# ZTF Science Data System Update

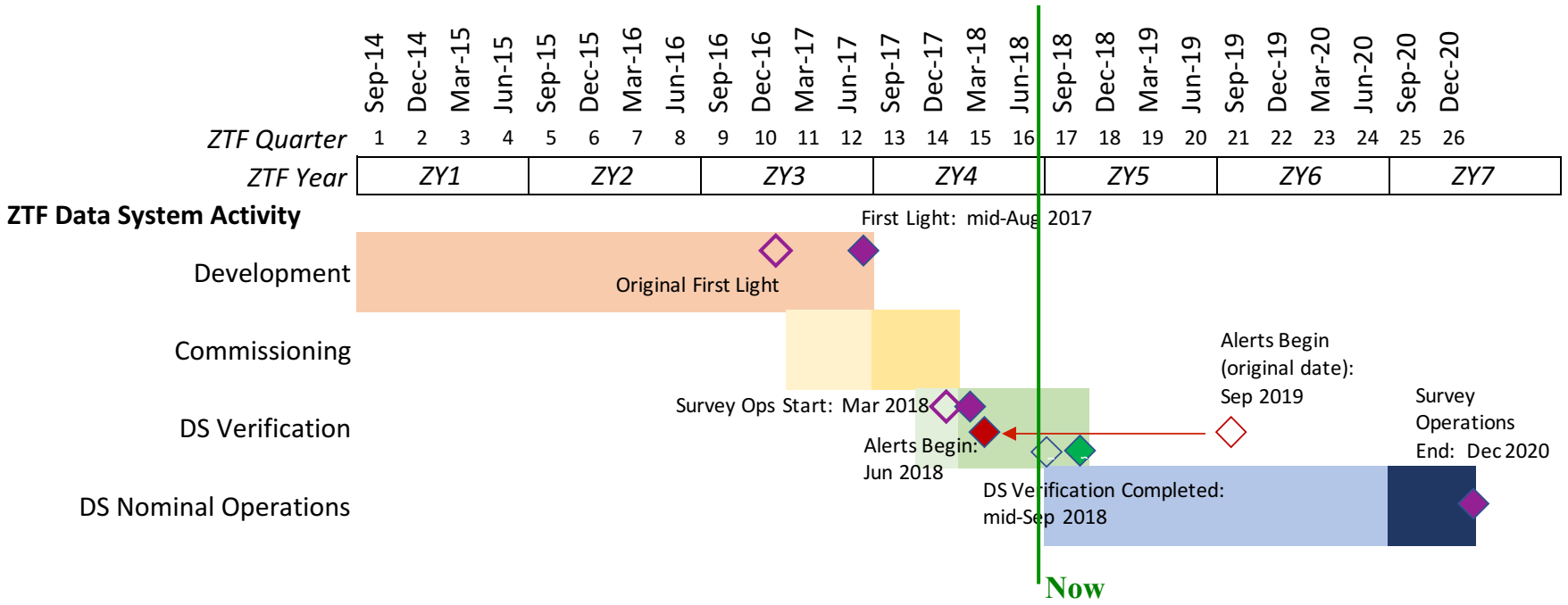
Frank Masci & the Caltech-IPAC ZTF Team

Presented by David A. Imel

ZTF Team Meeting, Stockholm, August 2018

## Contents

# ZTF Data System Timeline




- Although ZTF Survey Operations started March 17, the Data System is currently in the "DS Verification" stage:
  - Pipeline tuning based on on-sky data
  - Accumulation of epochal images for reference images.
- DS Verification will complete in September 2018.
- **Major additional task injected to the Verification phase: Alert Production.**

# Recent Accomplishments

(Since Last ZTF Team Meeting, March 2018)

# Pipeline

- **ALERTS:** Implemented—described later in this talk.
- Ghost prediction and masking (both co-moving and counter-moving types)
- Automated reference-image generation: addressing issues described in this talk.
- ZSTREAK: Optimized Fast-Moving Object (streak) detection pipeline
- ZMODE:
  - Optimized Moving Object detection pipeline that links events to create tracklets
  - More accurate reporting of known asteroids and comets to associate with alert streams and *ZMODE* output

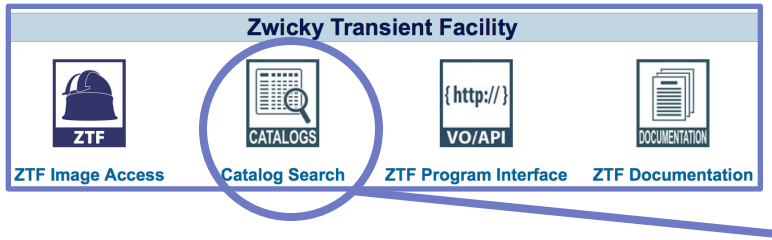
# QA & Data Products

- Better real-time reporting of QA metrics, pipeline status and failures for Observing System team
- All ancillary file products now downloadable through GUI
- **Lightcurves (Matchfiles):**
  - Products from linking epochal image extractions now routinely made
  - Products containing only partnership data now also made
  - Query GUI and Time Series Visualization Tool now ready for partnership
- Data System documentation up to date; paper submitted to PASP

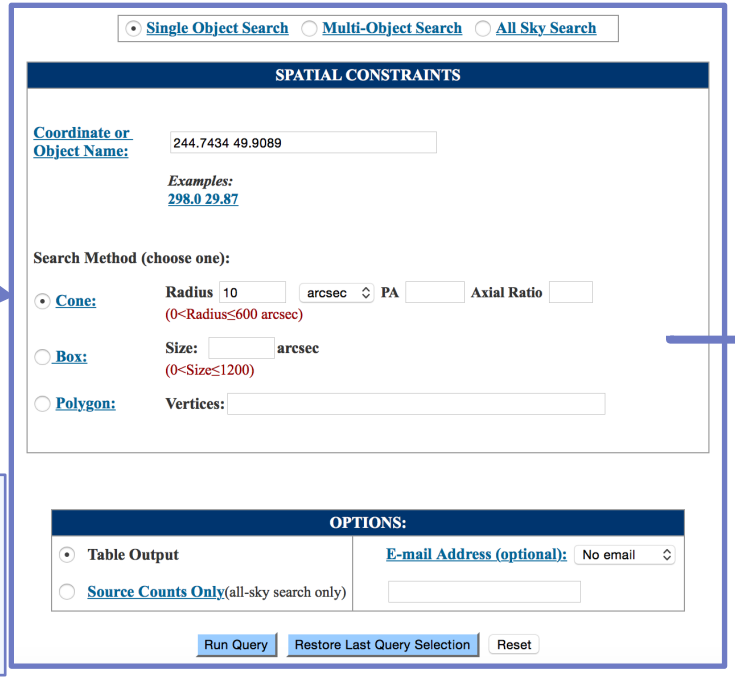
# Lightcurve retrieval GUI / Time Series Tool

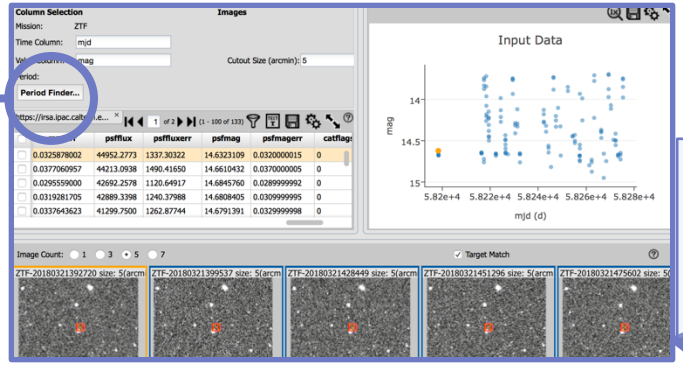
Now available through IRSA (*https://irsa.ipac.caltech.edu/Missions/ztf.html*).

**See talk by Emmanuel Joliet on Tuesday.**




**First query Objects Table**

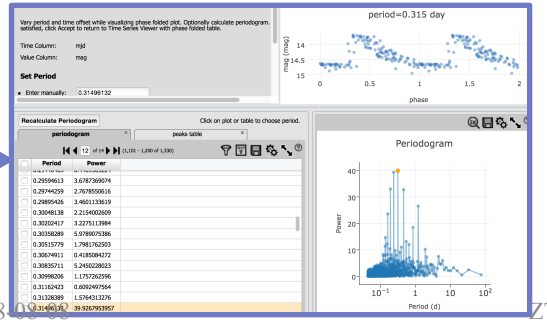


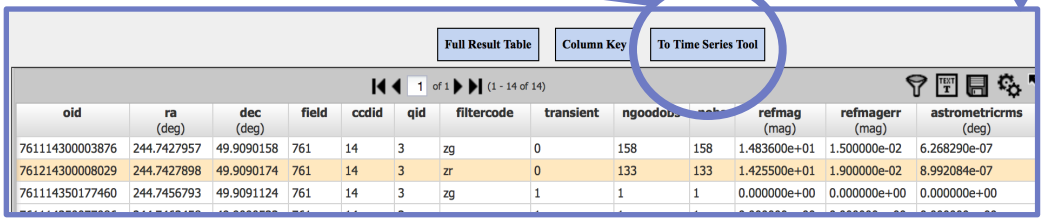




**Retrieve lightcurves; optionally download or send to Time Series Viewer**

**Results**

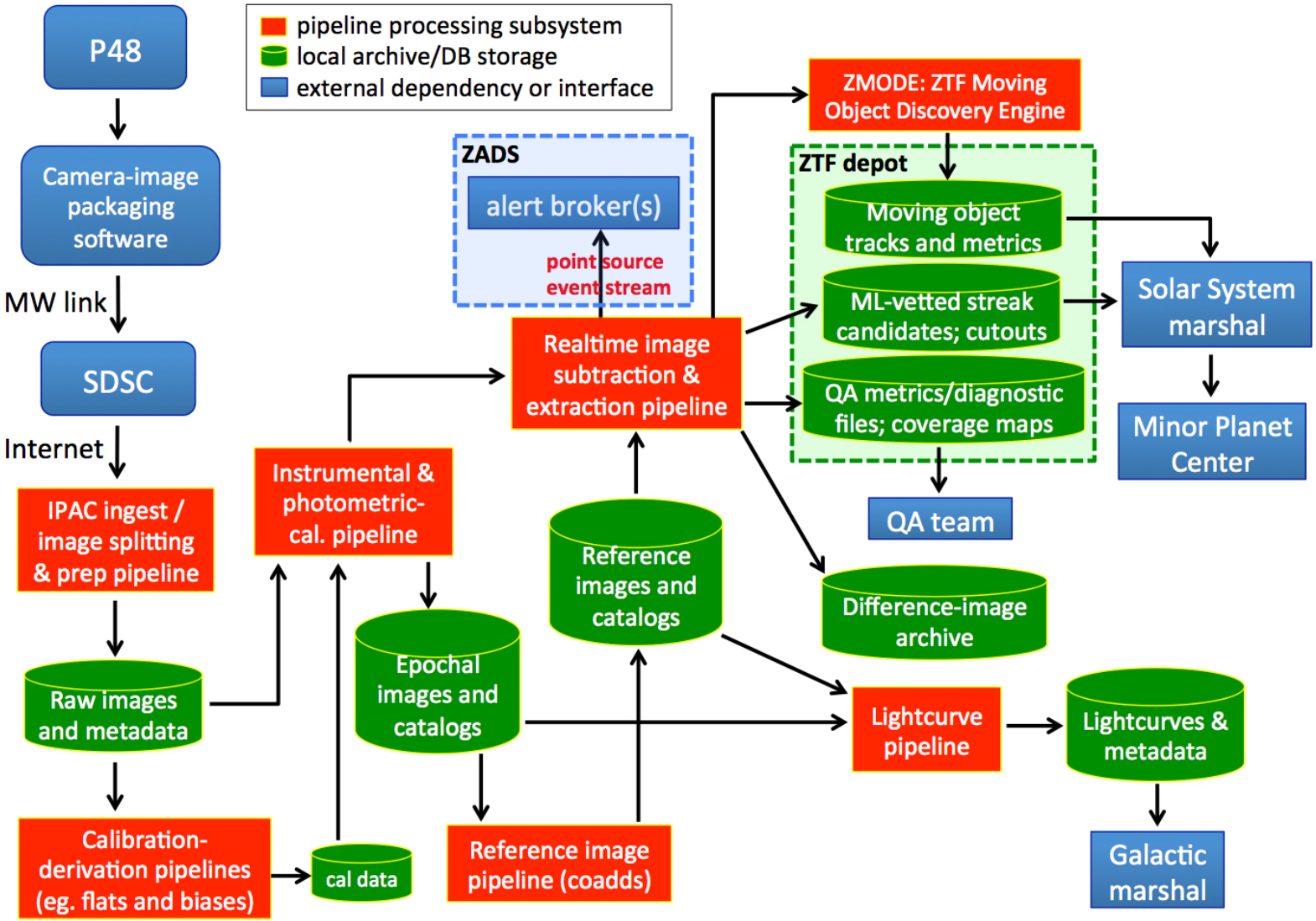
**Period finding**



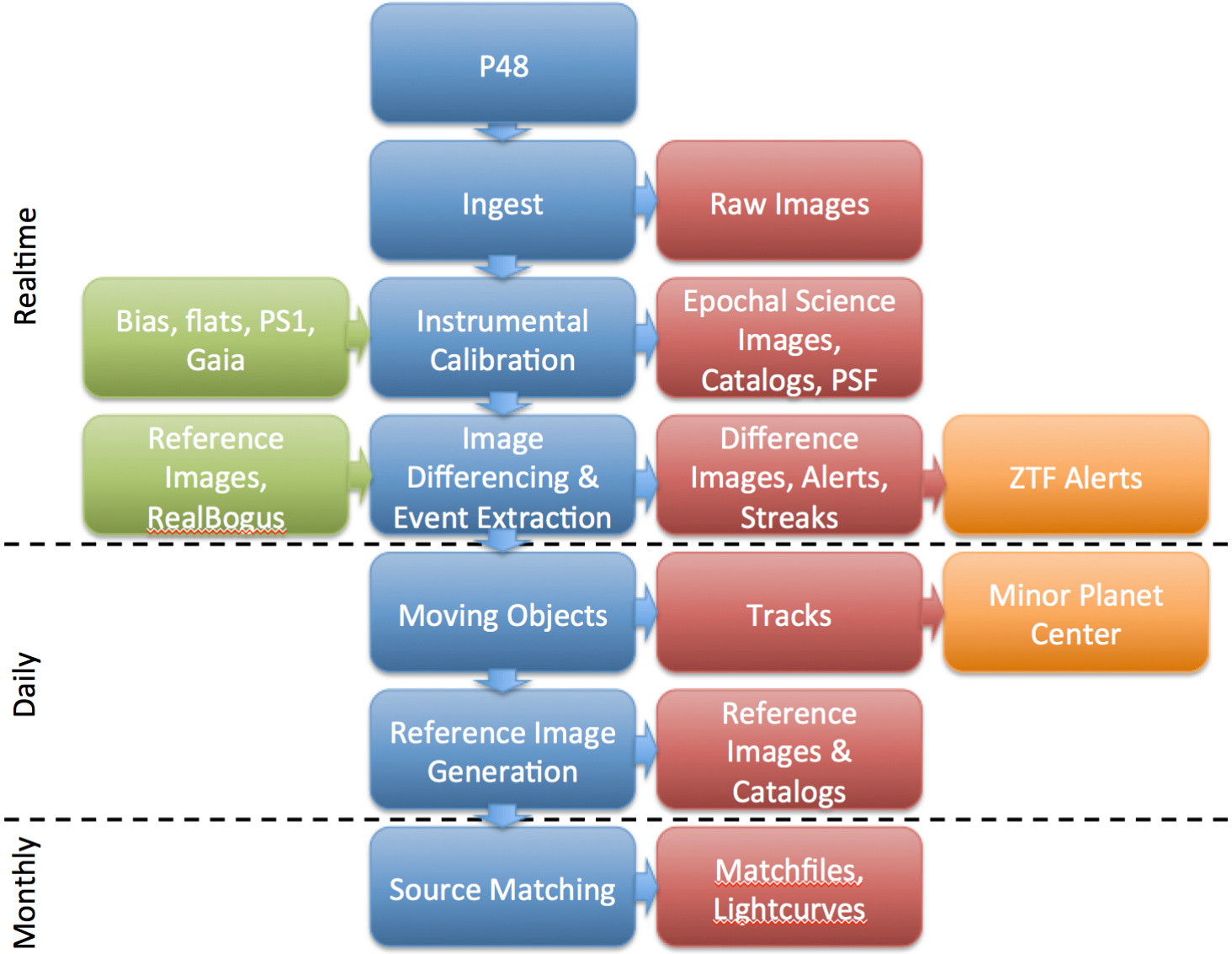



# Development Status

# Architecture



# Data Flow


P48
Ingest
Raw Images
Realtime
Bias, flats, PS1, Gaia
Instrumental Calibration
Epochal Science Images, Catalogs, PSF
Reference Images, RealBogus
Image Differencing & Event Extraction
Difference Images, Alerts, Streaks
ZTF Alerts
Moving Objects
Tracks
Minor Planet Center
Daily
Reference Image Generation
Reference Images & Catalogs
Monthly
Source Matching
Matchfiles, Lightcurves

# DS Development Status

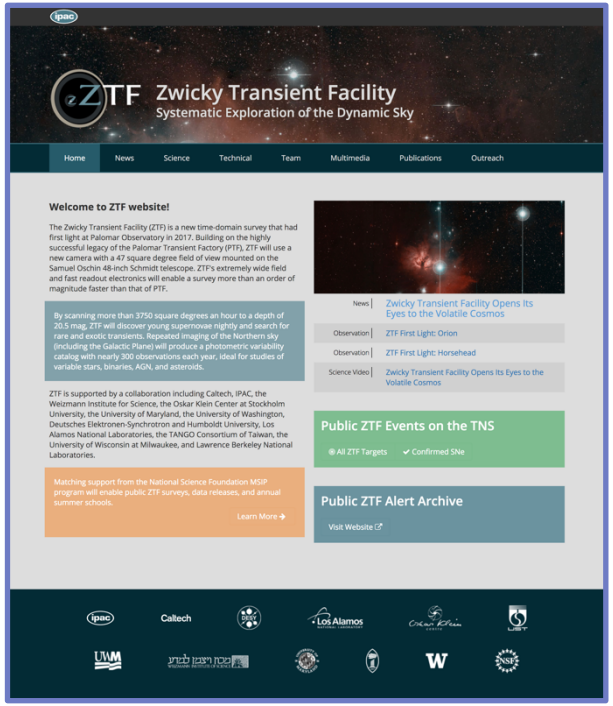
| *Component Group* | DS Component | Status 2017-05 | Status 2018-08 | Notes |
|---|---|---|---|---|
| | | | | |
| *Transfer* | Data Transfer Software Protocols from P48 to IPAC | 100% | 100% | |
| | | | | |
| *Pipeline* | Ingest, CCD quadrant-splitting, floating bias correction | 100% | 100% | |
| | 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 by start of commissioning |
| | Throughput testing: algorithm & cluster optimization | 5% | 100% | Completed by end of commissioning |
| | | | | |
| *Archive* | Image and catalog file product server | 30% | 100% | Completed by 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 by start of commisioning. |
| | | | | |
| *Alerts* | Transient alert distribution infrastructure & interfaces | 2% | 100% | Alerts generated and available for brokers starting June 4, 2018. |

# Baseline Deliverables (1/2)

1. **Instrumentally calibrated, readout-quadrant based epochal image products:**
   - images with photometric zero-points from PSF-fit photometry; bit-mask images; PSF templates
   - two source catalogs per image: PSF-fitting and aperture photometry
   - difference images, accompanying PSFs and QA metadata
   - ➢ Archived at IPAC. Accessible via GUI or API; can interface with Moving-Object Search Tool
2. **Raw image data (CCD-based files with metadata) and image calibration products used in pipelines**
   - ➢ Archived at IPAC. Accessible via GUI or API.
3. **Reference images (co-adds), coverage, unc maps, and two source catalogs per image:** PSF-fitting and aperture
   - ➢ Archived at IPAC. Accessible via GUI or API.
4. **Alert (point-source event) stream** from real-time image-differencing pipeline: packetized with metadata, 30 day photometric histories, upper limits, ML-scores, provisional names, cutouts on new, reference and diff images, …
   - ➢ Mirrored at UW using Kafka; access is through specific program ID channel
5. **Products to support Solar System / NEO discovery and characterization:**
   - streaks (fast moving objects) from difference images: metrics, ML-scores, and cutouts
   - moving object tracks from linking point-source events; known objects are tagged
   - ➢ Available at ZTF-depot webserver (restricted audience)
   - ➢ Moving object metadata (ephemeral astrometry & photometry) delivered to IAU MPC

# Baseline Deliverables (2/2)

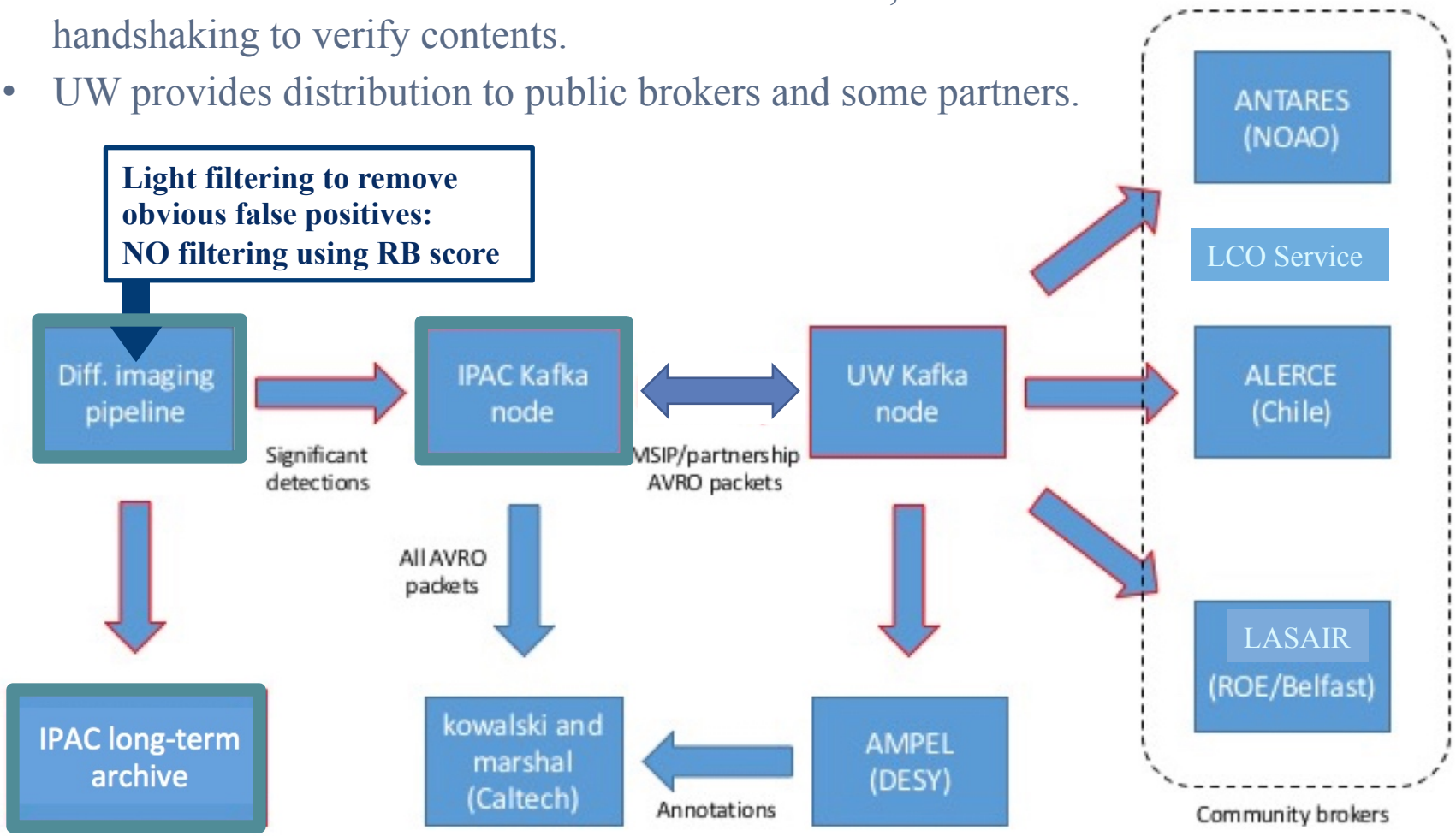
6. **Quality assurance metrics, summary statistics, and coverage maps for performance monitoring**
   - ➢ Available via ZTF-depot webserver (restricted audience)
7. **Matchfiles: all lightcurves per readout-quadrant:** from source-matching of epochal PSF-fit extractions
   - ➢ Available, but restricted (galactic marshal)
8. **Service to query lightcurves & metrics using "object-based" position searches, extracted from matchfiles**
   - ➢ Archived at IPAC. Accessible via GUI or API; can interface with time-series viewer/analyzer with period estimation
9. **Data System Documentation:** pipeline descriptions, recipes, and tutorials on data-retrieval
   - ➢ Progressively updated; paper submitted to PASP.
10. **ZTF Public Website:**
    - ➢ http://www.ztf.caltech.edu
    - ➢ Designed by IPAC Communications and Education (ICE) team
    - ➢ Updated by project team members with privileges



# Alerts

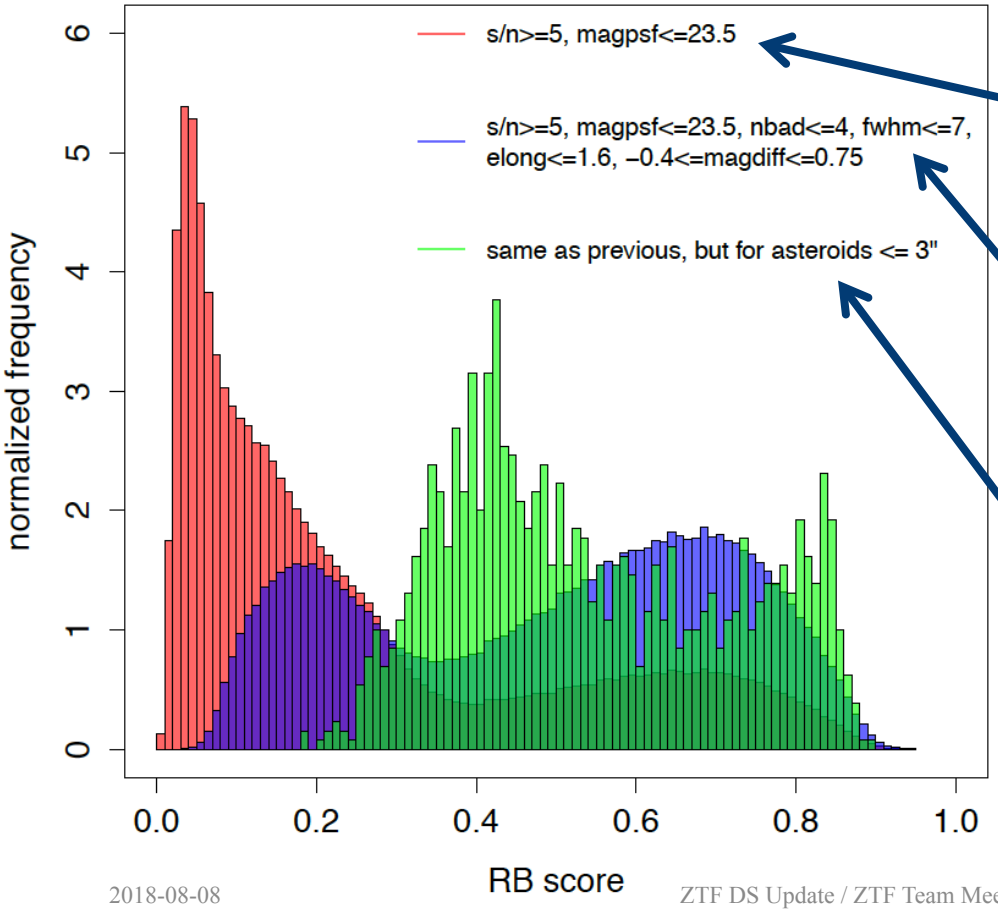
# Alert Packet Distribution

- The ZTF Data System at IPAC generates alerts, which are available via its Kafka Node, and are also archived at IPAC.
- The UW Kafka node mirrors the IPAC Kafka node, with handshaking to verify contents.
- UW provides distribution to public brokers and some partners.



# 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)

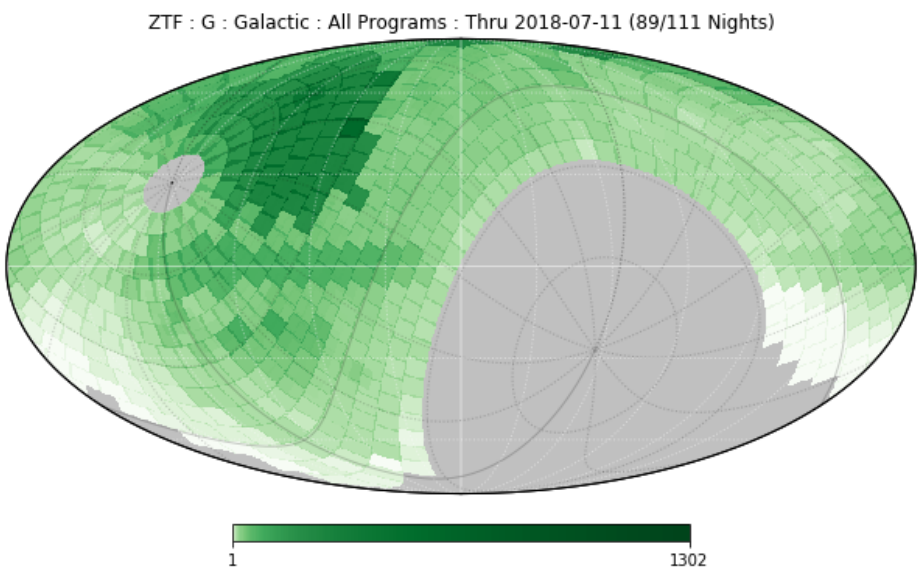
Alerts associated with known asteroids (number = 1300)

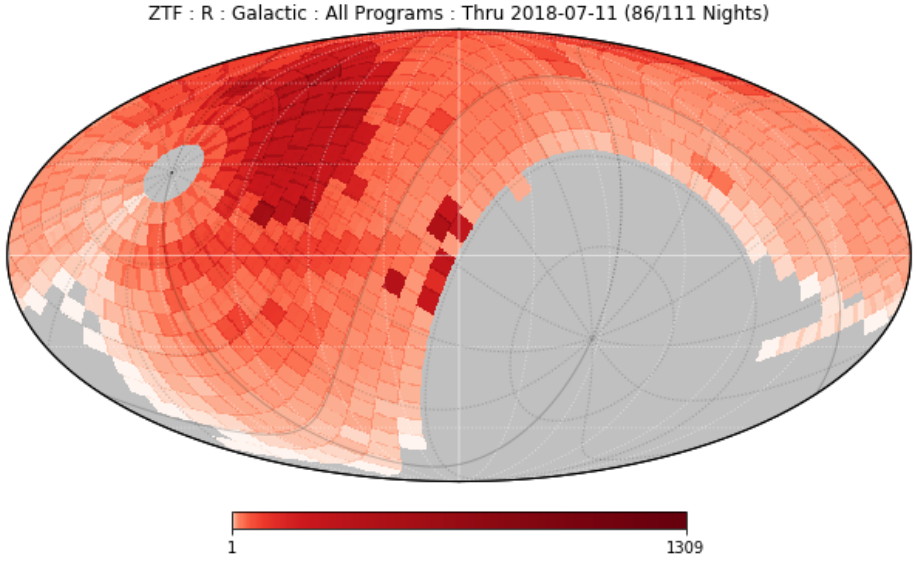
# Alerts: Activities and Accomplishments

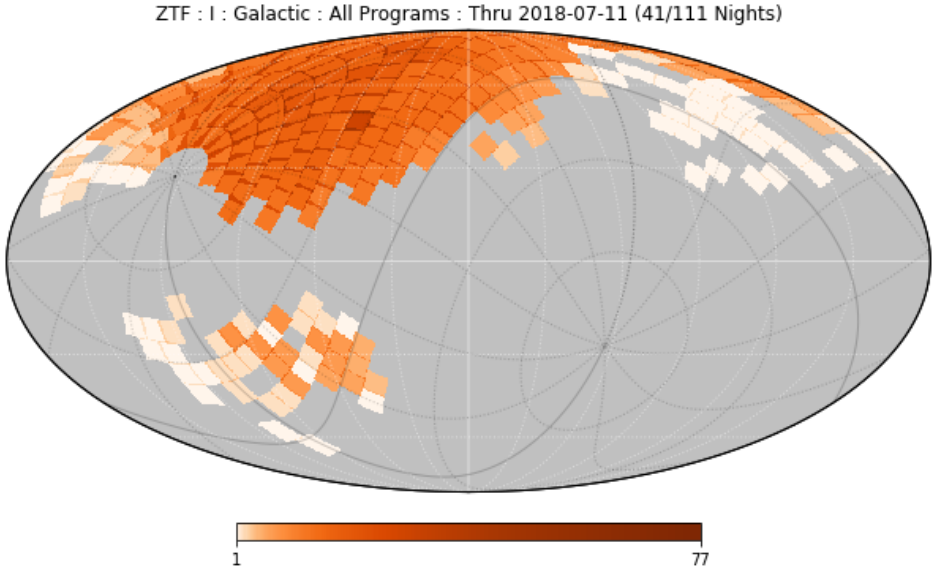
- Infrastructure and software for distributing Alert Packets: includes Kafka; hardware; UW interfacing.
  - Interface tuning drastically improved event-writing speed to Kafka.
- Alert packet schema and contents stabilized following feedback from science working groups.
- R&D on filtering of difference-image 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

# Coverage and Data Holdings

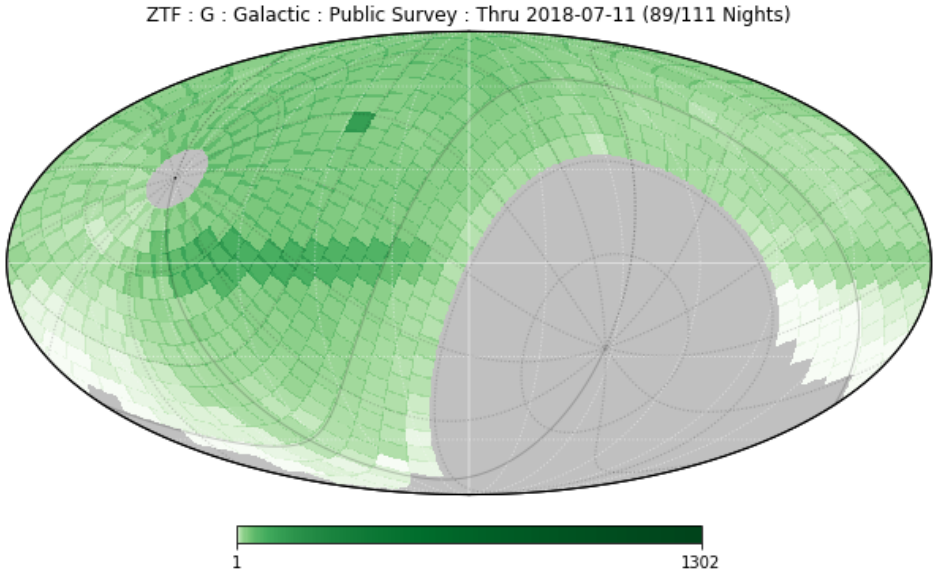
# Sky coverage: all programs
# Mar 17 (science ops start) – Jul 11, 2018

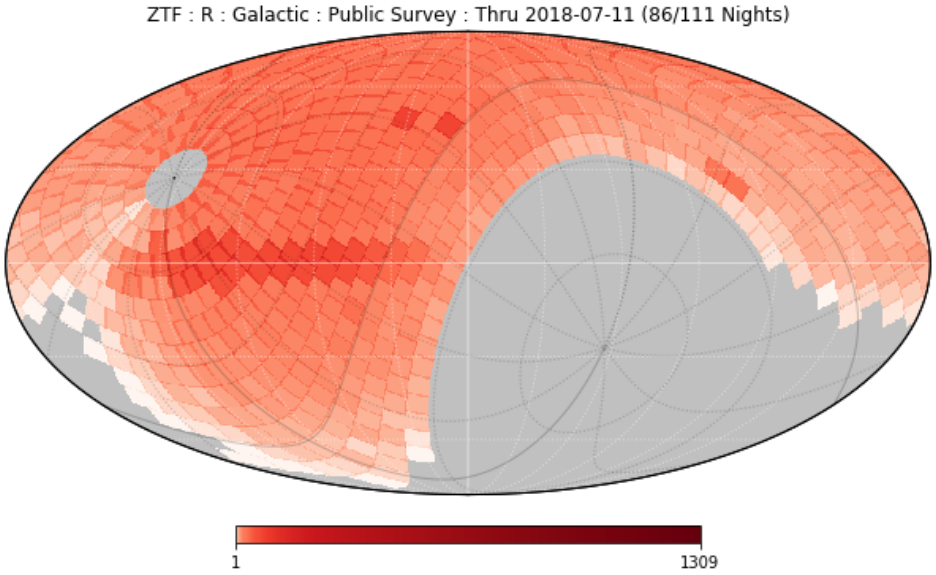


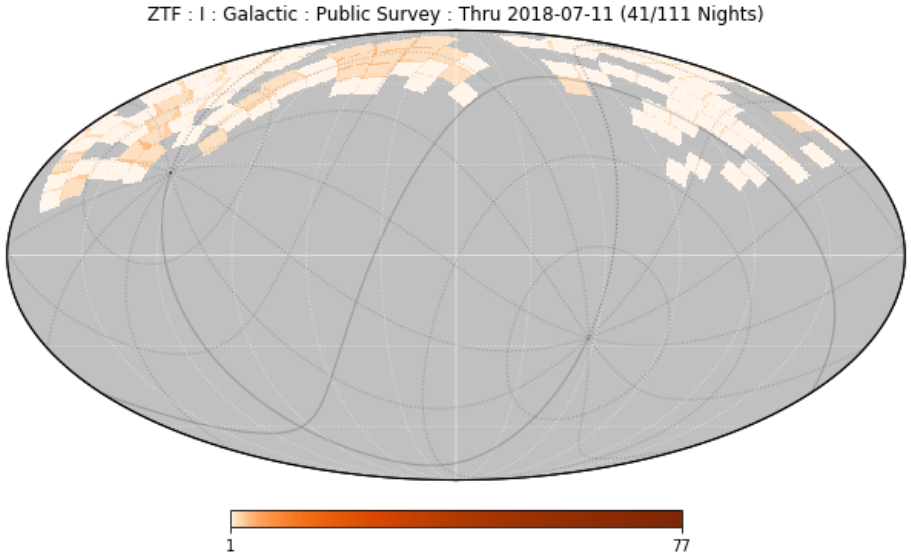



# Sky coverage: public only
# Mar 17 (science ops start) – Jul 11, 2018



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

| **Exposure Metric** | **g** | **r** | **i** |
|---|---|---|---|
| Raw on-sky | 25,149 | 27,403 | 2,126 |
| Survey-ready quadrant-based reference images | 10,205 | 17,675 | 1,001 |
| Lightcurve matchfiles | 5,708 | 13,848 | 233 |
| Epochal science image products archived (all CCD quadrants) | 3.2M (238 TB) | | |

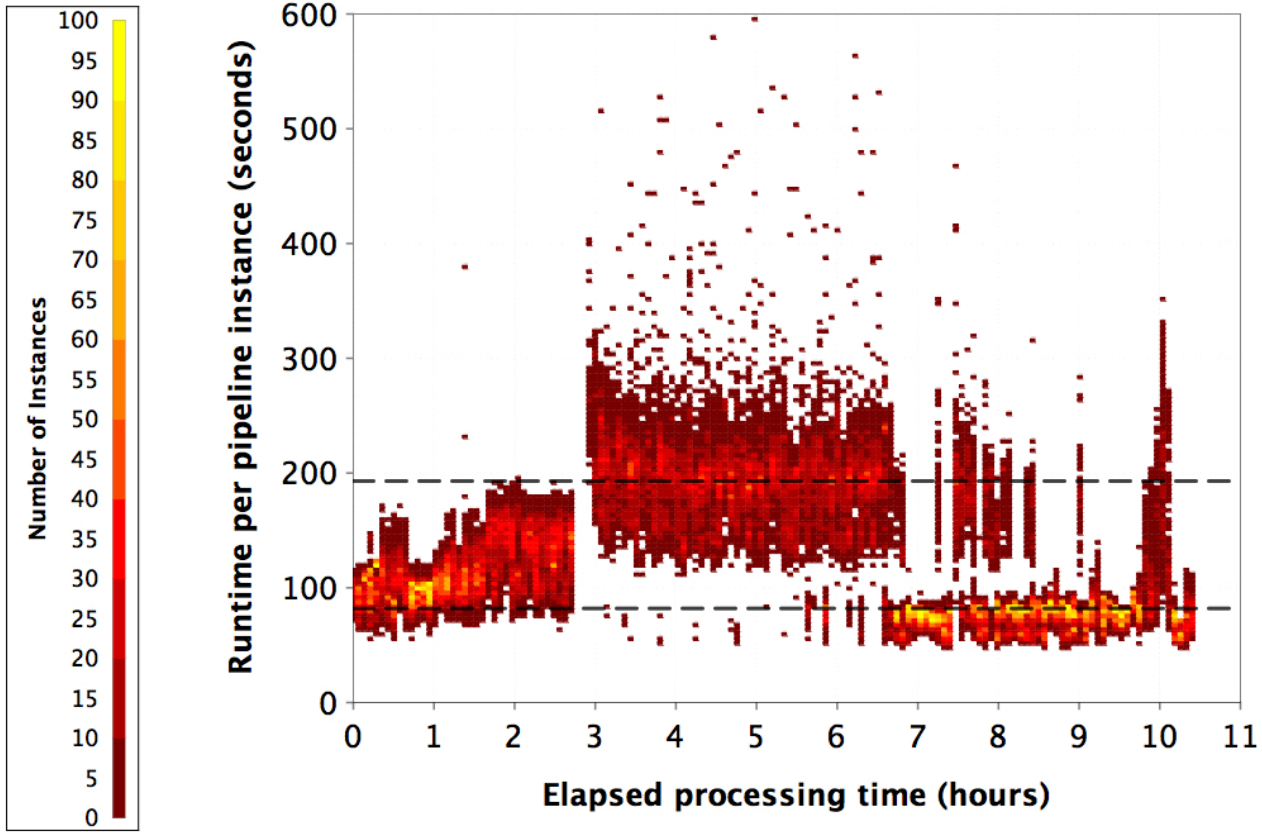
| Extraction Metric | Number |
|---|---|
| Epochal science image PSF-fit extractions | 45 B |
| Epochal science image aperture-based extractions | 27 B |
| Reference image PSF-fit extractions ("seeds" for lightcurves) | 1.3 B |
| Reference image aperture-based extractions | 0.4 B |

| Event Metric | Number |
|---|---|
| Raw candidate events from all difference images | + 46M<br>– 22M |
| Alert packets generated from difference images | + 11M<br>– 7.8M |
| Alert packets associated with known solar system objects (≤3 arcsec) | 447K |
| Streaking objects not associated with known objects | 13 |
| Streaking objects associated with known objects | >5K |
| Moving object tracklets not associated with known objects, delivered to the MPC | 1,724 |
| Moving object tracklets associated with known objects, delivered to the MPC | >300K |

# Pipeline Performance

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

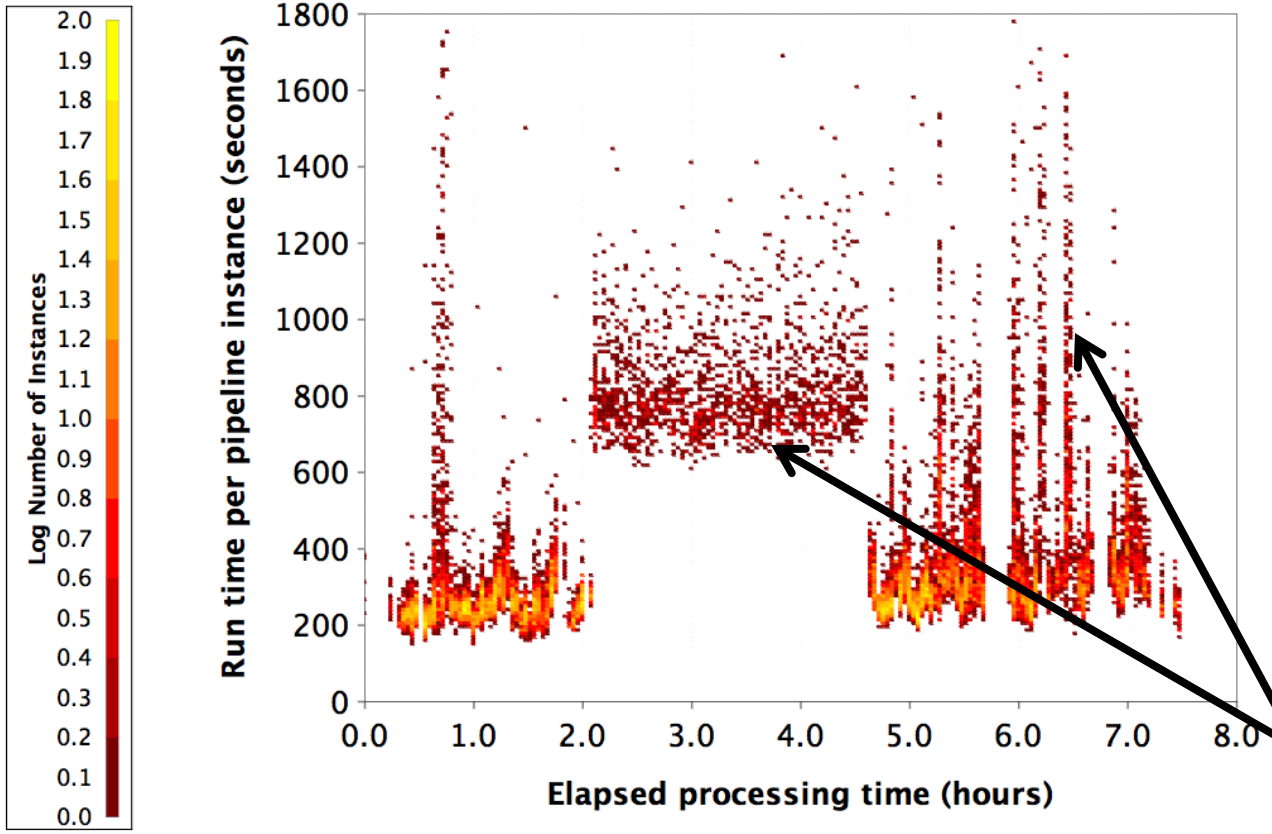
**Most exposures in this night were at high Galactic latitude**




- 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

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

**~ 40% of exposures in this night were in dense galactic plane fields**



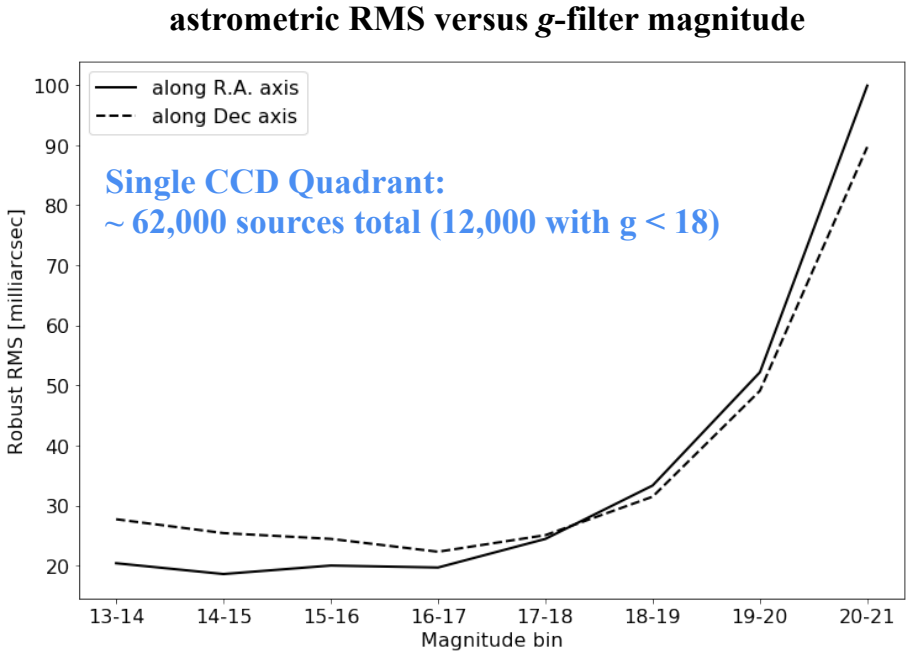

- Based on fields processed on night of July 13, 2018 (UT)
- Reference images were available **for all** quadrants
- 95$^{th}$ percentile in runtime for entire night is ~14 min per quad
- A few % of exposures in deep-drilling fields exceed 20 min.

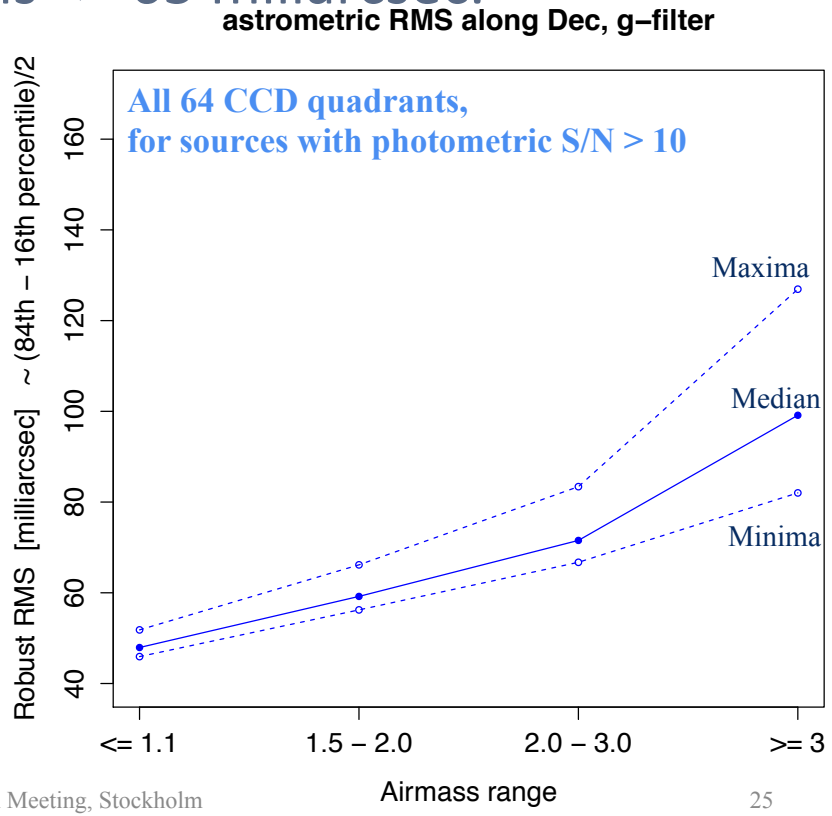
# On-Sky Performance

# Absolute Astrometric Performance

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.

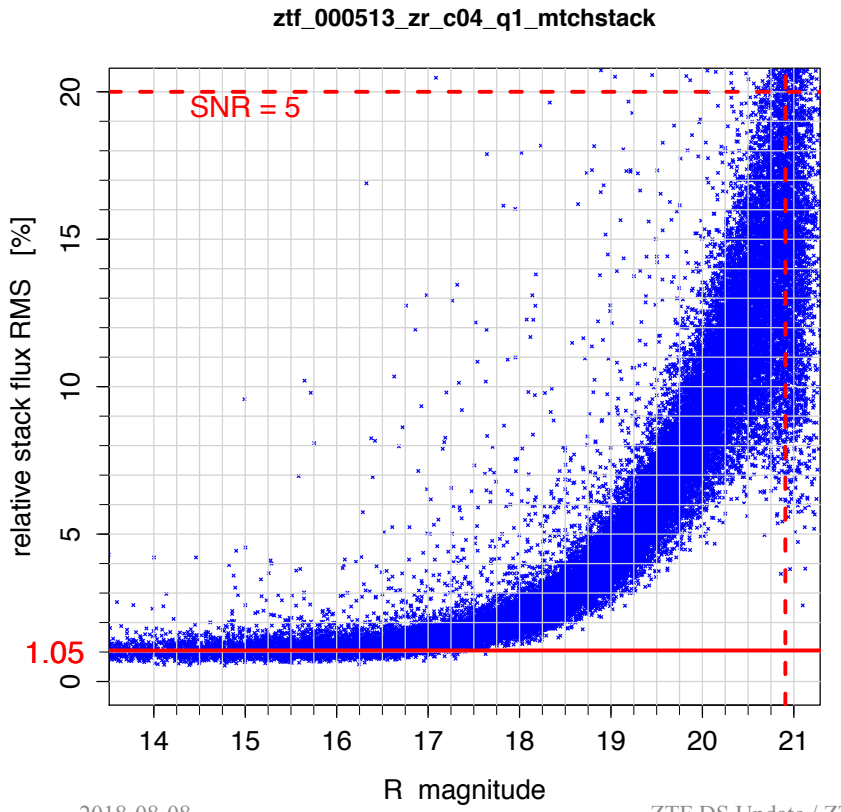


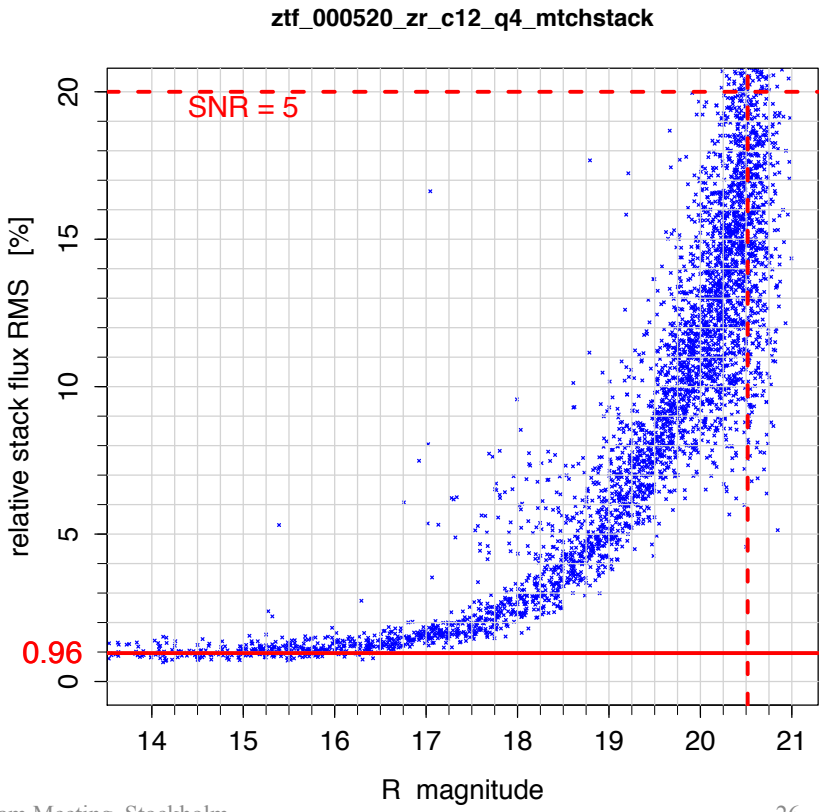
# Photometric Precision (Repeatability)

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

*galactic plane field*

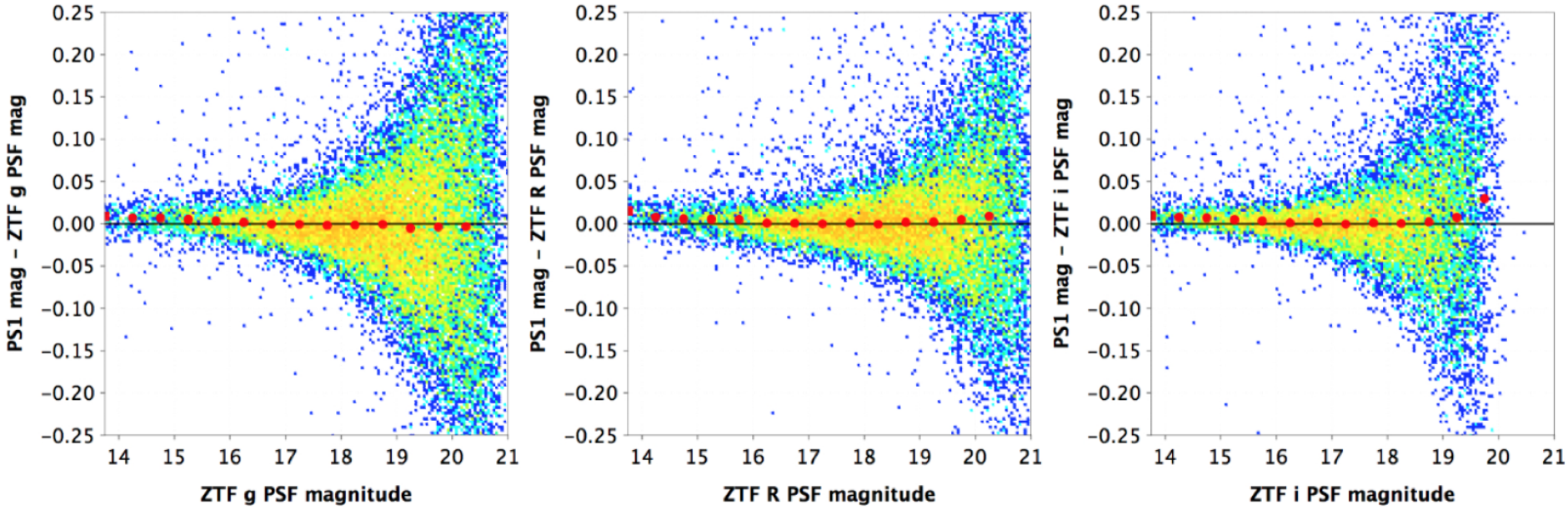



*high galactic latitude field*



# Photometric calibration check

Following calibration, magnitude dependent biases of < 0.03 mag are still seen in PSF-fit and aperture-based catalogs with respect to Pan STARRS1. This is variable across fields.

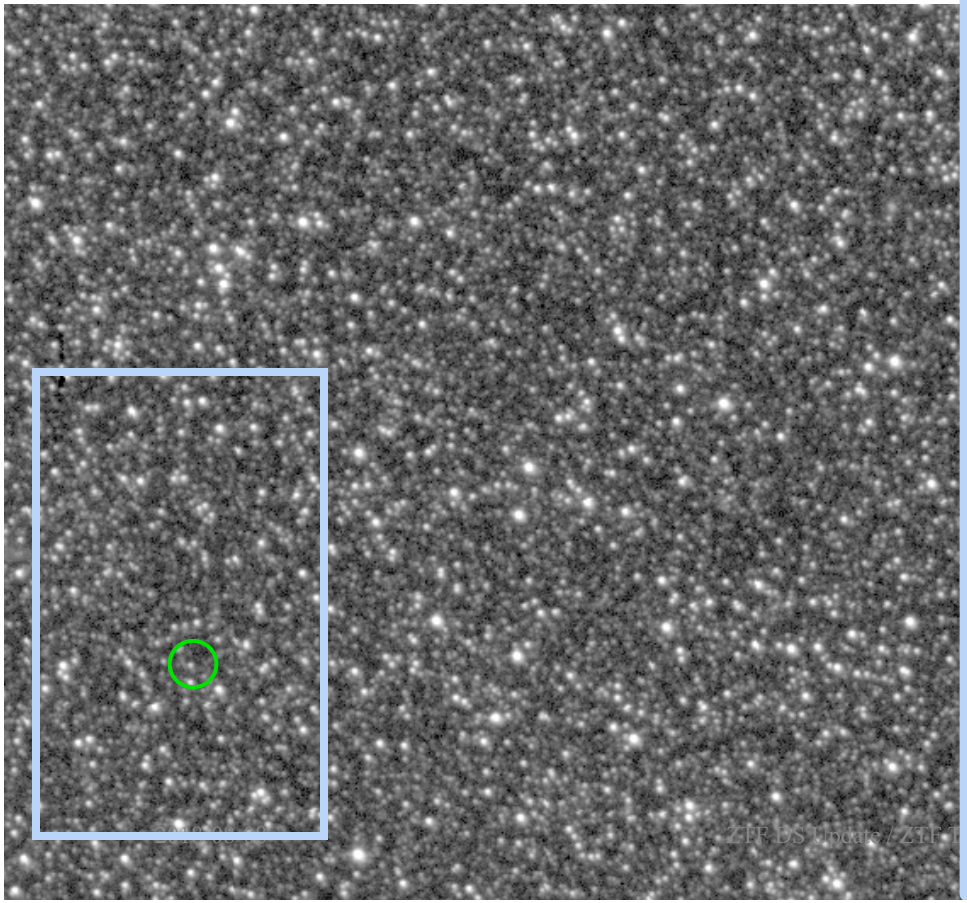


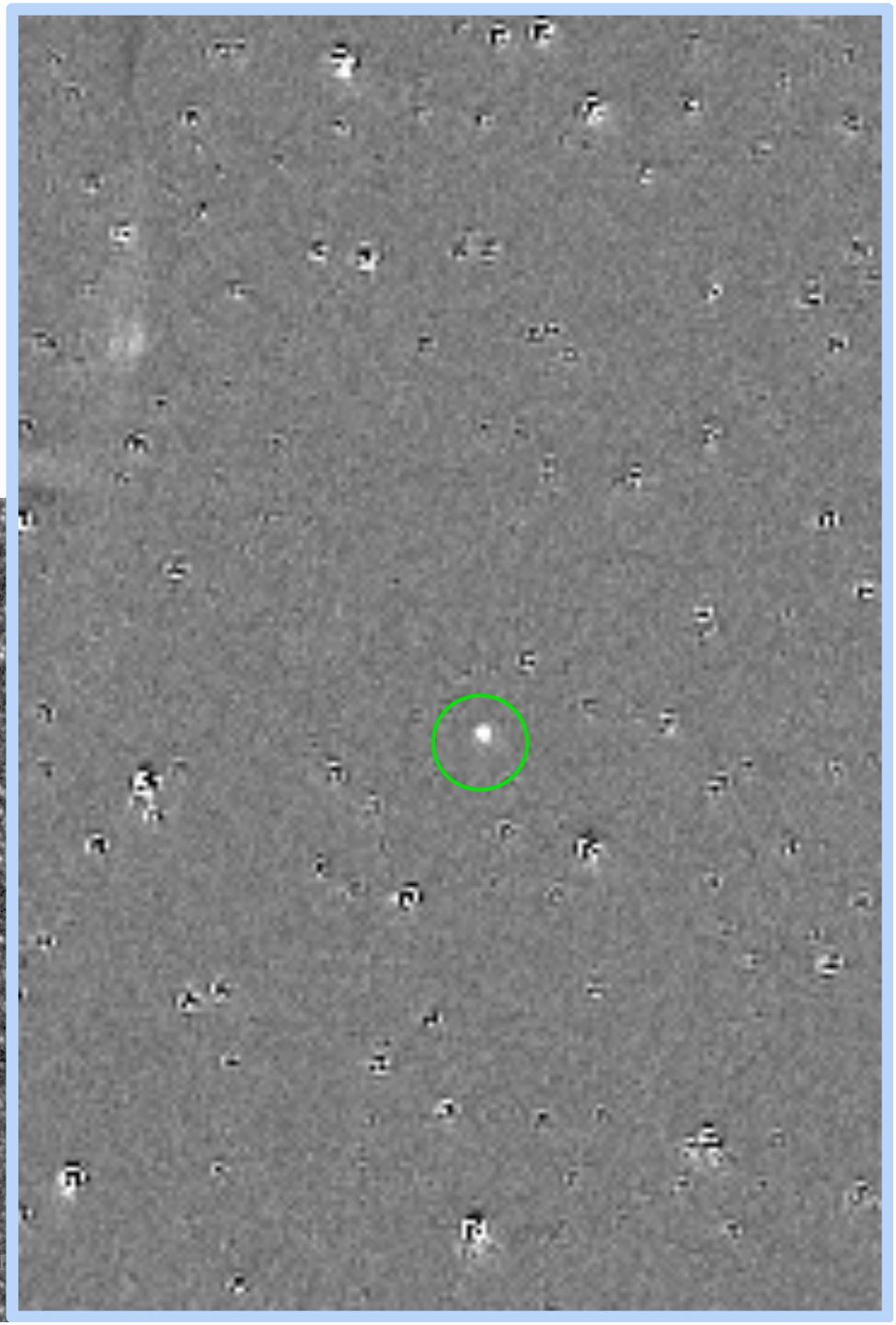

*These are from quadrant-based PSF-fit catalogs; all in the galactic plane ( |b| <~ 8° )*

# Issue: Image Subtraction Quality

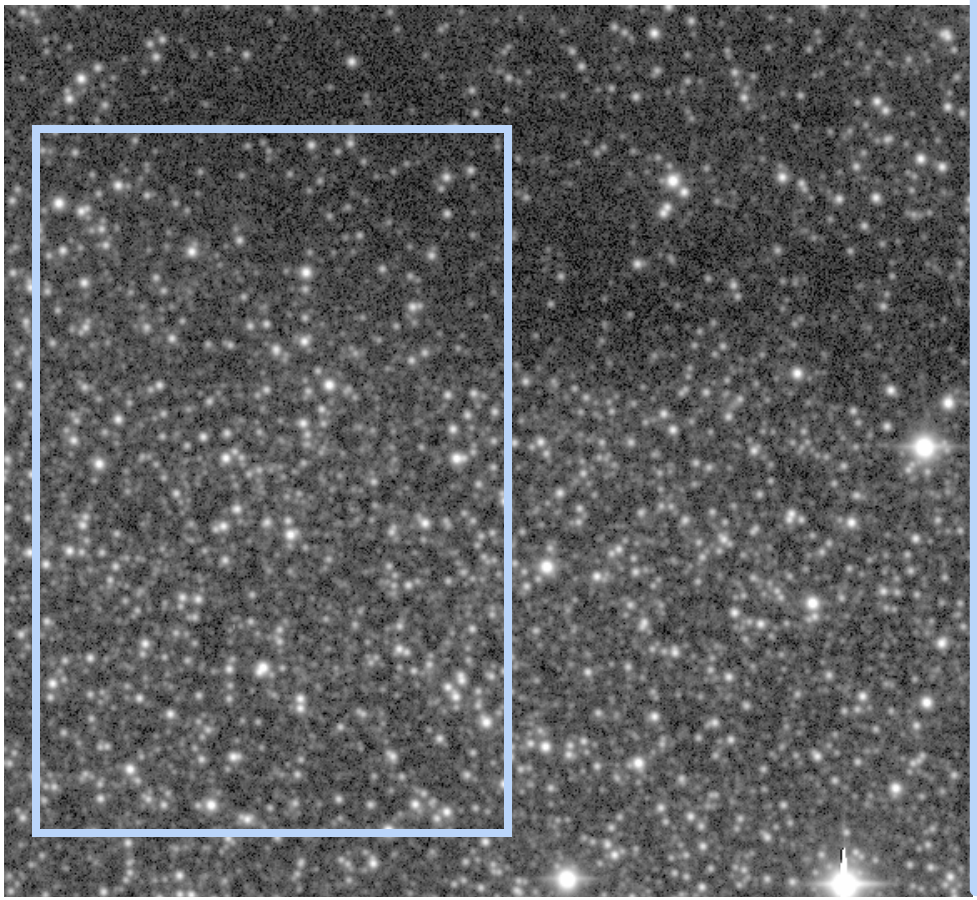
- Some survey fields in the Galactic plane present a challenge for automated image subtraction.
- Usually occurs when large variations in background/source-density are present – related to gain-matching.
- 30 fields have been identified where several quadrants therein repeatedly give bad subtractions.
- Currently, as a stop-gap, we do not plan to generate alerts from these quadrants if the number of raw events is > 500.
- The per-event (single-epoch-based) RB scores are not effective here; these events are all PSF-like.
  - ➢ We advise examining the lightcurves from merged alert packets (or histories therein) to weed out false events.

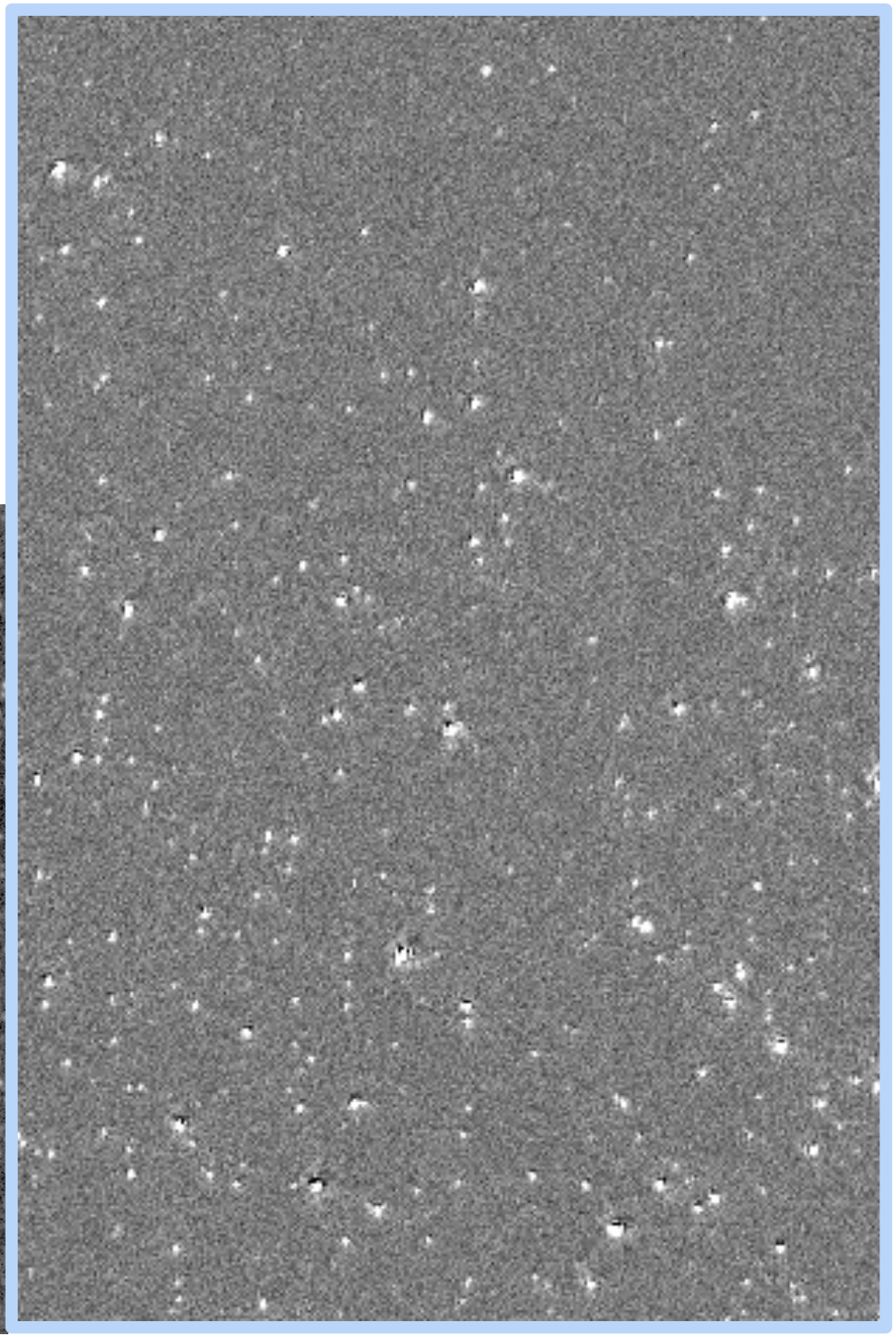
Example of "reasonable" (usable) subtraction from a Quadrant in field 331 that yielded a few **real** events.





Example of a “bad” subtraction from another quadrant in the same field that yielded thousands of false events.





# Issue: Reference Image Coverage

| | | | |
|---|---|---|---|
| Quadrants with N ≥ 15 archived science images | 22,366 (g) | 29,869 (r) | 7521 (i) |
| Current Reference Images (2018-07-16) | 10,610 (g) | 17,876 (r) | 1006 (i) |
| Reference Images with new provisional cuts | 16,815 (g) | 26,083 (r) | 5242 (i) |

- **Problem diagnosed:** the criteria used to select good quality science images for reference-image generation were too tight.
- **We will regenerate all references to a higher uniform depth following this meeting (40 images deep).**
- *Issue will be discussed further by Russ Laher on Wednesday.*

```
obsdate ≥ 2018-02-05
1.75 ≤ FWHM ≤ 4.5
25 ≤ ZP ≤ 27
0.02 ≤ color_coeff  ≤ 0.15
maglim ≥ 19 mag
global_median ≤ 1000 DN
robust_spatial_rms ≤ 80 DN
```

*Example (r-filter) of new provisional cuts for generating reference image.*

# To-Do

**In progress, for end Sep'18:**

- Improve subtractions in challenging fields in the Galactic plane
- Regenerate reference images to higher (uniform) depth to support LIGO / Virgo runs in October 2018
- Automated generation of all-sky coverage maps for reference images
- Synopsis of reference image holdings
- Enable image-cutouts on archived (compressed) difference images – IRSA service (now available for other images)

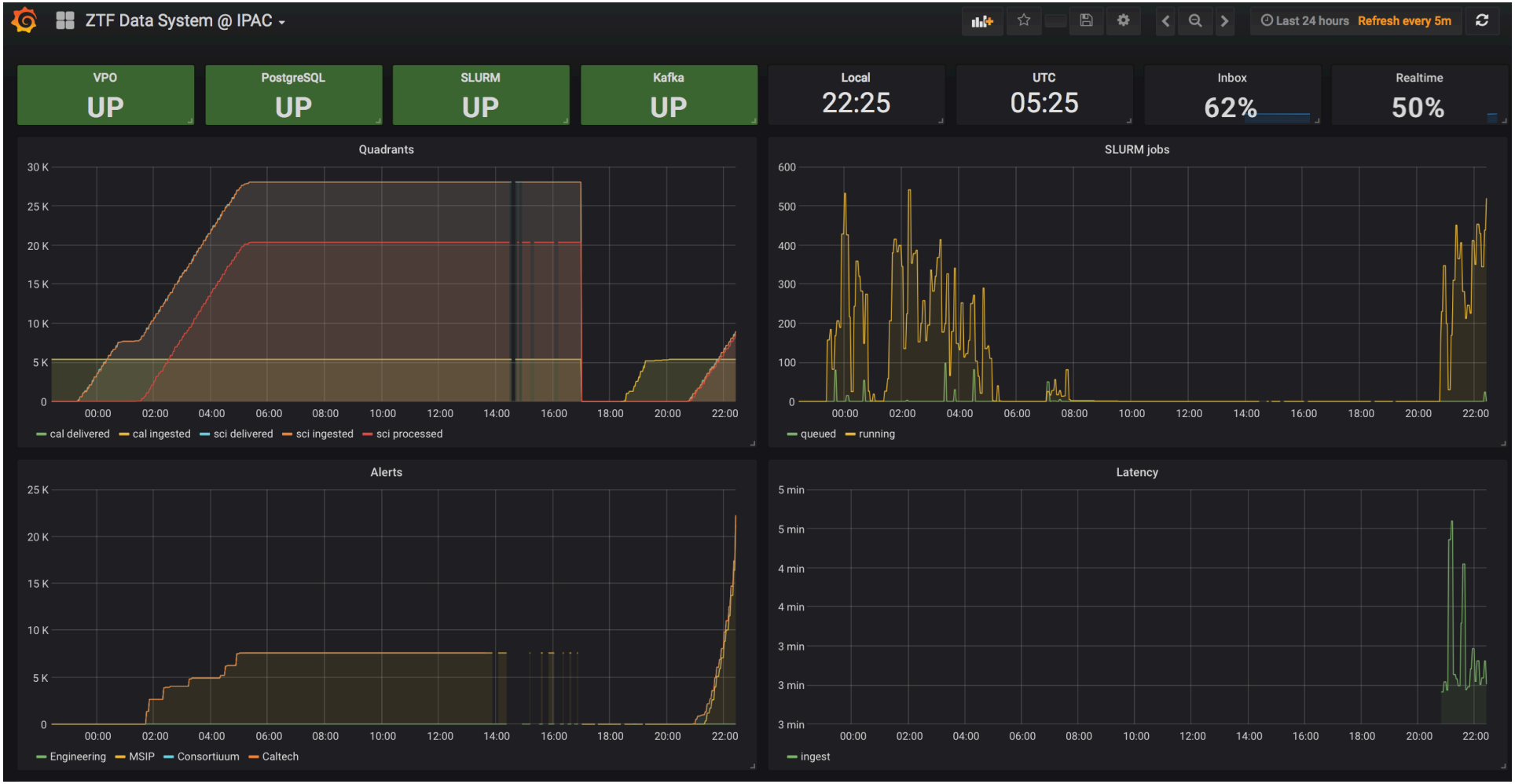
**Ongoing / as-needed:**

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

**Approved, not yet scheduled:**

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

# Real-Time ZTF DS Dashboard In-Prep



# Requests We've Received

*These are subject to review and approval by the ZTF Data System Change Control Board, Matthew Graham, Chair*

- Forced photometry service using image archive: would implicitly include more accurate estimation of upper-limits for non-detections. (Prohibitive in production.)
- Simulated transient injection pipeline and infrastructure
- Sentinel service (for monitoring targets of interest using archived products)
- Set-up of "sandbox" environment for analysis, testing, and prototyping; pending MOU on usage/data-access

# Reference Material

# In Closing …

- A number of limitations and deficiencies have been identified and are well understood
- Please continue reporting possible issues to the Project Scientist. We will explore if it is a “feature” or a bug
- It is in our best interest to document these as “cautionary notes” for the community
- The Data System staff value your feedback!

# ZTF Pipelines and run frequency

Overall, there are 9 interdependent pipelines, grouped into four categories.

- All implemented and tested on simulated camera-image data; some pipelines also tested using real camera data.
- All baseline archival products, formats, and methods for access are finalized.

**Raw data ingestion and initial processing:**

1. Raw data ingest, archival of raw images and storage of metadata in database [*realtime*]
2. Raw-image decompression, splitting into readout-quadrant images, floating bias correction, QA metrics [*realtime*]

**Calibration-image generation:**

3. Bias-image derivation from stacking calibration images acquired in afternoon [*before/after on-sky operations*]
4. High-v flat (pixel-to-pixel responsivity) from stacking illum. flat-screen exposures [*before/after on-sky operations*]

**Real-time science-level processing:**

5. Instrumental calibration of readout-quadrant images: includes astrometric and photometric calibration [*realtime*]
6. Image-subtraction with transient-event extraction (point sources & streaks), alert packets & distribution [*realtime*]

**Ensemble-based (collective-image/catalog) processing:**

7. Reference-image generation (co-addition of epochal images from 5) [*when sufficient good quality data available*]
8. Source-matching/lightcurves with relative photometric refinement; inputs from 5 & 7 [*every month, TBD*]
9. Moving object tracks, orbit-fitting, QA; from linking point-source events from 6 [*end of night, 3-4 day window*]

# ZTF Lightcurve Pipeline (*matchfile* creation)

- All sources detected in epochal images are matched against the reference-image source catalog for a given field, CCD quadrant, and filter
- The "cleanest" least variable sources are used as anchors for the relative photometric calibration
- Individual image gain-correction factors are computed using a global least-squares fit across all epochs
- These gain-corrections are applied the image photometric zero-points
- The refined zero-points *are expected* to improve relative photometry to a few millimag for bright sources
- This pipeline will be triggered on timescales of typically one month (TBD), contingent on data accumulated
- All lightcurves for a single CCD quadrant and filter are stored in a "matchfile" (hdf5 pytable format)
- Accompanying each lightcurve is a set of >100 metrics: RMSs, Skews, Stetson indices …
- All lightcurves and metrics are seeded by an object ID; these objects are loaded into a database to support spatial searches; associated lightcurve is retrieved from the "matchfile" containing that object position
- Expect of order 1.3 billion objects (individual lightcurves) for ZTF
  - There will be multiple (disjoint) lightcurves per object due to the two overlapping science grids
  - There is no plan at present to splice lightcurves belonging to the same object

# Data Access / visibility policy

- Observing time during science operations will be split between three categories:
  - **Public** (NSF-funded MSIP survey: 40%)
  - **Private collaboration** (40%)
  - **Caltech TAC** (20%)
- Managed per exposure (epoch) using a *programID* propagated from scheduler to raw-image metadata

---

- Private/Caltech observers can access their data in near-realtime, soon after archive ingestion. This includes all calibration products and lightcurves from epochs tagged by their respective *programID*s queried via archive GUI.
- Public data will only be available at the public release times for general access by all.
  - raw images, processed epochal images, accompanying source-catalog files, difference images
  - reference images and catalog files
  - lightcurves constructed from public epochal data only
  - calibration data products
- Public alert packets (triggered from events detected in public exposures) will only contain public data. This includes their 30 day event histories.
- Private alert packets (triggered from events detected in private exposures) will contain public data in their 30 day event histories.
- Caltech alert packets (triggered from events detected in Caltech exposures) will contain data from all three programs in their 30 day event histories.
- No restriction on input data used to generate products for Solar System science: streaks & moving-object tracks; selected (human-vetted) products will be delivered to MPC.
- No restriction on input data used to generate reference image (co-add) products.
- No restriction on input data used to generate source match-files (lightcurve files):
  - MOU in place with the only customer of these products: Galactic Marshal
  - only privately-tagged and *already-released* public data therein to be ingested by Marshal

# Reminder on documentation

- **Primary document:** design, deliverables, product usage and access, cautionary notes, performance;
  Linked from ZTF public website under:
  *https://www.ztf.caltech.edu/page/technical#science-data-system*

- **Science Data System paper** (high-level overview); submitted to PASP:
  *http://noir.caltech.edu/twiki_ptf/bin/viewfile/ZTF/ZtfPapers?rev=8;filename=zsds.pdf*

- **Archive access and services:**
  *https://irsa.ipac.caltech.edu/Missions/ztf.html*

- **Public alert archive and usage:**
  *https://ztf.uw.edu/alerts/public/*