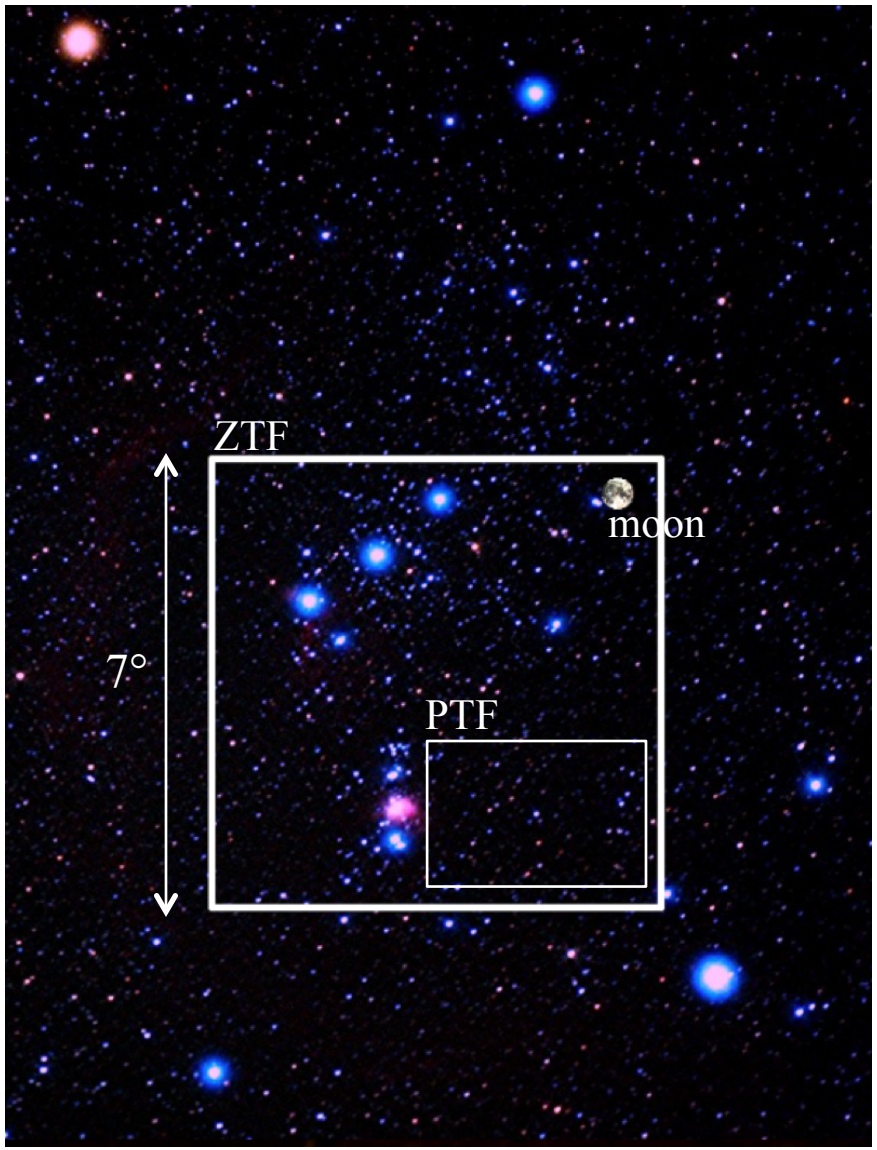
ZTF Pipelines and Deliverables

Frank Masci & the IPAC-Caltech ZTF Team

iPTF/ZTF Workshop, May 2016

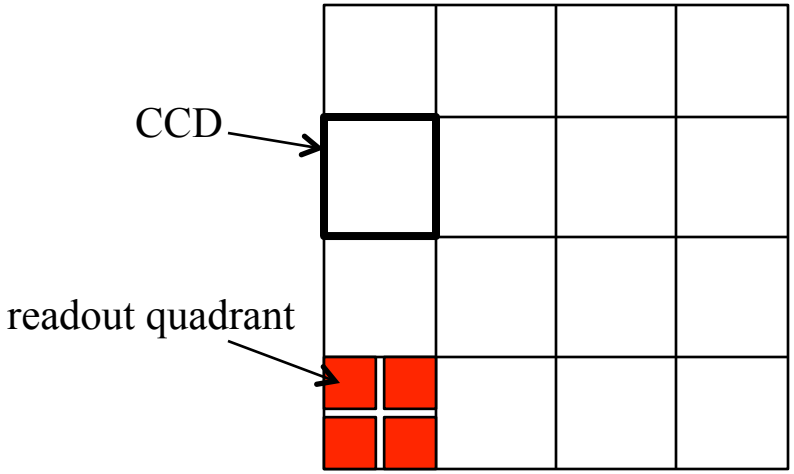
ZTF Field-of-View



Survey rate is ~ 3900 deg^2 / hour
Faster than the sky rotates!

ZTF Raw Camera Image Data

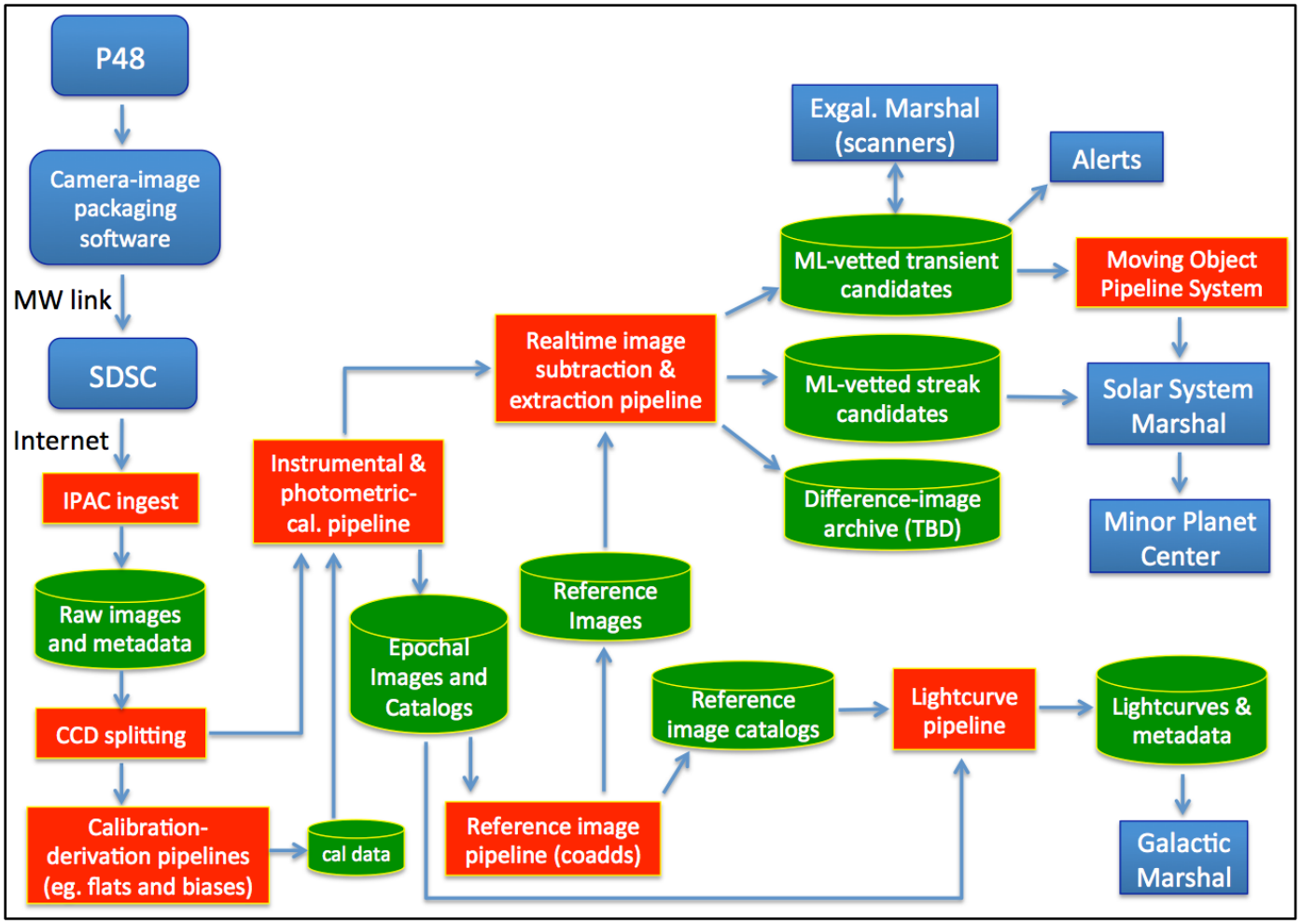
- One camera exposure: 16 CCDs; each ~ 6k x 6k pixels
- Image data packet transmitted is one CCD (four readout-quadrant images)
- 16 CCD-based image files are transmitted every 45 sec.
- Full camera exposure: ~ 1.3GB uncompressed.
- Require ~ 3x lossy compression to accommodate achievable bandwidth; ostensibly ~ 100 Mbits / sec.
- Inbound data rate after compression: ~ 80 Mbits / sec.



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

Data Flow in the ZTF Science Data System (ZSDS)

- The ZSDS will be housed at the Infrared Processing and Analysis Center (IPAC), Caltech
- Consists of data processing pipelines (red), data archives (green), and user-interfaces



Deliverables & Products

1. Instrumentally calibrated, readout-quadrant based epochal images, masks, and two source catalogs per image: PSF-fitting and aperture photometry. Only PSF-fit photometry will be absolutely calibrated. ZPs derived therefrom.

2. Reference images (co-adds of epochal images) and two source catalogs per image: PSF-fitting and aperture.

3. Lightcurve database: based on PSF-fit photometry of sources matched across all epochs. Photometry refined.

4. Products to support near-realtime discovery: database of (thresholded) transient candidates, image cutouts.
   - Access via marshal-driven scanning interface(s): based on canned/standard queries periodically executed by real-time pipeline and staged for external access. Avoids query bottlenecks on database.
   - Archival of image subtractions is TBD -- necessary for archival research!

5. To commence 12 months after survey start: transient alert stream extracted from real-time pipeline products (TBD)

6. Solar system/NEO support: moving object tracks from linking transients extracted from image-subtractions, with single-exposure streak detections: stored in DB, human vetted and delivered to the IAU's Minor Planet Center.

ZTF data product volumes / source counts

Per night:

Assuming average length of night at Palomar is ~ 8h:40m (summer: ~6h:20m, winter: ~ 11h), we expect ~ 700 camera exposures per night on average => 44,800 readout quadrant images.

- raw data (including calibrations): ~ 367 GB compressed (3x)
- instrumentally-calibrated epochal images, masks, and metadata: ~ 3.1 TB
- aperture photometry (epochal) catalogs: ~ 140 GB
  - ~ 310 million sources per night
- PSF-fit photometry (epochal) catalogs: ~ 44.8 GB
  - ~ 900 million sources per night
- image-subtractions and metadata ~ 2 TB

Total per night: ~ 5.65 TB

For three-year survey:

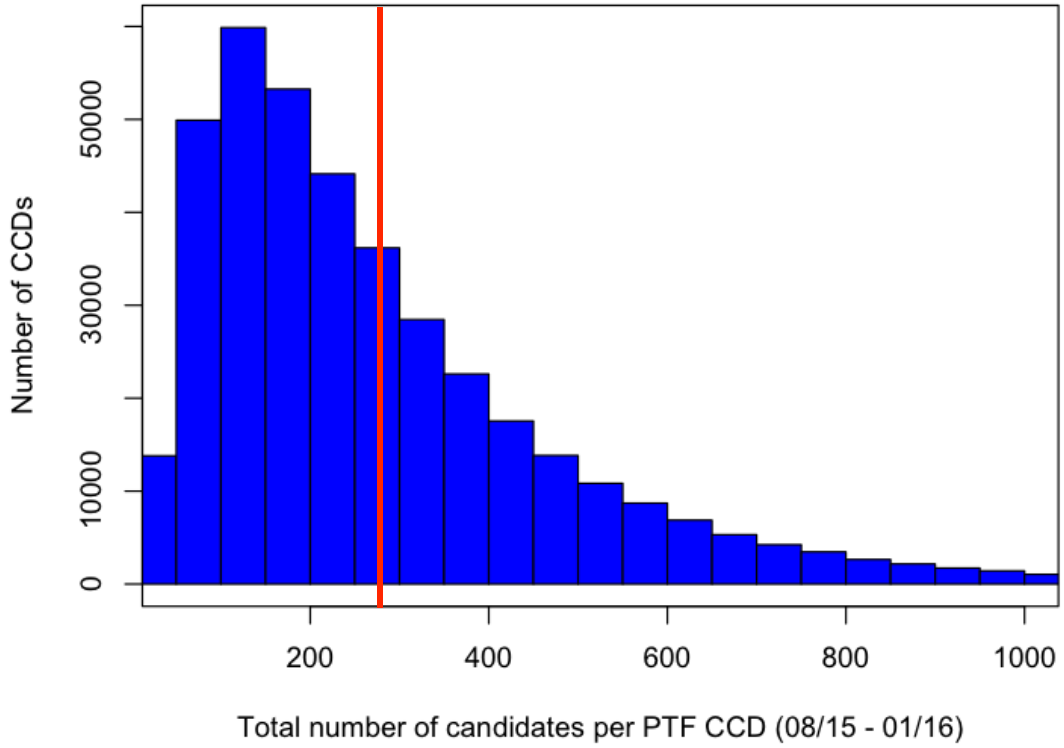
Assuming ~ 250 to 280 "good" nights per year (from PTF),

Total image/catalog file products: ~ 4.2 to 4.7 PB

*** Includes storage of image-subtractions (not in baseline budget).
Excludes database storage for raw transients, other metadata, and epochal lightcurve database.

Number of (raw) transient candidates

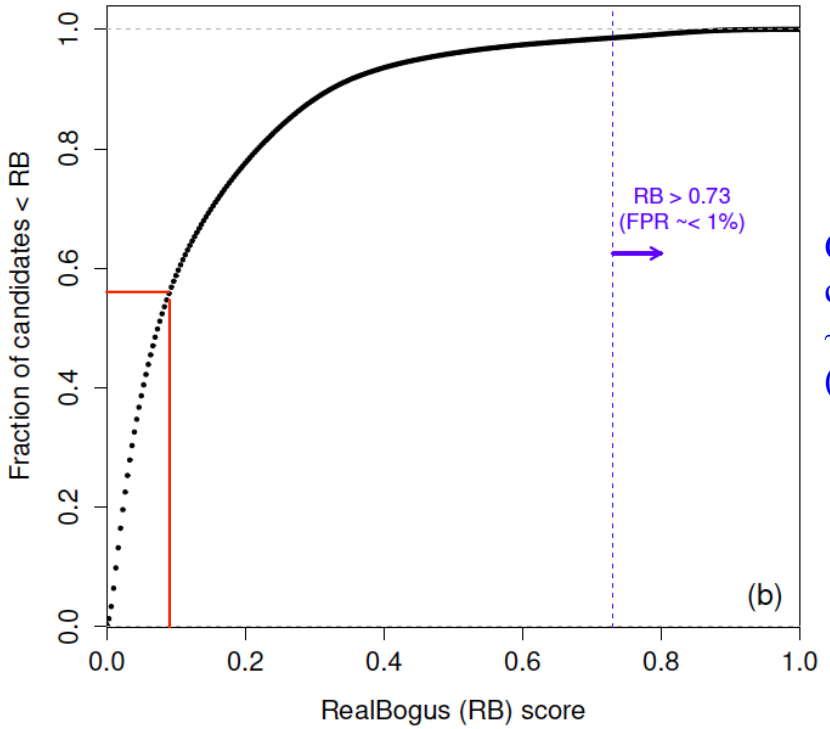
- From **PTF**, encounter ~ 260 raw, **<u>non</u>** machine-learned vetted candidates per CCD at > 4σ using PTFIDE.
- One ZTF CCD readout quadrant covers ~ one PTF CCD + ~ 10%. Hence we can extrapolate to ZTF.
- Have ~ 700 exposures * 64 readout quads: ~ 44,800 positive subtractions per night on average.
- Implies **~ 13 million transient raw candidates** per night for ZTF. Includes all transients (+ variables + asteroids)



or ~ per ZTF readout quadrant

Machine-learning to the rescue!

- Use the *RealBogus* (RB) quality score from a machine-learned classifier: crucial for PTF (down to 4σ).
- If avoid everything with a RB score < 0.1, only need to store ~ 6 million candidates per night in DB for ZTF.
- If use RB > 0.73 (< 1% false-positive rate) found for PTFIDE subtractions, need to scan <~ 400,000 cands/night.
- Translates to <~ 10 candidates per ZTF quadrant image or <~ 14 candidates/deg^2 on average (<u>all transients</u>).



Cumulative fraction of transient candidates versus RB score from ~22,000 PTFIDE subtractions (Masci et al. 2016).

ZTF Pipelines

Overall, there will be 10 inter-dependent pipelines (one is TBD):

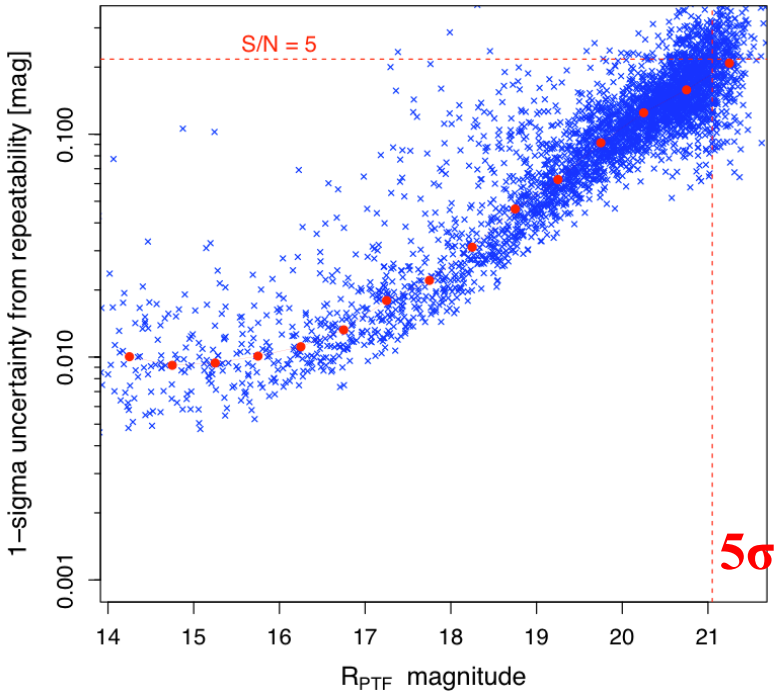
1. Raw data ingest, archival of raw images and storage of metadata in database [*realtime and continuous*]
2. Raw-image uncompression and splitting into readout-quadrant images, w/ simple QA [*realtime and continuous*]
3. Bias-image derivation from stacking calibration images acquired in afternoon [*made before on-sky operations*]
4. High-ν flat (pixel-to-pixel responsivity) from stacking calibration images [made *before on-sky operations*]
5. **TBD:** Low-ν flat from either long-term ZPVM or dithered-star observations [*every week, month or longer?*]
6. Instrumental calibration of readout-quadrant images: astrometry and absolute phot. cal [*realtime and continuous*]
7. Image subtraction and transient discovery with metadata and cutouts [*realtime and continuous*]
8. Reference-image generation (co-addition of epochal images from 6) [*as needed: when good quality data available*]
9. Source-matching with relative photometric refinement for lightcurves; inputs from 6 [*every two weeks or longer?*]
10. Moving object pipeline system (MOPS): both tracklets and streaks from 7 [*every 3 or 4 hours during night*]

Basic Photometric Calibration

- Photometric calibration will be performed with respect to an external catalog (e.g., Pan-STARRS1) using PSF-fit extractions on a readout-quadrant image basis:

$$m_i^{PS} - m_i^{ZTF} = ZP + b\left(g_i^{PS} - R_i^{PS}\right) + \varepsilon_i \quad \Rightarrow \quad \text{solve for } ZP,\ b \text{ per image}$$

- Expect an *absolute* precision of ~ 2 - 3%.
- *Relative* photometric precision using PSF-fitting on PTF images ~ 1% (no refinement of ZPs across epochs)
  - ➢ Biggest limitation is flat-fielding!



5σ limit ~ 21 mag

Development Status

- Have a small set of simulated ZTF image data to facilitate development. Follows camera specs.
- Operations and Transients database schemas in place
- Quality Assurance metrics identified across all pipelines
- Have an ingest pipeline in place that loads metadata into DB
- Have a CCD-splitting pipeline that also performs floating-bias corrections with QA metrics
- Prototype instrumental calibration pipeline is ~80% complete. Uses mock calibration inputs.
- Image-subtraction pipeline prototyping in progress: exploring ZOGY algorithm
- Data-flow / processing model in place: operations file-system and archive interfaces defined.

- Currently testing on a 32-node compute cluster with 32 CPU cores/node (1024 concurrent threads)
  - Need more throughput testing to decide if more CPUs needed (~ August 2016)

Near-term schedule

Many components are being developed in parallel (including archive services)

- **May 30, 2016:** prototype instrumental calibration pipeline (with basic astrom./phot. calibration)
- **June 10, 2016:** prototype image-subtraction / transient-discovery pipeline
- **June 20, 2016:** prototype reference-image (co-addition) pipeline
- **June 30, 2016:** prototype source-matching pipeline to support lightcurve generation
- **June 30, 2016:** calibration software in place: high-ν flats and bias maps
- **July, 2016:** need plan in place for low-ν flat generation/delivery
- **August 2016:** initial throughput testing using prototype pipelines on larger simulation set; includes interfacing with all databases and archive
- **September 2016:** software to collate QA metrics for external monitoring
- **Sep - Dec 2016:** moving-object pipeline improvements; including streak detection
- **Sep - Dec 2016:** Continued pipeline improvements / refinements
- **Jan 2016:** need Gaia and PS1 catalogs in house for integration with pipelines

Concerns, discussion points …

- Data transfer bandwidth from P48 to San Diego: can we achieve a sustained 100 Mbits / sec?

- Will be a challenge to accommodate lightcurve DB based on single-epoch PSF extractions
  - if use PSF-fit extractions with no filtering, get ~ 900 million sources / night (>5σ)
  - difficult to load a single night's data fast enough before next night
  - ~ 700+ billion row database after 3 years; will become very difficult to query
  - can indeed support ~ 300 million extractions per night

- Baseline budget doesn't include storage of difference-image products (~ 1.7 PB / 3 years)

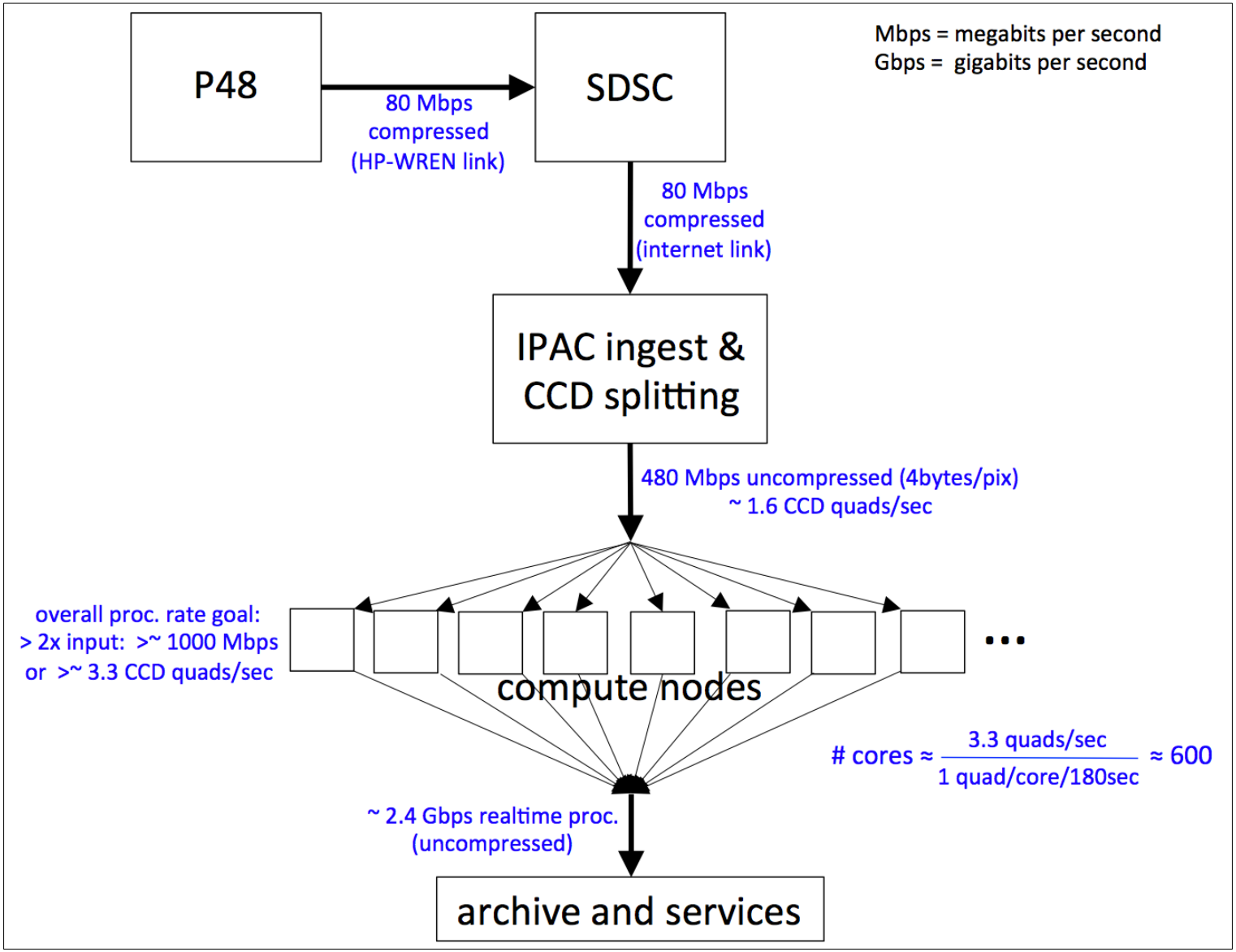
- Flat-fielding plan: whether low-spatial frequency responsivity maps are needed to achieve best *relative* photometric precision; currently a placeholder in ZTF pipeline.

- Sky-tiling geometry (iterate on Eran's proposed grids). Some pipeline functions assume static grids.

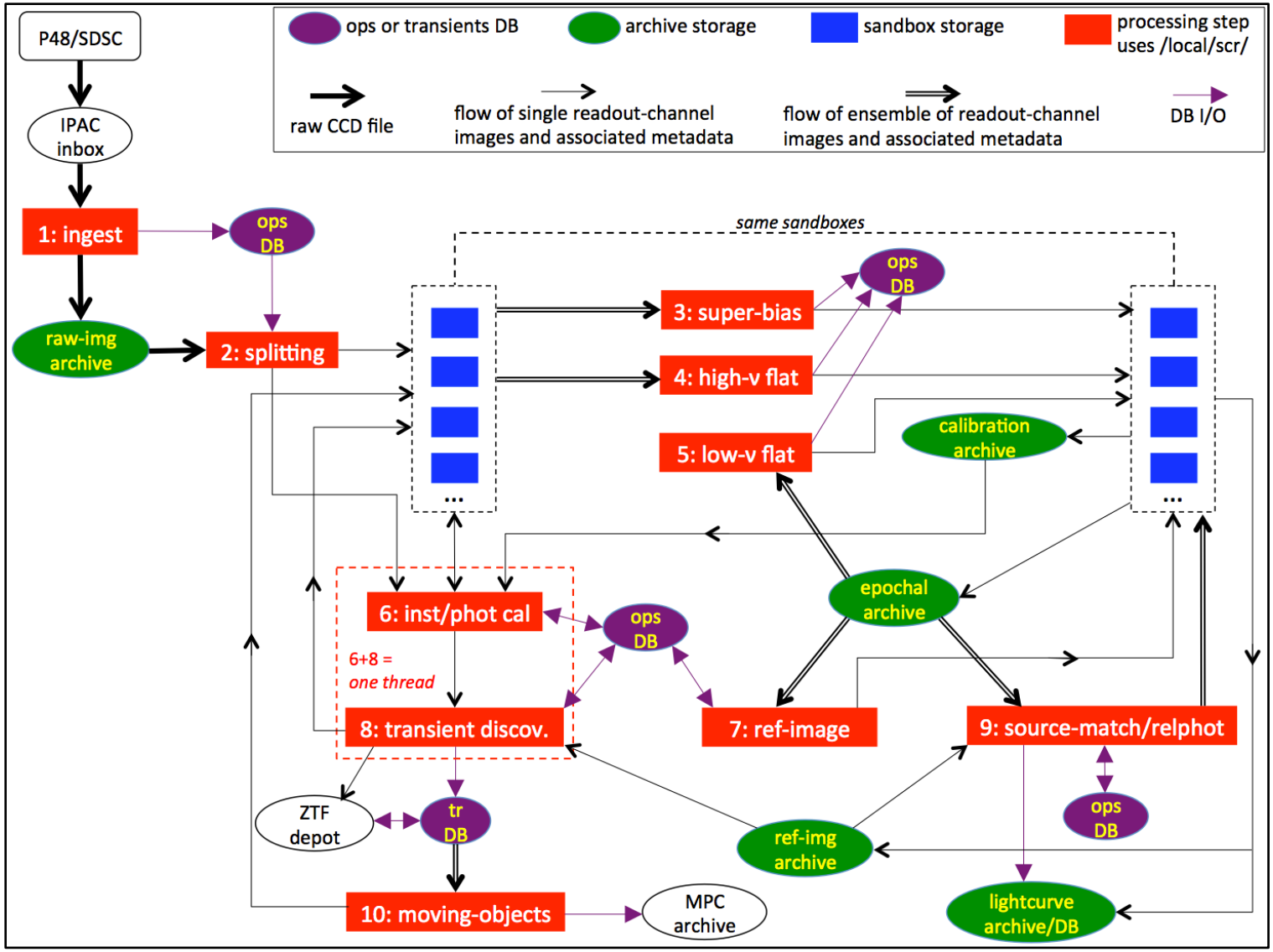
- External catalogs to support astrometric and photometric calibration in pipeline:
  - Gaia first release is expected in September 2016
  - Pan-STARRS1 catalogs expected when? Need by Jan 2017 to allow for R&D / pipeline integration

Back up slides

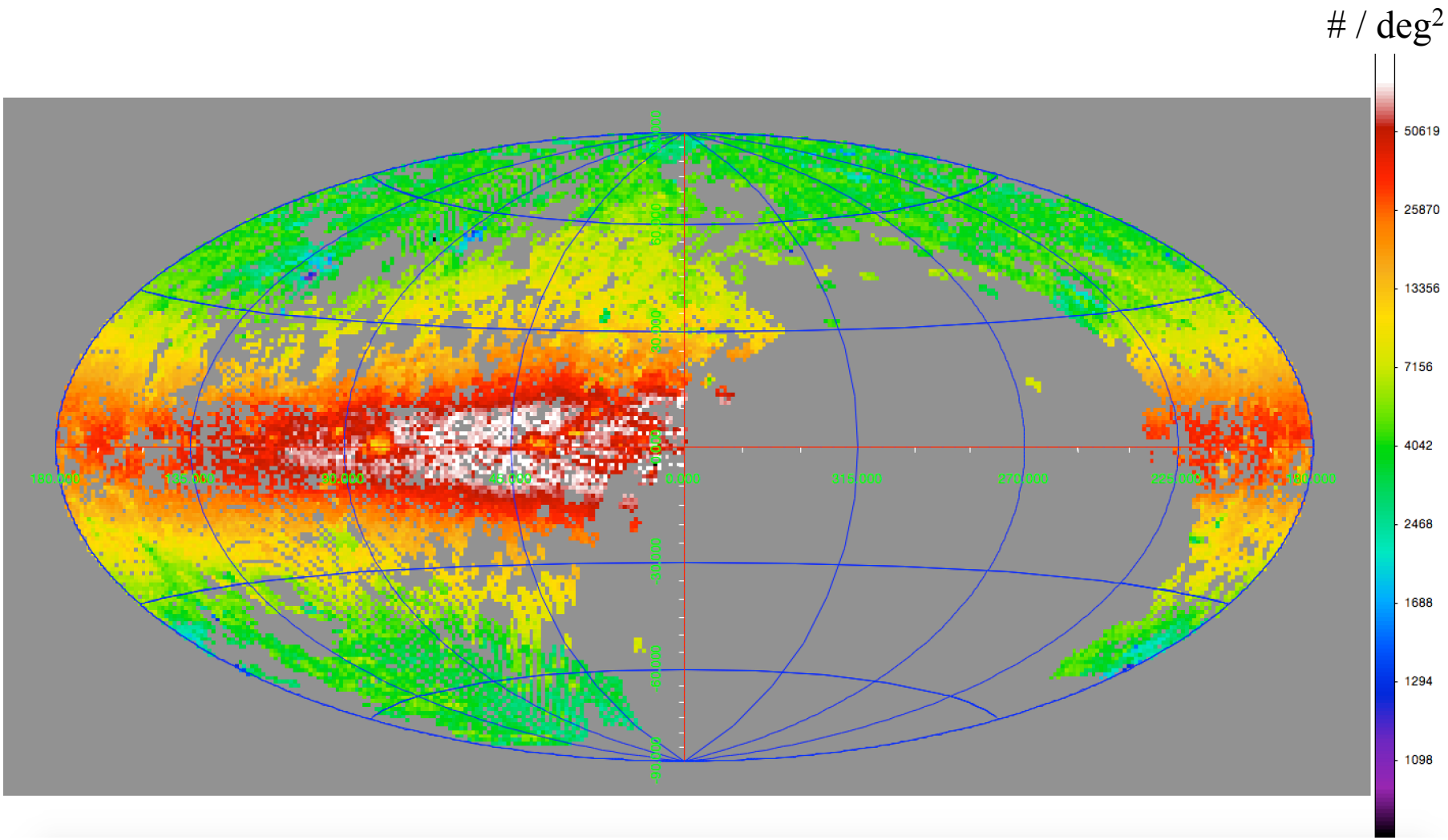
Cluster Processing Throughput



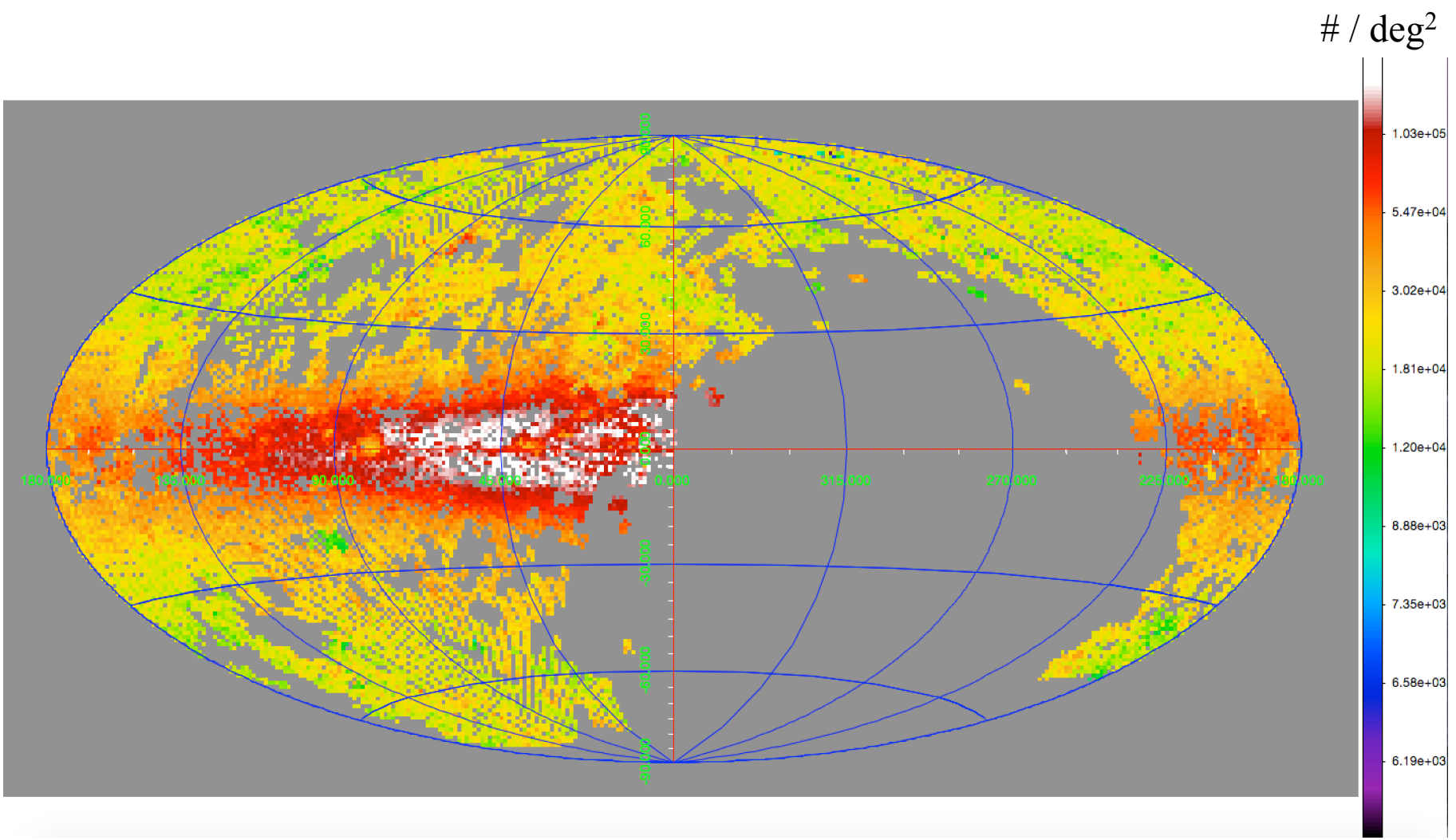
Data Flow / Processing Model (preview)



Density of aperture (SExtractor) extractions from PTF CCDs

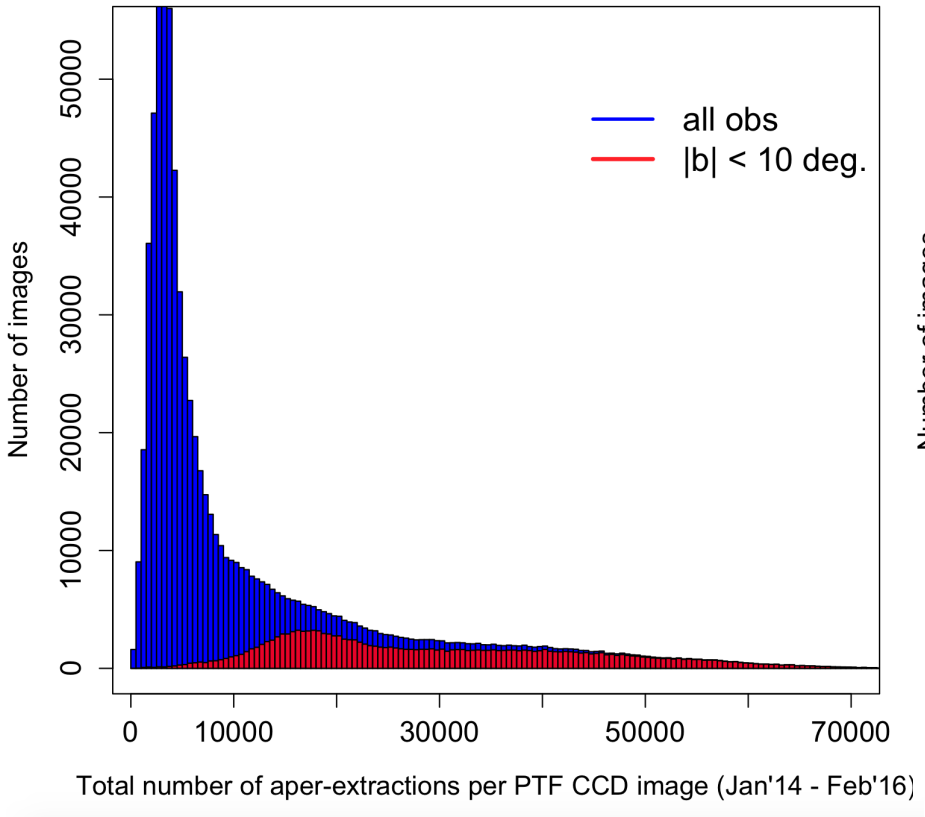


Density of PSF-fit extractions from PTF CCDs



Number of sources extracted from PTF CCDs

Aperture (SExtractor)



PSF-fitting (DAOPhot)