QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Grist: Grid Data Mining for Astronomy

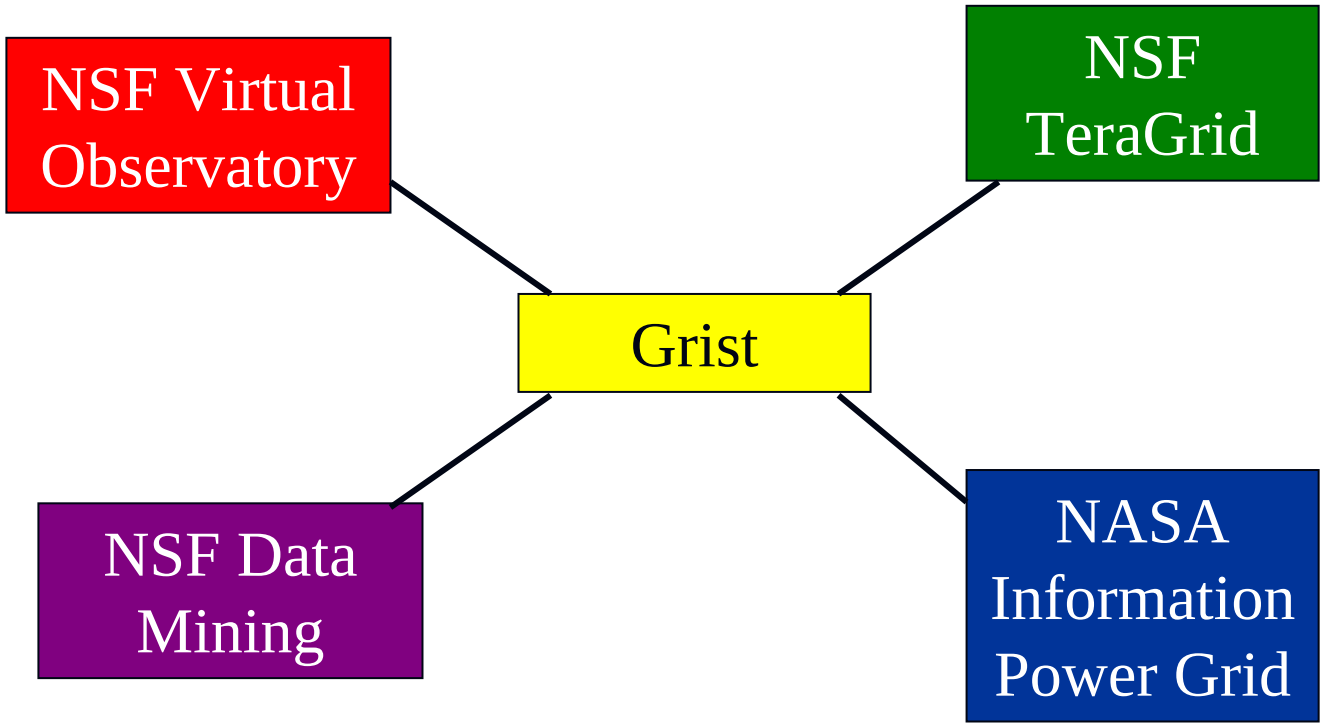
Joseph Jacob, Daniel Katz, Craig Miller, and Harshpreet Walia (JPL)
Roy Williams, Matthew Graham, and Michael Aivazis (Caltech CACR)
George Djorgovski and Ashish Mahabal (Caltech Astronomy)
Robert Nichol and Dan Vandenberk (CMU)
Jogesh Babu (Penn State)

http://grist.caltech.edu/

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Pasadena, CA, July 12-15, 2004

Grist Relation to Other Projects



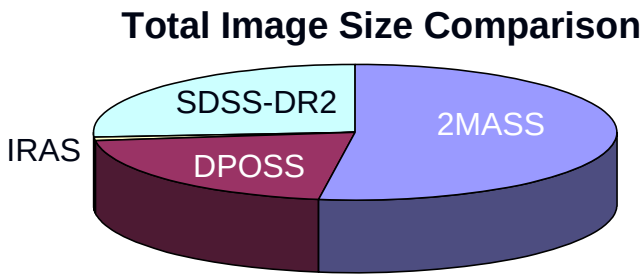
Motivation

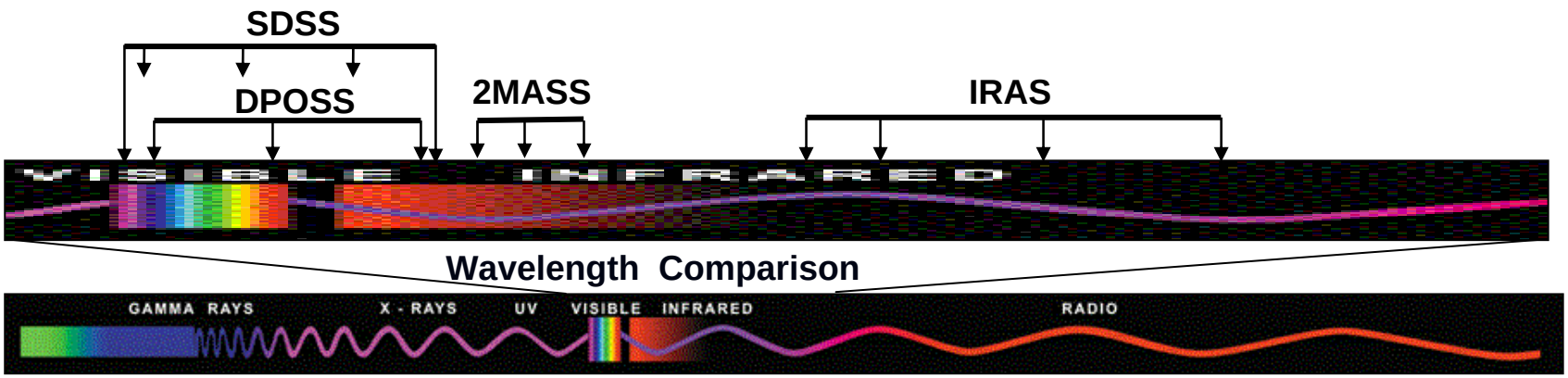
- Massive, complex, distributed datasets
  - Exponential growth in data volume; In astronomy the data volume now doubles every 18 months
  - Multi-terabyte surveys; Petabytes on horizon
  - Need to support multi-wavelength science
- Distributed data, computers, and expertise
  - Need to enable domain experts to deploy services
  - Need framework for interoperability

grist
GRID DATA MINING FOR ASTRONOMY

Selected Image Archives

IRAS	1 GB	1 arcmin	All Sky	4 Infrared Bands
DPOSS	4 TB	1 arcsec	All Northern Sky	1 Near-IR, 2 Visible Bands
2MASS	10 TB	1 arcsec	All Sky	3 Near-Infrared Bands
SDSS-DR2	5 TB	0.4 arcsec	3,324 square degrees (16% of Northern Sky)	1 Near-IR, 4 Visible Bands





grist
GRID DATA MINING FOR ASTRONOMY

Grist Objectives

Building a Grid and Web-services architecture for astronomical image processing and data mining.

- Establish a service framework for astronomy
  - Comply with grid and web services standards
  - Comply with NVO standards
- Deploy a collection of useful algorithms as services within this framework
  - Data access, data mining, statistics, visualization, utilities
- Organize these services into a workflow
  - Controllable from a remote graphical user interface
  - Virtual data: pre-computed vs. dynamically-computed data products
- Science
  - Palomar-Quest exploration of time-variable sky to search for new classes of transients
  - Quasar search
- Outreach
  - “Hyperatlas” federation of multi-wavelength imagery: Quest, DPOSS, 2MASS, SDSS, FIRST, etc.
  - Multi-wavelength images served via web portal

Approach

- Building services that are useful to astronomers
- Implementing NVO protocols on TeraGrid
- Processing and exposing data from a real sky survey (Palomar-Quest)
- Workflow of SOAP-based Grid services with GUI manager

Broader Impact

- Connecting the Grid and Astronomy communities
- Federation of big data in science through distributed services
- Managing a workflow of services

Why Grid Services?

The Old Way: Developer sells or gives away software to clients, who download, port, compile and run it on their own machines.

Grid Services provide more flexibility. Components can:

- Remain on a server controlled by the authors, so the software is always the latest version.
- Remain on a server close to the data source for efficiency.
- Be on a machine owned by the client, so the client controls level of service for themselves.

Grid Paradigms Explored in Grist

- Services replace programs
- Separation of control from data flow
- Separation of metadata from data
- Database records replace files
- Streams replace files

grist
GRID DATA MINING FOR ASTRONOMY

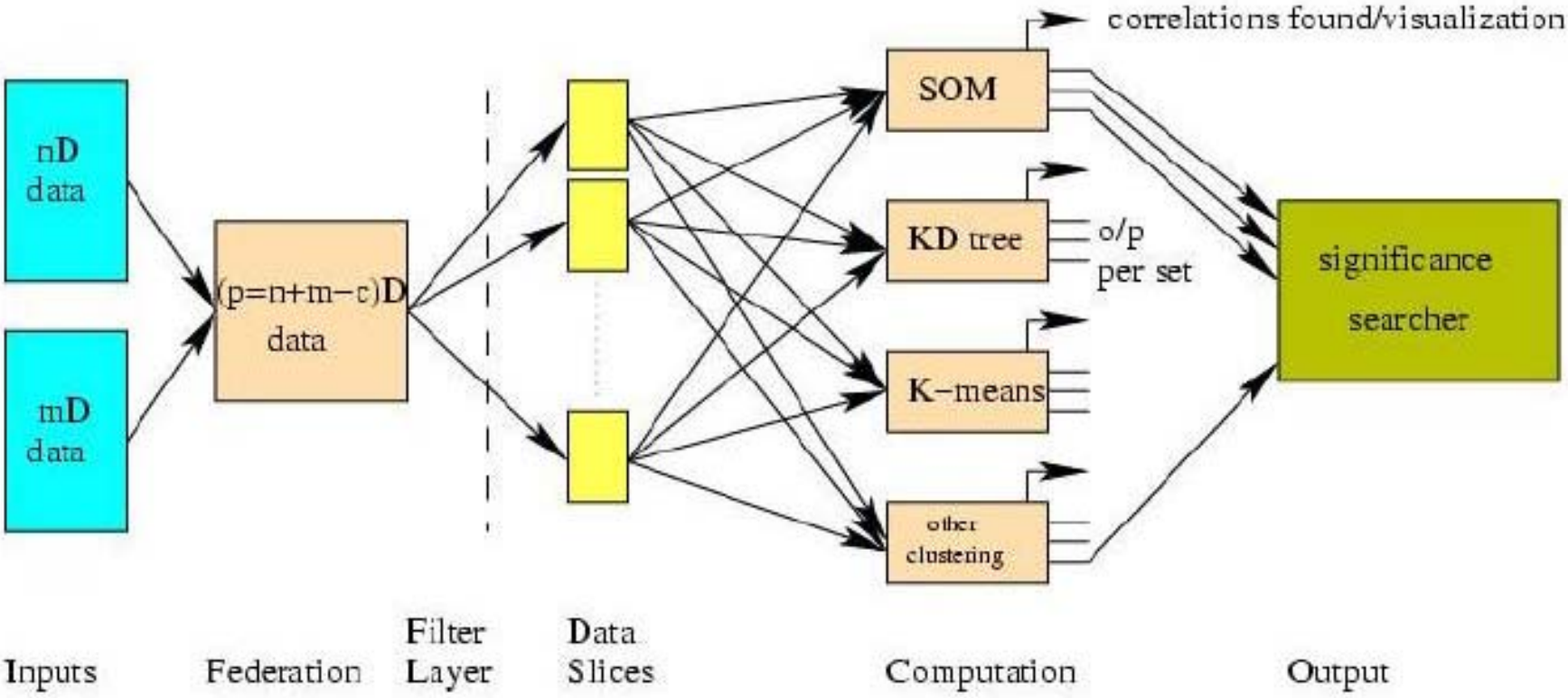
Grist Services

- Data Access (via NVO)
- Data Federation (Images and Catalogs)
- Data Mining (K-Means Clustering, PCA, SOM, anomaly detection, search, etc.)
- Source Extraction (Sextractor)
- Image Subsetting
- Image Mosaicking (yourSky, Montage, SWarp)
- Atlasmaker (Virtual Data)
- WCS transformations (xy2sky, sky2xy)
- Density Estimation (KDE)
- Statistics (R - VOStatistics)
- Utilities (Catalog and image manipulation, etc.)
- Visualization (Scatter plot, etc.)
  - Computed on client
  - Computed on server; image sent back to client

grist
GRID DATA MINING FOR ASTRONOMY

Example Workflow Scenario

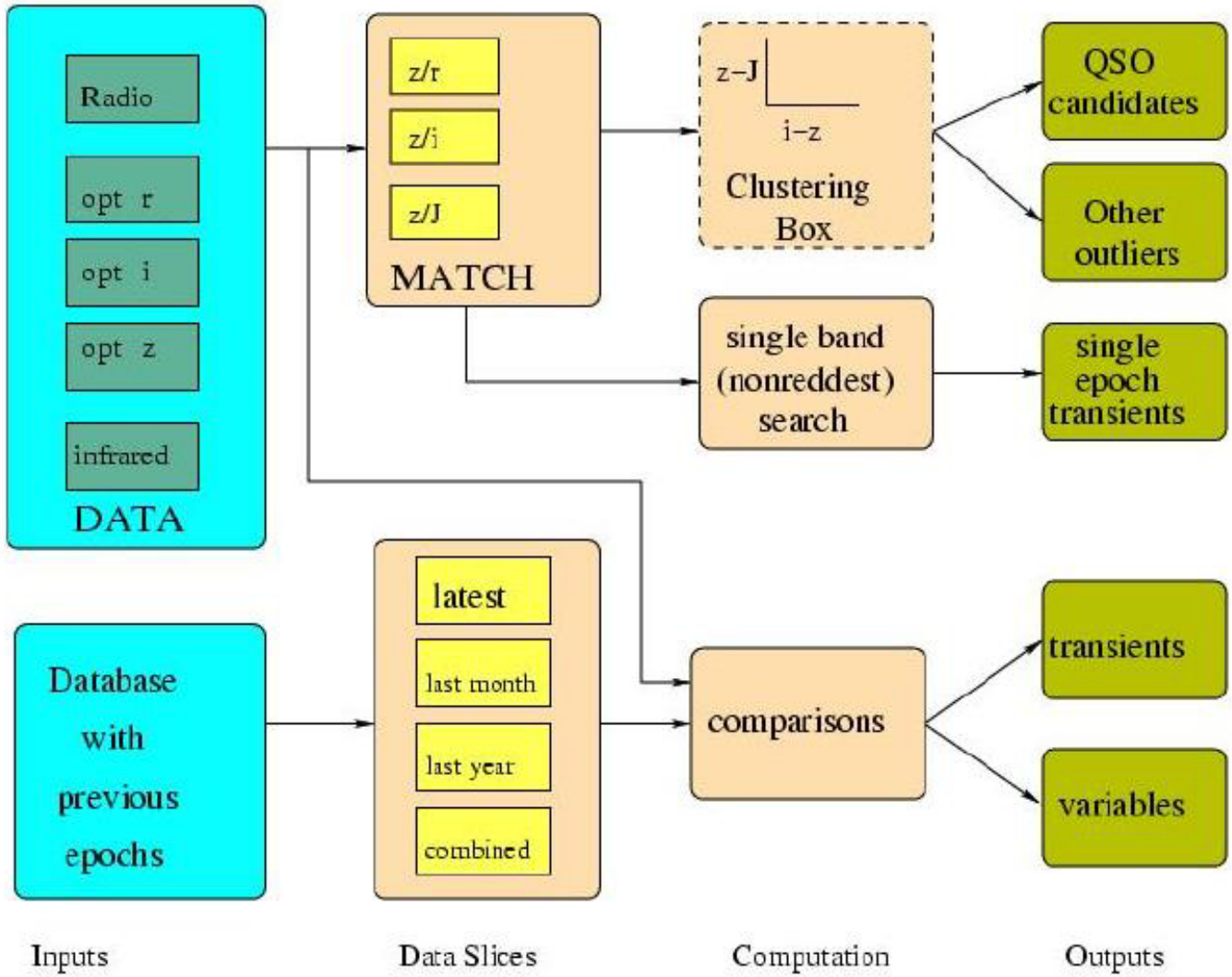
Dimensionality Reduction and Subsetting



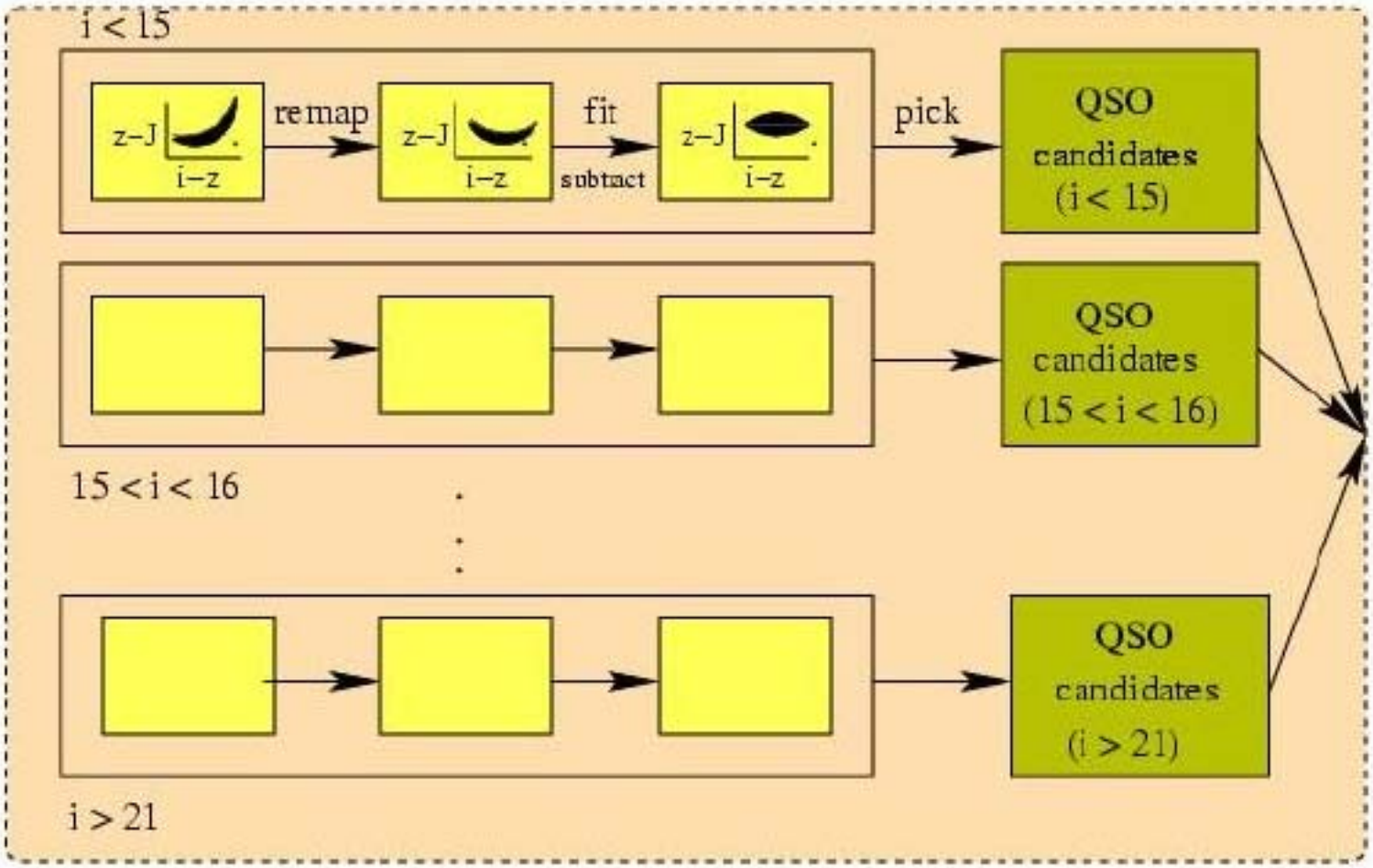
~billion sources, ~hundred dimensions

Another Example Workflow Scenario

Quasar and Transient Search

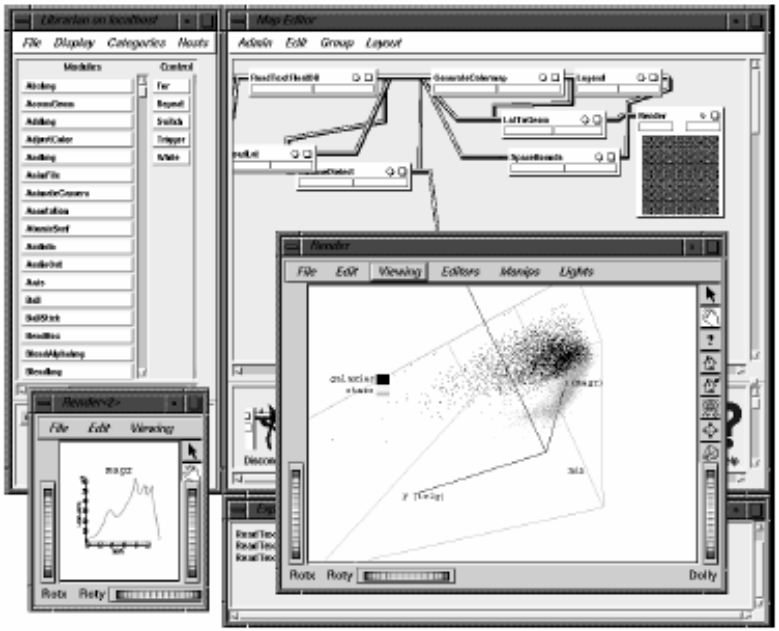


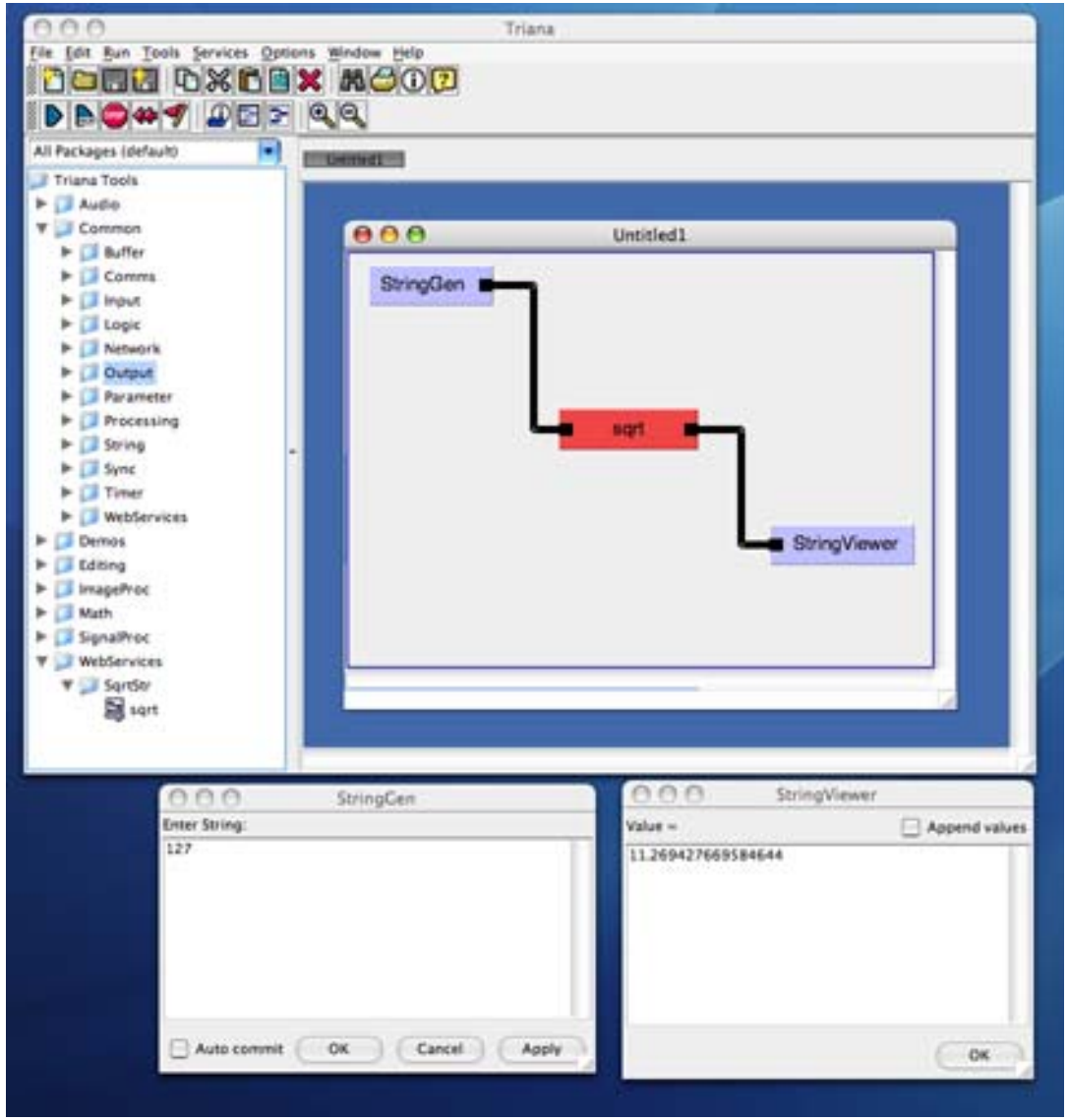
Quasar and Transient Search (cont.)



grist
GRID DATA MINING FOR ASTRONOMY

Workflow GUI Illustrations

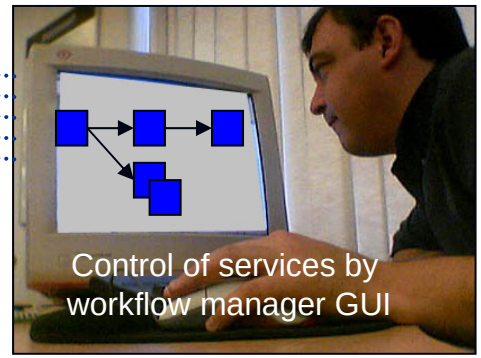




grist
GRID DATA MINING FOR ASTRONOMY

Service Composition

Palomar-Quest Survey
Time-domain astronomy
50 Gbyte every clear night



NCSA

Yale

TERAGRID

Caltech

NVO

Data exposure
through NVO spigots

Data Mining
- Fast transient search
- Stacked multi-temporal images for faint sources
- Density estimation

Astronomy knowledge
- z >4 quasars
- Transients

Source detection

Resampling

Scaling

Compression

Browse, Education

virtualsky.org

Federated images

Sloan (optical)

2MASS
(infrared)

fusion

NVO Standards

- National Virtual Observatory (NVO) defines standards for astronomical data
  - VOTable for catalogs
  - Datacube for binary data
    - 1D: time series, spectra
    - 2D: images, frequency-time spectra
    - 3D+: volume (voxel) datasets, hyper-spectral images.
- NVO defines standards for serving data
  - Cone Search
  - Simple Image Access Protocol (SIAP)

Grid Implementation

- Grid services standards are evolving. Fast!
  - Infrastructure based on Globus Toolkit (GT2)
    - Enables secure (user authenticated on all required resources with a single “certificate”) remote job execution and file transfers.
  - Open Grid Services Architecture (OGSA/OGSI) combines web services and globus (GT3)
  - WSRF: WS-Resource Framework (GT4)
- Grid Services converging with Web Services Standards
  - XML: eXtensible Markup Language
  - SOAP: Simple Object Access Protocol
  - WSDL: Web Services Description Languages
- Therefore, as a start, Grist is implementing SOAP web services (Tomcat, Axis, .NET), while we track the new standards

grist
GRID DATA MINING FOR ASTRONOMY

Grist Technology Progression

- Standalone service
- Service factory - single service
- Control a single service factory from workflow manager
  - Start, stop, query state, modify state, restart
- Service factory - multiple services
- Chaining multiple service factories
- Control multiple service workflow from workflow manager
  - Handshaking mechanism for when services are connected in workflow manager (confirm services are compatible, exchange service handles, etc.)
- Service roaming

Open Issues in Grist (1)

- Stateful Services
  - Client starts a service, goes away, comes back later
  - Where to store state?
    - Distributed: On client (cookies?)
    - Centralized: On server
- Service Lifetime
  - Client starts a service then goes away and never returns
- Receiving Status
  - Client pull
  - Server push
    - But what if client goes away?
- Maintaining State
  - On client
  - On server

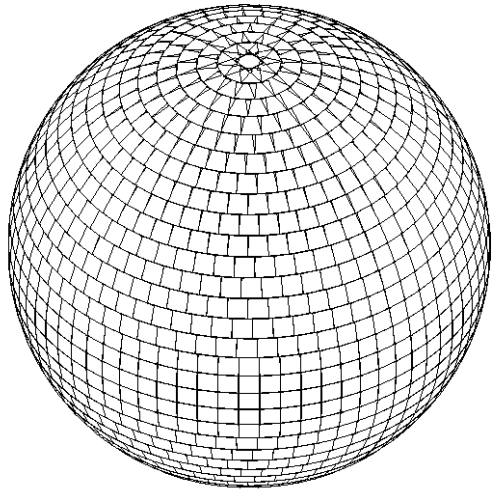
Open Issues in Grist (2)

- Chaining Services
  - Controlled by central master (the client)
  - No central control once chain is set up
- Service Roaming
- Protocols for shipping large datasets (memory limitations, etc)
- Authentication

grist
GRID DATA MINING FOR ASTRONOMY

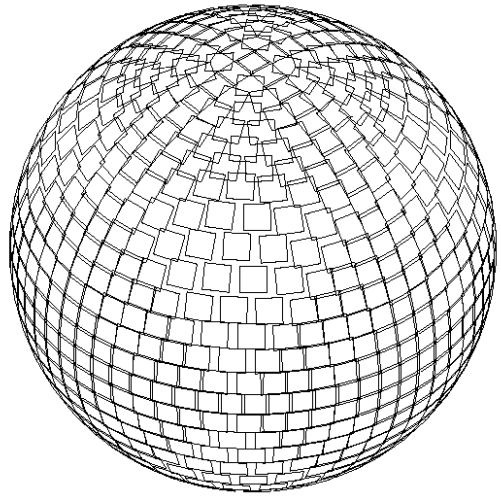
“Hyperatlas” Partnership

Collaboration between Caltech, SDSC, JPL, and the Astronomy domain experts for each survey.



TM-5 atlas

4.869147607046481 degree chart width



HV-4 atlas

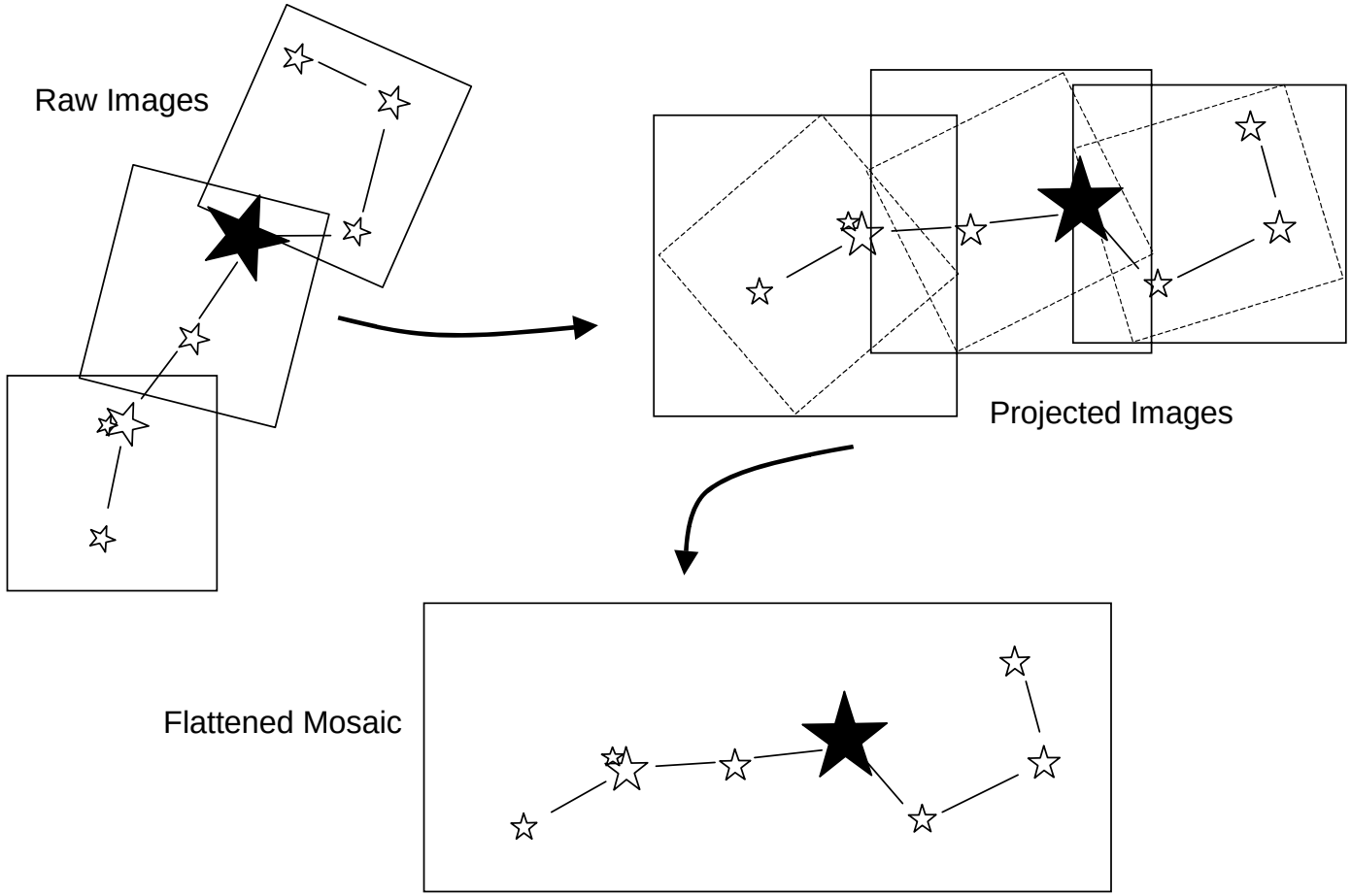
6.340943507916159 degree chart width

Objectives:

- Agree on a standard layout and grid for image plates to enable multi-wavelength science.
- Share the work and share the resulting image plates.
- Involve the science community to ensure high quality plates are produced.

Image Reprojection and Mosaicking

FITS format encapsulates the image data with keyword-value pairs that describe the image and specify how to map pixels to the sky



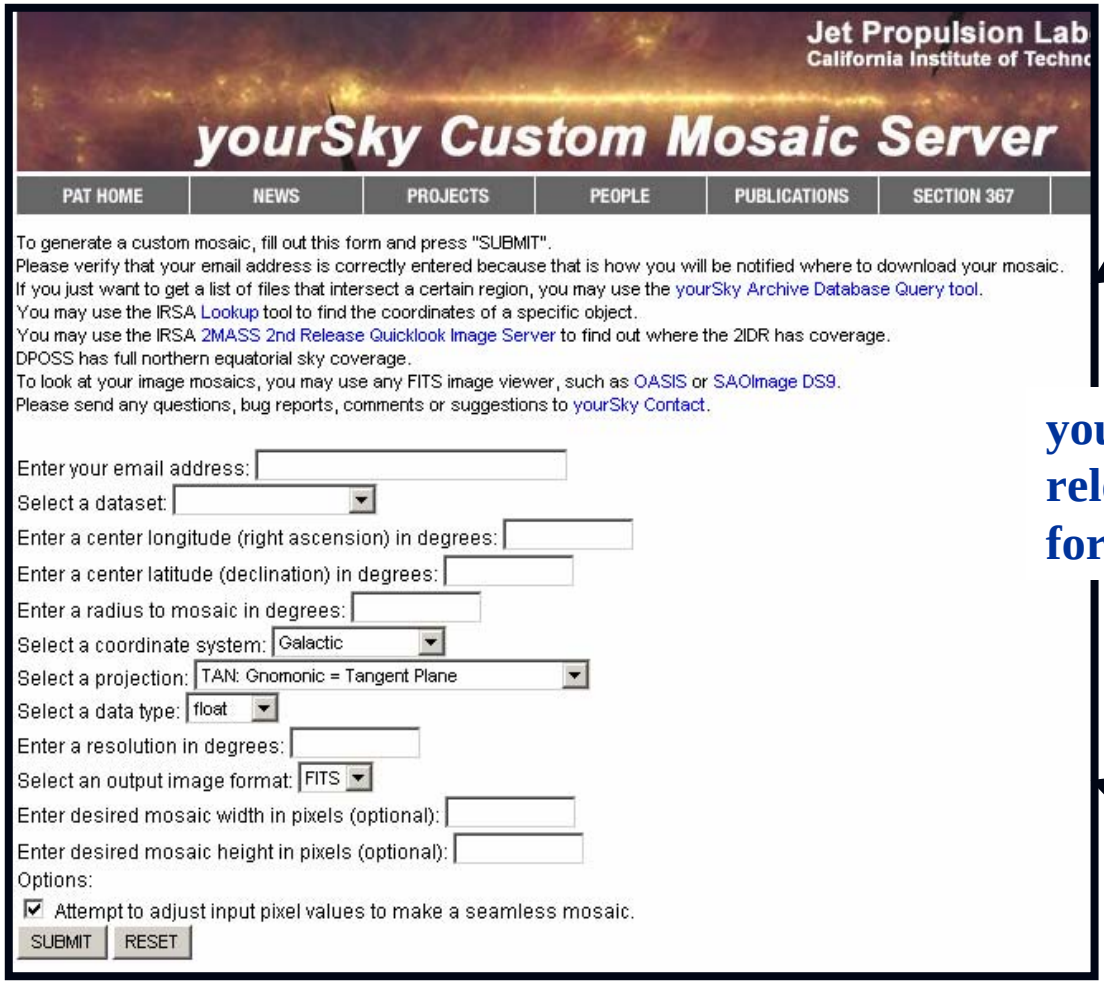
Science Drivers for Mosaics

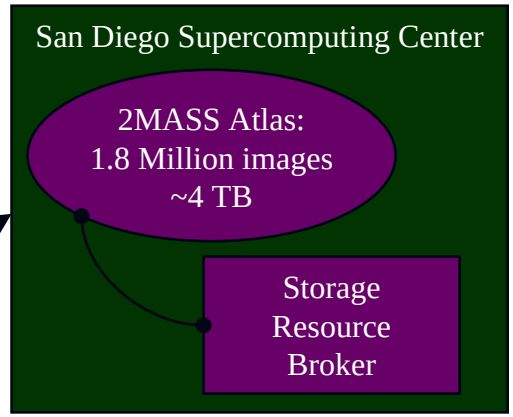
- Many important astrophysics questions involve studying regions that are at least a few degrees across.
  - Need high, uniform spatial resolution
  - BUT cameras give high resolution or wide area but not both => need mosaics
  - required for research and planning
- Mosaics can reveal new structures & open new lines of research
- Star formation regions, clusters of galaxies must be studied on much larger scales to reveal structure and dynamics
- Mosaicking multiple surveys to the same grid – **image federation** – required to effectively search for faint, unusual objects, transients, or unknown objects with unusual spectrum.

grist
GRID DATA MINING FOR ASTRONOMY

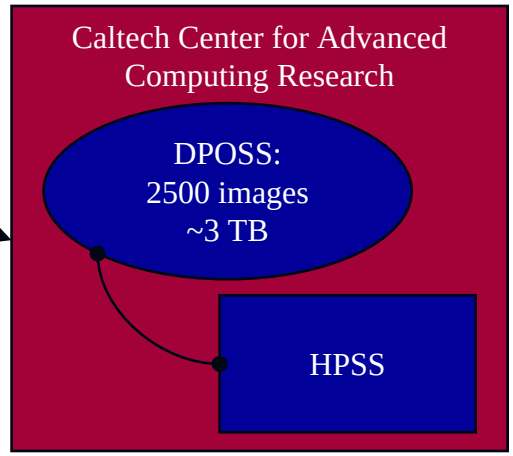
yourSky Custom Mosaic Portal

http://yourSky.jpl.nasa.gov/





yourSky can access all of the publicly released DPOSS and 2MASS images for custom mosaic construction.



grist
GRID DATA MINING FOR ASTRONOMY

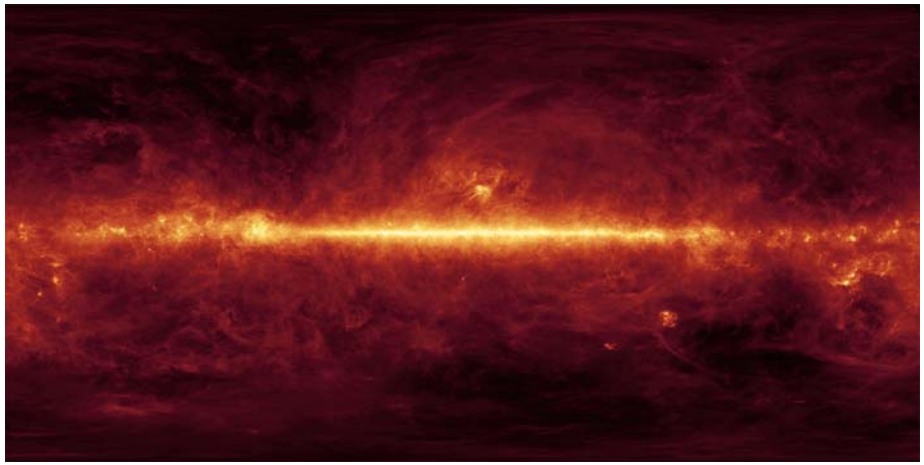
Montage: Science Quality Mosaics

http://montage.ipac.caltech.edu/

- Delivers custom, science grade image mosaics
  - User specifies projection, coordinates, spatial sampling, mosaic size, image rotation
  - Preserve astrometry & flux
  - Background modeled and matched across images
- Modular “toolbox” design
  - Loosely-coupled engines for Image Reprojection, Background Matching, Co-addition
    - Control testing and maintenance costs
    - Flexibility; e.g custom background algorithm; use as a reprojection and co-registration engine
  - Implemented in ANSI C for portability
- Public service will be deployed on the *TeraGrid*
  - Order mosaics through web portal

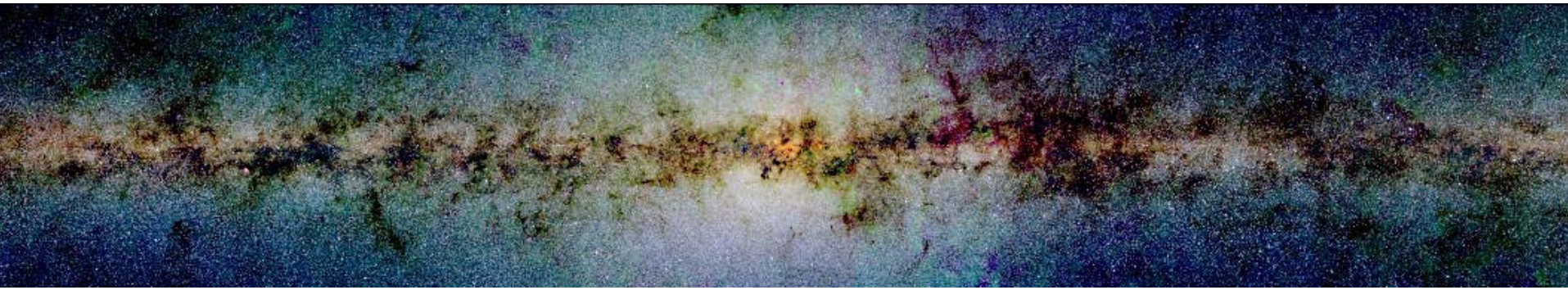
grist
GRID DATA MINING FOR ASTRONOMY

Sample Montage Mosaics



100 µm sky; aggregation of COBE and IRAS maps (Schlegel, Finkbeiner and Davis, 1998)

- 360 x 180 degrees; CAR projection



2MASS 3-color mosaic of galactic plane

- 44 x 8 degrees; 36.5 GB per band; CAR projection
- 158,400 x 28,800 pixels; covers 0.8% of the sky
- 4 hours wall clock time on cluster of 4 x 1.4-GHz Linux boxes

Summary

- Grist is architecting a framework for astronomical grid services
- Services for data access, mining, federation, mosaicking, statistics, and visualization
- Track evolving Grid and NVO standards
- http://grist.caltech.edu/