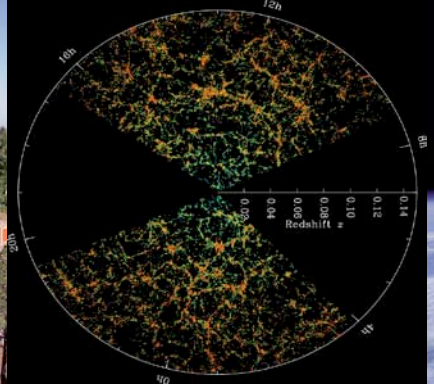
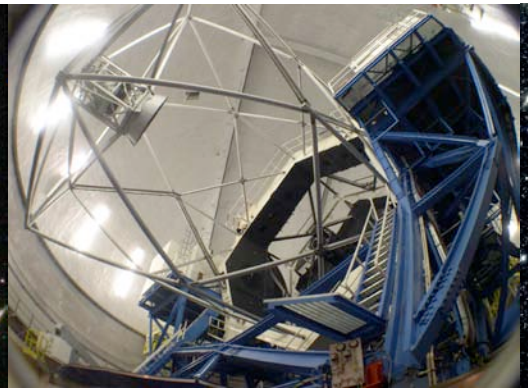


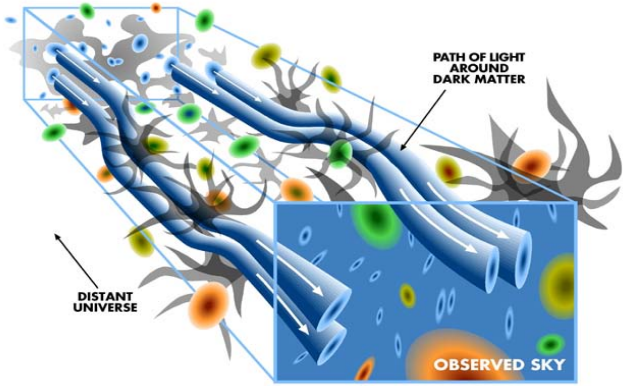
Complex Scientific Analytics in Astrophysics at Extreme Scale

Andrew Connolly
University of Washington



Science of Big Data Sets

- **Nature of the Universe**
  - **Big Questions**
    - Dark Energy, Dark Matter, New Physics
  - **Small Effects**
    - Large volumes and data
      - Petabytes (surveys and simulations)
      - Need to store, move, analyze
      - Data are complex, gappy, noisy
      - Computationally complex
    - Systematics are important
      - Poisson noise often doesn't dominate
      - Need approximation algorithms
  - **Large Projects, Small Research Teams**
    - Astronomy is collaborative in nature
    - Science with small distributed groups
    - Sharing information is critical



What is the Science We Want to Do?

- **Finding the unusual**
  - Billion sources a night
  - Nova, supernova, GRBs
  - Instantaneous discovery
- **Finding moving sources**
  - Asteroids and comets
  - Proper motions of stars
- **Mapping the Milky Way**
  - Tidal streams
  - Galactic structure
- **Dark energy and dark matter**
  - Gravitational lensing
  - Slight distortion in shape
  - Trace the nature of dark energy

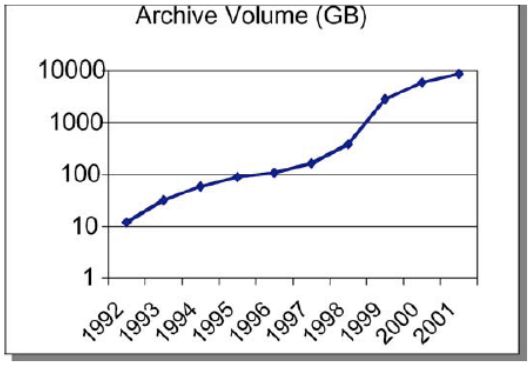
What are the Operations We Want to Do?

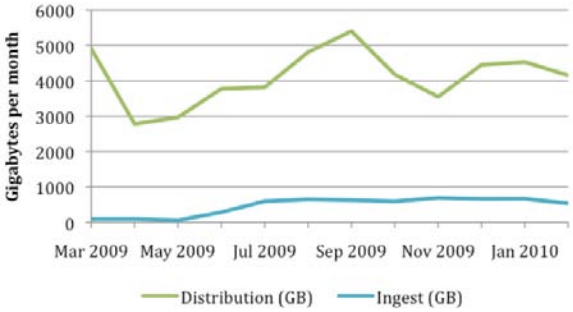
- **Finding the unusual**
  - Anomaly detection
  - Dimensionality reduction
  - Cross-matching data
- **Finding moving sources**
  - Tracking algorithms
  - Kalman filters
- **Mapping the Milky Way**
  - Density estimation
  - Clustering (n-tuples)
- **Dark energy and dark matter**
  - Computer vision
  - Weak Classifiers
  - High-D Model fitting

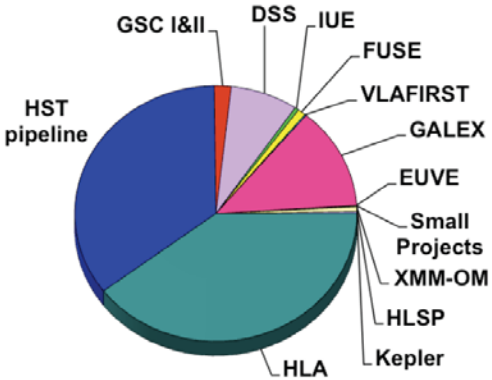
Not just volume also complexity

Data in Astrophysics

- **Archives**
  - **Heterogeneous**
    - Distributed across US and the world
    - Archive centers tend to be wavelength specific (HST, Chandra, XMM, GALEX, 2MASS, WISE)
  - **Image Data**
    - Standards defined for images (FITS)
    - Metadata for images critical but not always complete or well defined
  - **Catalog data**
    - Typically SQL databases, also flat files for small data sets
    - No standard for catalogs (native format is tabular and array based), no common vocabulary, mainly floats
    - APIs for access proposed, partially implemented

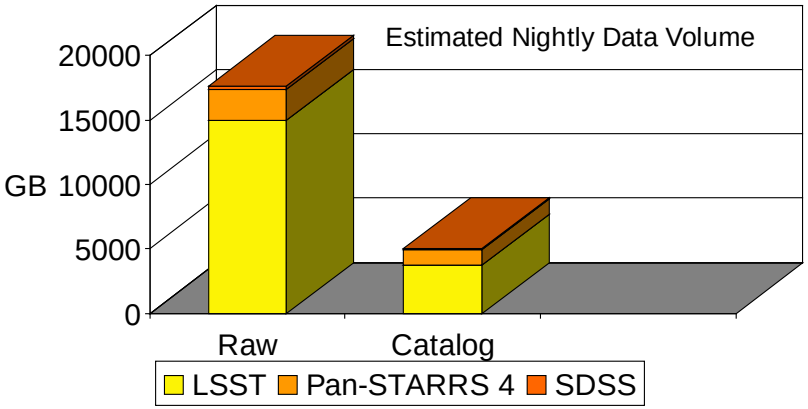






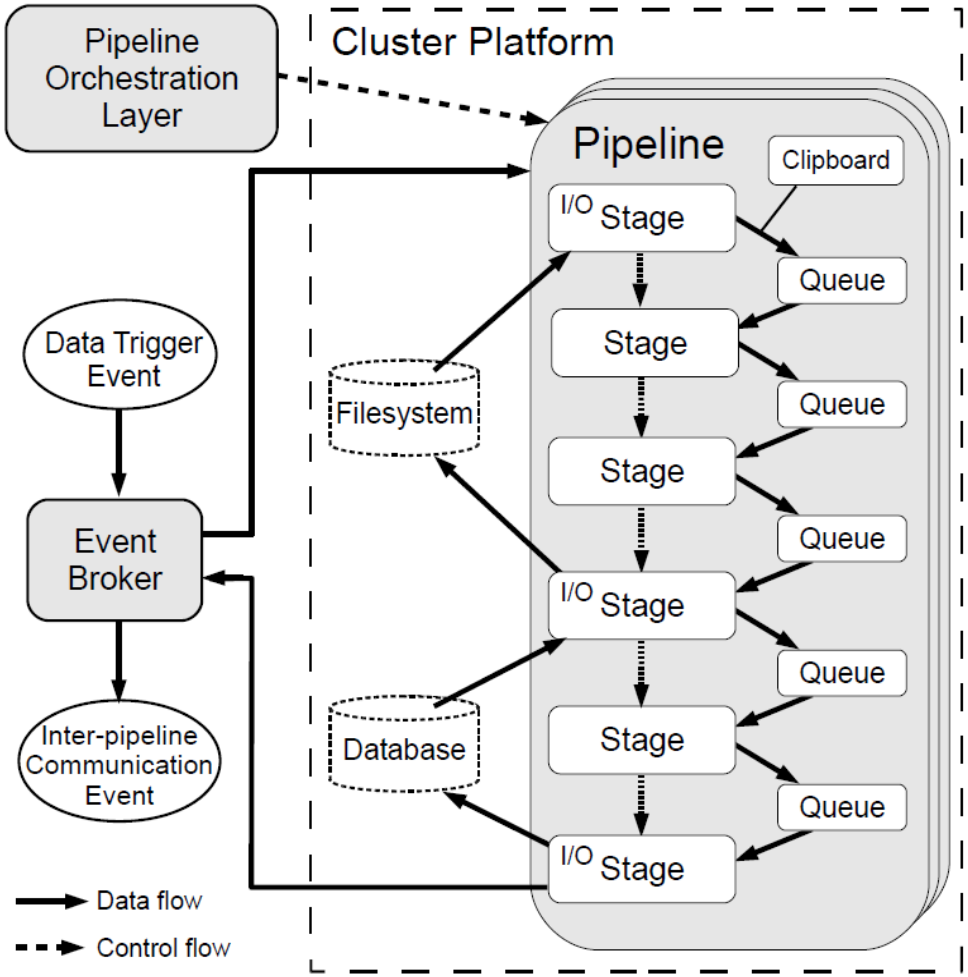
The Big Surveys

- **2000-2010**
  - **Sloan Digital Sky Survey (SDSS)**
    - 120 Mpixel camera, (0.08 PB in 10 years)
    - 300 Million unique sources (4 TB)
    - Typical access through SQLServer
- **2010-2014**
  - **PanSTARRS (PS1)**
    - 1.4 Gpixel camera (0.4 PB per year)
- **2017-2027**
  - **Large Synoptic Survey Telescope (LSST)**
    - 3.2 Gpixel camera (6 PB per year)
    - 1000 observations of every source
- **Simulations (the gorilla in the room)**
  - **TBs per run generated today**
  - **TBs per hour in the next 5 years**



Can't calculate everything a priori (can't just store the results)

- **Real-time processing pipelines**
  - Calibrate and transform images (pixel operations)
  - Real-time analysis (60s)
  - 5000 cores at base
- **Data Release pipelines**
  - Source detection and characterization (pixel operations)
  - Archive 10^8 detections per night
  - 20,000 cores at archive
- **Science Pipelines**
  - Pixel and catalog operations
  - Scans of 10s PB catalogs
  - Reanalysis of images
  - Cross matching of different data sets
  - 40 Tflops available

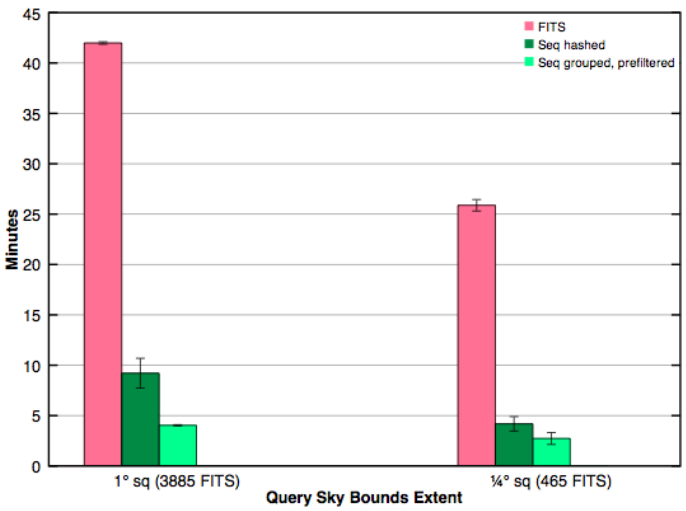


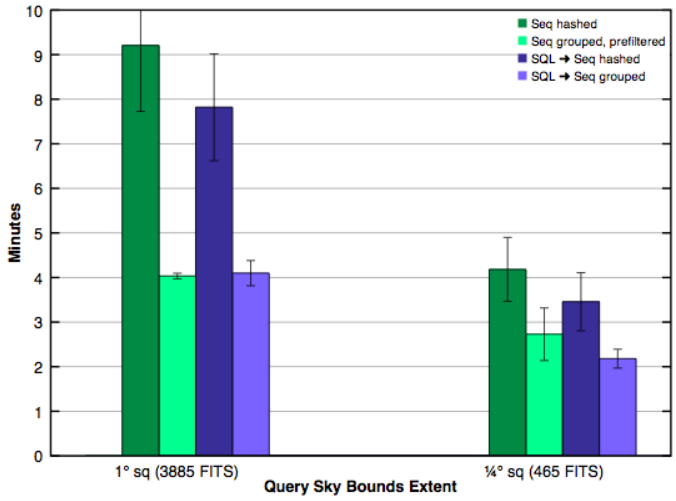
Science from Querying Data

- **Meet the scientist**
  - **Broad range of abilities and requirements**
  - **Mathematically sophisticated (but not necessarily computationally)**
  - **Good at scripting (IDL, Python)**
  - **Code is often throw away (but this is changing)**
  - **Good at learning new approaches (e.g. SQL, AWS)**
    - But needs to see fast returns if an early adopter
    - Community driven
    - Pretty tolerant…
- **Provenance in science from large surveys**
  - **Large data sets and analysis codes**
  - **Reproducibility is critical (validation and publishing)**

Analysis and Data in the Cloud

- **Cloud related operations a natural fit**
  - **Move analysis to data**
    - 80% of operations simply parallel
    - Large sweeps through the data
- **Details impact design**
  - **Textual vs binary data**
  - **Hadoop application to stack images**
    - Stack images (80+) to detect faint galaxies (register, transform, blur, add)
    - Data distributed not “ideal” fashion (many small files)
    - RPCs drive the clock time
    - Packaging of data to improves efficiency (sequence files, metadata database, storing where images reside on disk)





Astro Summary

- Broad range of data sets
  - Power-law of sizes with big surveys in the tail
- Large data sets homogeneous (1000-fold scaling over 20 years)
  - Surveys: PB next year, 100s PB in 10 years
  - Simulations: PB this year
- Lots of defined formats for storage of images (but not catalogs)
- Storage is in a number of forms
  - Images (raw and processed) on file systems
  - Catalogs (read only to scientist) on databases
  - Scientist defined data sets (to be joined with existing catalogs)
- Applications are complex (supervised and unsupervised machine-learning) and broad
  - Hand tuning individual applications or general scalable frameworks
  - Cloud-based applications are arriving
- Silver bullets tend to get tarnished in the light of day