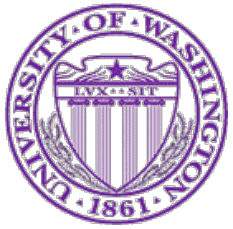


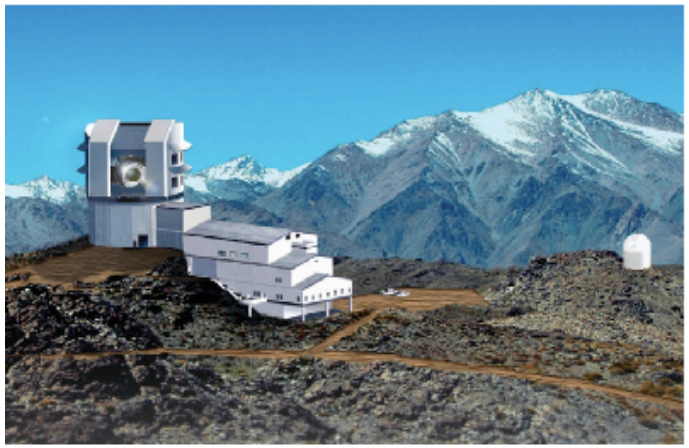
SCALING UP AND SCALING OUT

Andrew Connolly

University of Washington

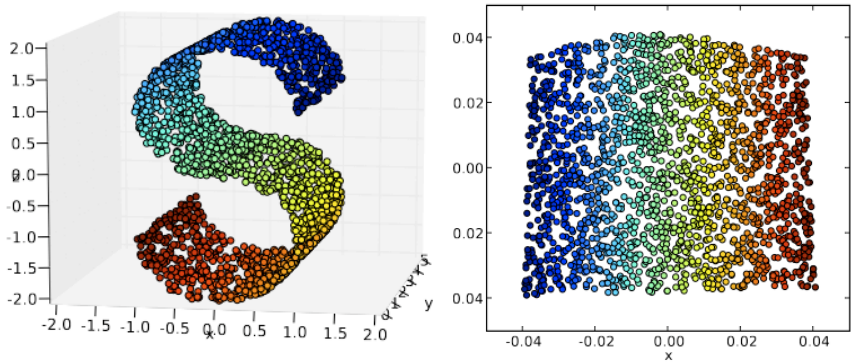
Scaling up to the Sky

- **Size of data (LSST)**
  - 30 TB of images per night
  - 10^{10} stars and galaxies
  - 10^{9} sources per night (10^{3} transients)
  - 1-10 PB in a database (catalogs)
- **Dimensions**
  - 100 attributes per source
  - Temporal information (1000 visits)
  - Variable sky (moving and transient)
  - Poorly defined basis functions for classification
  - Incomplete and noisy observations

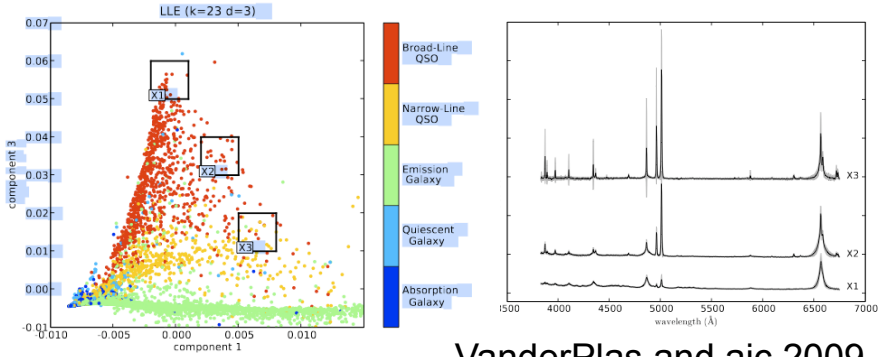


Looking for structure

- **Reducing dimensionality**
  - Global and local measures
  - PCA vs LLE
  - Local structure within 4000 dimensional space of 100K spectra
- **Learning structure**
  - Controlled by neighbors and projected dimensionality
  - Without feature extraction – outperforms SDSS pipelines
  - Learning the rules is slow



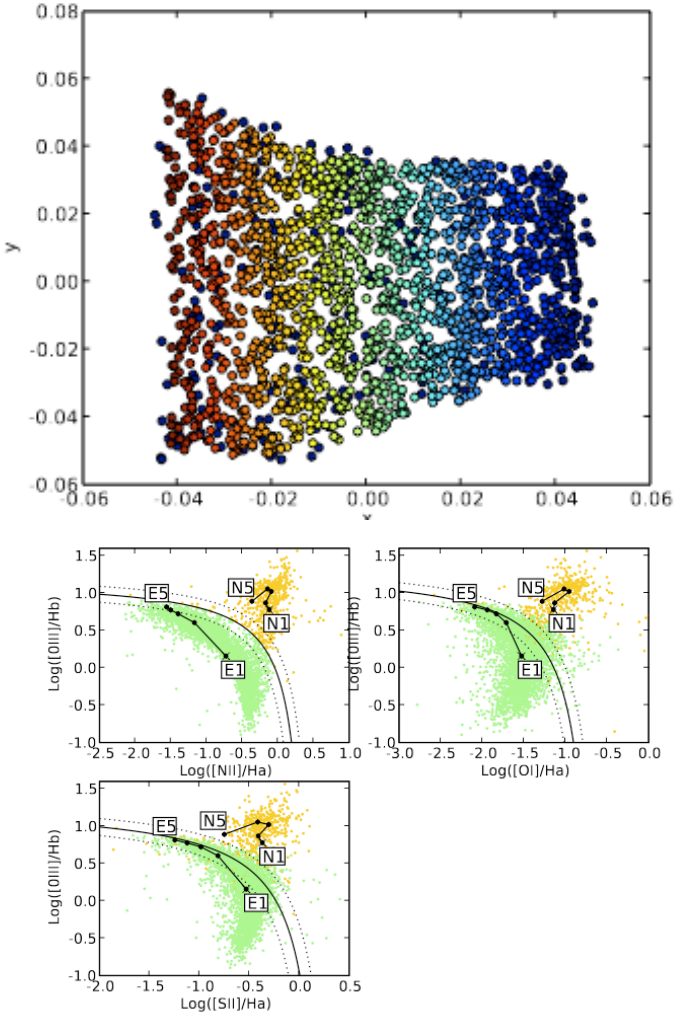
Roweis and Saul 2000



VanderPlas and ajc 2009

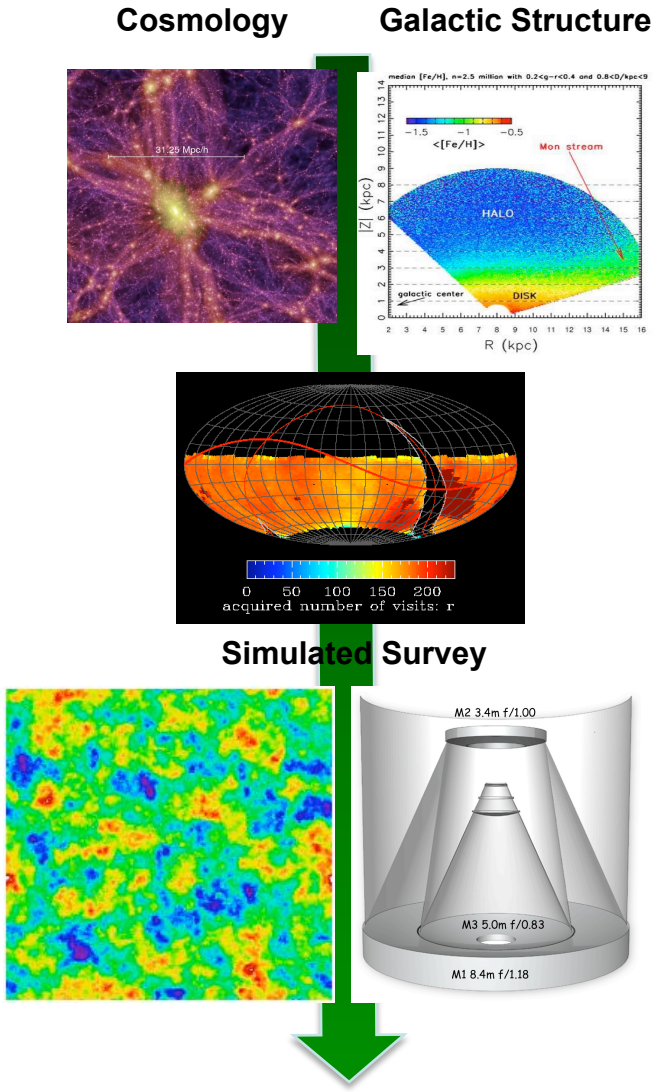
Cost of searching for structure

- **Slow aspects**
  - **Searching for neighbors**
    - Parallel/tree searches
  - **Presence of noise and missing data**
  - **Numbers of dimensions**
- **Sampling strategies**
  - **How many sources is enough?**
  - **Brute force vs stratified sampling vs physical sampling**
  - **Sample based on variance of the local structure**
  - **100-fold reduction in training sample**
  - **Many applications: classification, photometric redshift calibration, photometric calibration**



Scaling up simulations of the sky

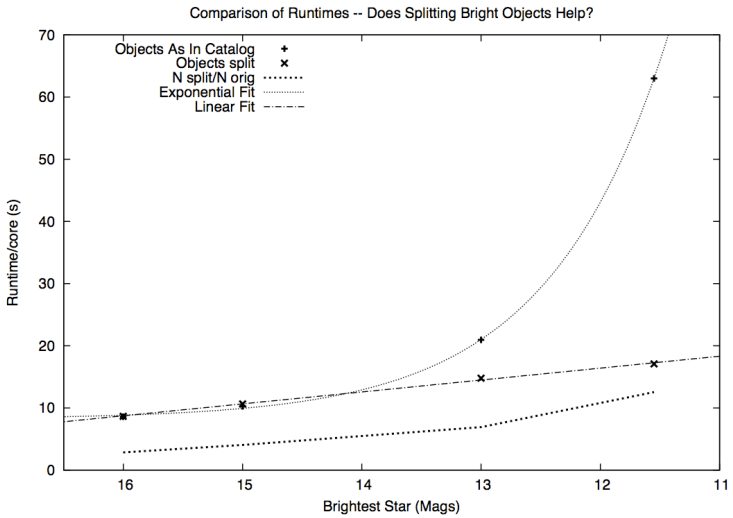
- **Simulating the LSST**
  - **One LSST focal plane**
    - 189 2Kx4K CCDs
    - 3.2 Gpixels (6.4 GB)
    - 10^6 Sources (r<24)
    - 15 seconds
  - **One simulated Focal plane**
    - 10^7 Sources (r<28)
    - Each source has a spectral energy distribution (λ effects)
    - 10^{11} photons
    - 2000 CPU hrs



Simulating the universe

Scaling out with MapReduce/Hadoop

- **Simply parallel applications**
  - Metric is wall clock not cpu clock
  - 1 CCD takes 10 hrs on a 1000 core cluster – so does 1000 CCDs
  - Mapreduce approach to improve balance (adaptable granularity)
  - Simulating a source at a time
- **Granularity of operations**
  - 10hrs to 20 mins (40 processors)
  - Many levels of parallelization: Focal plane, CCD, amplifier, source, photon
  - Trade off between overhead and cpu time



Future: simple scalable algorithms

- **The real life 80-20 split**
  - **Do I need all of my data I memory at the same time**
    - Large memory machines or message passing
  - **Most astronomers live in the 80% regime**
    - Simply parallel and simply scalable
    - Cosmology codes, npt statistics don't reside here
  - **Simple distributed processing paradigms work**
    - Focus on serial applications
    - Allow the analysis to grow with the data
- **Map reduce paradigm**
  - **Our bread and butter**
  - **Enables science not programming**
  - **Scales (?) to the next generation of surveys**

But what do we teach…..

- IDL
- IDL
- IDL
- Oh and a little bit of Python….