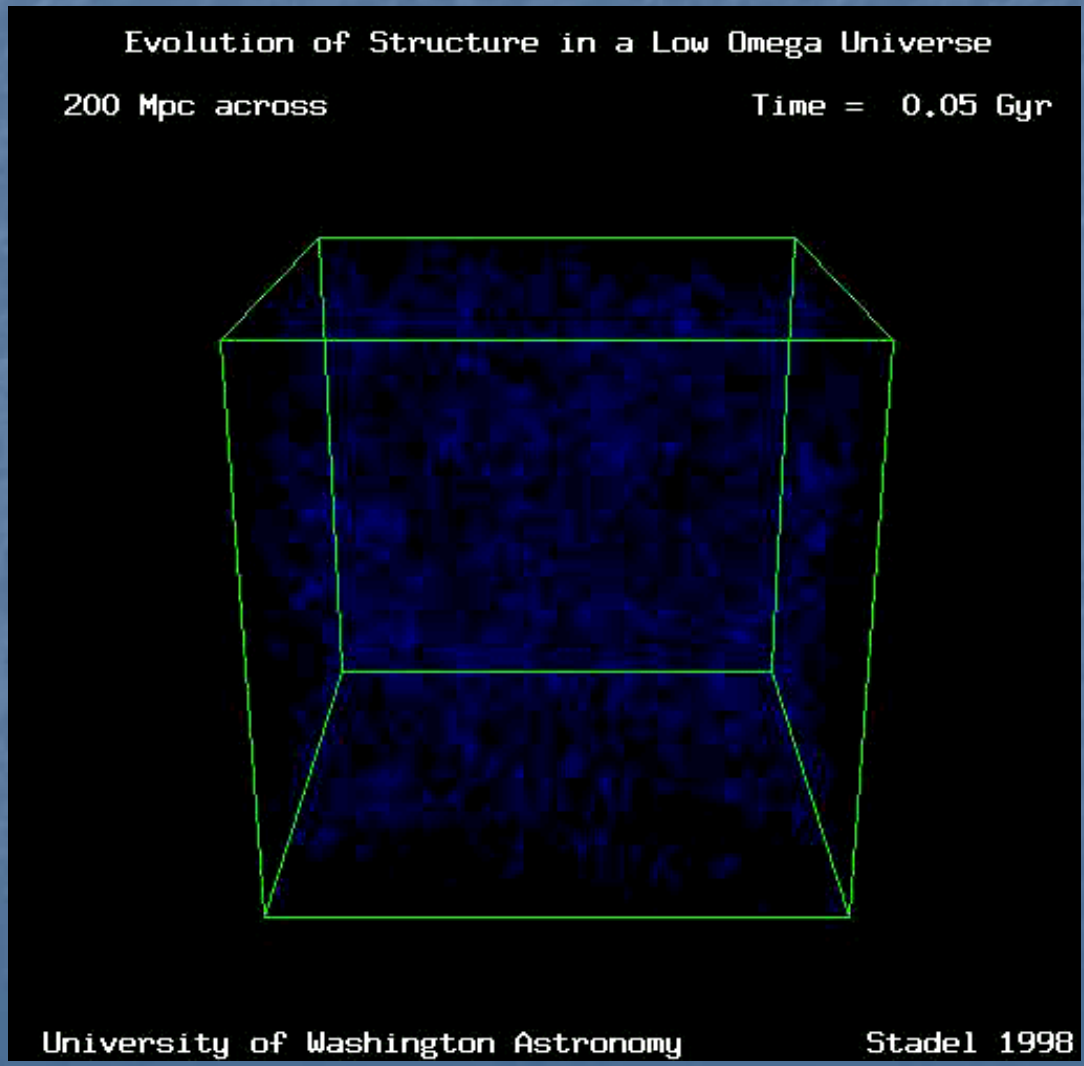


What are we trying to extract?

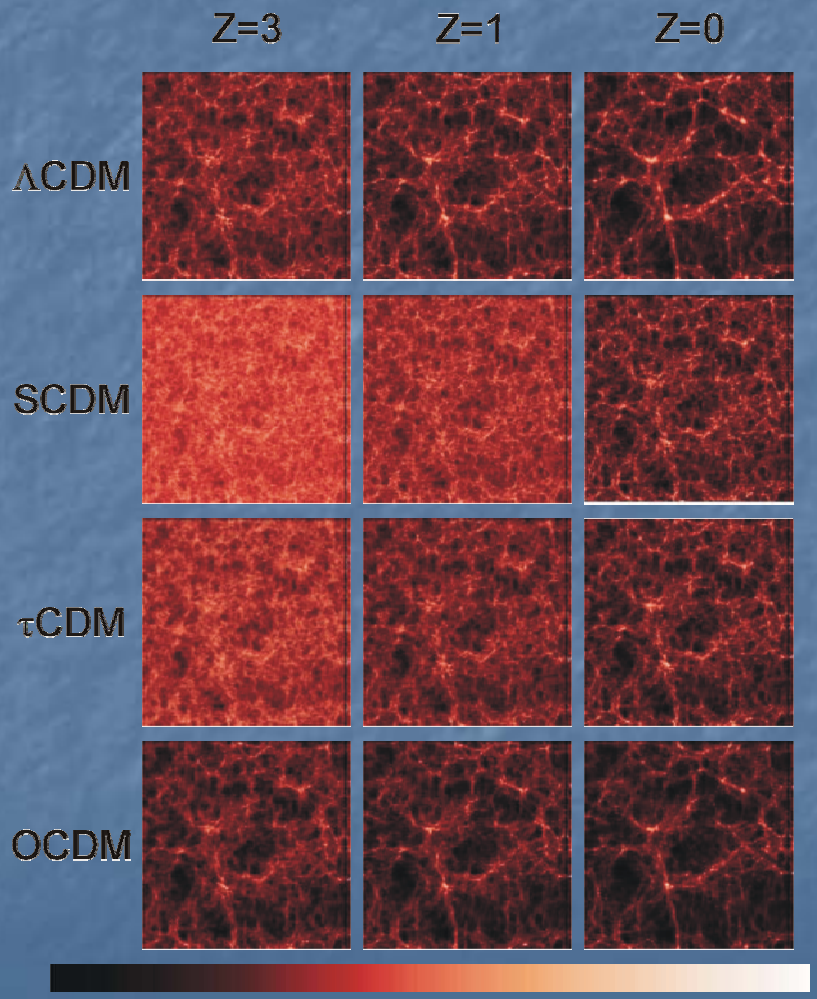
- The universal questions
  - What gives rise to the coherent structure we see
  - How are the luminous and dark matter correlated
  - What classes of source exist in the universe
  - How do sources change with time

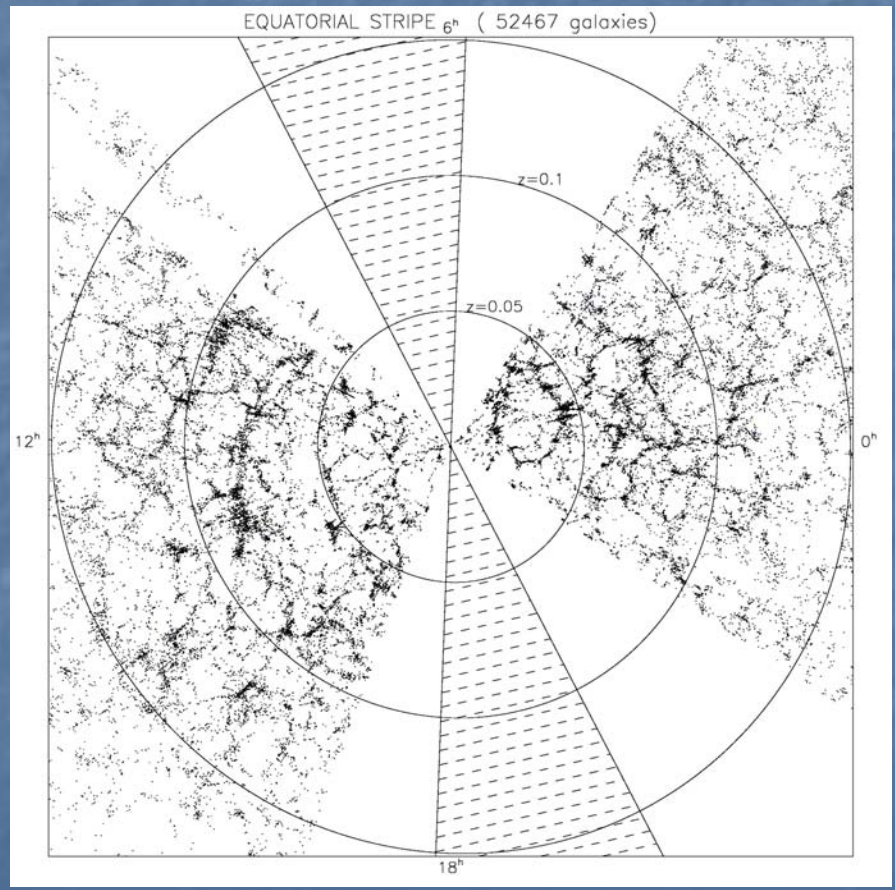


What do we want to learn?



Constraining Cosmology

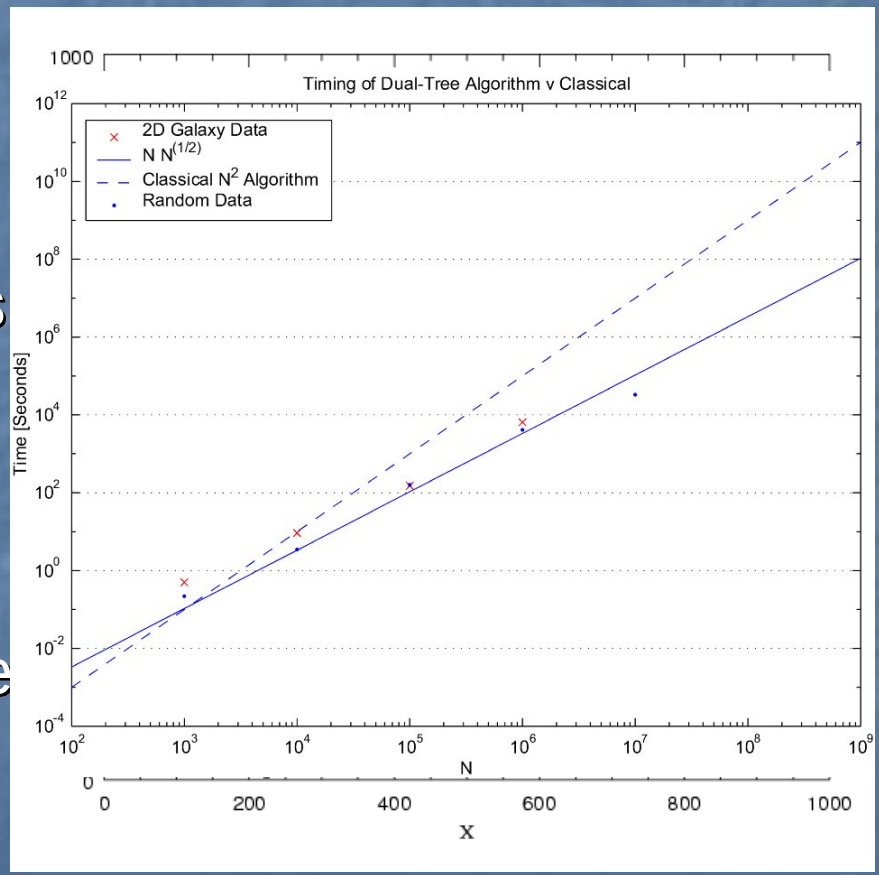




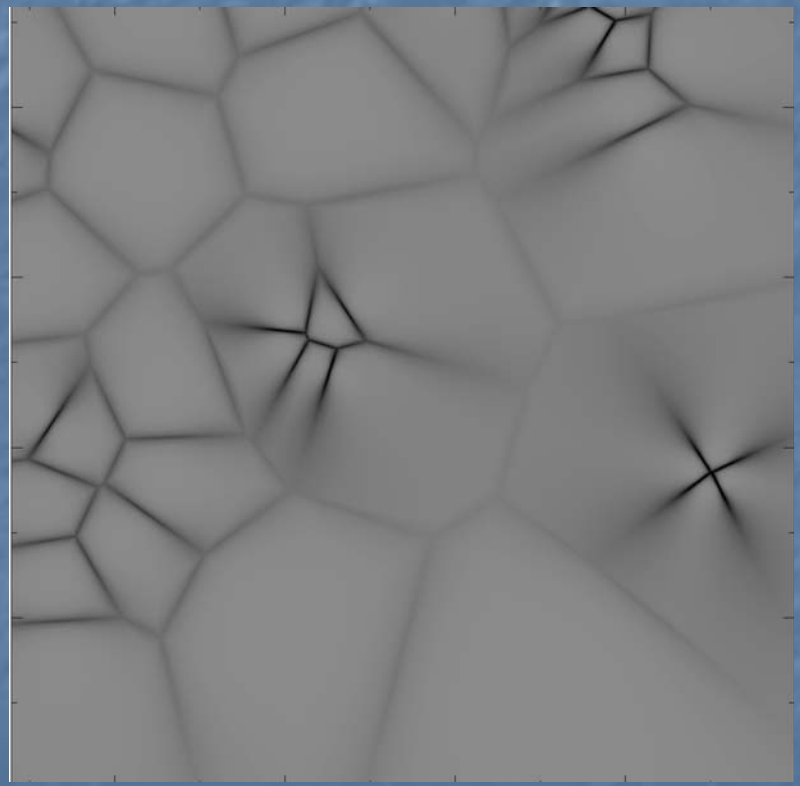
Virgo Consortium

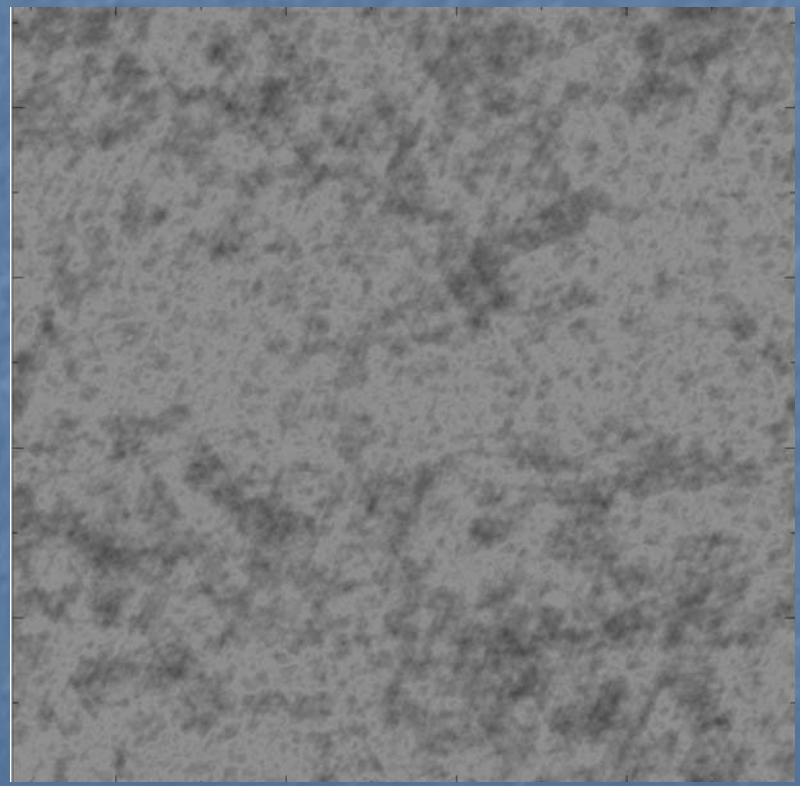
The Statistics of Clustering

- Condense the point processes to a “single” number
- Two-point correlations
  - Find the distribution of pairs
  - Compare to random
  - Scales as N^2 naively
- Three point correlations
  - Find the distribution of triple
  - Scales a N^3 naively



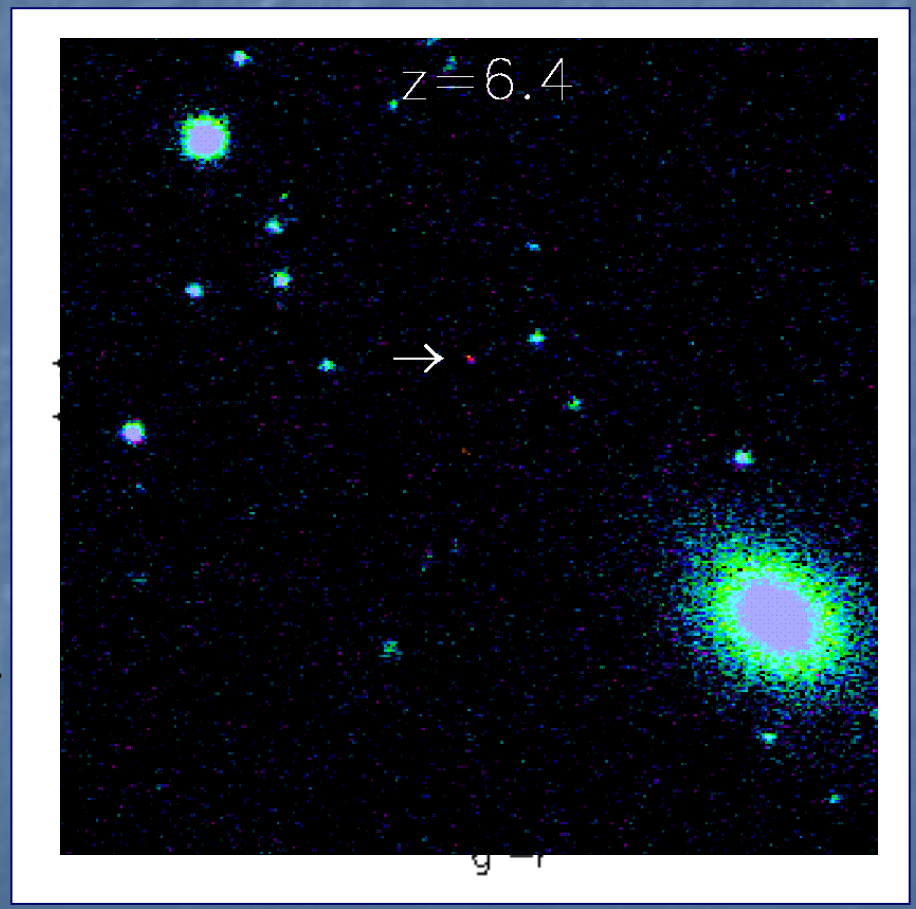
Why higher order statistics?



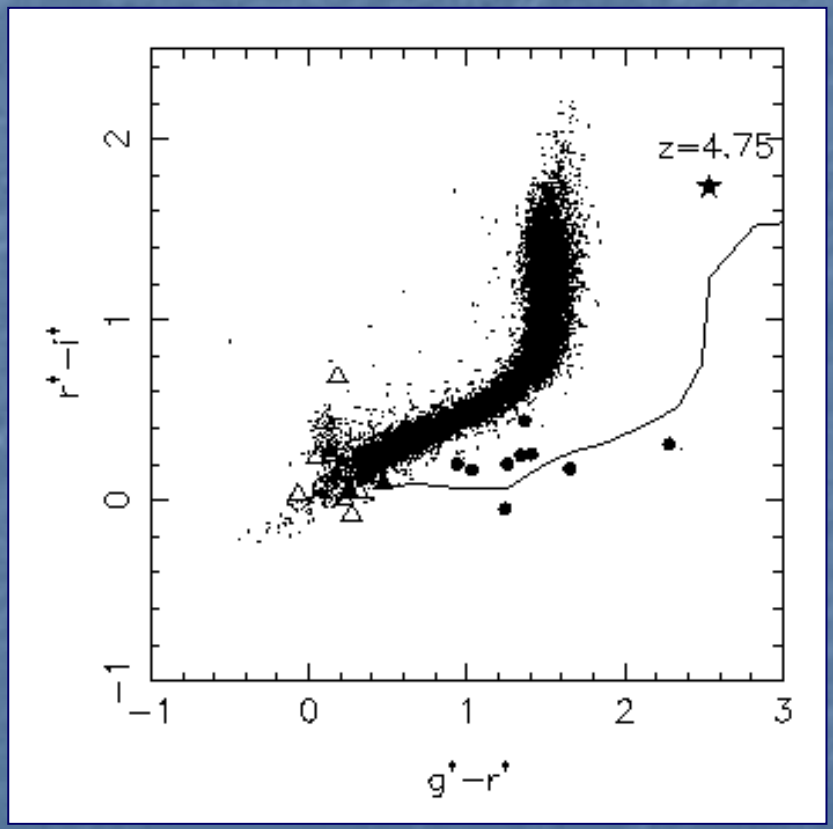


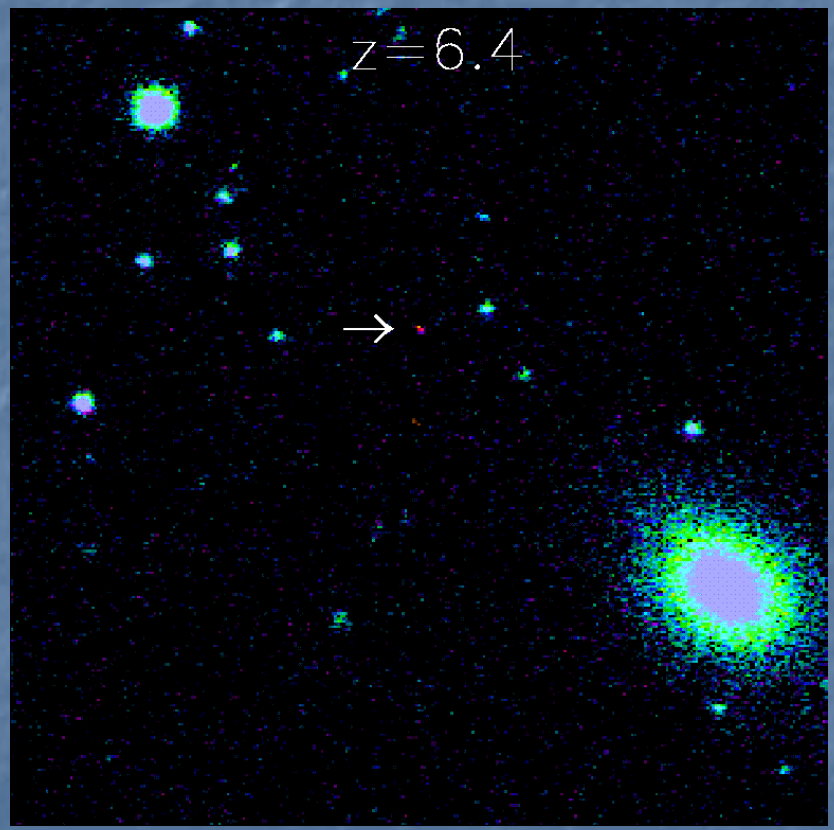
Classification and Anomalies

- Classification leads to new physics
  - Classification of neutron stars, QSOs, high redshift galaxies
- Anomalies from these classification
  - Photometric variables
    - Supernovae
    - Gamma-ray bursters
  - Astrometrically variable
    - Proper motions
    - Asteroids (PHAs)
- Anomalies can swamp any hope to follow them up
  - LSST provides 1000+ variables per night (plus false detections)
  - Need robust statistics
  - Need robust metrics for classification



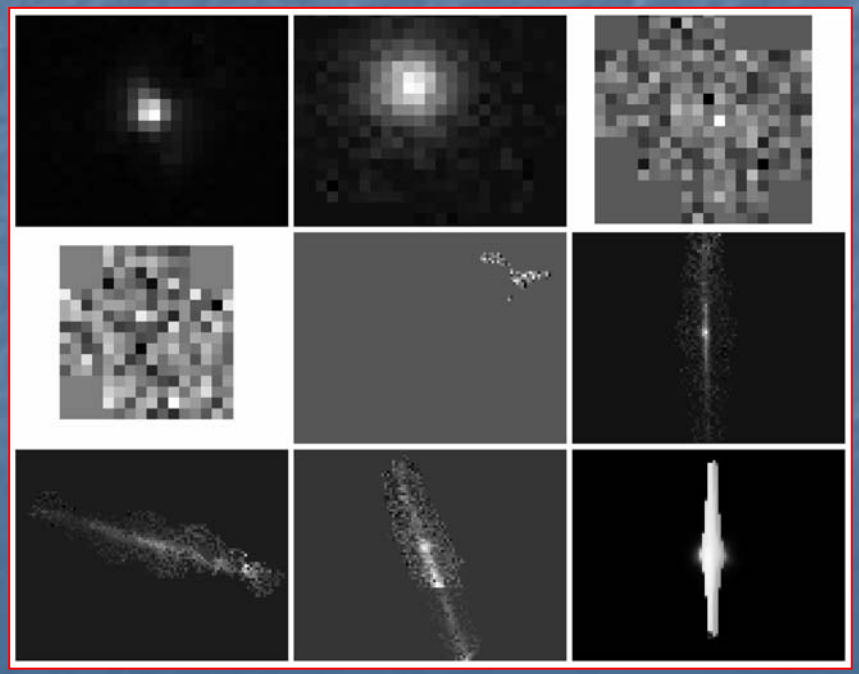
Classic classification

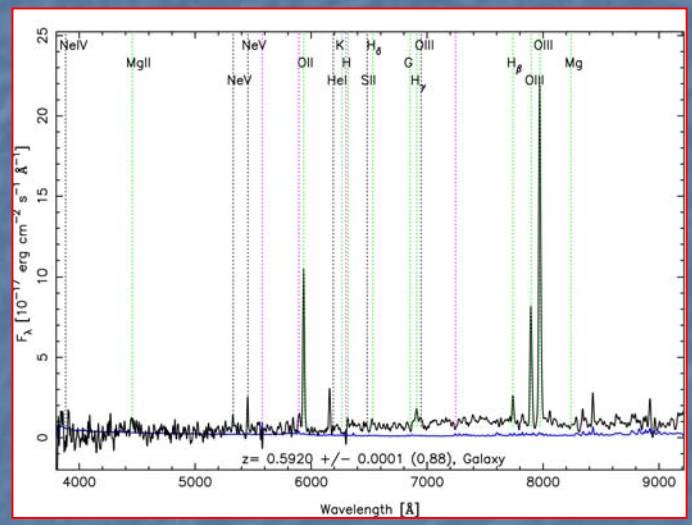


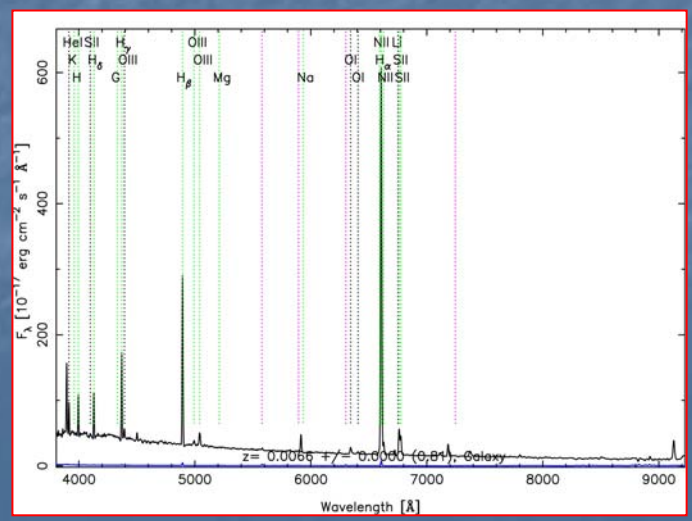


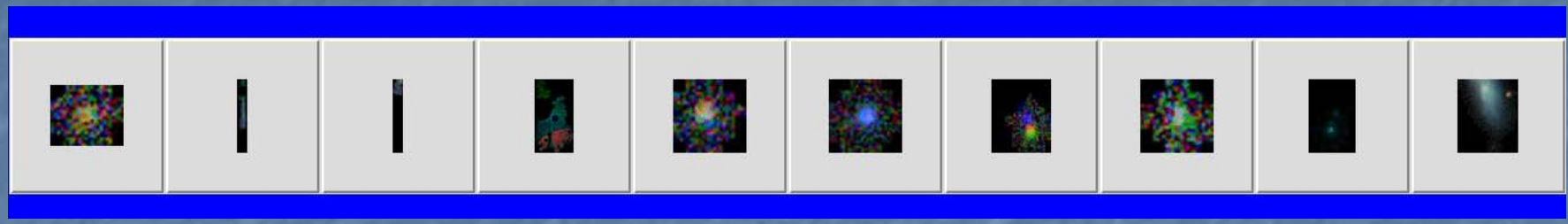
General Anomaly Finding

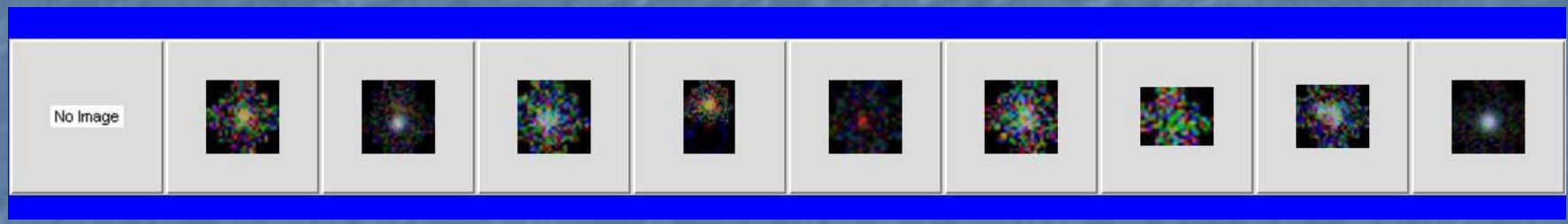
- Bayes Nets and Dependency Trees
  - Trades between linear (fast) and general (slow) correlations
  - Need to learn classifications










No Image

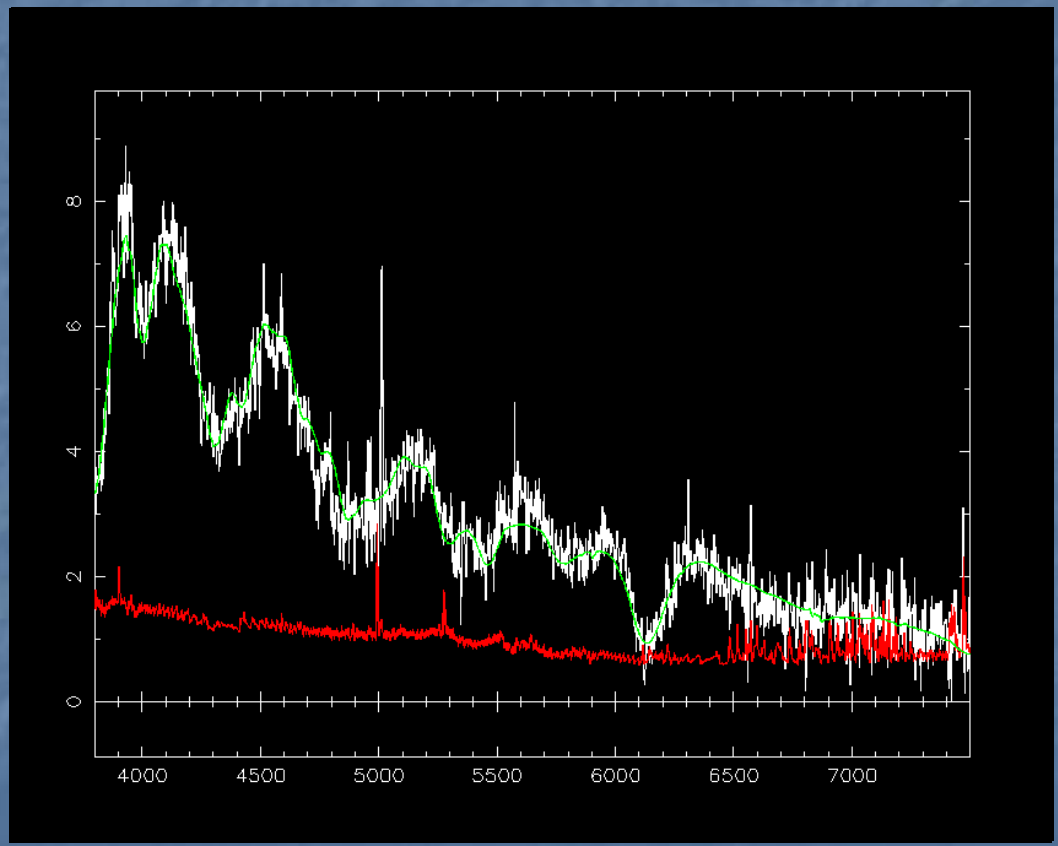

No Image

Variable sources: static and moving

- Searching for the unusual
  - Subtracting a galaxy to identify SNe
  - Variable seeing conditions
  - Variable astrometry (DCR)
  - Need to work at the S/N limits of the surveys not at the S/N=100
  - Need to process in real time and with few false positives

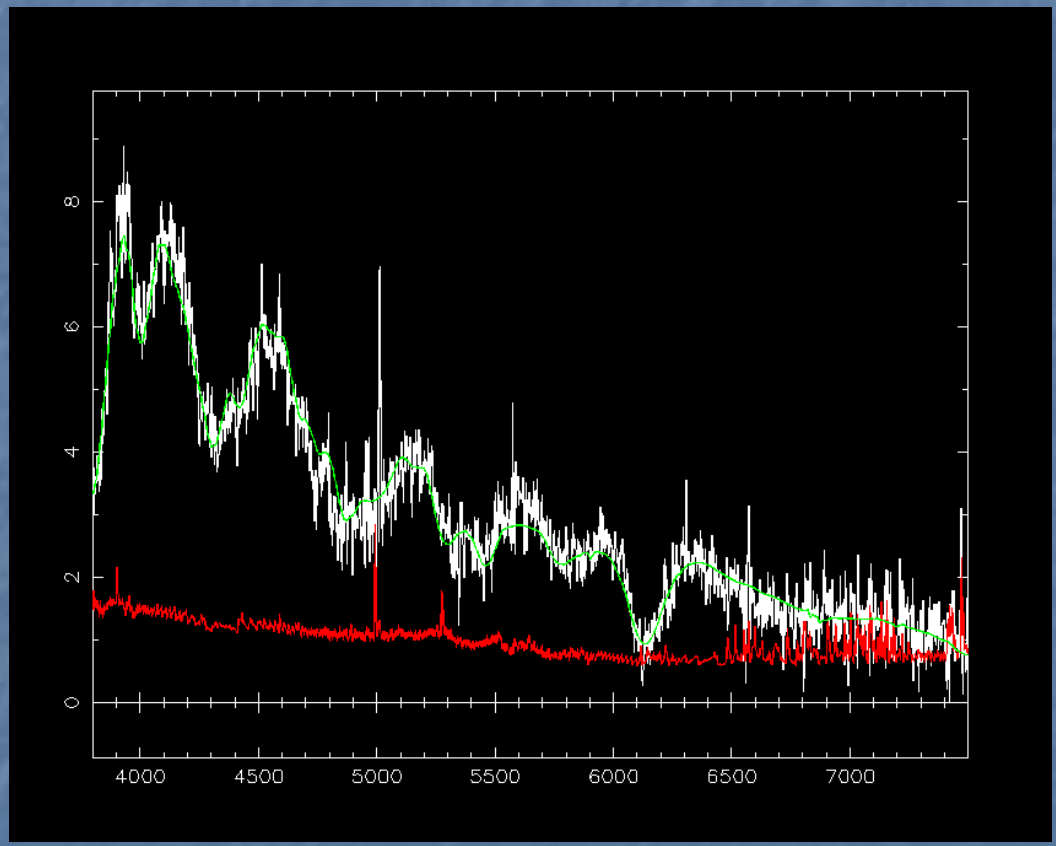
Correlated Features

- Times series as spectral data
  - Decomposing onto compact bases
  - Isolate variable sources (subtle and distinct)
- Principal Component Analysis
  - Correlations present in spectral (time) domain
  - Linear and non-linear relations
  - PCA and ICA necessary to extract these information



Variable Sources

- Isolating correlated signatures
  - Fits for given models
  - Extending to the unknown
  - Can we learn a signature
  - Can we define the statistics for the significance



Future Directions

- Era of the PB database
  - Multifrequency data (all-sky)
  - 400+ times steps
  - >10^9 sources (200 parameters per record)
- New services
  - Image matching and coaddition
  - Clustering statistics
  - Anomaly finding, classification
  - Compact descriptions of data
  - Wire formats for data
  - Accessibility and extensibility
  - Searching time domain for 10^{10} sources