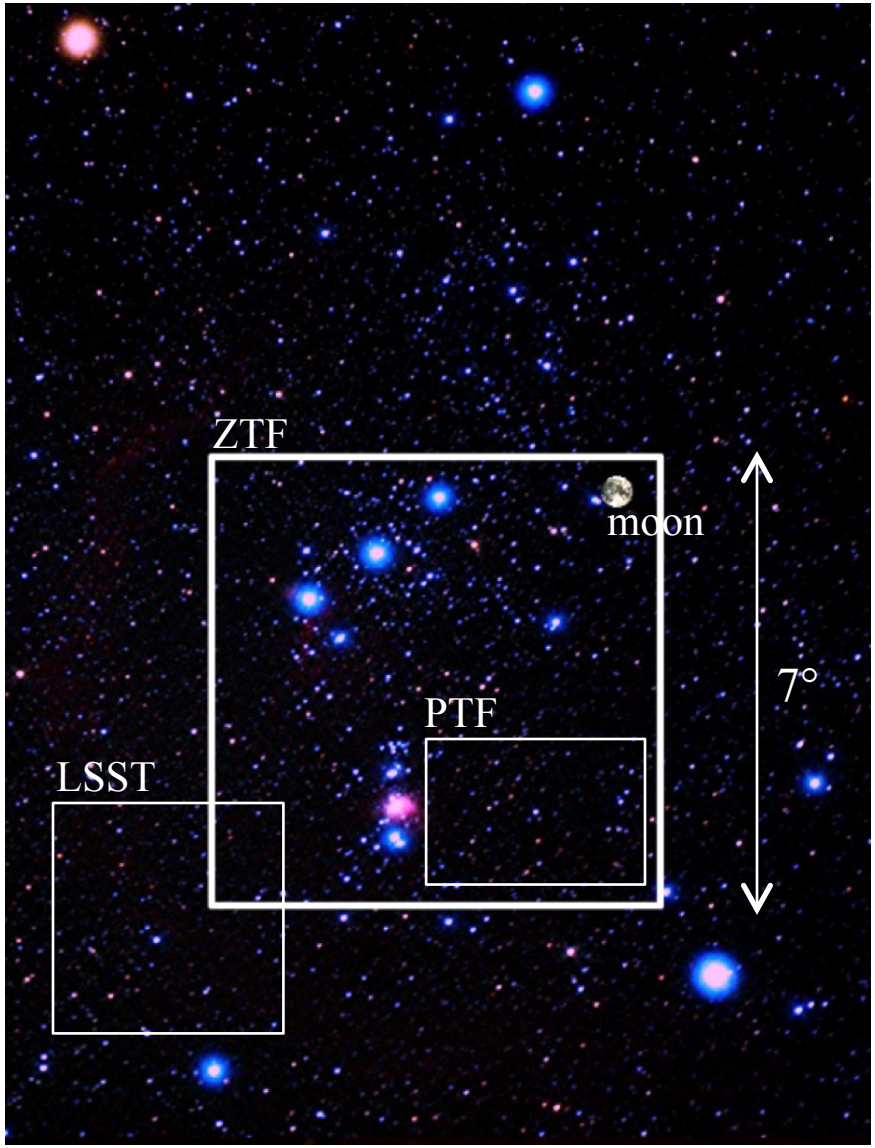
ZTF Science Data System

Frank Masci & the IPAC-Caltech ZTF Team

NSF Site Visit, July 2018

ZTF at a glance



- A fast, wide-area time-domain survey:
  - fast, young, and rare flux transients
  - counterparts to gravitational wave sources
  - low-*z* Type Ia SNe for cosmology
  - variable stars & eclipsing binaries
  - Solar System objects
- Active detector area: ~ 47 deg^2
- Areal survey rate: 3760 deg^2 / hour
- Single exposure depth (5σ): $r \sim 20.5$ mag.
- Median image quality (*r*): ~ 2.2” (FWHM)
- Nominal survey duration: 3 years
- Number of filters: 3 (*g, r, i*)
- Survey entire Northern visible sky to $\delta \sim -30°$

https://www.ztf.caltech.edu

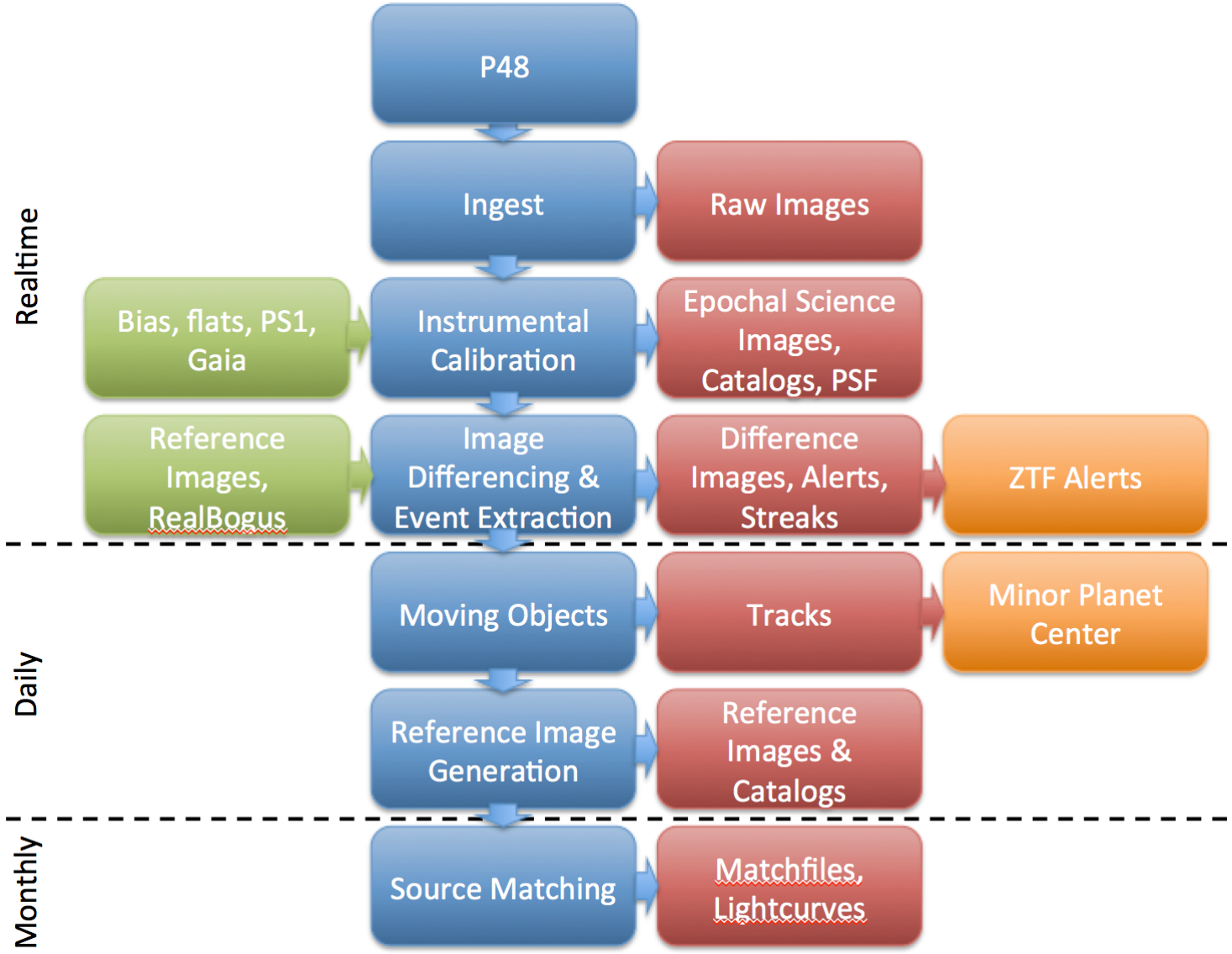
Key Project Dates

- Engineering commissioning start: **Oct 14, 2017**
- First light announcement: **Nov 14, 2017**
- Science validation period: **Feb 3 – Mar 16, 2018**
- Start of science operations: **Mar 17, 2018**
- Commencement of public alerts: **June 4, 2018**
- Dedication of ZTF at Palomar Science Meeting: **July 10, 2018**
- First ZTF Summer School: **Jul 18 – 20, 2018**
- First public data release: **Mar 2019**
- Second ZTF Summer School: **Jul 2019**
- Second public data release: **Sep 2019**
- Third public data release: **Mar 2020**
- Third ZTF Summer School: **Jul 2020**
- Fourth public data release: **Sep 2020**
- End of science operations: **Dec 2020**

The ZTF Science Data System (ZSDS)

- The ZSDS is housed at IPAC, Caltech
- IPAC is a multi-mission science center (IRAS, 2MASS, *Spitzer*, WISE/NEOWISE, LSST, Euclid, WFIRST …)
- Responsibility for ZTF:
  - raw data ingestion; all processing pipelines; quality assurance metrics
  - long-term data archiving, curation, user-interfaces to retrieve data
  - generation of transient alert packets to support near real-time discovery
  - maintenance of operations, databases, fileservers, and archive infrastructure
  - documentation and user-support

Data flow: timeline view



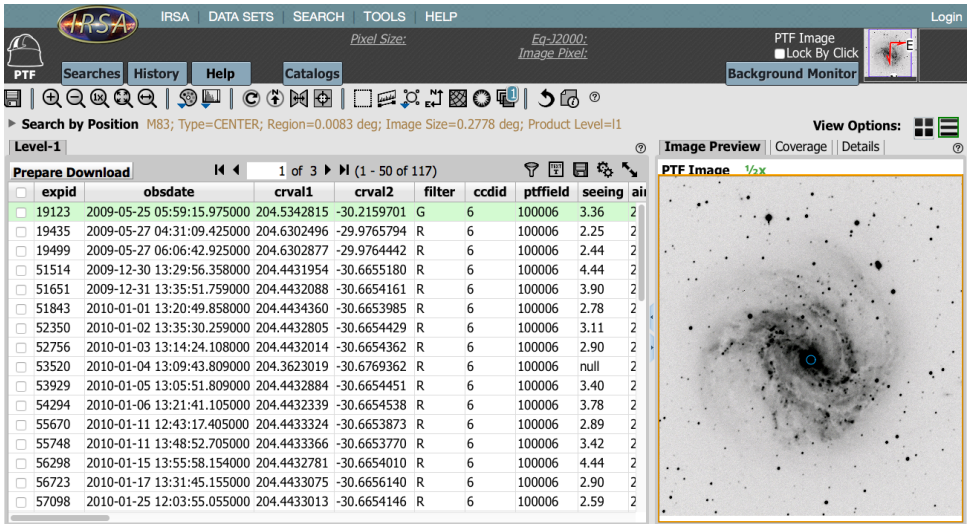
Primary Deliverables

1. **Instrumentally calibrated, epochal image products, bit-masks, source catalogs, PSFs, and difference images**
2. **Raw image data and image calibration products used in pipelines**
3. **Reference images (co-adds) from combining (1):** coverage maps, uncertainty maps, and source catalogs
4. **Alert (point-source event) stream** from real-time image-differencing pipeline: packetized with metadata
5. **Products to support real-time Solar System / NEO discovery and characterization:** both streaks and tracks
6. **Lightcurves & metrics from matching sources across individual epochs using (1) to beginning of survey**
7. **Quality assurance metrics, summary statistics, and survey coverage maps:** for performance monitoring
8. **Documentation:** pipeline descriptions, recipes, and tutorials on data-retrieval and analysis

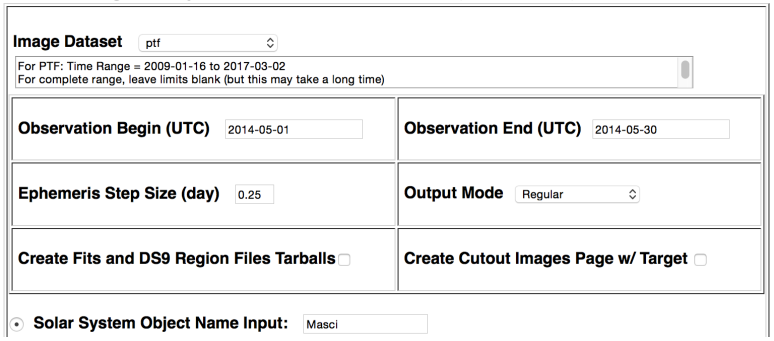
User interfaces to retrieve/analyze archive products

- Will allow search by position, time-windows, filtering on metadata, object name, interactive manipulation, catalog overlays, visualization, basic analysis of lightcurves with periodogram service.
- Accompanying APIs (command-line driven retrieval) also available

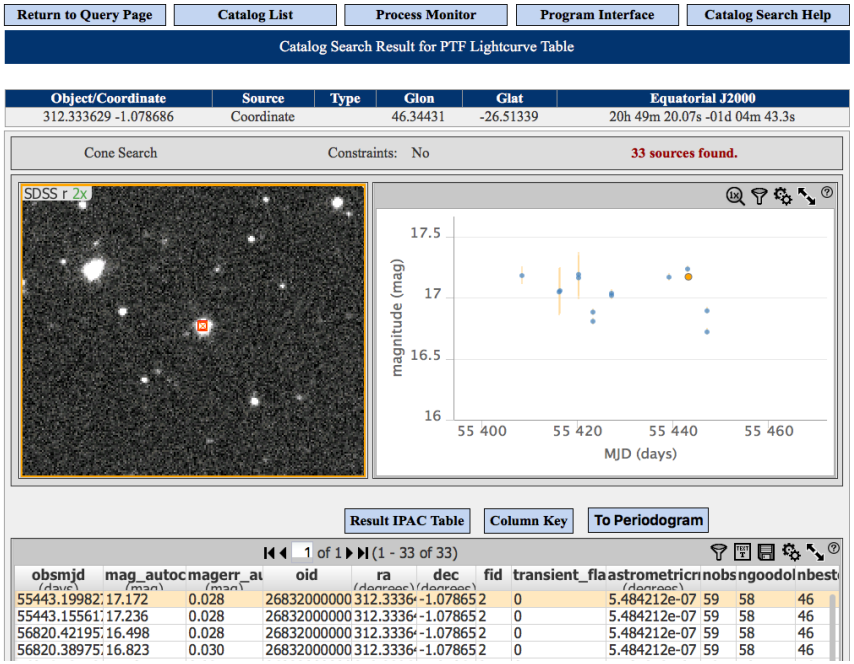
Image viewer and file-product retrieval



Moving Object Search Tool (MOST)



Lightcurve viewer/analyzer and retrieval



Status of sub-systems

Component Group	DS Component	Status 2017-05	Status 2018-07	Notes
Transfer	Data Transfer Software Protocols from P48 to IPAC	100%	100%	
Pipeline	Ingest, CCD quadrant-splitting, floating bias correction	100%	100%	
	Calibration generation: biases, high & low-v flats	100%	100%	
	Insrumental calibration (astrometric & photometric)	100%	100%	
	Reference image generation (co-addition)	100%	100%	
	Source-matching & photometric corrections for lightcurves	100%	100%	
	Transient event discovery	100%	100%	
	Machine-learned vetting of transient events	0%	100%	ML Module integrated. Parameters being tuned.
	Pipeline executive: job scheduling/task orchestration	30%	100%	Completed start of commissioning
	Throughput testing: algorithm & cluster optimization	5%	100%	Completed end of commissioning
Archive	Image and catalog file product server	30%	100%	Completed start of survey operations
	Lightcurve retrieval service w/metadata	5%	100%	Completed July 2018
Depot	Event metadata	100%	100%	
	Stamp-image cutouts	100%	100%	
	Pipeline QA metrics	100%	100%	
	User access/server setup	0%	100%	Completed start of commisioning.
Alerts	Transient alert distribution infrastructure & interfaces	2%	100%	Alerts generated and available for brokers starting June 4, 2018

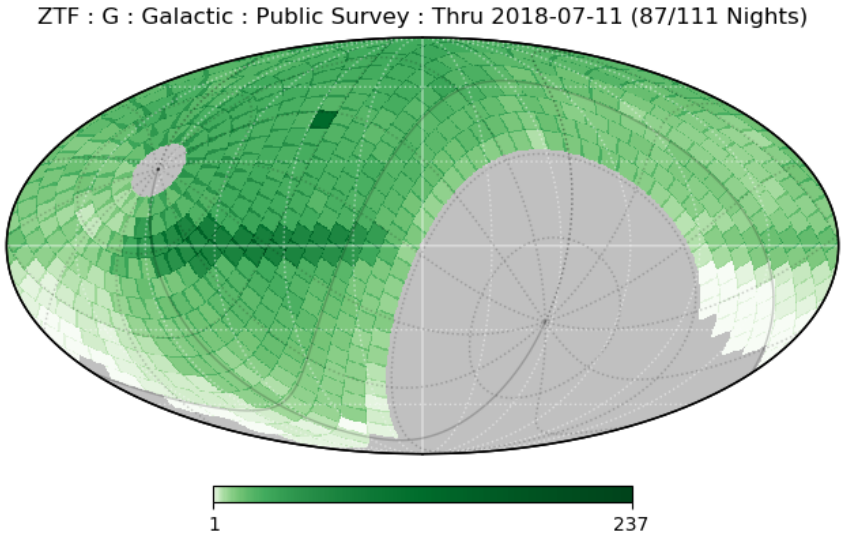
Processing architecture & hardware

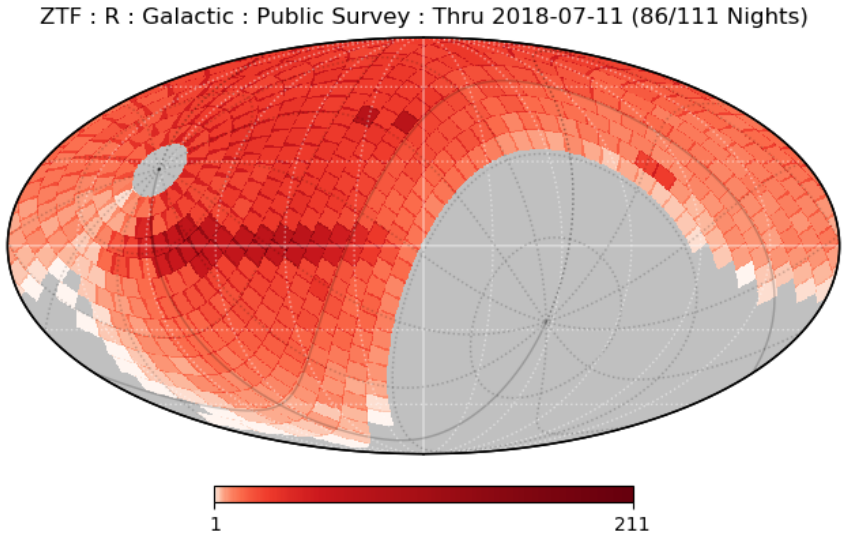
- Compute cluster consists of 66 2.5GHz nodes to support parallel processing: 1192 processor cores
- Expect ~ 3.3 PB in data products at end of survey
- *Postgres* databases are used throughout
- Parallel fileservers also used to distribute I/O load

Racks containing 66 compute nodes

Archive fileservers/disk arrays (holds up to 4 PB)

Sky coverage: public program only
Mar 17 (science ops start) – Jul 11, 2018





Accumulated data volumes and statistics
Mar 17 (science ops start) – Jul 11, 2018

- Number of raw *on-sky* camera exposures ingested: 25,149 (*g*), 27,403 (*r*), 2,126 (*i*) [~ 71 TB uncompressed]
- Number of archived epochal science image products from all CCD quadrants: 3,223,995 [~ 238 TB]

- Number of epochal science image PSF-fit extractions: ~ 45 billion
- Number of epochal science image aperture-based extractions: ~ 27 billion

- Number of reference (co-added) image PSF-fit extractions ("seeds" for lightcurves): ~ 1.27 billion
- Number of reference (co-added) image aperture-based extractions: ~ 0.4 billion

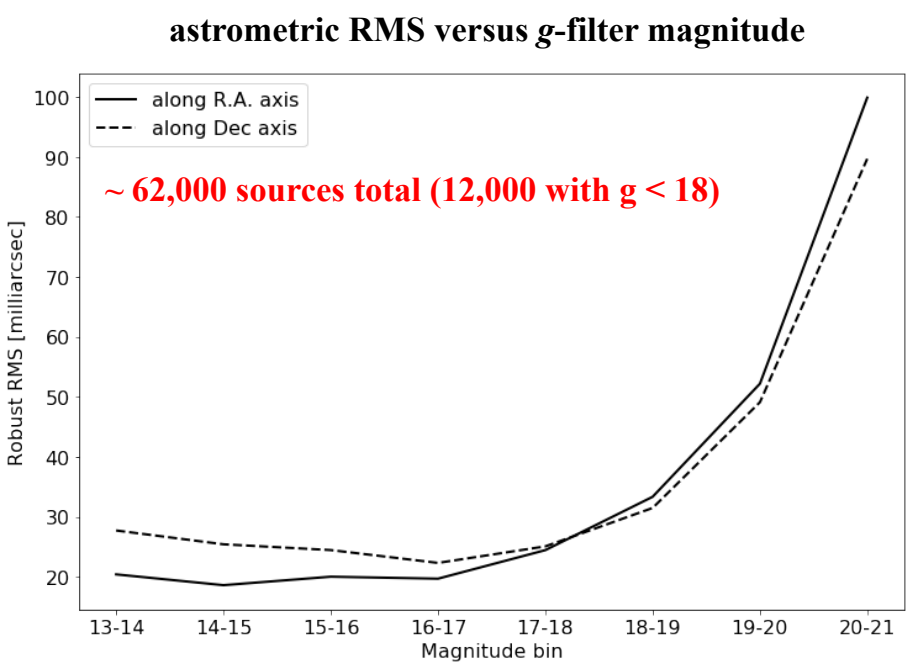
For nominal three-year survey:

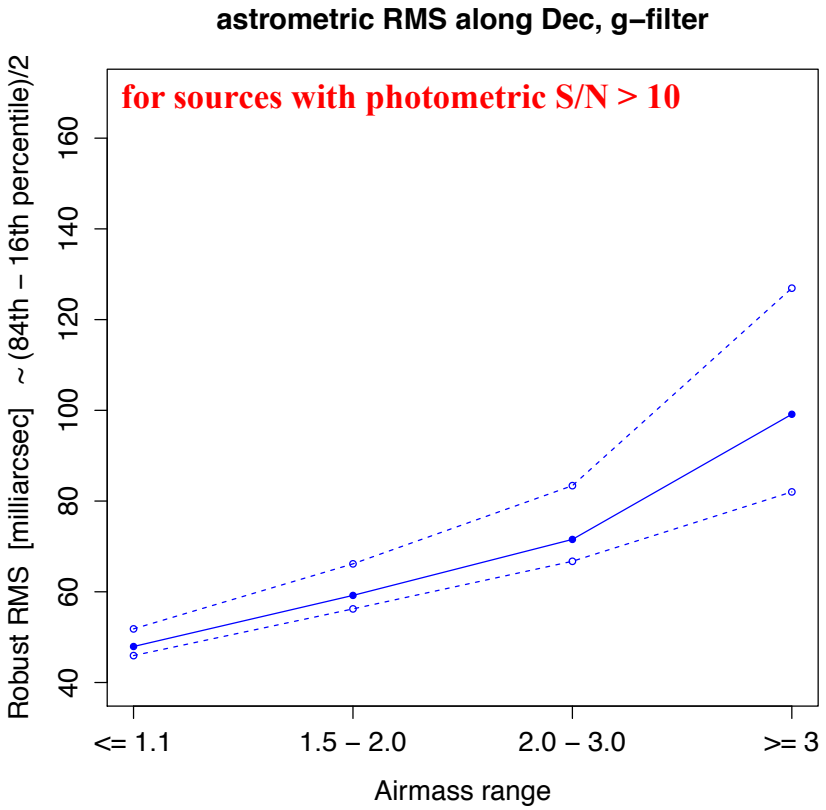
(number of "good" observing nights / year: ~ 260)

- Volume of data products: ~ 3.3 PB
- Number of single-exposure extractions: ~ 600 billion (PSF-fit based)
- Number of reference images (co-adds in static library): ~ 282,000 (~ 55 TB)

Astrometric performance relative to Gaia

- Astrometric precision of bright stars with *r, g* < 18 mag at airmass < 1.2 is <~ 30 milliarcsec (RMS per axis)
- Accuracy for sources with S/N > 10 (*g, r* < 20 mag) at airmass < 2 is <~ 65 millarcsec

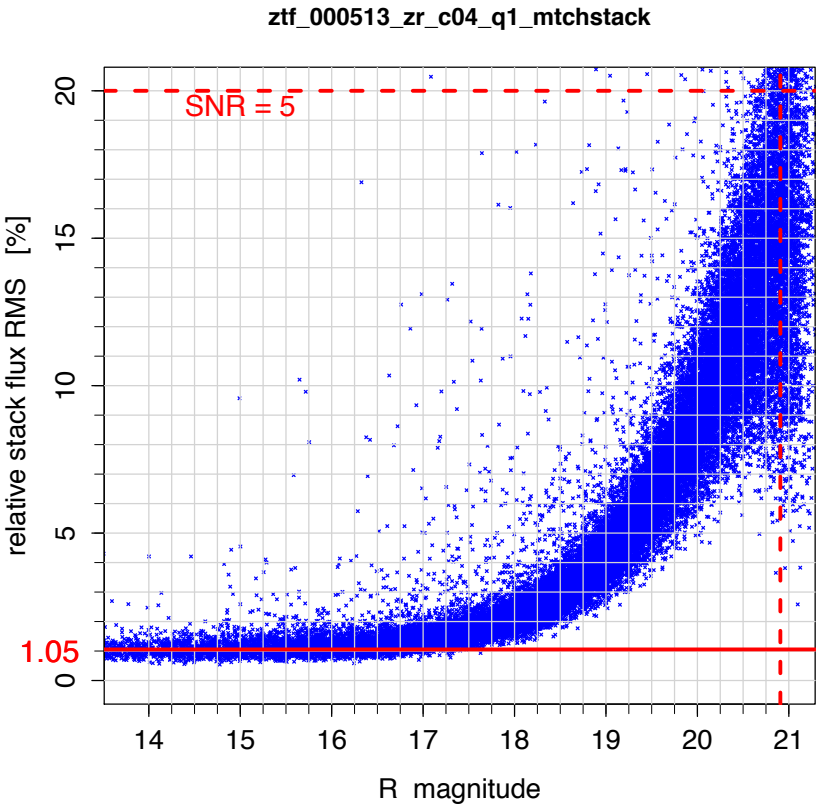




Photometric precision (repeatability)

- From matching epochal PSF-fit source catalogs: typical range is ~ 8 to 20 millimag; depends on airmass
- 5-σ limiting depths are consistent with expectations and photometric uncertainties in PSF-fit catalogs

galactic plane field



high galactic latitude field

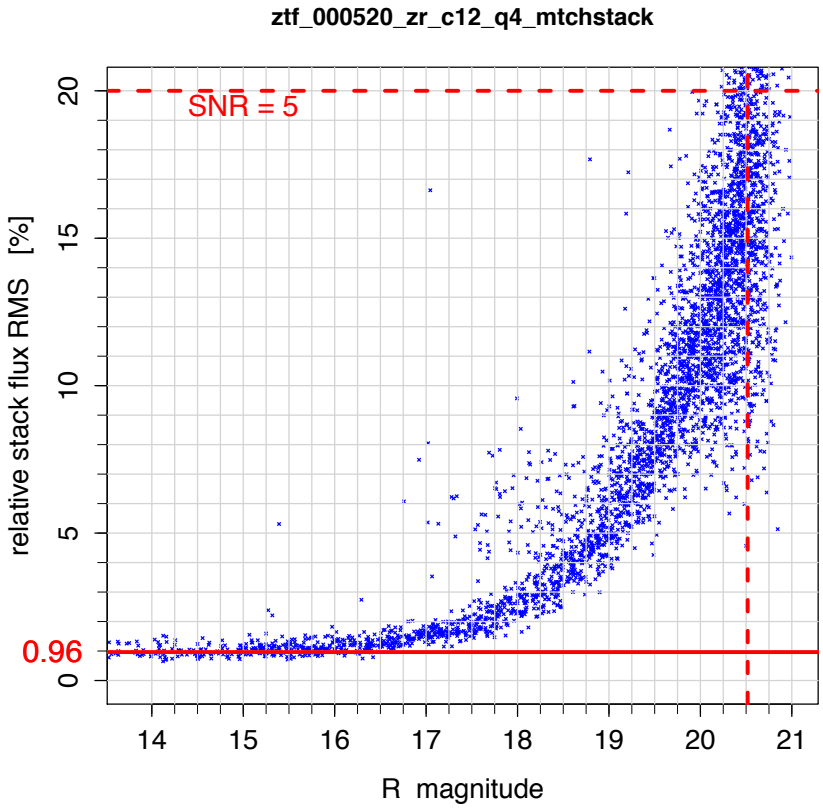
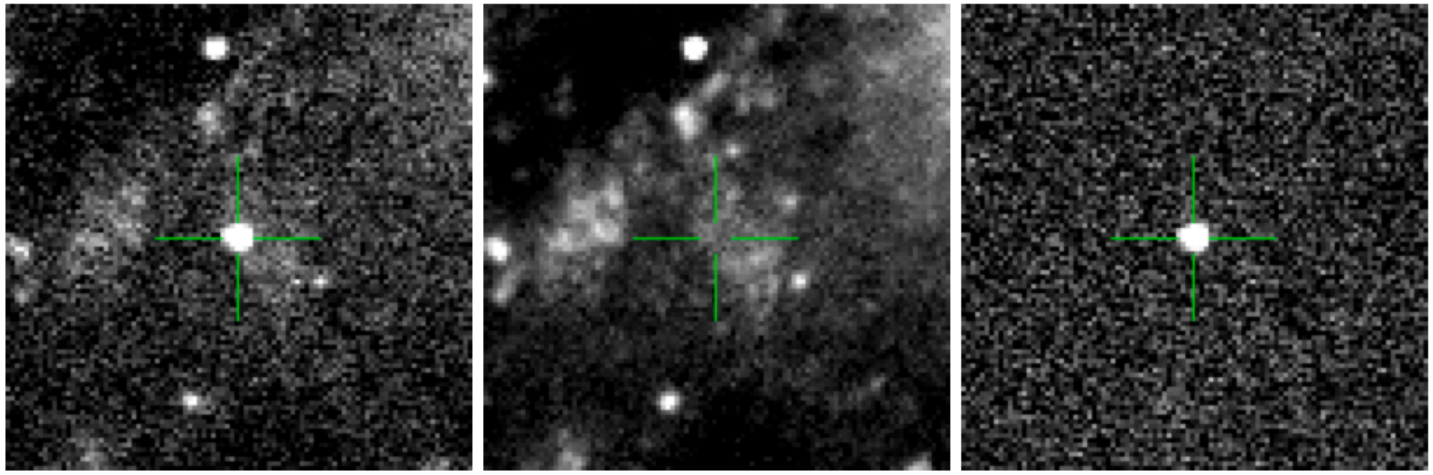


Image Differencing & Event Extraction

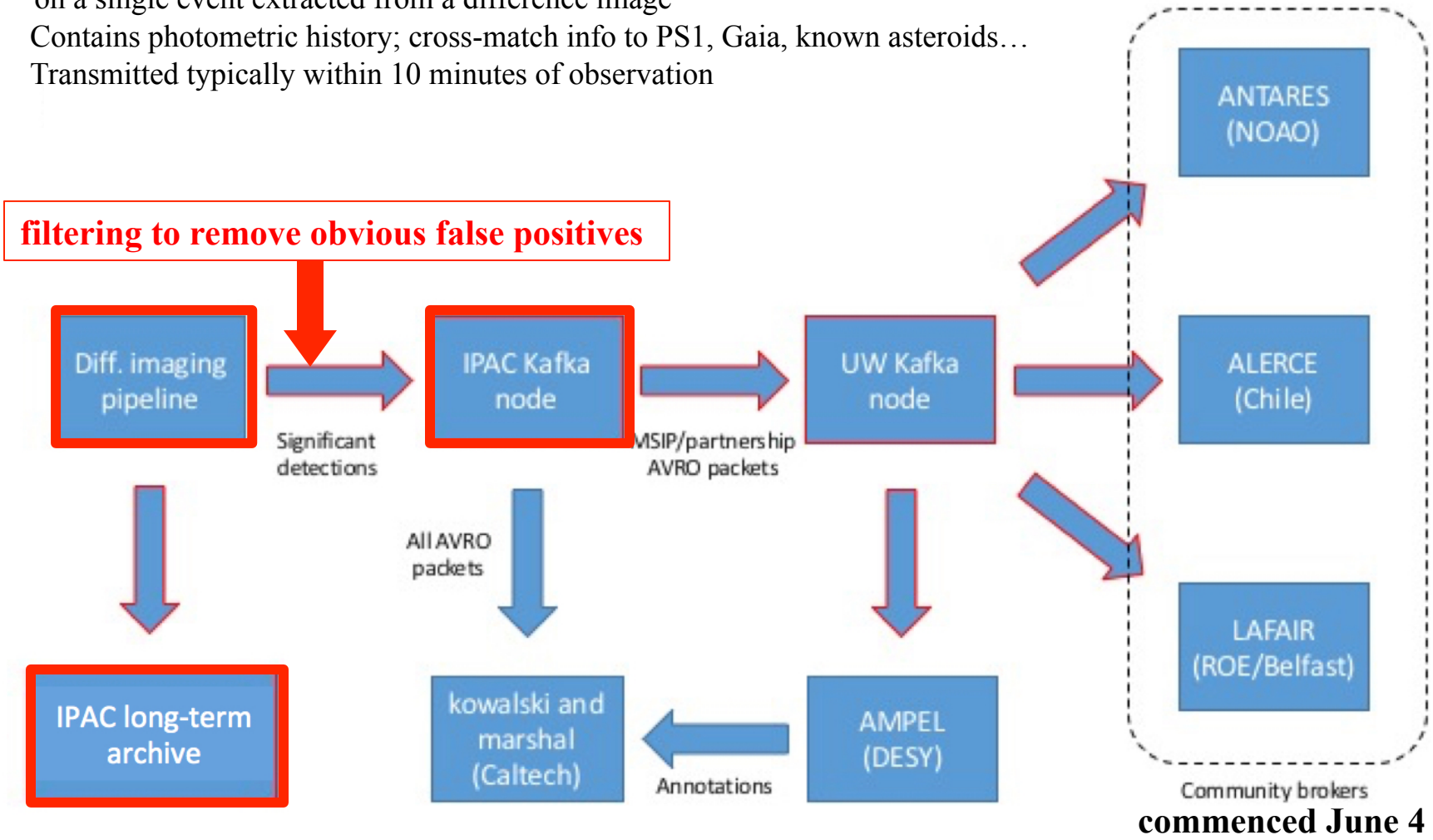
- We difference repeated images of the sky against a reference image (= co-add of historical images)
- Events are then extracted above a signal-to-noise threshold of 5-σ
  - ➢ can be triggered from any flux-transient/variable source, moving object, or occasional artifact
- Events are filtered to remove obvious false-positives (image-artifacts)
- Then used to generate "alert packets" (next slide)

New image **Reference image** **Difference image**



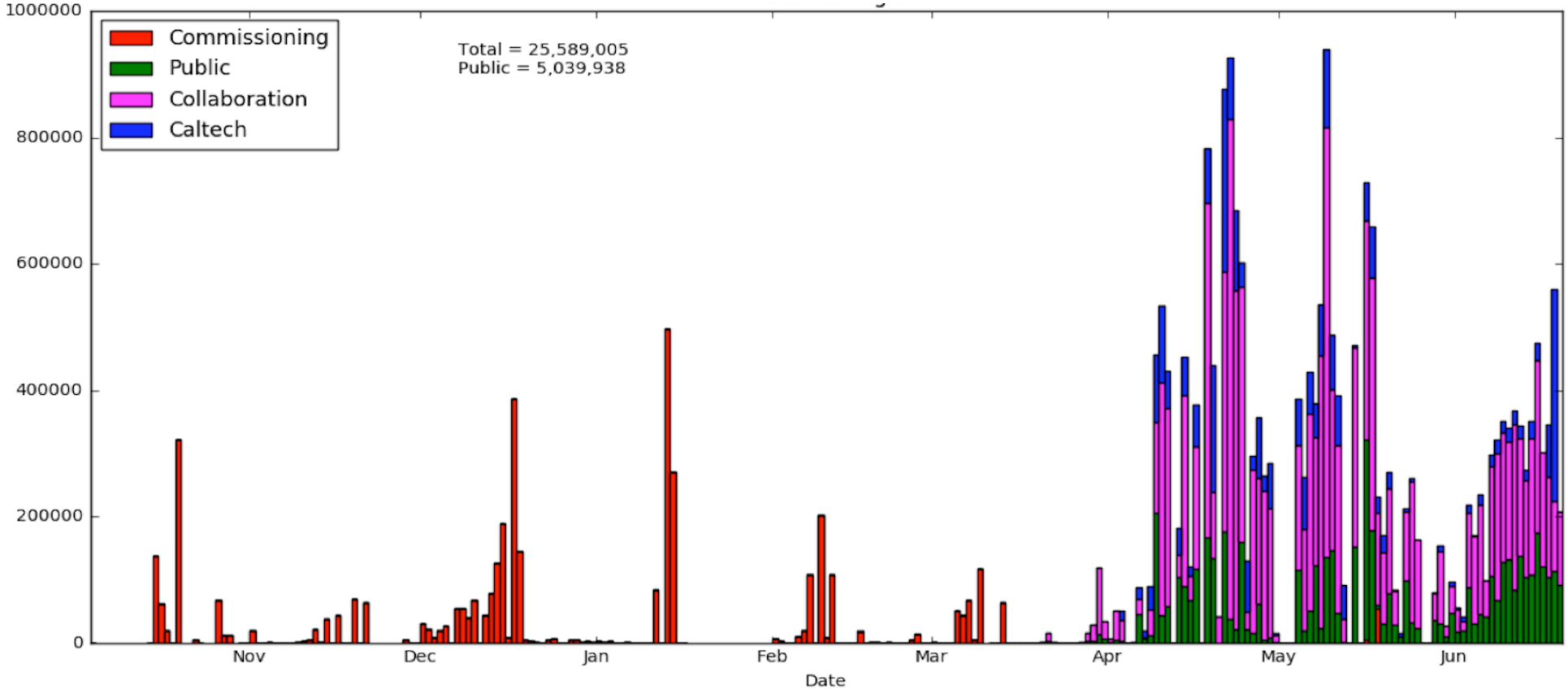
Alert Packet Distribution

- **Alert packet:** self-contained file containing metadata and contextual information on a single event extracted from a difference image
- Contains photometric history; cross-match info to PS1, Gaia, known asteroids…
- Transmitted typically within 10 minutes of observation



commenced June 4

Alert History (number per night)

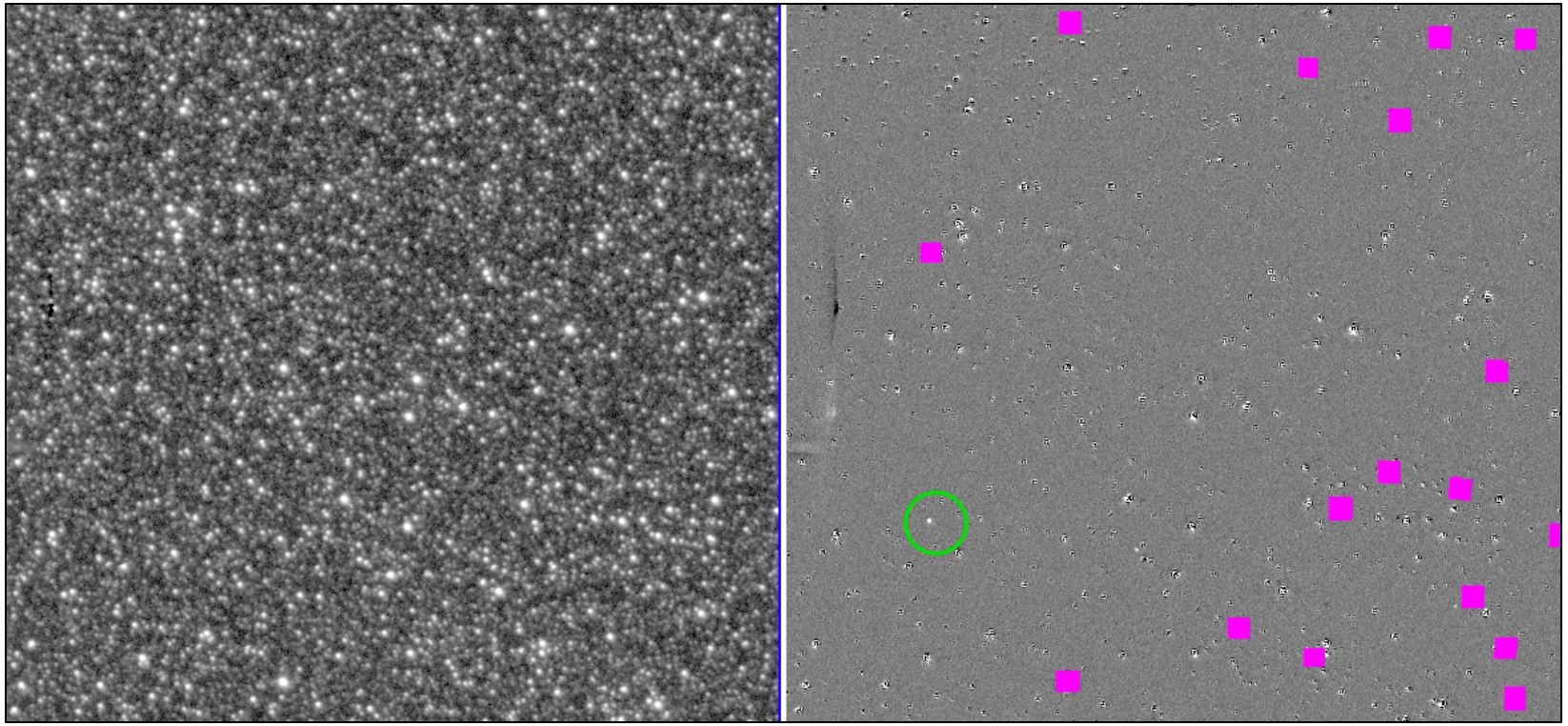


- Since June 4 (start of public alerts), total number of public alerts generated = 3,775,116
- Number of alert packets associated with known Solar System objects (≤ 3 arcsec): ~ 172,000
- In accord with design expectations

Image differencing / alerts from galactic plane

- Very little known over large swath of galactic plane from previous surveys
- Alerts are mostly from variable stars, eclipsing binaries, novae, and asteroids that happen to cross

New image **Difference image**



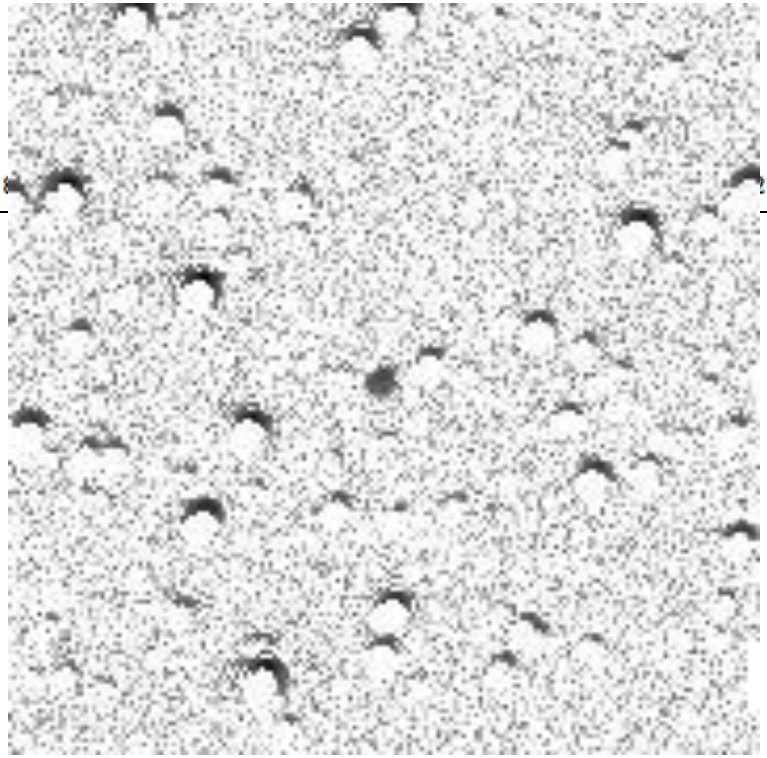
- One typically combines the photometry from many alerts to create lightcurves in order to characterize variability

Asteroids can dominate alerts(!)

2018

Code	# PHA	# NEA			#Atens			#Apollos			#Amors		
		All	1km	H<22	All	1km	H<22	All	1km	H<22	All	1km	H<22
G96	23	442/	0/	101	40/	0/	4	251/	0/	46	151/	0/	51
F51	8	204/	1/	60	10/	0/	3	98/	1/	30	96/	0/	27
703	5	124/	2/	22	12/	1/	1	89/	1/	13	23/	0/	8
C51	2	10/	0/	8	0/	0/	0	6/	0/	5	4/	0/	3
T08	1	15/	0/	4	2/	0/	0	9/	0/	1	4/	0/	3
T05	1	13/	0/	4	0/	0/	0	9/	0/	2	4/	0/	2
I41	1	10/	0/	1	3/	0/	0	7/	0/	1	0/	0/	0
D29	1	2/	0/	1	0/	0/	0	2/	0/	1	0/	0/	0
568	0	6/	0/	0	0/	0/	0	3/	0/	0	3/	0/	0
309	0										3/	0/	1
F52	0										1/	0/	1
Q66	0										0/	0/	0
Y00	0										0/	0/	0
T14	0										0/	0/	0
807	0										1/	0/	1
Total	42										290/	0/	97

ZTF



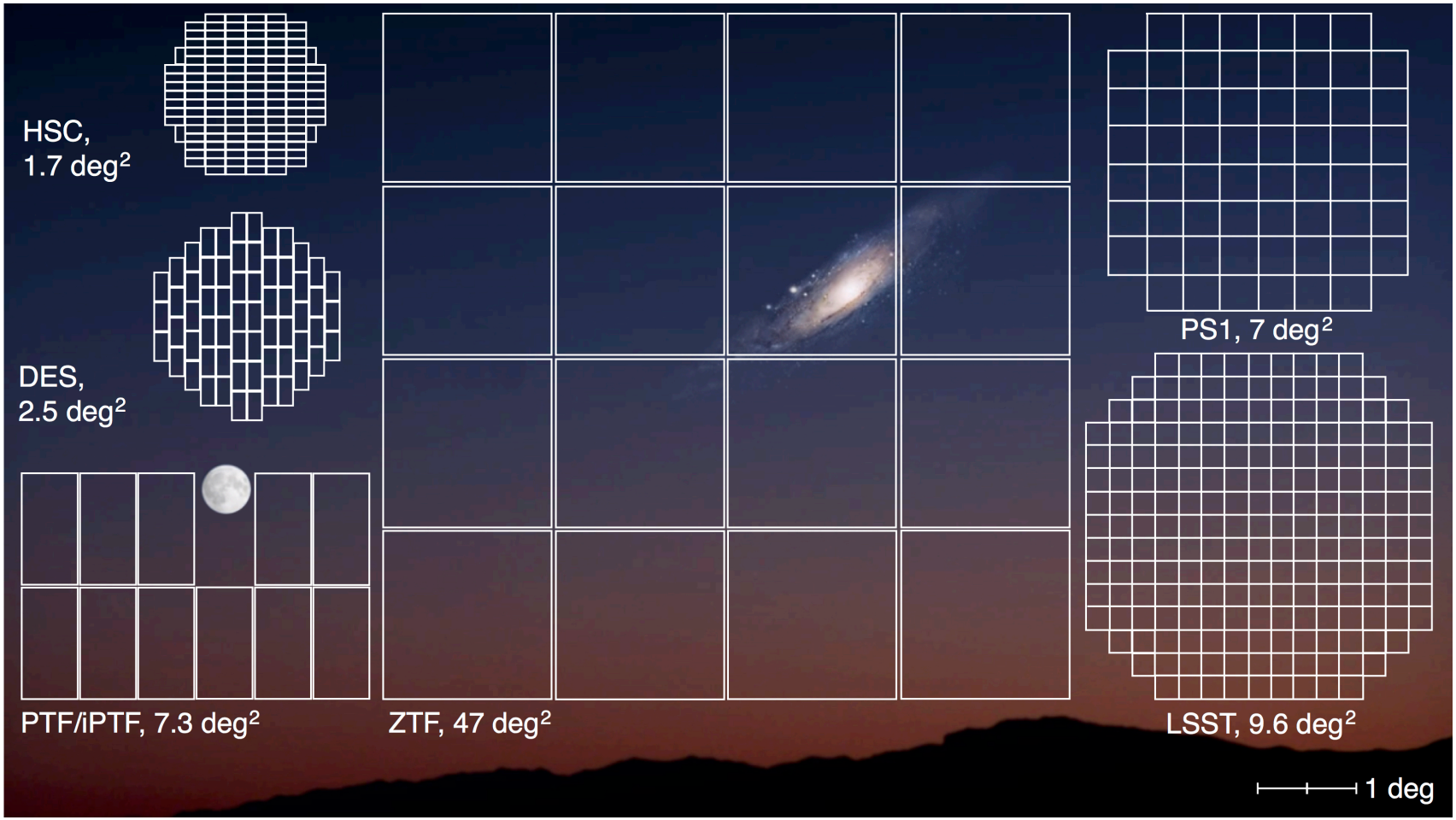
Movie of 80 difference images from a deep drilling Galactic plane field

Summary

- ZTF is on the sky; science operations commenced on March 17, 2018
- Feedback received from Science Working Groups in partnership was crucial for development
- Now settled into steady-state automated operations
- IPAC is satisfying primary requirement: processing > 95% of images in < 10 minutes
- Astrometric accuracy with respect to Gaia is < 65 milliarcsec (S/N > 10)
- Photometric precision from repeatability is ~ 8 to 20 millimag
- Archive system and services for retrieval all in place; ready to support first public data release
- Public alerts now available for dissemination by community brokers (since June 4)
- **Reminder:** public alerts commenced ~2.5 years earlier than planned; developed on short timeframe
- Scientific results are starting to proliferate

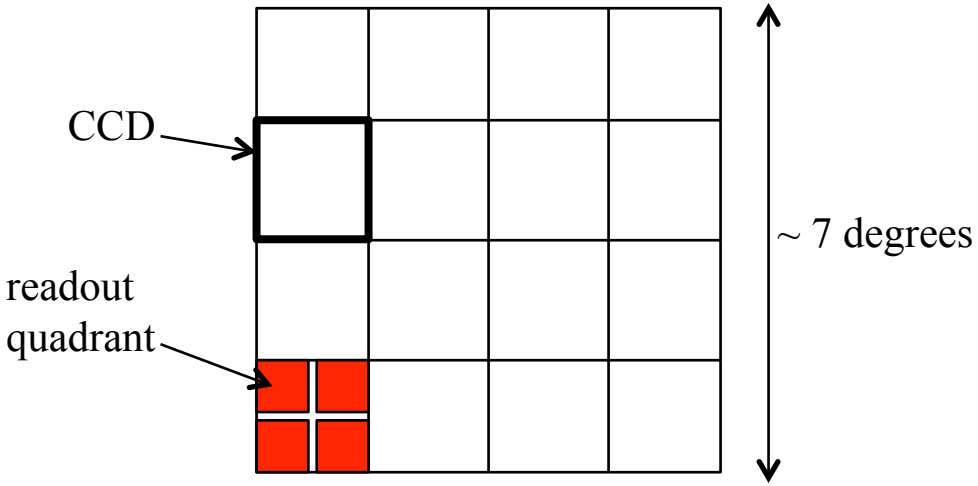
Back up slides

Comparison to other survey cameras



Raw Camera Image Data & terminology

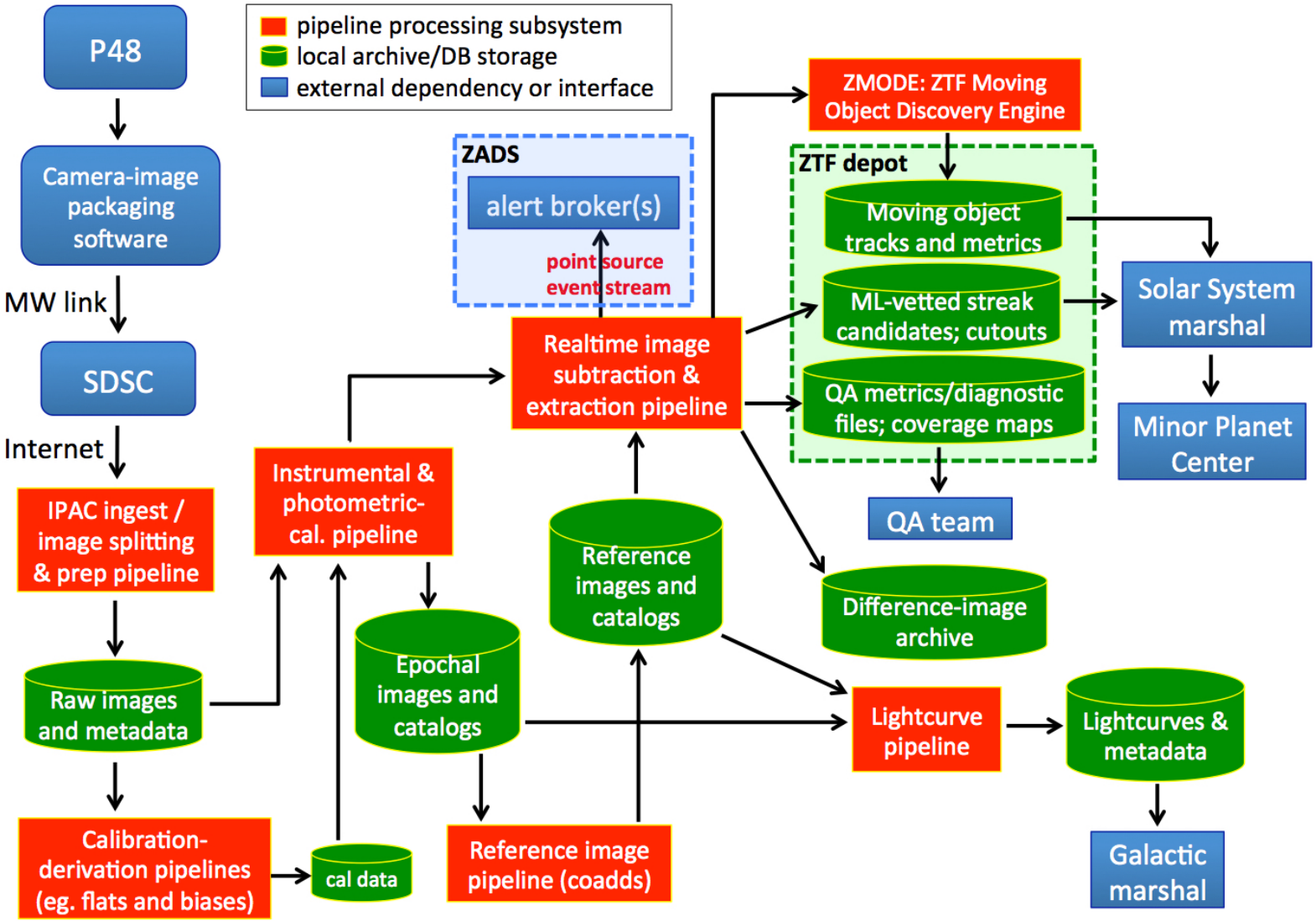
- One camera exposure: 16 CCDs; each ~ 6k x 6k pixels
- Image data packet transmitted is per CCD (four readout-quadrant images & overscans)
- 16 CCD-based image files are acquired every 45 sec (30s exposures)
- Full camera exposure = one survey field on sky: ~ 1.3 GB (no compression)



Basic image-unit for pipeline processing from which all file-based products are derived is a ~3k x 3k readout quadrant:

- processed (epochal) science images
- reference image (co-adds)
- difference images
- source extraction table files
- lightcurve (source-match) files

Data and processing flow



Accomplishments over last six months

- Infrastructure and software for distributing Alert Packets: includes Kafka; hardware; UW interfacing
- Alert packet schema and contents stabilized following feedback from science working groups
- R&D on filtering of difference-image events for alert packets to mitigate obvious false-positives
- Improved quality of differential photometry in alert packets with meaningful uncertainties
- Solved depth-issue for alerts generated from deeper (300 sec) exposures to support ToOs
- Refinements to PS1 Star/Galaxy scores for associating with alerts
- Long-term archiving of alert packets at IRSA – now subject to same user/*programID* access policies
- Tuned/optimized Moving Object detection pipeline that links events to create tracklets (*ZMODE*)
- Tuned/optmized Fast-Moving Object (streak) detection pipeline (*ZSTREAK*)
- More accurate reporting of known asteroids to associate with alert streams and *ZMODE* output
- Reporting of known comets to associate with alert stream and *ZMODE* output
- Ghost prediction and masking (both co-moving and counter-moving types)
- All ancillary file products now downloadable through GUI
- Lightcurve (*matchfile*) products from linking epochal image extractions now routinely made
- Lightcurve (*matchfile*) products containing only partnership data now also made
- Lightcurve query GUI and Time Series / visualization tool now ready for partnership
- Automated reference-image generation (more later)
- Better real-time reporting of QA metrics, pipeline status and failures for Observing System team
- Data System documentation up to date and paper submitted to PASP

Task List and Refinements

In progress (tied to baseline design), completion: ~ end of Sept 2018:

- Enable image-cutouts on archived (compressed) difference images – IRSA service (now available for other images)
- Synopsis of reference image holdings: what are we missing and why?
- Regenerate reference images to higher (uniform) depth to support aLIGO / Virgo runs in October 2018
- Automated generation of all-sky coverage maps for reference images
- Improve subtractions in "challenging" (deep-drilling) fields in Galactic plane (previous slide)
- Faster delivery of data quality metrics to Observing System team in realtime (in lieu of processing latencies)

Near future (pending analysis and contingent on resources):

- Correct dome flats for edge / scattering / CCD-etching effects prior to stacking
- Star-flat assessment and application (DESY group input)
- Exposure-time correction map (flat augmentation, ~ 0.2% at edges)
- *i*-filter fringe correction (DESY group input)
- Astrometric corrections at high airmass (>~ 3) to support ToOs

Ongoing / ad-hoc:

- Continued refinements to point-source and streak real-bogus classifiers
- Improved S/G classification scores for PS1 to associate with alert streams
- Update to Gaia DR2 (for both astrometric calibration and alert association)

Additional functionality (pending approval)

- Forced photometry service using image archive
  - will implicitly include more accurate estimation of upper-limits for non-detections (prohibitive in production)
- Fake transient injection pipeline and infrastructure
  - design specifications received
- Sentinel service (for monitoring targets of interest using archived products)
  - can the lightcurve query service accommodate this?
- Set-up of "sandbox" environment for analysis, testing, and prototyping; pending MOU on usage/data-access
  - primarily to support calibration-related tasks on previous slide

Expectations prior to survey start: data volumes and statistics

Estimates per night:

(based on an average on-sky duration of ~ 8hr 40min)

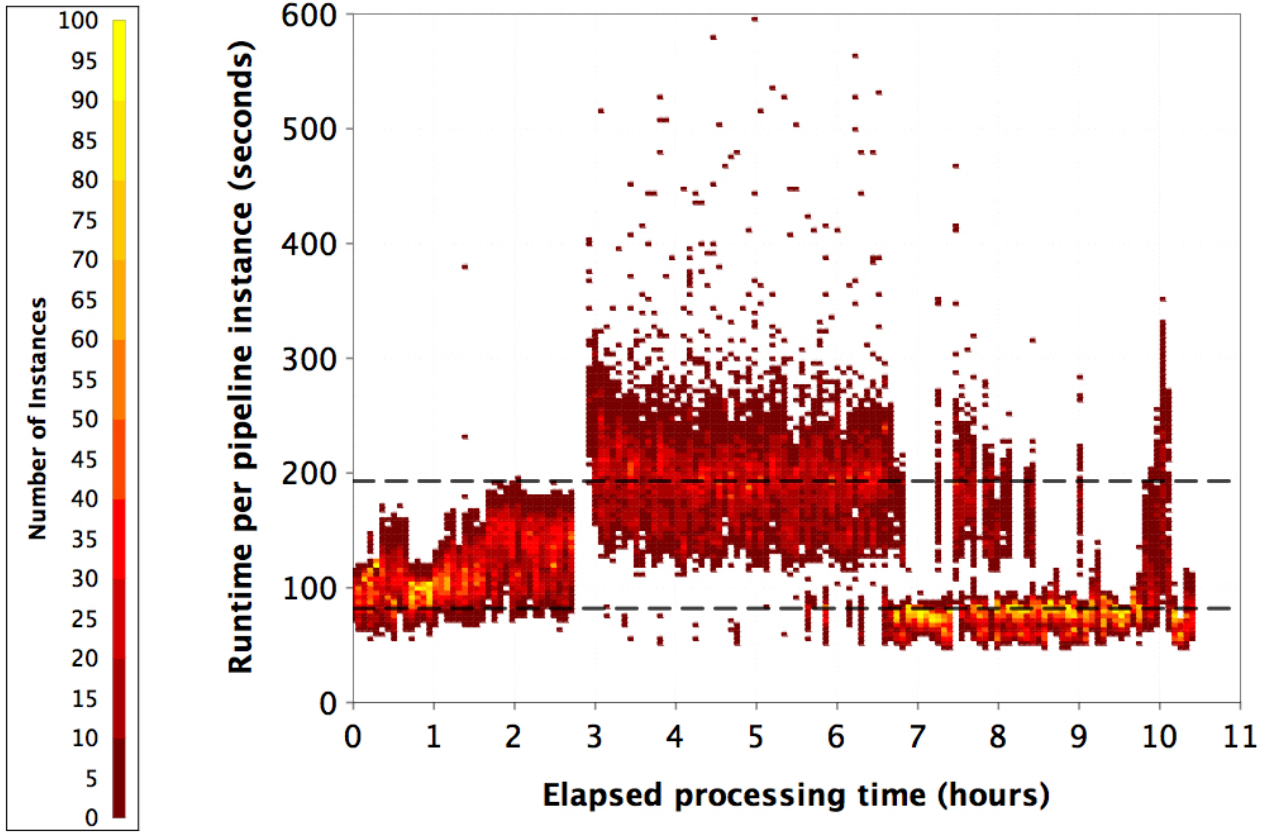
- Number of on-sky camera exposures per night: ~ 700
- Raw image data volume: ~ 0.8 TB (no compression)
- Raw incoming data rate: ~ 230 mega bits per second
- Data product volume: ~ 3.5 TB per night (real-time products only; seasonal)
- Number of unvetted **point source** difference-image alerts (flux and motion-induced): <~ 1 million
- Number of streaks (candidates for "fast-moving" objects following ML vetting): <~ 1000
- Number of single exposure extractions: ~ 700 million (PSF-fit based): sky location dependent.
- Number of single CCD-quadrant image products (science, difference, mask, catalogs): ~ 230,000

For nominal three-year survey:

(number of "good" observing nights / year: ~ 260)

- Volume of data products: ~ 3.3 PB
- Number of single-exposure extractions: ~ 600 billion (PSF-fit based)
- Number of reference images (co-adds in static library): ~ 282,000 (~ 55 TB)

ZTF real-time pipeline runtime
processing unit = one CCD quadrant image

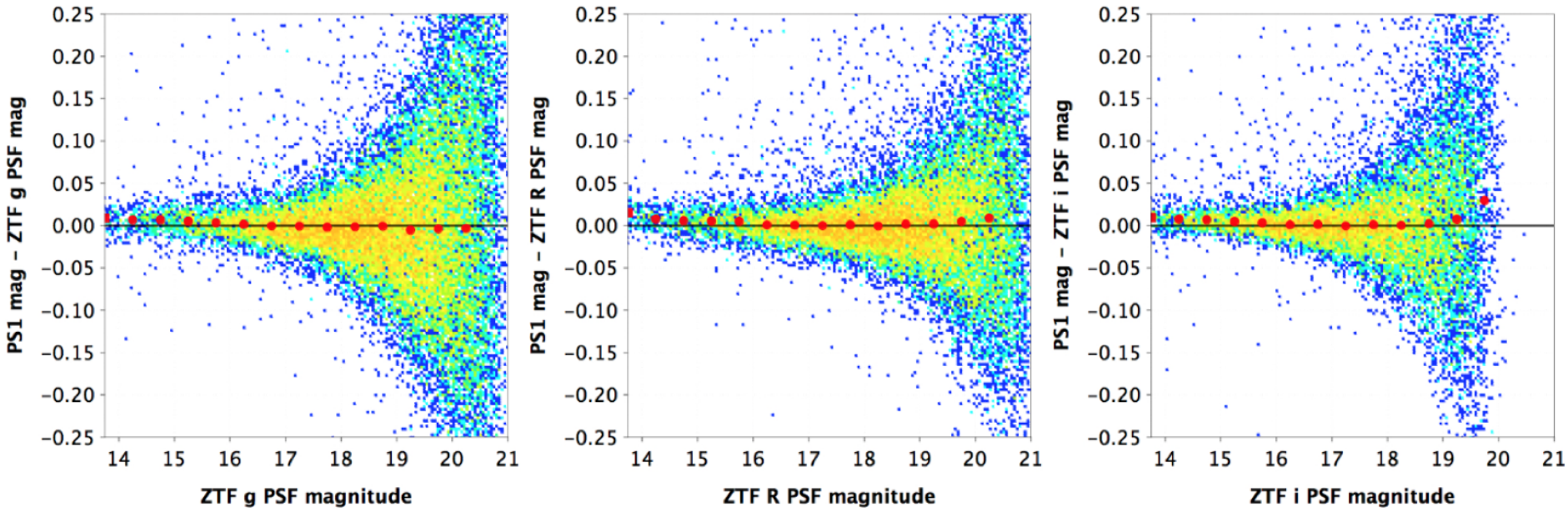


- 66 machines × 8 jobs each
- Based on fields processed on night of March 9, 2018 (UT)
- High tail: with image-diff pipeline, alert generation etc since ref images were available: <~ 4 minutes
- Low tail: science image processing only (no ref images available): <~ 2 minutes

Photometric calibration

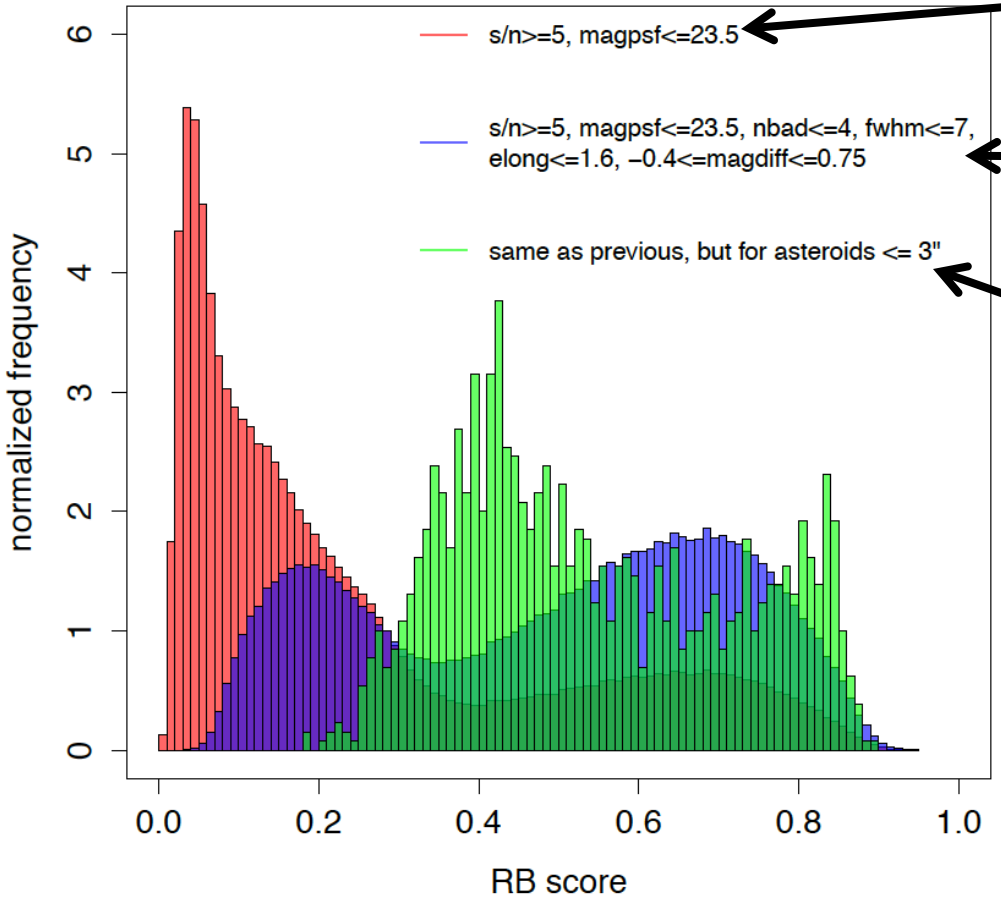
- Performed using the Pan STARRS1 catalog
- Achieved accuracy for bright sources is <~ 2.5% (RMS) relative to PS1

Below are from quadrant-based PSF-fit catalogs; all in the galactic plane (|b| <~ 8°)



Alert filtering and *RB*-score distributions

Below are distributions of *RB* score for difference-image-detected events from only the **public survey** from three recent nights (more on this in next presentation).



All raw events (number = 643860)

"Light" filtering prior to all alerts leaving IPAC (number = 228287)

Alerts associated with known asteroids (number = 1300)

Staffing Plan, Tasks and FTE breakdown

- We are currently in the Data System Verification phase; this will continue until **July 31, 2018**
- This phase will include possible low-level pipeline development, bug fixes, and tuning in response to science analyses, pending CCB approval (see slide 9)

Sep 2017 → Jul 2018

Data System Task	Dev	Commissioning, SV & DS Verification	Nominal Ops
Task management and reporting to project; respond to help-desk; budgeting; costing; documentation;	0.50	0.40	0.30
Pipeline upgrades, optimization tweaks, tuning	2.70	0.35	
Archive development, user-interfaces, and services	1.50		
Simulation, QA, on-sky performance trending with feedback to pipeline developers	0.15	0.15	
Database administration (archive and pipeline DBs)	0.20	0.30	0.10
Ongoing PTF / iPTF reprocessing	0.20		
Pipeline maintenance and operations: specifically pipeline operator tasks, reprocessing, monitoring		1.0**	1.50
Archive ingest and IRSA-related operations: manage archive volumes, tools, services, docs, help-desk		1.00	1.00
System admin: maintenance, monitoring, install/ patching of hardware & system software; backups	0.50	0.50	0.25
TOTAL	**5.75**	**3.70**	**3.15**

** was 1.5 FTE prior to Dec 31, 2017

ZTF Science Data System Staff

- **Ben Rusholme:** data link from P48 to IPAC; pipeline job scheduling/executive; optimization; software/configuration management; hardware config.; alert distribution infrastructure (Kafka)
- **David Shupe:** astrometric calibration; source-matching and relative photometry pipeline
- **Russ Laher:** pipeline infrastructure; integration and testing; data ingest; pipeline executive; database schemas and stored procedures; bias- and flat-generation pipelines;
- **Steven Groom (and staff; IRSA Lead):** pipeline/archive interface design; system engineering; hardware shopping/costing and provisioning.
- **Frank Masci (ZSDS Lead):** instrumental and photometric calibration; reference-image generation; image-subtraction; extraction; moving-objects; algorithms; analysis; documentation…
- **David Flynn (and staff; ISG Lead):** system-engineering and hardware
- **Ed Jackson:** database management
- **Jason Surace:** image simulation; data analysis
- **Ron Beck:** pipeline operations
- **David Imel (IPAC manager):** budgeting and personnel
- **George Helou (IPAC director)**

The ZSDS is not a clone of PTF. Developed from scratch & optimized to handle 15x higher data rate