

# Data Pipelines & Archives for Large Surveys

Peter Nugent (LBNL)

# Overview

Major Issues facing any large-area survey / search:

- Computational power for search - data transfer, processing, storage, databases and the manpower to keep it all flowing.
- Above coupled with historical record of what you know about that part of the sky so you don't *repeat* a discovery.
- Algorithms to determine which objects to follow.
- Conversely, which objects that you will reject.
- Optimization of the above coupled with scheduling to maximize the scientific gain from your search and follow-up.

# SN Factory & DeepSky

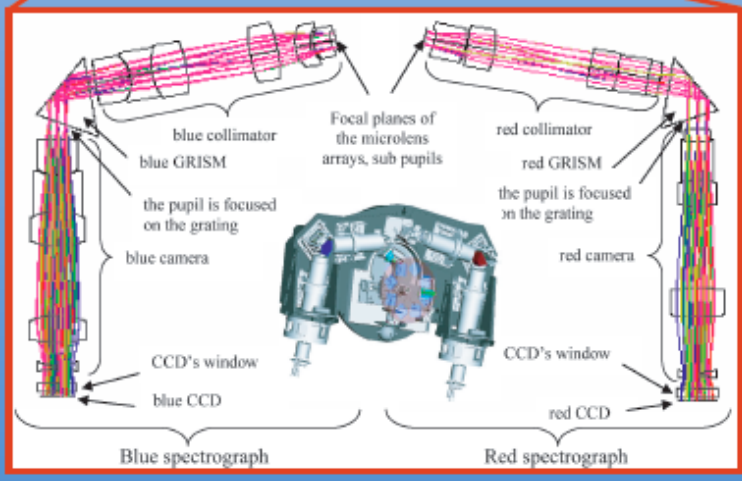
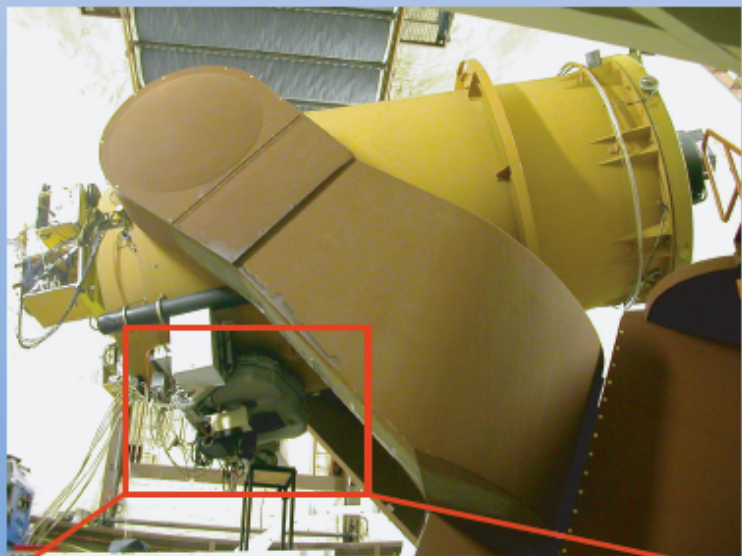
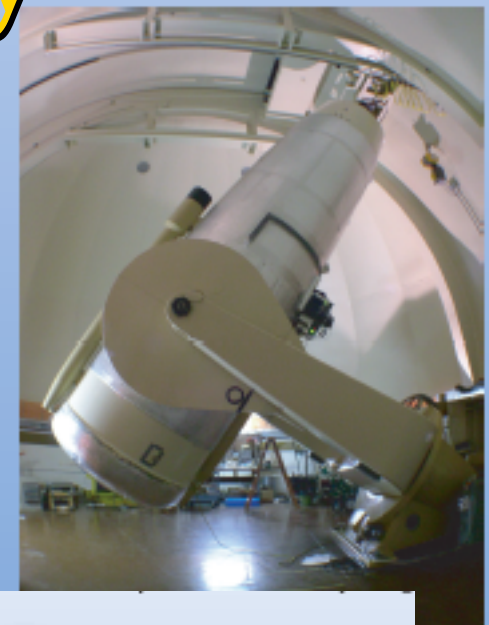
Two parts to this talk:

- (1) Introduction of real-time transient pipeline for SN Factory
- (2) Introduction to DeepSky archive

# Nearby SN Factory

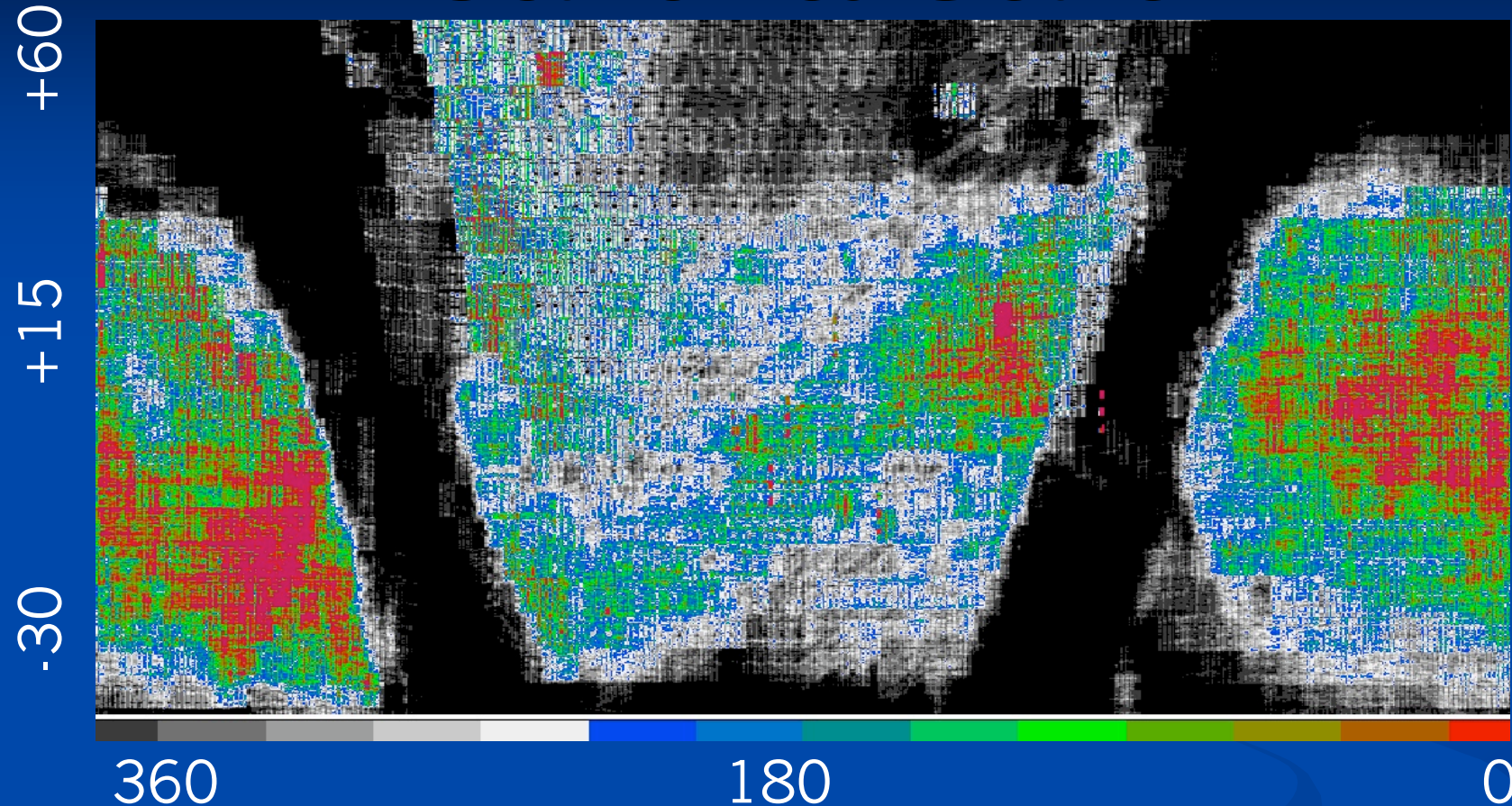
Searching on Palomar  
Oschin Schmidt

Spectrophotometric  
Follow-up w/  
UH 2.2-m





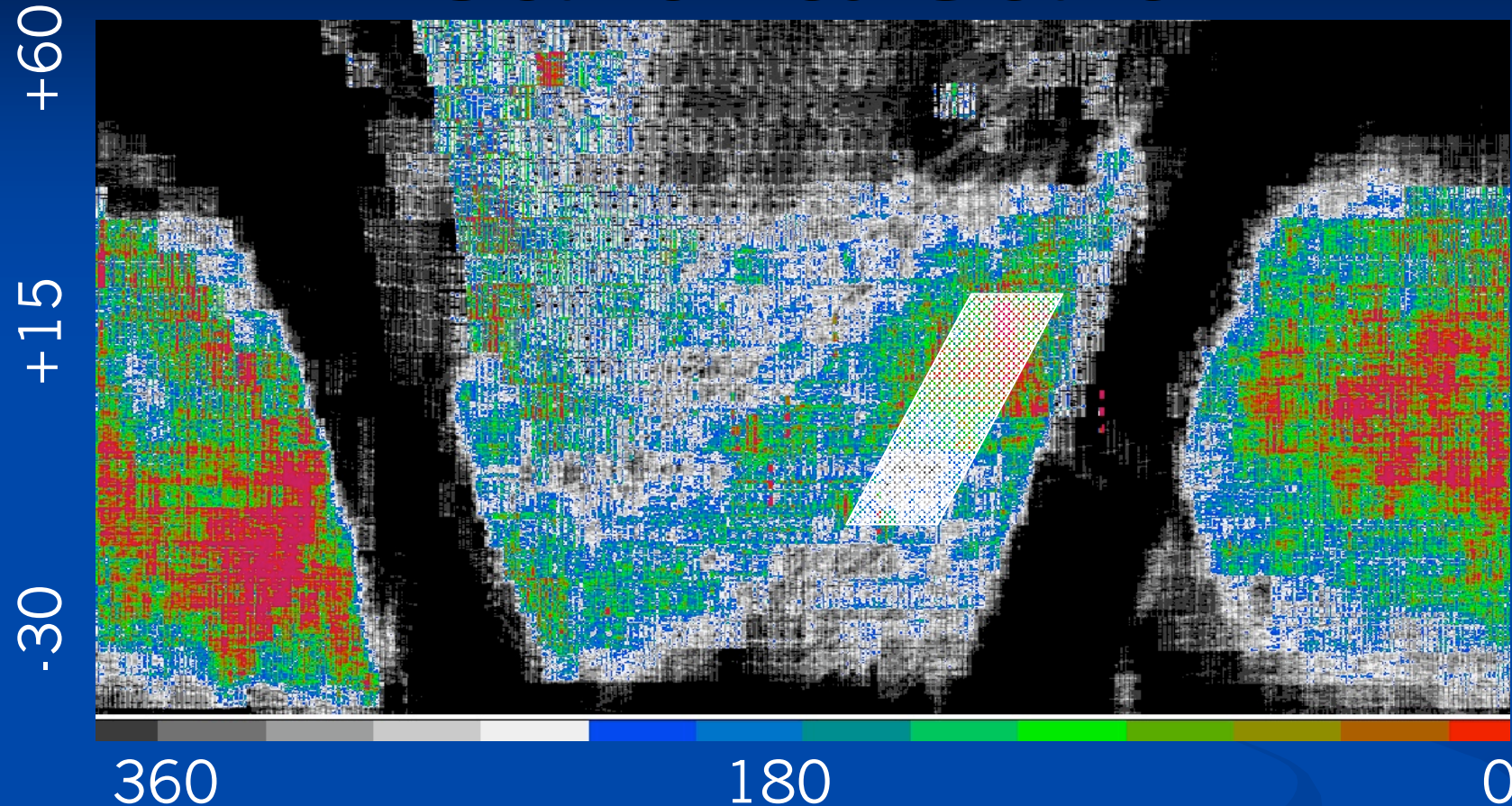
# Search & Goals



Asteroids: One person's garbage is another's gold!  
Pave the sky each night,  $\sim 600$  sq. deg. to mag 21.5  $\rightarrow$  2 SNe/night.  
Goal is 200 SNe Ia in the Hubble flow over the project lifetime.

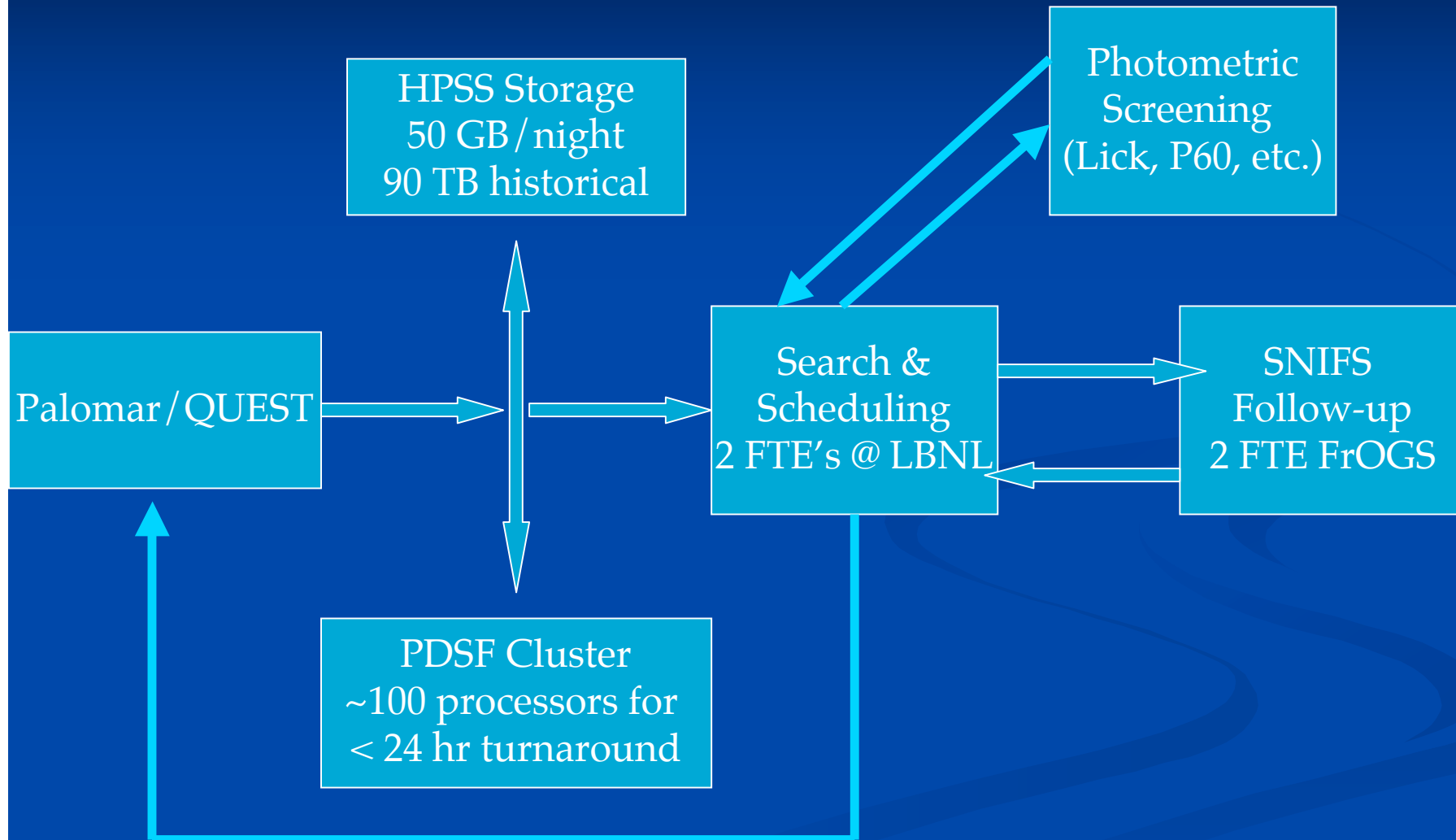


# Search & Goals



Asteroids: One person's garbage is another's gold!  
Pave the sky each night,  $\sim 600$  sq. deg. to mag 21.5  $\rightarrow$  2 SNe/night.  
Goal is 200 SNe Ia in the Hubble flow over the project lifetime.

# Data Flow...



# Searching the data

When the SN-only search is going on the Palomar-schmidt we cover about 110 pointings of the QUEST camera a night. These pointings consist of 2 images in one filter (RG610) separated by 1 hour.  $S/N \sim 7$  to  $R$  of 20.0 during full moon and 21.5 during new moon. This results in an effective area of  $\sim 600$  sq. deg.

This generates about 10,000 image-pairs in a night. 10% of these will have 5-sigma detections on them. Since 1000 subtractions is too much to scan, we use a boosted decision tree to knock this down to a more reasonable number  $< 100$ .



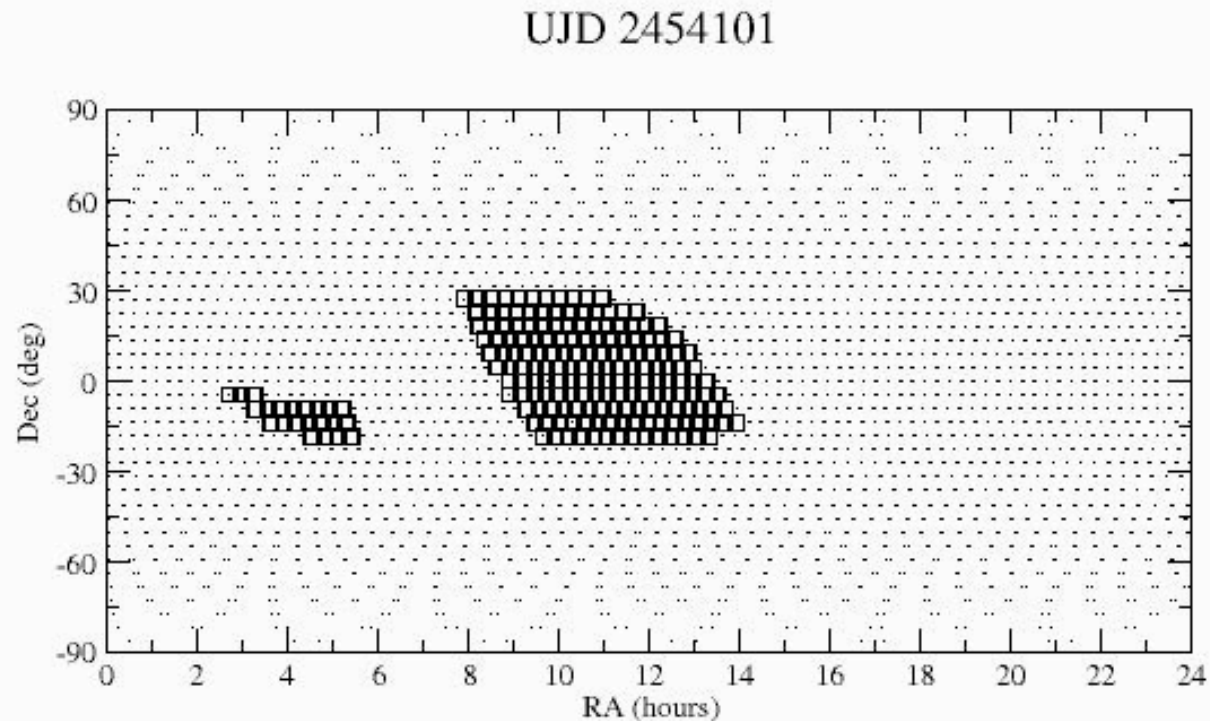
# Vetting & Scheduling

The vetting process links all candidates with known astronomical catalogs (MPECs, NED, SDSS, etc.) to aid in eliminating garbage. Furthermore we use our own database of historical images (typically  $> 50$  images over past 7 years anywhere in northern sky) to screen out anything that has varied in the past or is on the way down now (this eliminates half of our candidates that are unknown).

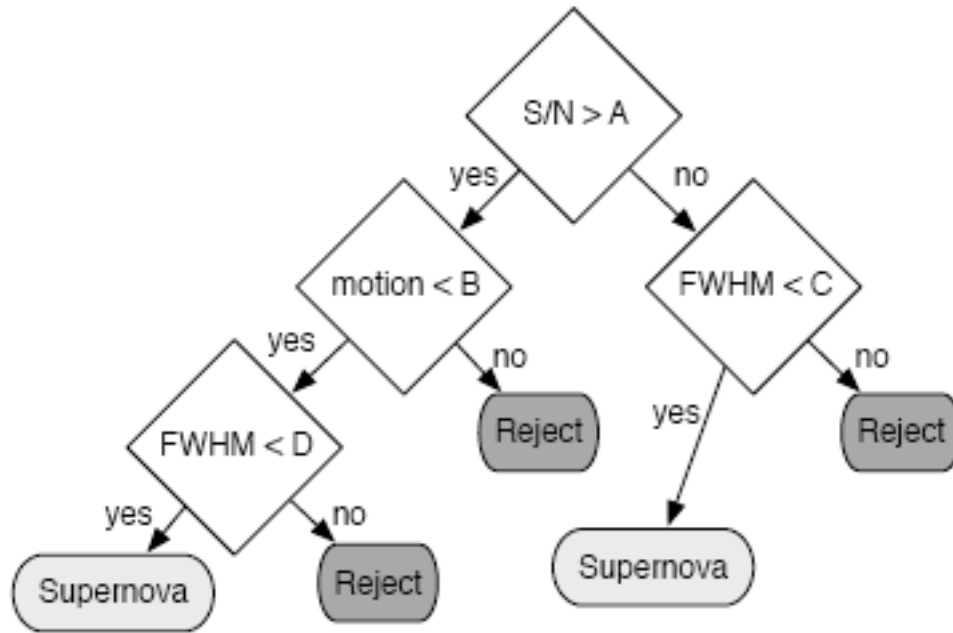
Scheduling requires prioritizing all candidates and follow-up objects. Weights are given to the candidate based on where it is on the sky, the number of constraining observations that identify this target as new, current follow-up load, etc. Weights are given to the current supernovae based on the number of observations obtained to date, desired cadence, redshift, interest level, etc. An optimized schedule is generated each day.

# Vetting & Scheduling

Example field visibility between a facility at La Silla and one in Greece. Requirements include visibility for 1 hour  $>$  60 days at both locations with  $\text{secz} < 2.0$ .



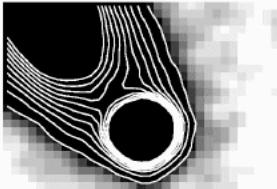
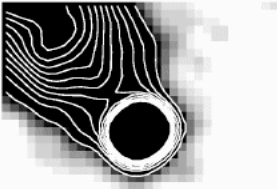
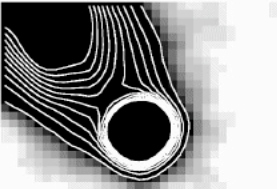
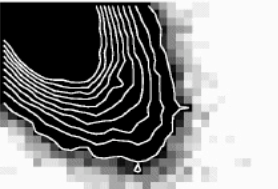
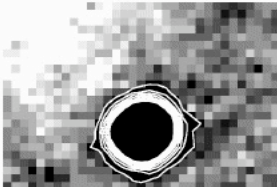

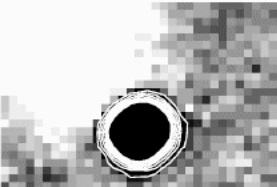
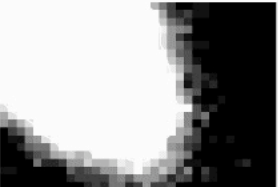
# Boosted Decision Tree



The SN Factory decision tree incorporates between 20 and 30 variables and is 85% effective. It cuts the scanning load by an order of magnitude. It has been trained on fake and historical SN.

Example decision tree that would treat high S/N objects differently than low S/N objects. In practice a real decision tree has many more branches and the same variable can be used at many different branches with different cut values (Bailey *et al.*, 2007)

# Discoveries...50Gb/night

NEW1 5.9415e+05	NEW2 3.9275e+05	NEW 5.0470e+05	REF 19496.	
				nb-255 "Caliban" REF feb699swatch7csg.fts feb699swatch263csg.fts feb1199swatch23csg.fts feb1199swatch151csg.fts feb1199swatch279csg.fts feb1999swatch39csg.fts feb1999swatch167csg.fts feb1999swatch295csg.fts NEW1 mar1099swatch24583csg.fts NEW2 mar1099swatch24839csg.fts
				
5.7466e+05	3.7325e+05	4.8521e+05	19496.	

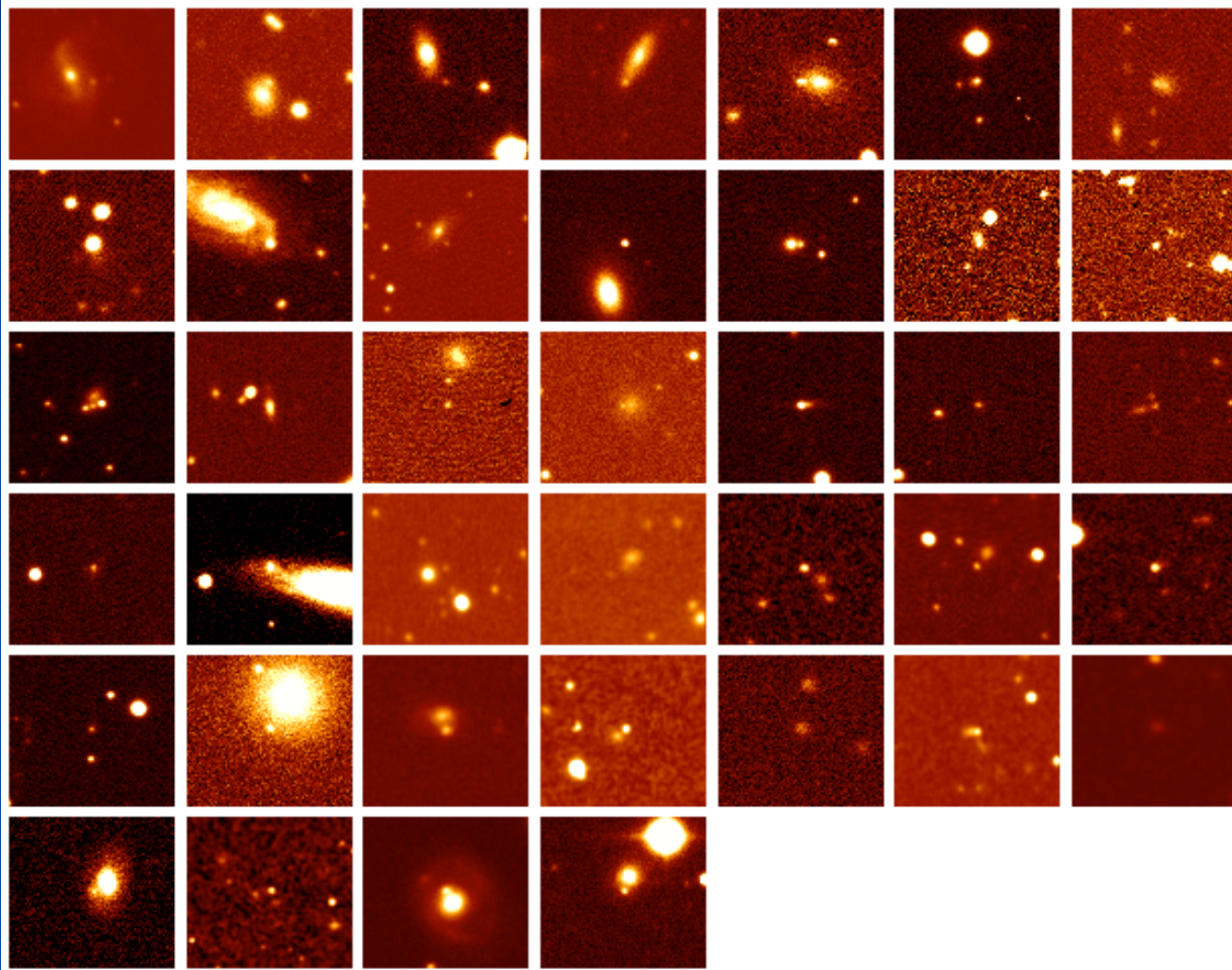
  

Ap.Sig = 490.0	Subtraction: swatch_3.91	Scanner: rknp
%Inc = 2489.	RA : 8:20:46.83	Subtractions :-3
PCyg.Sig = 247.8	Dec : +18:28: 3.1	Host : 5
MaxPixSig = 0.000	Refsys= feb699swatch7csg.fts	Shape : 5
MX Y = -6.543	Pos on refsys: ( 34.1, 721.1)	Position : 5
FWX = 3.566		Motion : 5
FWY = 3.582		Overall : 5
NeighDist = 24.47		
NeighMag = 17.00		
Mag = 15.90	The subtraction isn't great, but this one is OBVIOUS.	
Theta = 90.00	In fact, at mag 16, it may be known already.	
New1Sig = 466.6		
New2Sig = 296.7		
Sub1Sig = 499.9		
Sub2Sig = 321.6		
Sub2-Sub1 = 123.3		
DSub1Sub2 = 0.1200		

1999-03-24 09:01:56.00

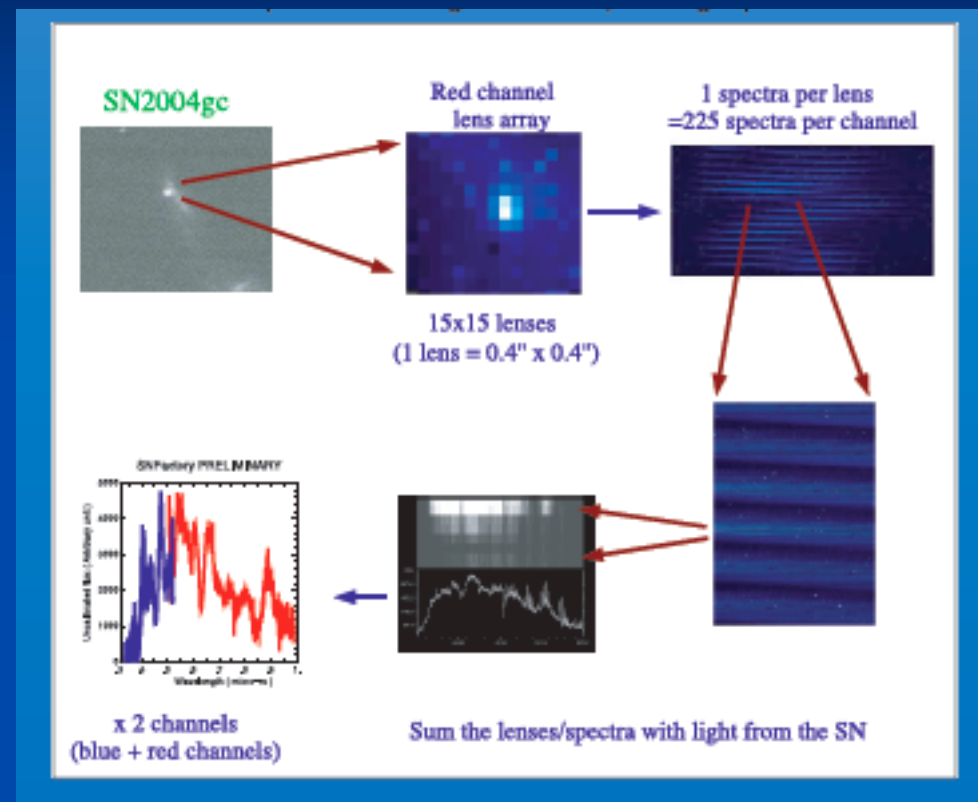
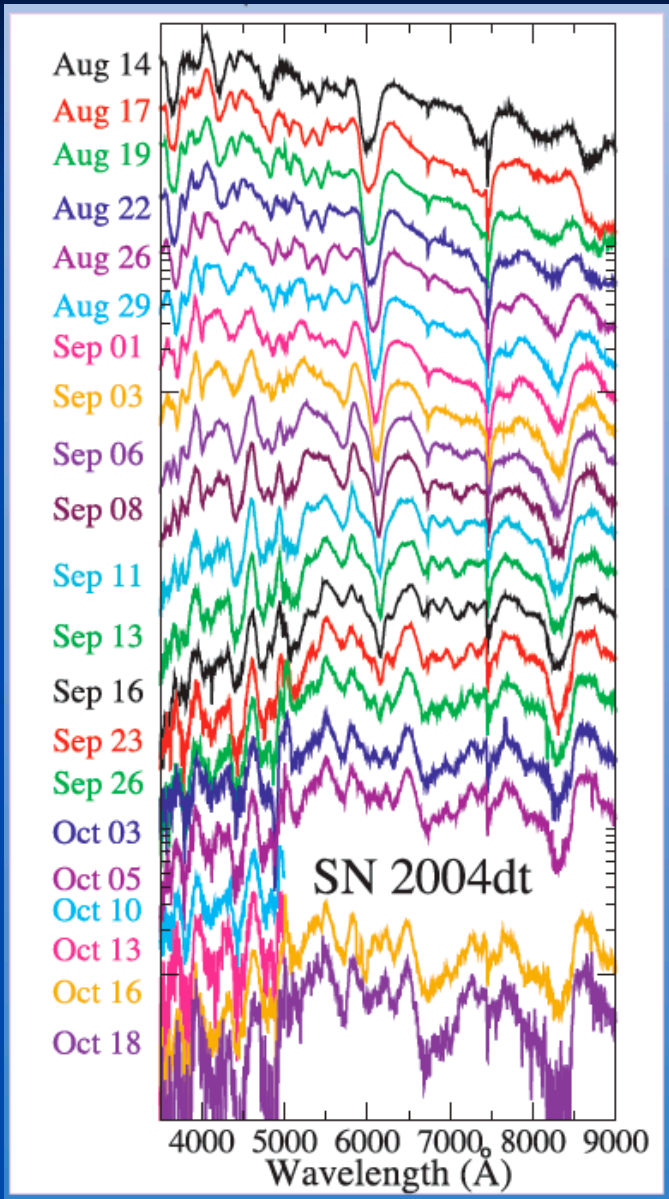


# Supernovae



August 2008's  
spectroscopically  
confirmed SN  
discoveries.

# SNIFS



Automated follow-up, overheads ~4min:  
 Leads to about 20 science spectra night  
 down to mag 20.0 on an 88" telescope.

# Discoveries 2008

Since June 1 of this year the search has had a cadence of 1 week covering ~2000 square degrees of sky and has yielded over 1000 transients with approximate breakdowns of (parens for number of screening spectra taken):

350 variable stars / novae / AGN (19)

218 supernovae (72 SN Ia, 50 core-collapse)

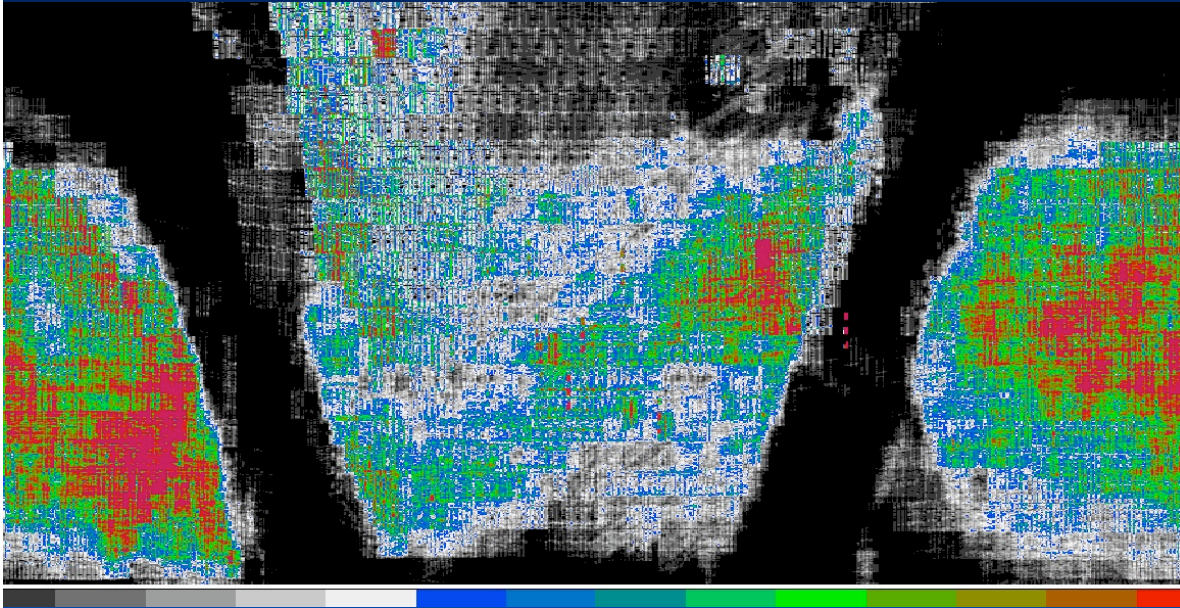
450 slow moving asteroids (1)

and the rest unclassified transients....

~30 SN Ia in the nearby Hubble flow ( $0.03 < z < 0.08$ ) and caught before peak brightness which we follow for 2 months. This combined with the screening spectra above completely fills our follow-up capabilities at the UH 2.2-m.



# DeepSky



This data spans 9 years and almost 20,000 square degrees, half from the NEAT 3-banger and half from the Palomar-QUEST camera on the Palomar Oschin schmidt telescope (all in R-band like filter). The entire dataset is 90 TB and will create both a temporal and static catalog of astrophysical objects. NERSC is re-processing and hosting this data on spinning disk (4% complete as of 1 hr ago - 6 months to go). Several historical pre-discoveries already!

See: <http://supernova.lbl.gov/~nugent/deepsky.html>



# DeepSky Processing

Most work being done on Davinci (SGI Altix - 32 Intel Itanium-2 processors with 192 GB of shared memory) and Jacquard (712-CPU Opteron cluster)

Both of these machines can see the NERSC Global Filesystem (an IBM GPFS), which is very fast and for which we have 70TBs of space.

Typically I process a months worth of data (500,000 6 MB images) in 2 days, making darks, flats, fringes and masks with a wcs solution for each image and loading them into the db. This averages out to 1 GB / minute which is within a factor of 3 of what SASIR needs to accomplish for a real-time pipeline. In terms of dedicated processors, I am typically running on ~20 constantly.

SASIR - a processor / chip (in today's chips) would be more than enough to keep up with the data-flow. Crucial is the access to a robust / fast filesystem.

See: <http://supernova.lbl.gov/~nugent/deepsky.html>

# Conclusions

I see no major hurdles for the SASIR telescope for the processing pipeline.

The weak link for us has always been the data-transfer - how to recover from a 48 hr downtime and get through it all and back to real-time. Requires redundancy and capability to ramp up processing power on-demand. 100 cpus - 300 cpus for 24 hrs.

Archiving and cataloging of large datasets is going on now, while SASIR is a step up in this, it is a manageable one given enough preparation time.