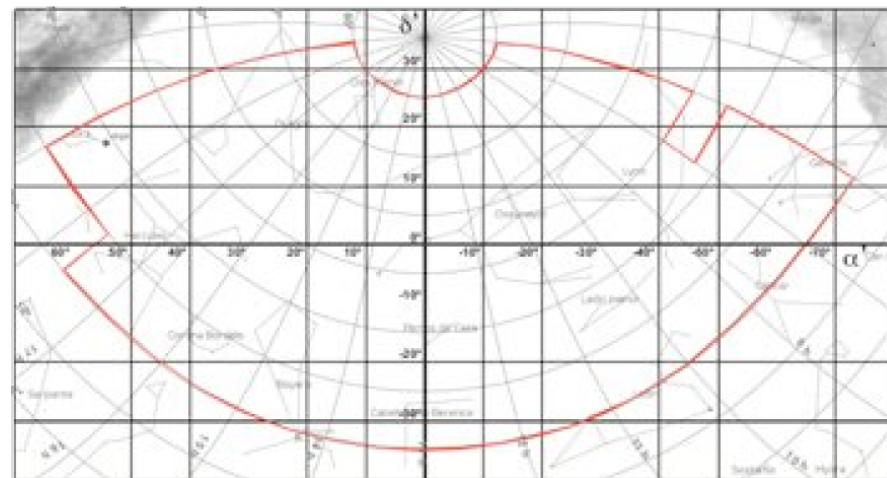
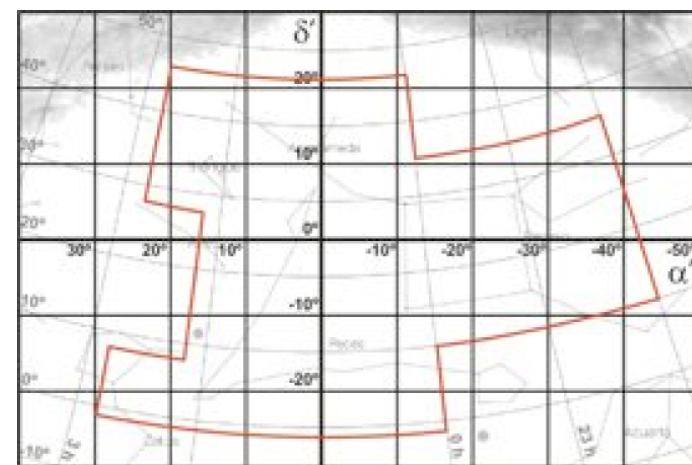
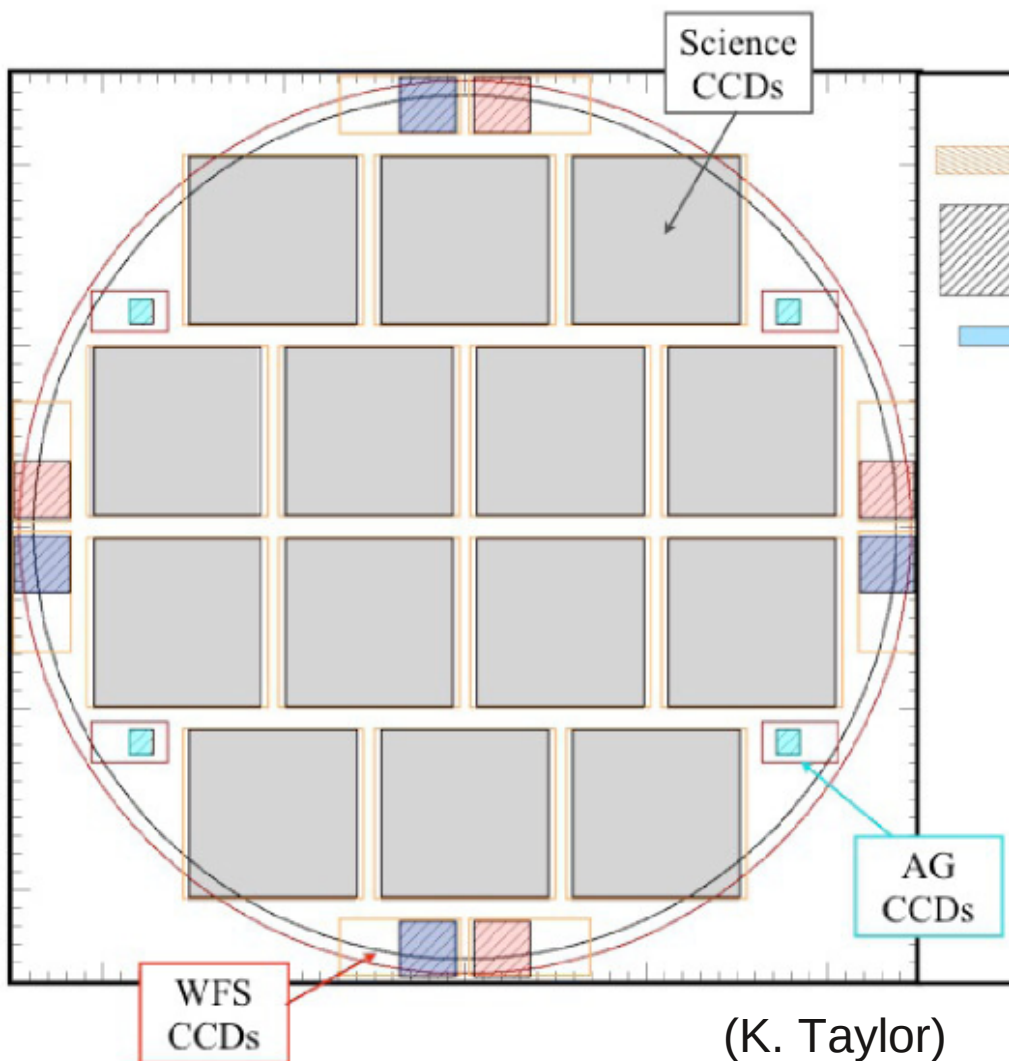


Software challenges in the implementation of large surveys: the case of J-PAS

Paulo Penteado - IAG/USP

pp.penteado@gmail.com

http://www.ppenteado.net/ast/pp_lsst_201204.pdf



(A. Fernández-Soto)

Outline

J-PAS overview

The problems:

- Going from survey strategy to taking daily observations.
- Going from daily observations to scientific data products.
- Going from scientific data products to scientific results.

Survey execution:

- Strategy / Scheduler
- Observatory control

Data products:

- Reduction Pipeline
- Generation of data products
- Archival, access and visualization

pp.penteado@gmail.com

http://www.ppenteado.net/ast/pp_lsst_201204.pdf

J-PAS overview

Javalambre PAU (Physics of the Accelerating Universe) Astrophysical Survey

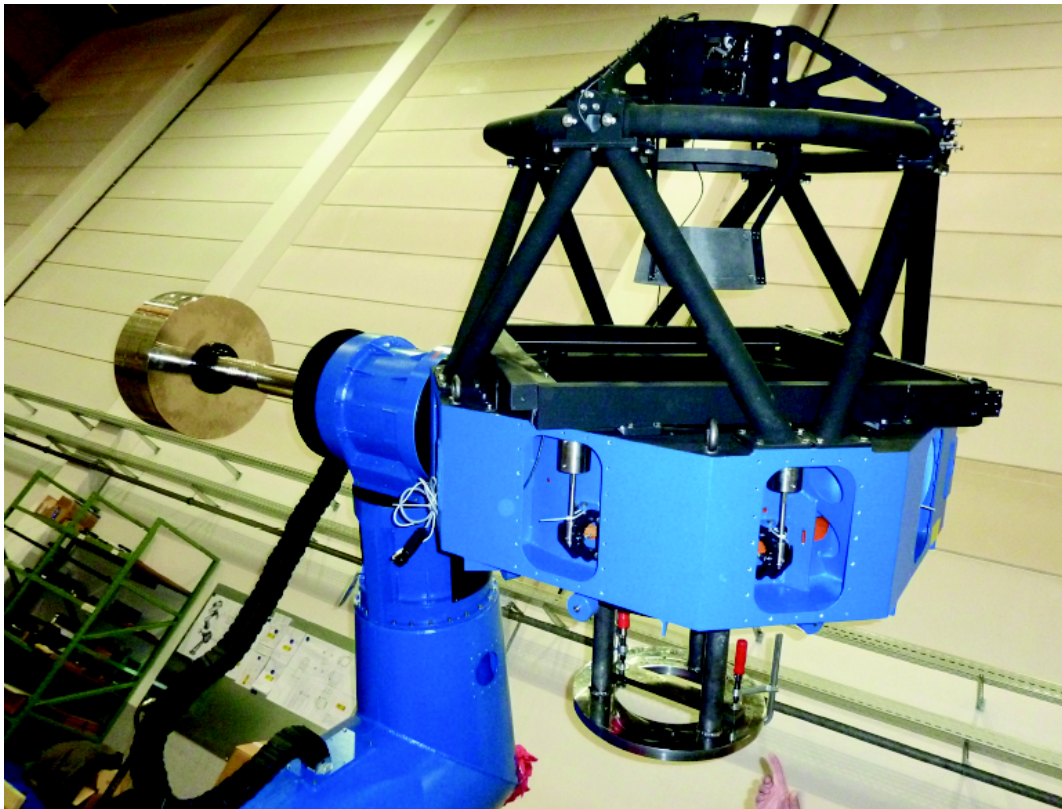
Two telescopes – T80 (80 cm) and T250 (2.5 m):

- 8000 $^{\circ}2$ photometric survey.
- 54-point spectrum over 3500-9000 Å
- redshift precision $\sigma_z \sim 0.003(1 + z)$ for over 14 million galaxies to measure BAOs (Baryon Acoustic Oscillations) to constrain cosmological parameters.
- High spatial resolution, low spectral resolution of a large area of the sky; useful for cosmology, study of galaxy clusters, stellar populations (MW and nearby galaxies), variable objects (supernovae, variable stars, Solar System).

J-PAS overview

T80 Telescope:

- 80 cm
- 1-CCD camera
- 2° FoV,
- 81 Mpixel
- 12 filters

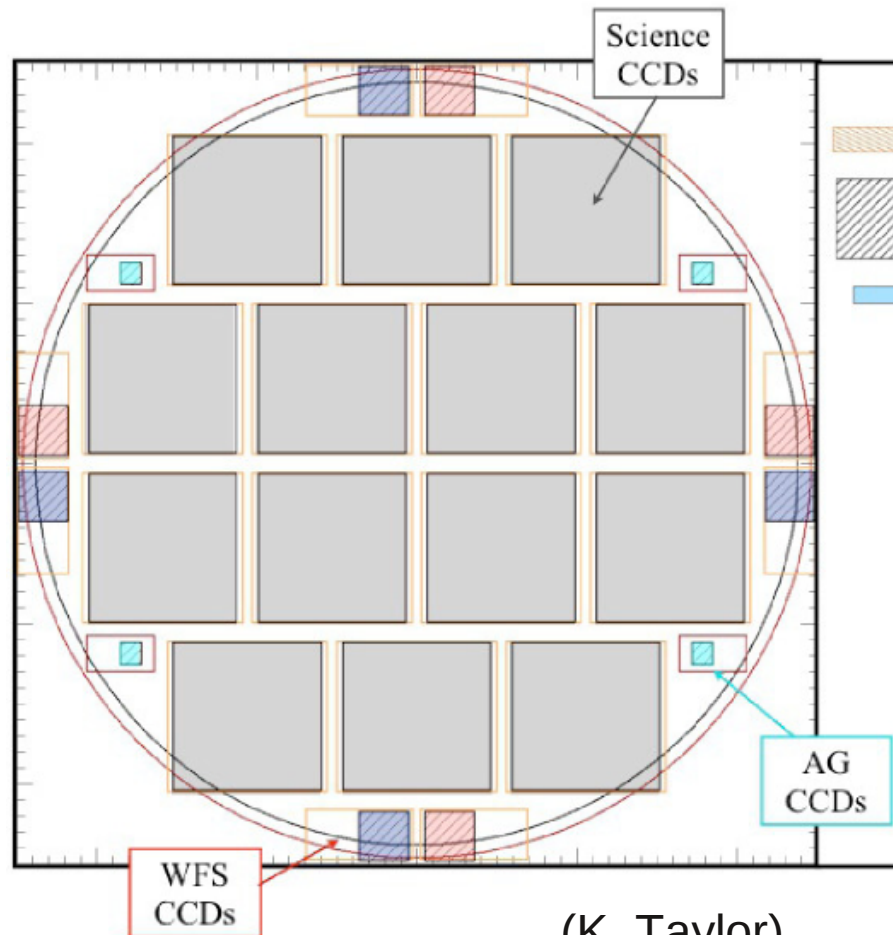


(J. Cenarro)

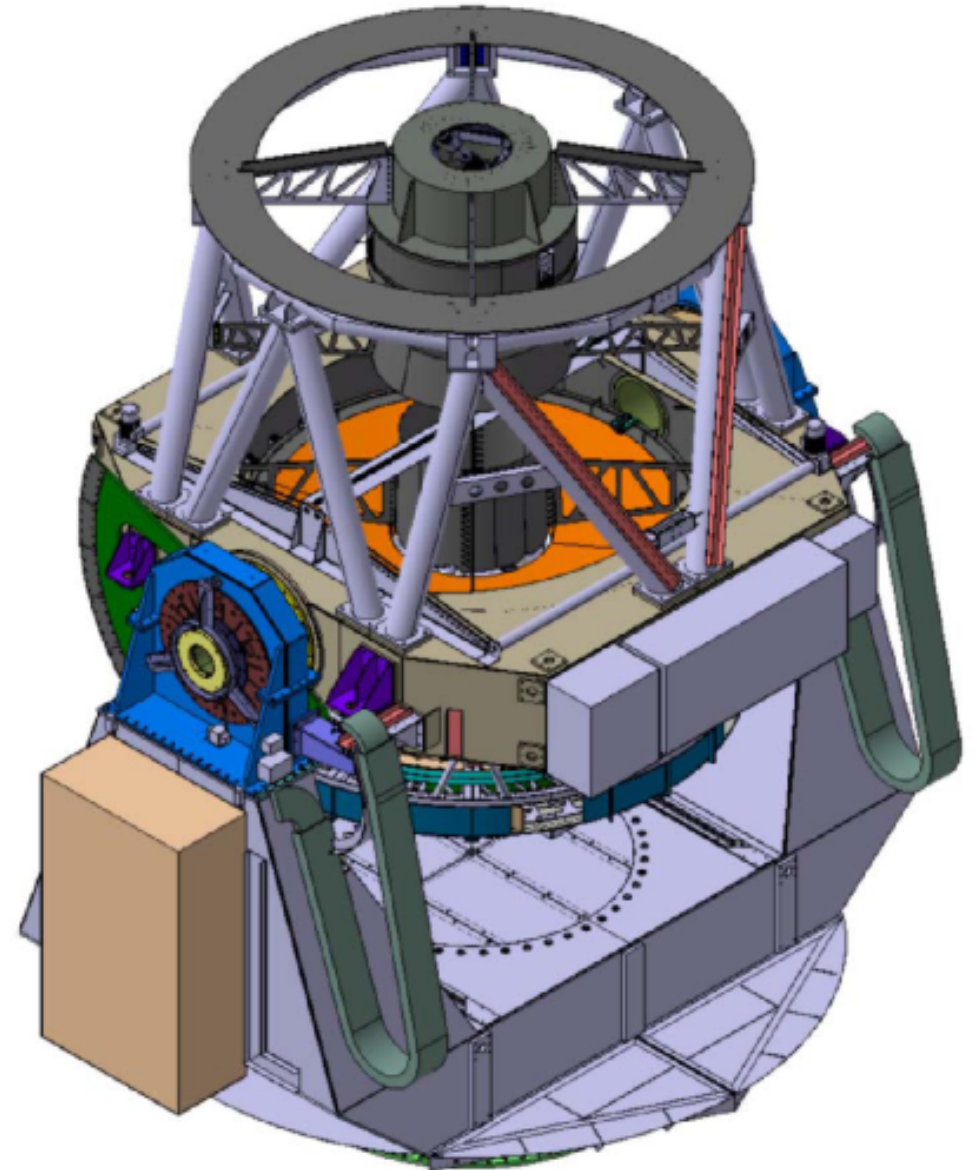
J-PAS overview

T250 telescope:

- 2.5 m
- 14-CCD camera
- 3° FoV
- 1.1 Gpixel
- 56 filters



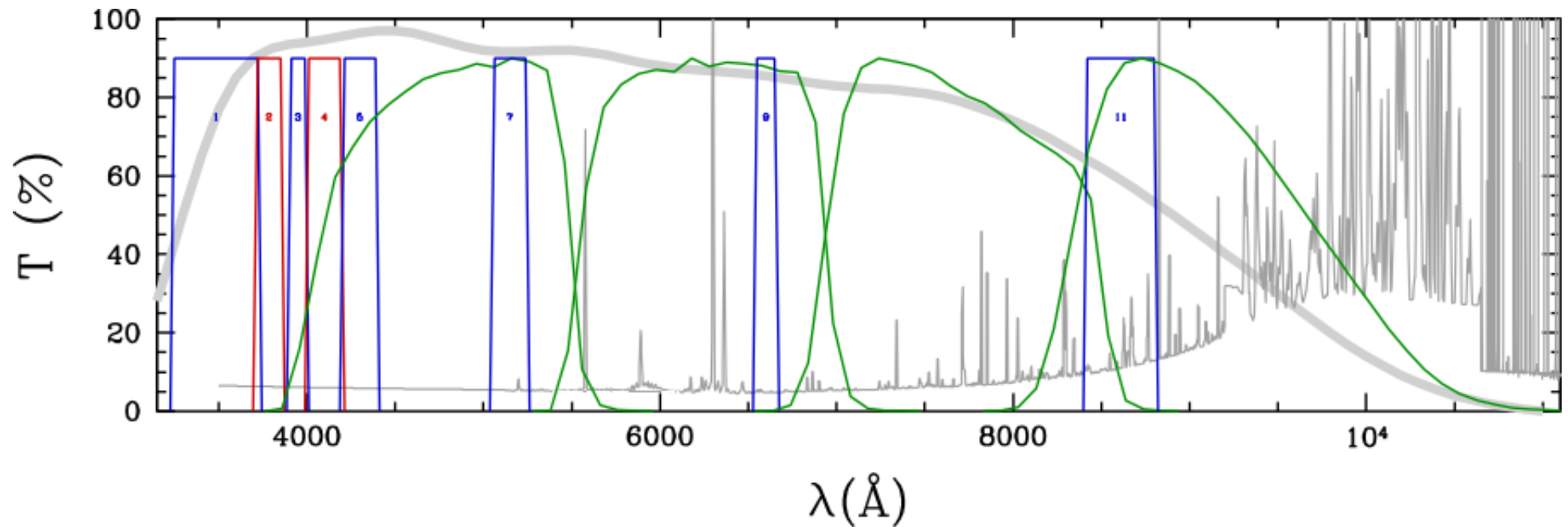
(K. Taylor)



(J. Cenarro)

J-PAS overview - Filter systems:

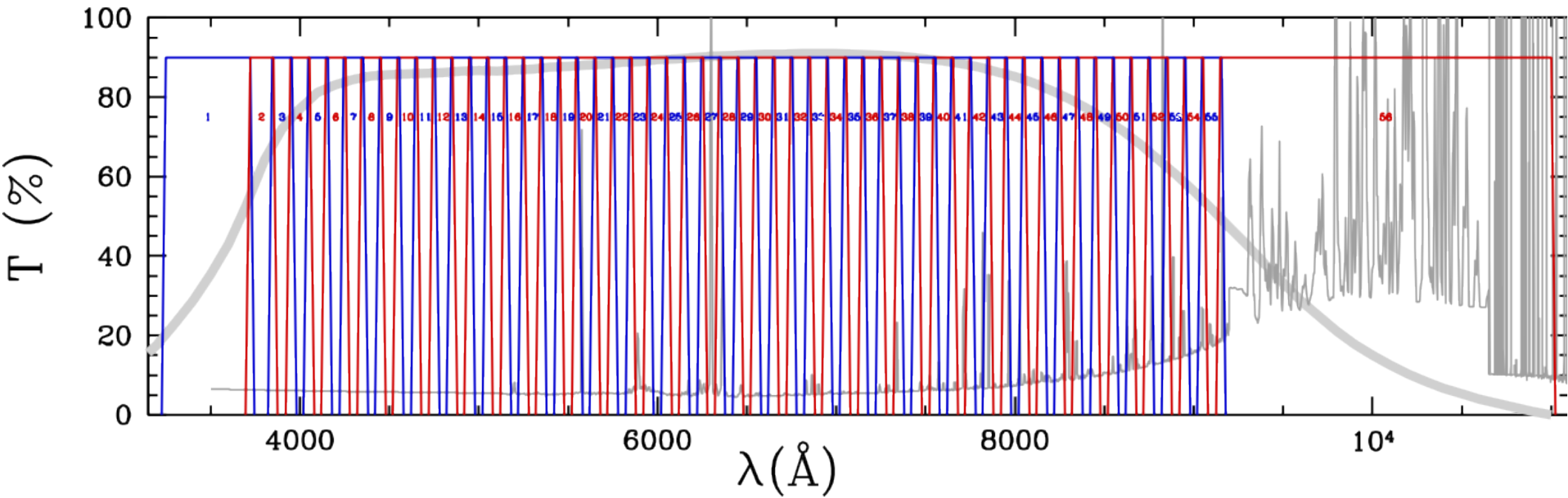
- T80: 12 filters:
 - 4 SDSS (g,r,i,z) + 7:



(A. Marín-Franch)

J-PAS overview - Filter systems:

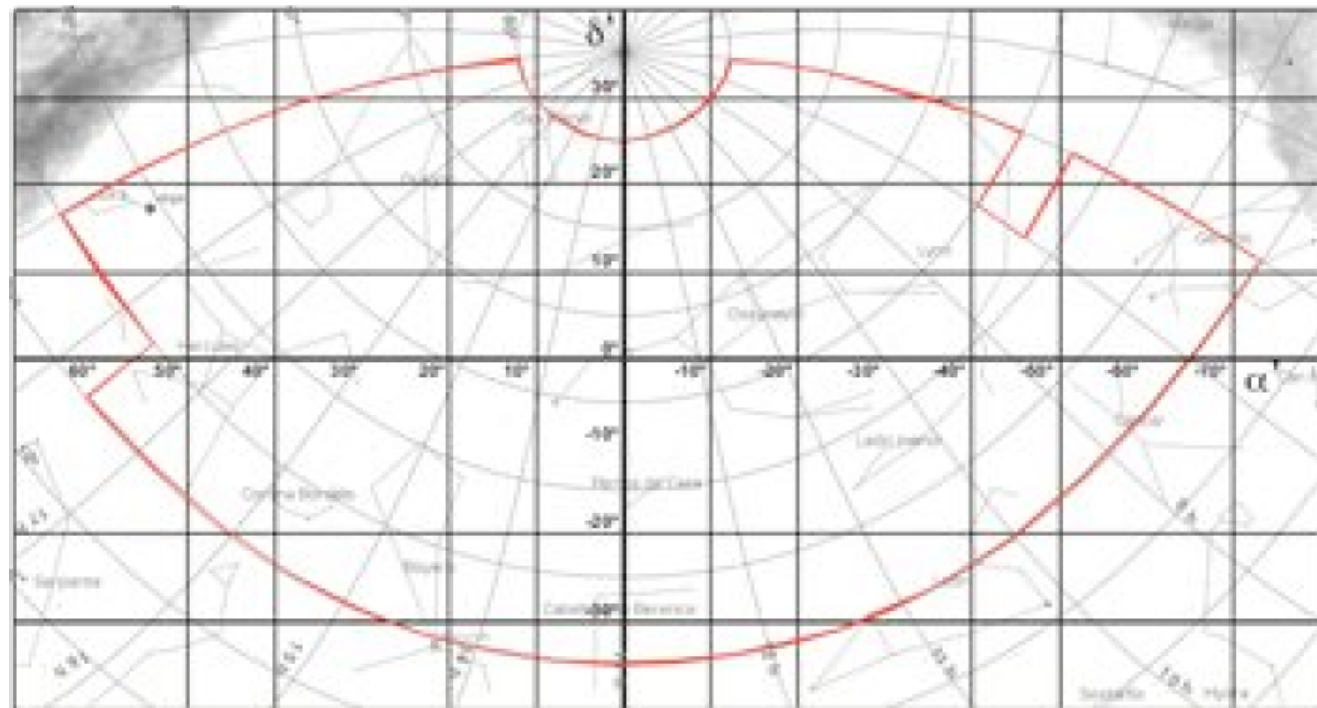
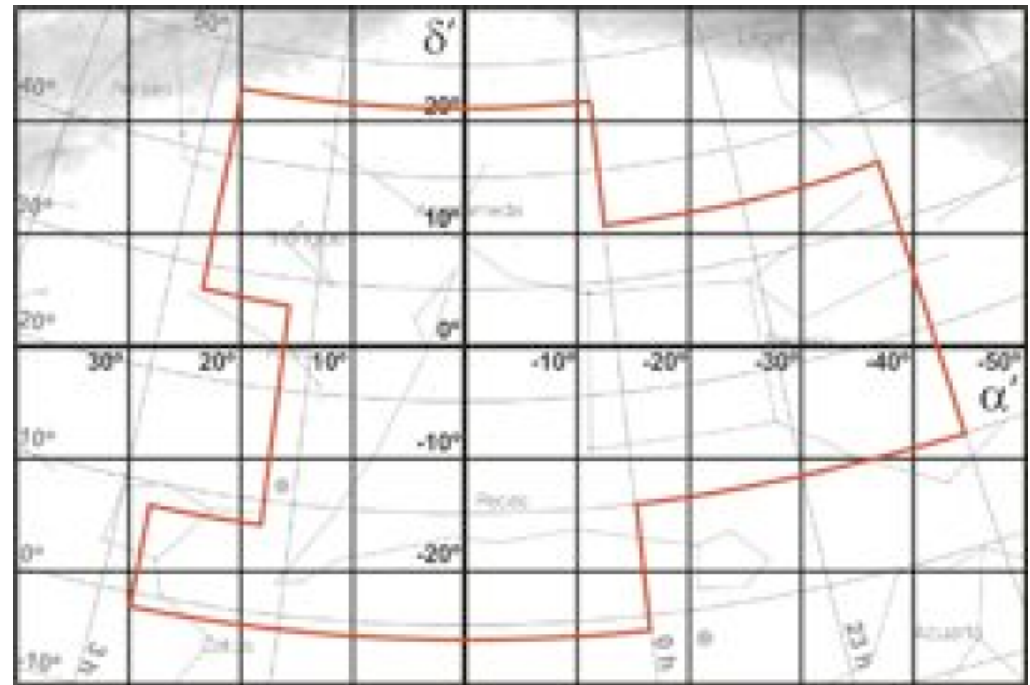
- T250:
 - 2 wide-band
 - 54 narrow-band filters:



(A. Marín-Franch)

J-PAS overview - Survey:

- Above 40° elevation;
- Away from the Galactic plane;
- Away from Galactic dust;
- Overlapping with SDSS;
- Wide shallow + Narrow deep coverage;
- 3 exposures over the whole area.



What is common between J-PAS and LSST?

Both surveys:

- Large-area coverage;
- Wide-field contiguous imaging;
- Multi-band imaging;
- Will produce large volumes of data to process, store and analyze.

All face the same main software challenges:

- **Going from survey strategy to taking daily observations.**
- **Going from daily observations to scientific data products.**
- **Going from scientific data products to scientific results.**

Going from survey strategy to taking daily observations.

Survey execution software: Strategy

Implements the strategy constraints (area x depth to cover, cadence, filter sequencing) to determine where to look each night:

J-PAS survey simulation (A. Fernández-Soto):



Simulations indicate where to change the strategy:

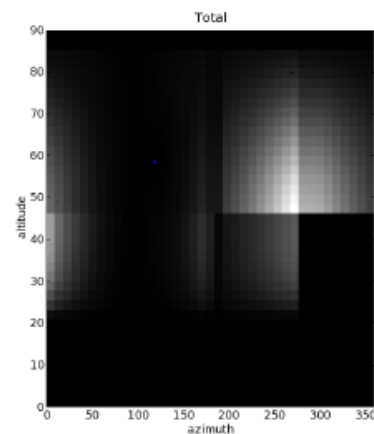
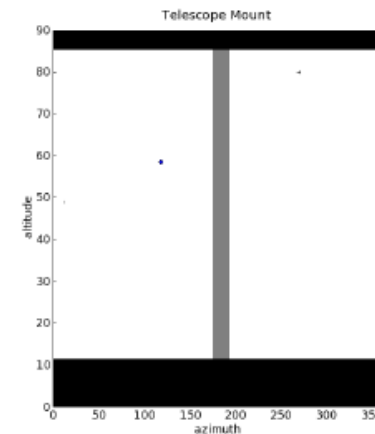
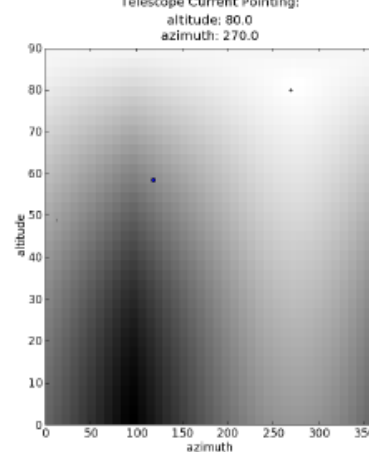
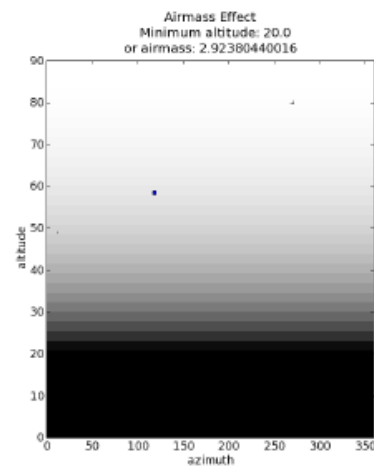
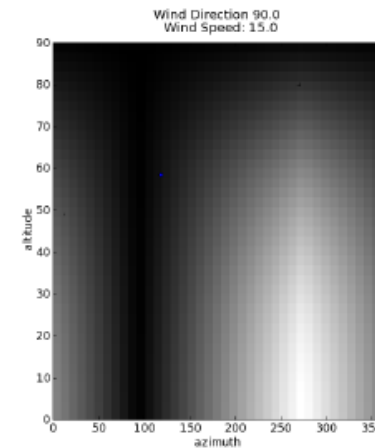
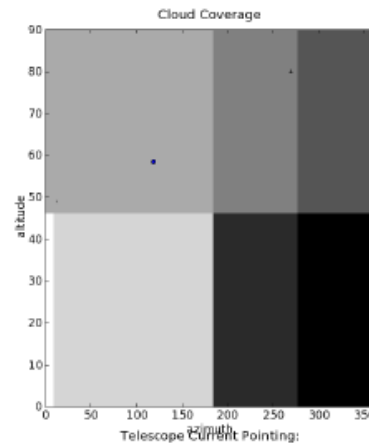
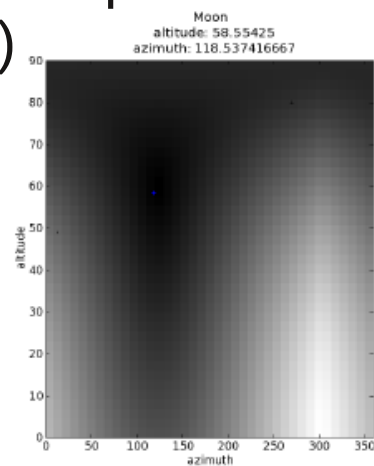
- Some regions can be observed deeper / more times
- Some strategies have too much overhead (pointing, instrument changes, etc.)

Survey execution software: Scheduler

Decides the exact observation sequence to take each night.

Needs to be dynamic:

- Account for downtime, failures, weather changes (**not binary**)
- Decide on the compromise between fields (not all fields can be observed at optimal time)

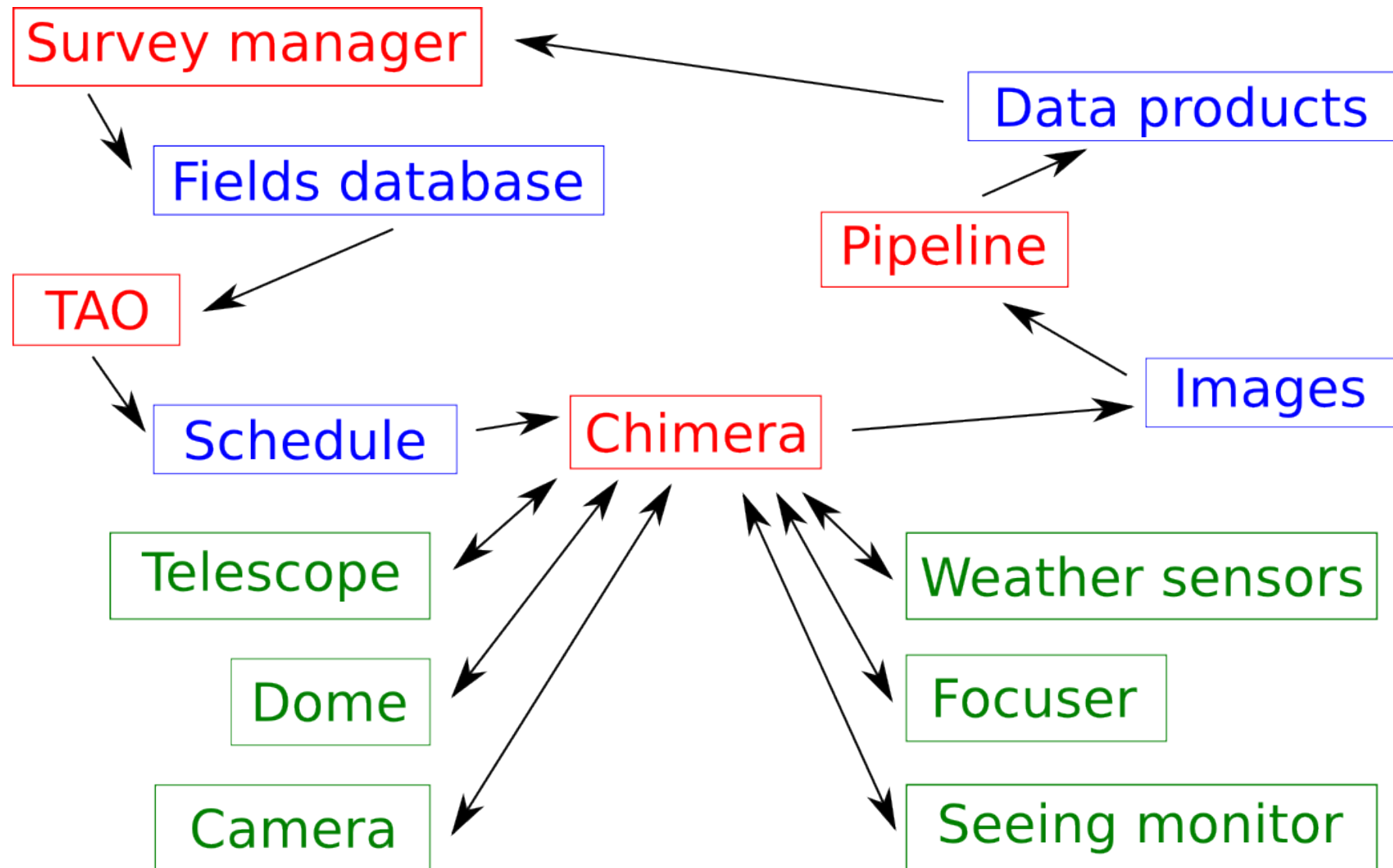


(A. Ederoclite)

Going from survey strategy to taking daily observations.

Survey execution software: Observatory control

Integrates all systems. Ex (for J-PAS / PAU-SUL):



Going from daily observations to scientific data products.

Data products: Reduction Pipeline

The pipeline is **not only** low level (radiometric, coadd, astrometric, etc.)

In large datasets, user cannot process everything:

- The pipeline must generate many scientific products that might be needed (and put them in the catalog).

Minimal set of necessary scientific products:

- Quality measurement (also used by the scheduler).
- Source location (in images) and identification (in source catalogs).
- Source classification (stars, galaxies, nebulae, Solar System).
- Basic measurements (aperture/PSF/isophotal photometry, morphology).
- Derived measurements (redshift, distance, spectral type, stellar type, temperature, mass).

Going from scientific data products to scientific results.

Data products: volume

- DSS - 1994 - 102 CDROMs - 66.3 GB (images)
- 2MASS - 2006 - 5 DVDs, 43 GB (catalog) 10 TB (images)
- SDSS - 3.4 TB (catalog) 11.6 TB (images)
- JPAS T80 - ??? (catalog) 87 TB (images) (~8 times SDSS)
- JPAS T250 - ??? (catalog) 2.3 PB (images) (~ 210 times SDSS)
(W. Schoenell)
- LSST: 200 PB
- Storing the data is not that hard:

LSST will never fill more than 22 hard drives. Individual investigators will be able to maintain their own data copies to analyze as they choose. With the data set increasing linearly with time, and storage doubling every 1.6 years, the peak storage cost occurs 2.3 years into any survey.

From ***A Letter to the NSF Astronomy Portfolio Review: LSST is Not “Big Data”***, David Schlegel (Lawrence Berkeley National Lab), 2012
<http://arxiv.org/pdf/1203.0591v1.pdf>

Going from scientific data products to scientific results.

Data products: volume

Storing the data is not that hard.

Neither is processing.

But accessing can be difficult:

Since 80's:

- Sequential read increased from 20MB/s to 80MB/s (fac of 4)
- Hard disk speed increased from 3k to 15k RPM (factor of 5)
- Storage cap. increased from 1GB to 2TB (factor of 2k!)

(W. Schoenell)

Storage (not only processing) will have to be massively parallel.

New software solutions necessary to do parallel processing on the distributed storage (cannot just use a big storage server).

Going from daily observations to scientific data products.

Data products: Archival, access and visualization

Archival is not just shoving all files in a hard drive.

The success of **any observation** depends on how the data being:

- Systematically processed (a **well-documented** pipeline);
- Well-organized (a **good database** is necessary if one has more than a few observations);
- **Accessible:**
 - The database must already have lots of fields with the variables that may be needed.
 - **VO protocols must be followed.** Not optional!
 - User-friendly access and visualization of queries is **essential** if the data is ever going to be used.

Going from scientific data products to scientific results.

Virgo - Millennium Database

Documentation

CREDITS/Acknowledgments

Registration

News

FAQ

Databases

millimil (context)

Streaming queries return unlimited number of rows in CSV for
Browser queries return maximum of 1000 rows in HTML format

```



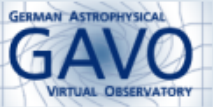
select power(10, .1*(.5+floor(log(g.np)/.1))) as
      avg(g.stellarMass) as stars_avg,
      max(g.stellarMass) as stars_max,
      avg(g.bulgeMass) as bulge_avg,
      max(g.bulgeMass) as bulge_max,
      avg(g.mag_b-g.mag_v) as color_avg
from millimil..DeLucia2006a g
where g.snapnum= 63
      and g.mag_b < 0
group by np
order by np
        
```

Maximum number of rows to return to the query form:

Demo queries: click a button and the query will show in the query form. Holding the mouse over the button will give a short explanation

Mainly Halos:

Mainly Galaxies:

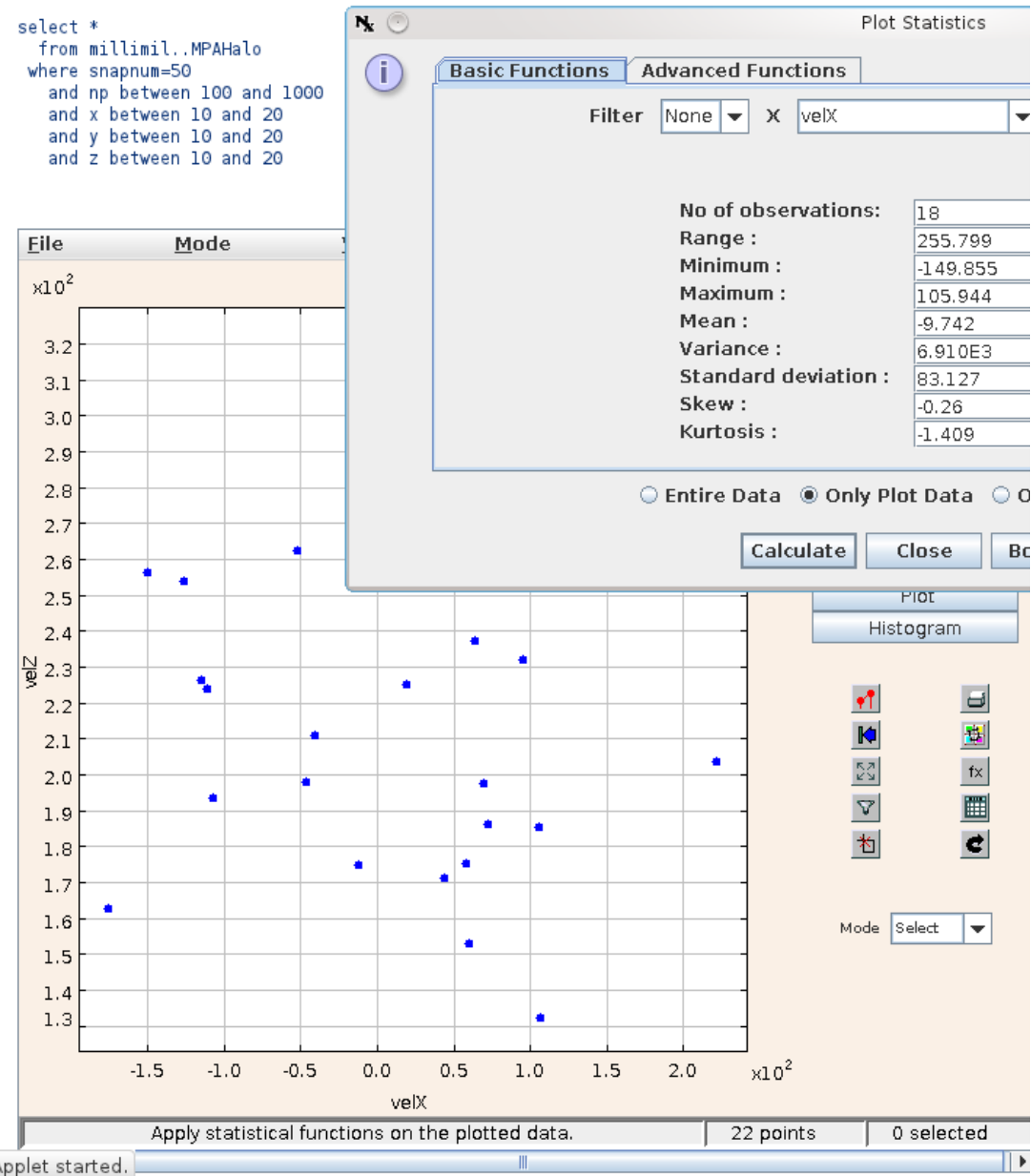




For help on the use of VOPlot follow this [link](#) to the VOPlot site.

The diagram shows the result of the following query:

```

select *
from millimil..MPAHalo
where snapnum=50
      and np between 100 and 1000
      and x between 10 and 20
      and y between 10 and 20
      and z between 10 and 20
        
```



Millenium simulation interface
GAVO (German Astrophysical VO).

<http://gavo.mpa-garching.mpg.de/Millennium/MyDB>

Going from scientific data products to scientific results

One (true) tale:

Cassini VIMS (Visual and Infrared Mapping Spectrometer) data:

$\sim 2 \times 10^7$ spectra of Titan recorded since 2004 (and still going)

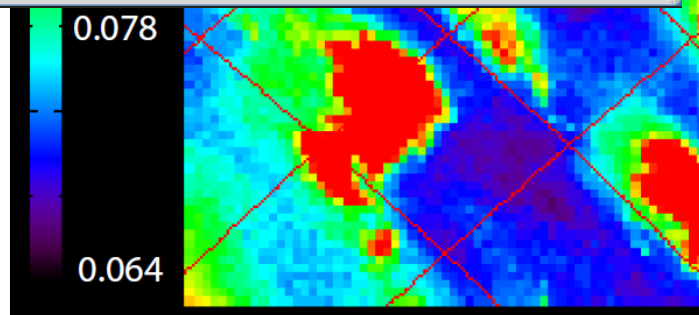
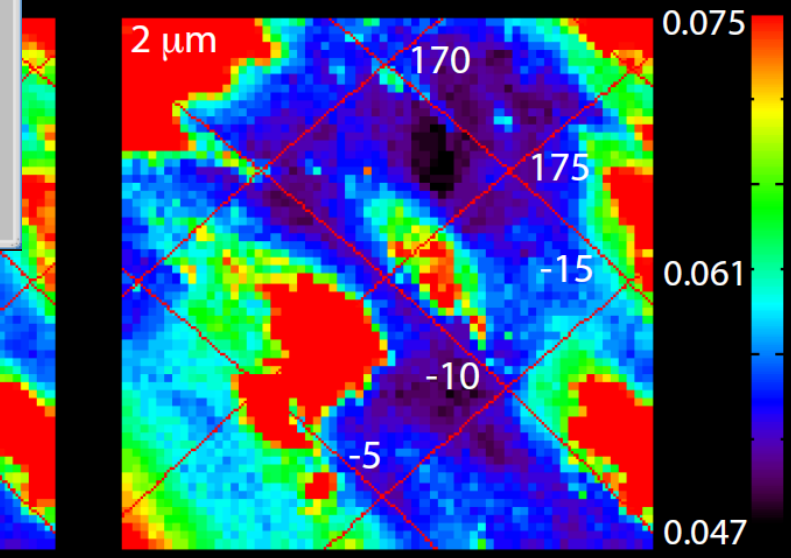
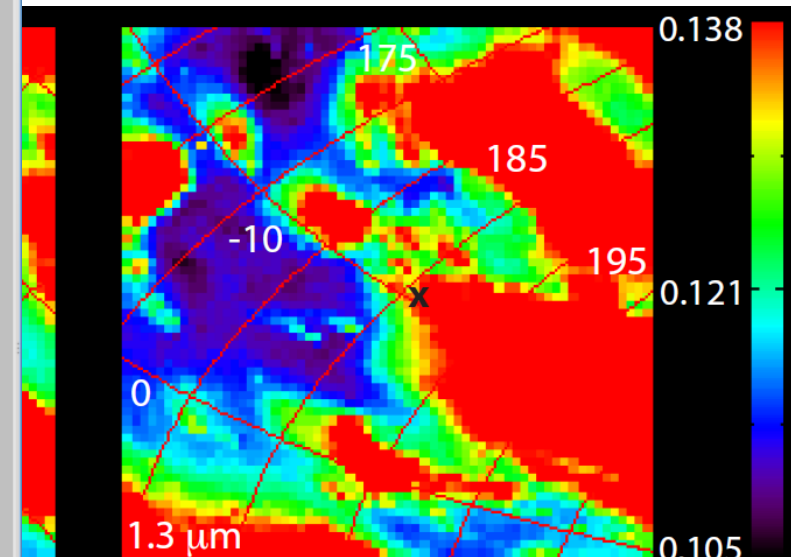
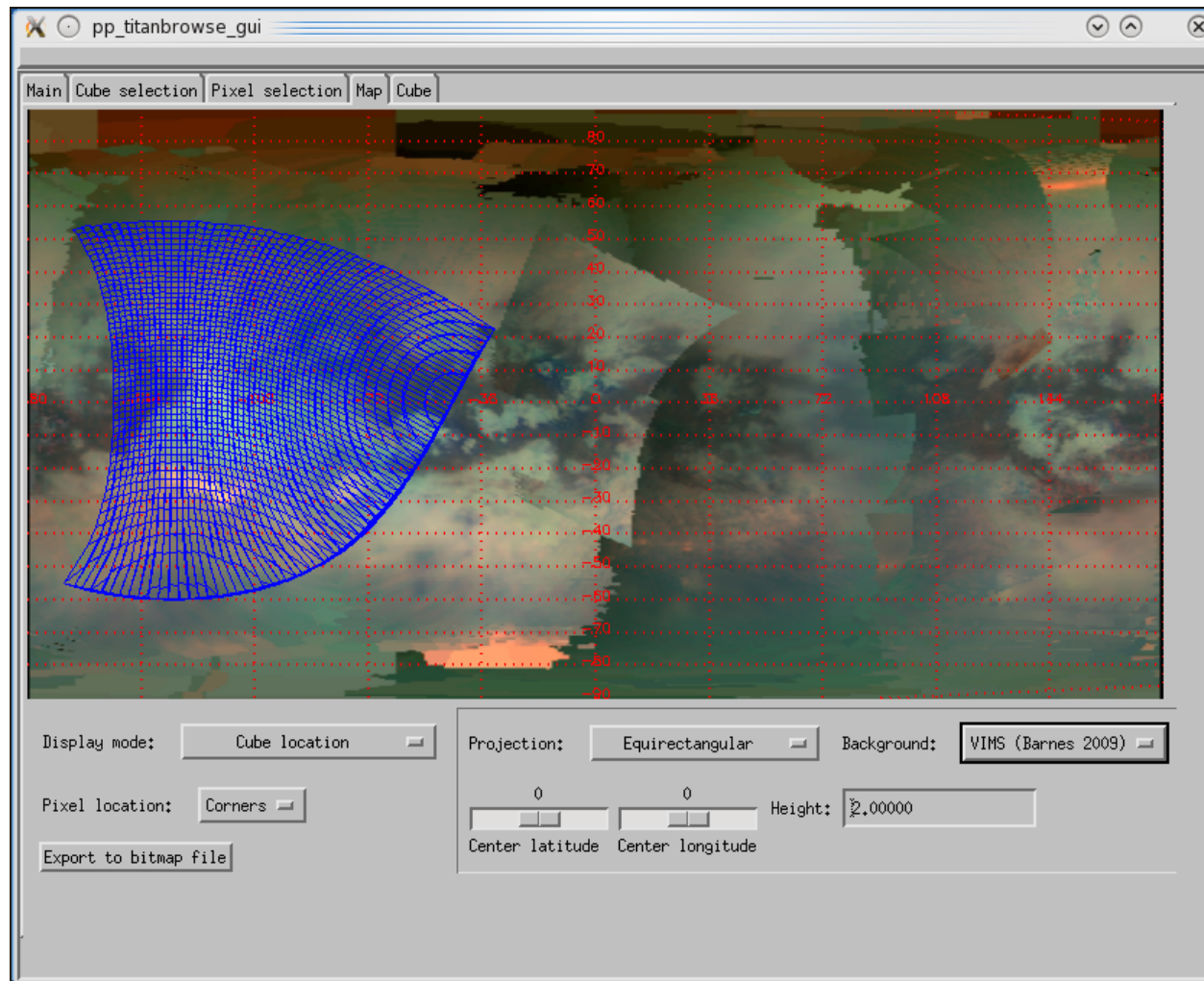
All processed data publicly accessible through NASA's PDS (Planetary Data System):

- All the usual metadata fields (exposure settings, geometry on target, etc.)

But PDS was insufficient:

- After several years in the public archive, the first tropical lake was found only with the use of a new system which integrates database (with more fields, both metadata and data), flexible query evaluation and visualization.

Going from scientific data products to scientific results



(Griffith, Lora, Turner, Penteado *et al.*, 2012; submitted to Nature)

(Penteado 2012, in prep.)

Going from scientific data products to scientific results

A promising new database system: SciDB (<http://www.scidb.org>)

- A data model based on multidimensional arrays, not sets of tuples.
- A storage model based on versions and not update in place.
- Built-in support for provenance (lineage), workflows, and uncertainty.
- Scalability to 100s of petabytes and 1,000s of nodes with fault tolerance.
- Support for "external" data objects so that data sets can be queried and manipulated without ever having to be loaded into the database.
- Open source in order to foster a community of contributors and to ensure that data is never "locked up" — a critical requirement for scientists..
- Mapping layer will have interfaces to at least C++, Java, and Python

(W. Schoenell)

Outline

J-PAS overview

The problems:

- Going from survey strategy to taking daily observations.
- Going from daily observations to scientific data products.
- Going from scientific data products to scientific results.

Survey execution:

- Strategy / Scheduler
- Observatory control

Data products:

- Reduction Pipeline
- Generation of data products
- Archival, access and visualization

pp.penteado@gmail.com

http://www.ppenteado.net/ast/pp_lsst_201204.pdf