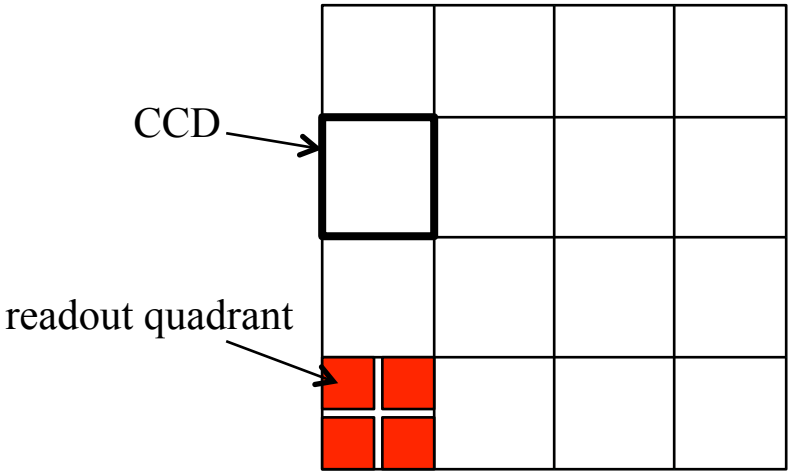
ZTF Data System and Deliverables

Frank Masci & the IPAC-Caltech ZTF Team

UW-ZTF joint meeting, April 2017

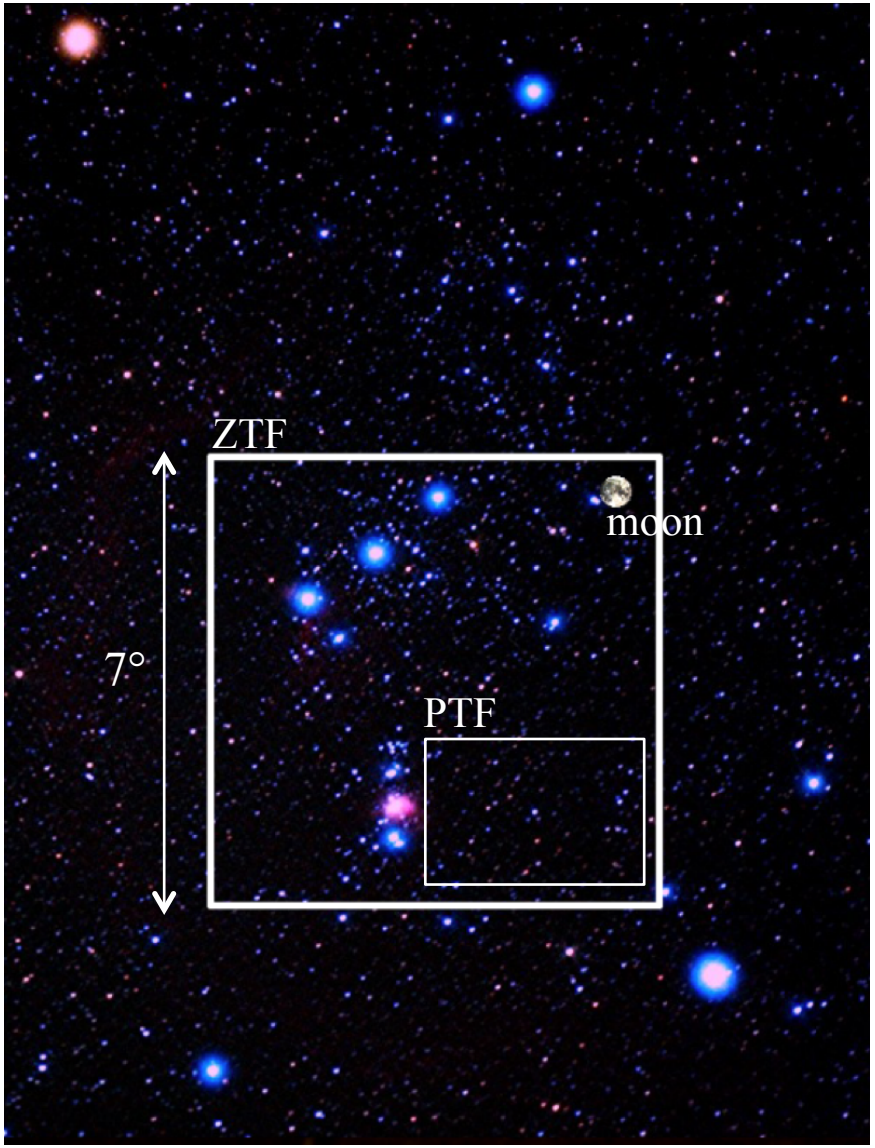
ZTF Raw Camera Image Data

- 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



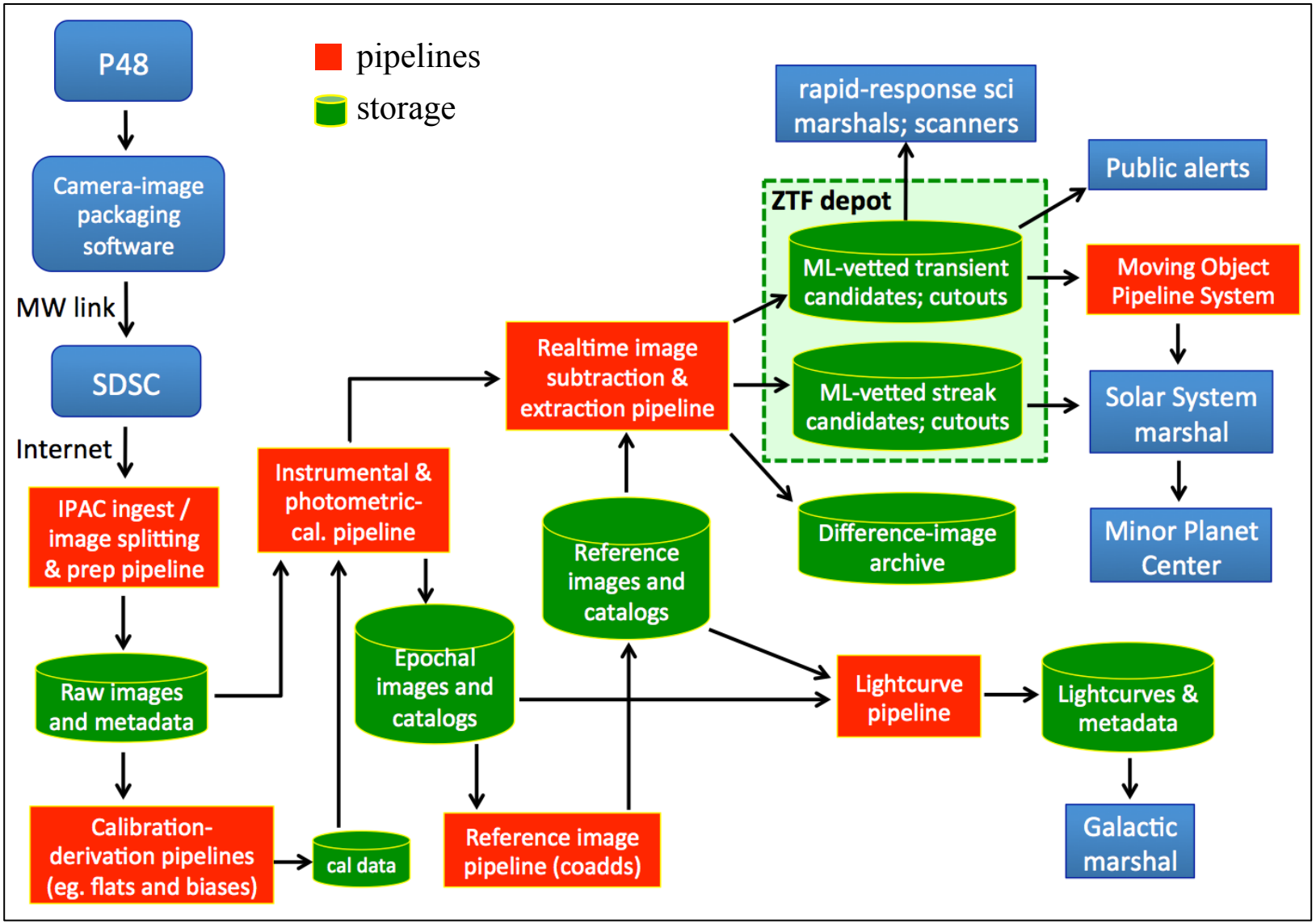
Basic image-unit for pipeline processing from which all products are derived is a ~3k x 3k readout quadrant image.

ZTF Field-of-View



- Survey rate is ~ 3760 deg^2 / hour
  Faster than the sky rotates!
- Depth: R ~ 20.4 mag AB (5σ)

Overview of the ZTF Data System



ZTF Pipelines

Overall, there are 10 inter-dependent pipelines:

Raw data ingestion/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, simple QA [*realtime*]

Calibration generation:

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. Low-ν flat from either long-term ZPVM or dithered-star observations [*every week, month or longer?*]

Real-time:

6. Instrumental calibration of readout-quadrant images: astrometry and photmetric cal [*realtime*]
7. Image subtraction and transient discovery (point sources / streaks), metadata and cutouts [*realtime*]

Ensemble-based processing:

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): tracklets from linking transients from 7 [*every 3 or 4 hours during night*]

Deliverables and Products

1. **Instrumentally calibrated, readout-quadrant based epochal image products:**
   - images with photometric zero-points derived from STF-fit photometry
   - bit-mask images
   - two source catalogs per image: PSF-fitting and aperture photometry
   - difference images with QA metadata

2. **Reference images (co-adds), coverage, uncert maps, two source catalogs per image**: PSF-fitting and aperture

3. **Match-files per readout-quadrant from source-matching of epochal extractions:**
   - based on epochal PSF-fit photometry catalogs: to support "object-based" lightcurve database:

4. **Products to support near real-time discovery:** *thresholded* transient candidates (point sources and streaks) with metadata and image cutouts: pushed to "ZTF-Depot"

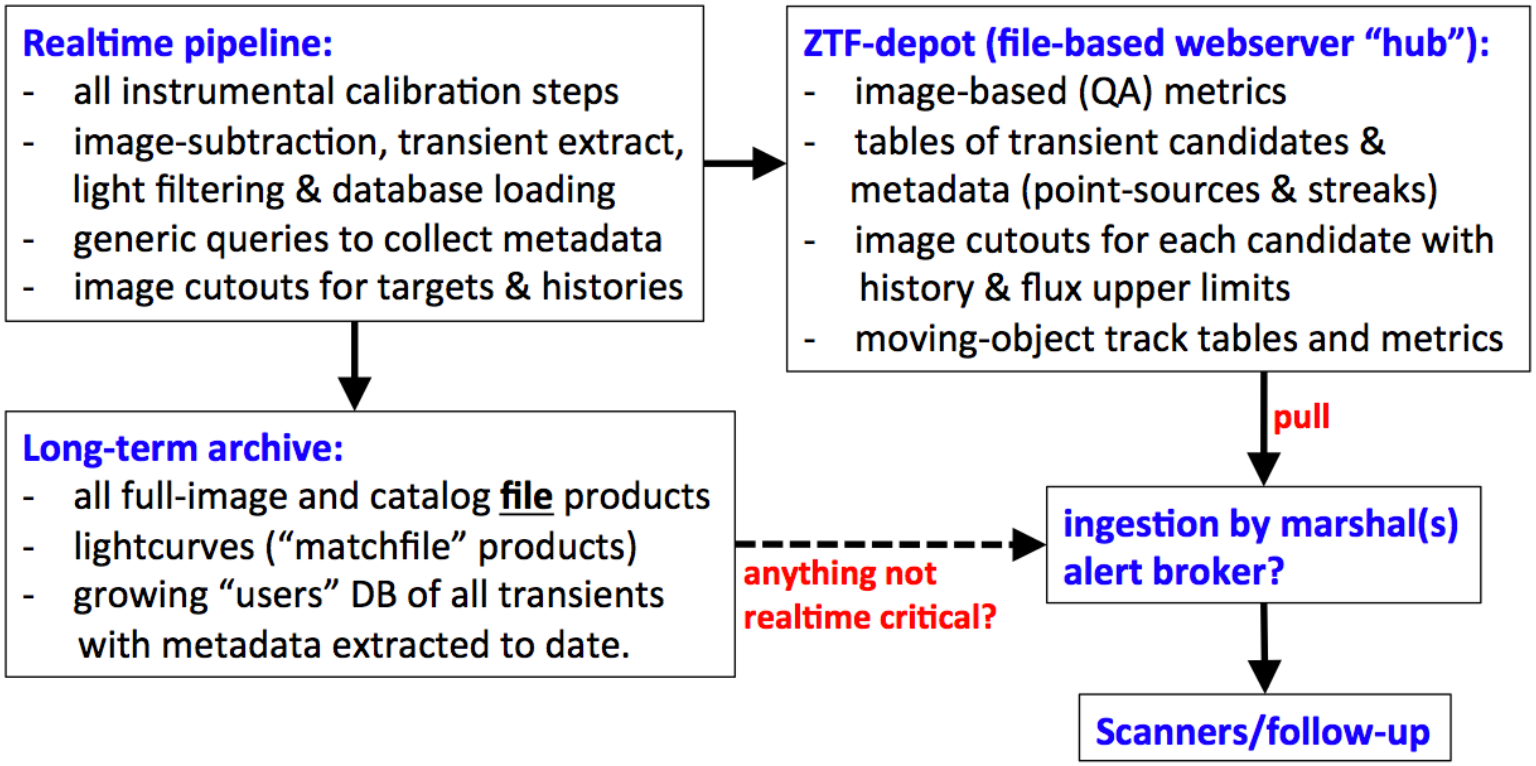
5. **To commence following survey start**: alert (event) stream extracted from real-time pipeline with metadata

6. **Historical (users) database of all transient candidates and metadata generated from real-time pipeline**

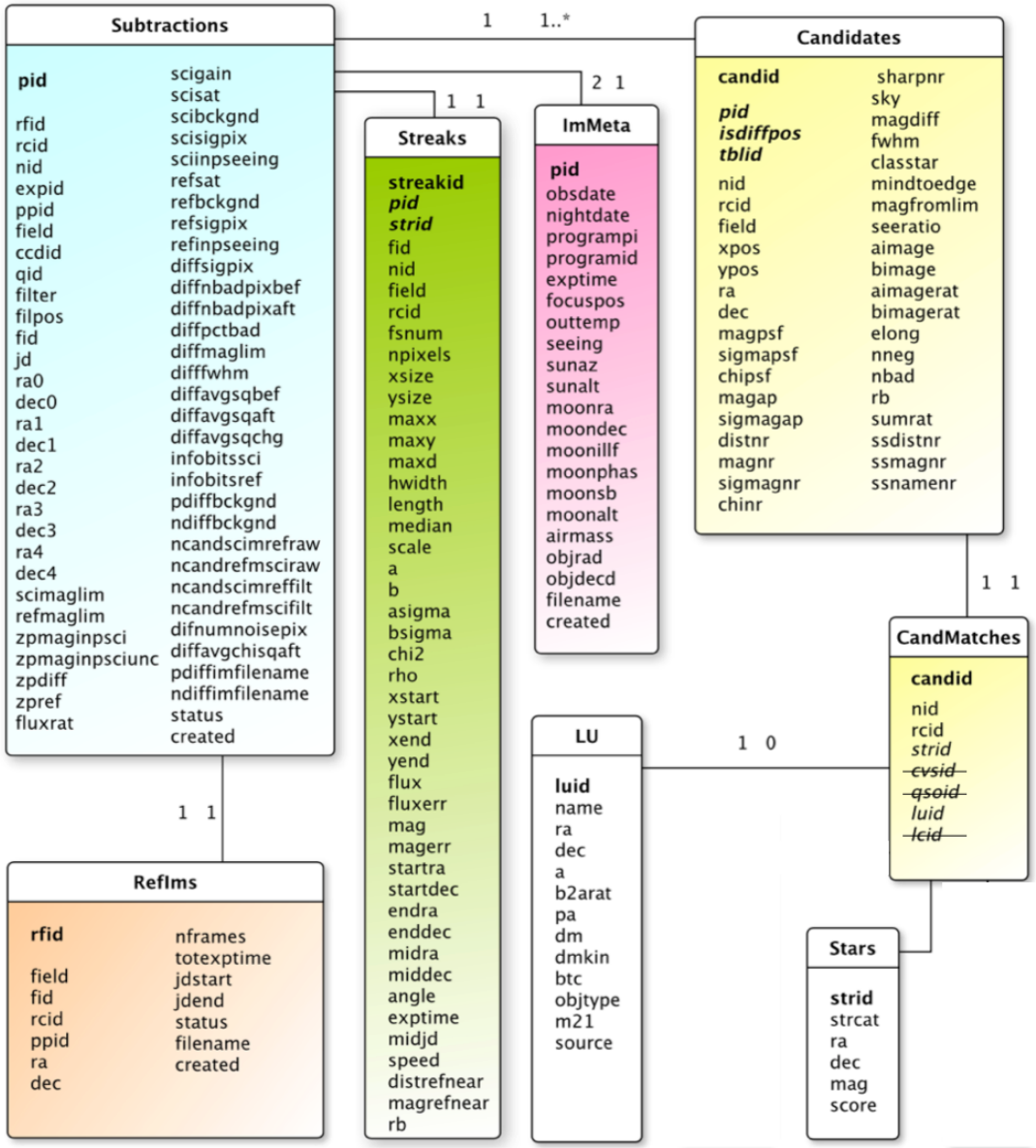
7. **Products to support Solar System/NEO discovery and characterization:**
   - moving object tracks from linking point-source transients; known objects are tagged: delivered to MPC.

ZTF Archive and Depot

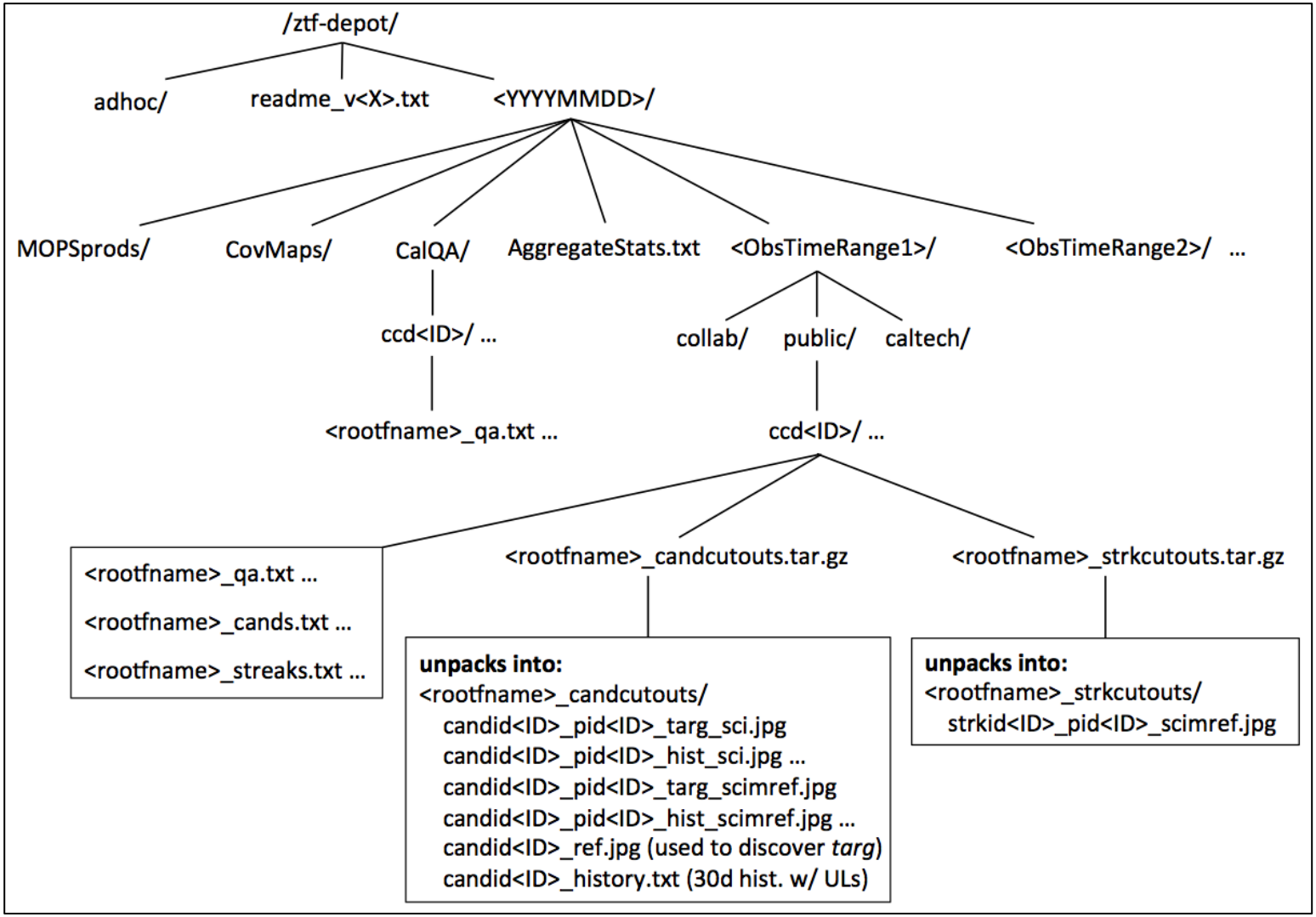
- **ZTF-Depot:** to support "fast response science": plan is to deliver a generic event stream (following automated vetting or light filtering in pipeline) to a webserver for collection by collaboration.
- **ZTF-Archive:** all products, including all extracted events from survey can be retrieved from archive



ZTF Transients/Events Database

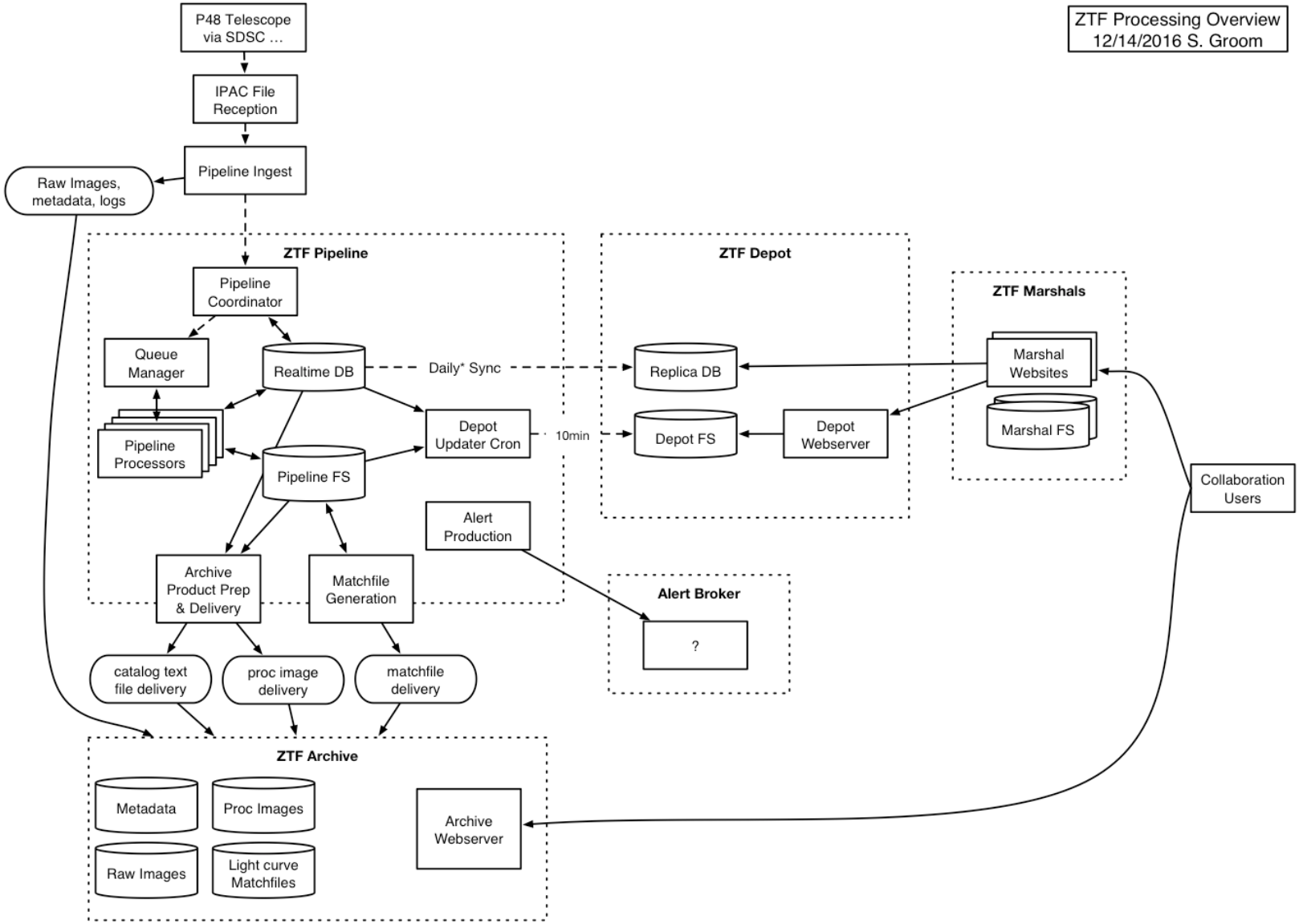


ZTF Depot file system

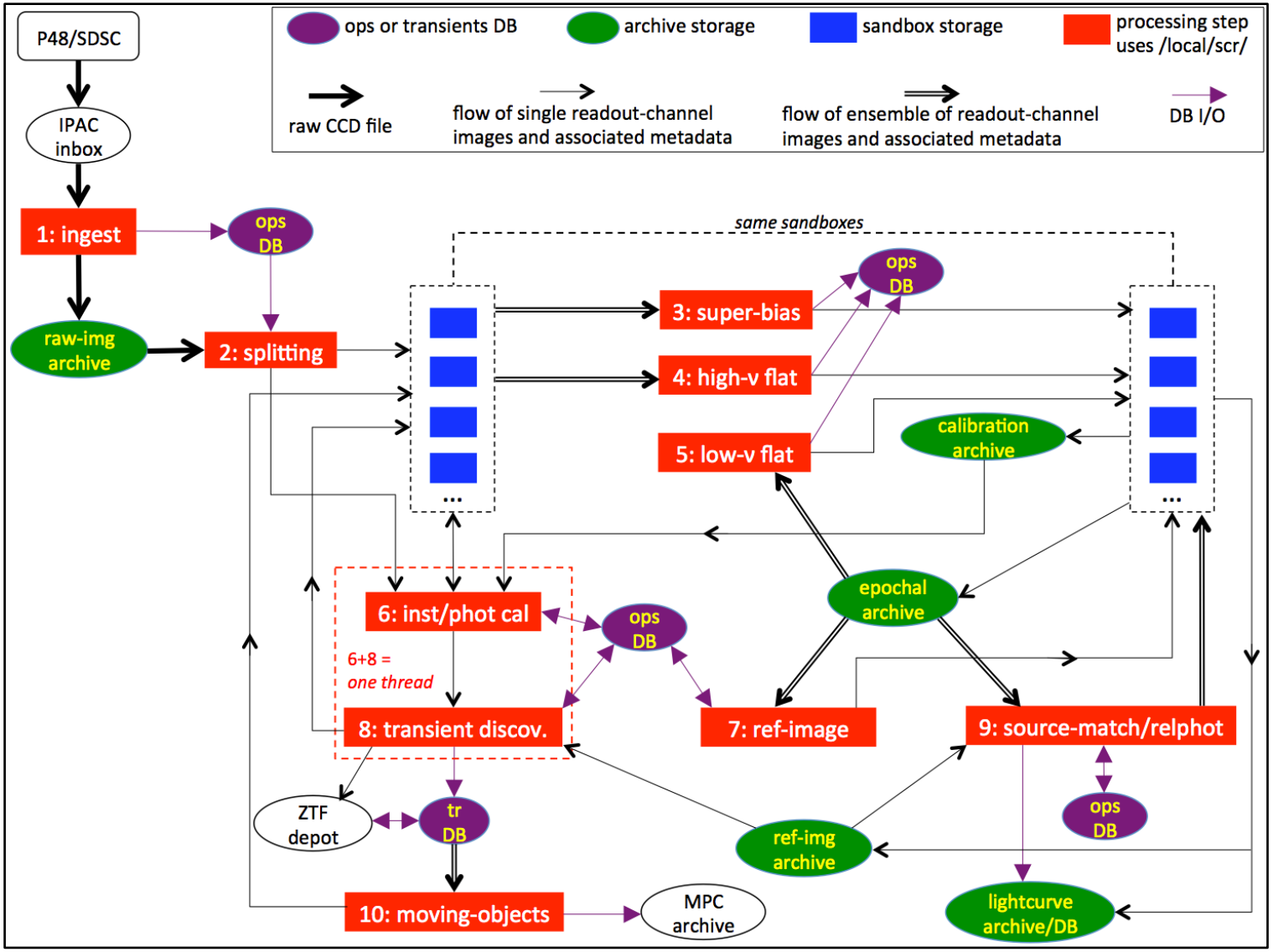


Back up slides

Architecture Details



Data & processing flow



ZTF Public Delivery Schedule (provisional)

- **First data release**: survey start + 12 months: ~ Dec 2018
- **Second data release**: survey start + 18 months: ~ Jun 2019
- **Third data release:** survey start + 24 months: ~ Dec 2019
- **Fourth data release:** survey start + 30 months: ~ Jun 2020
- **Fifth data release:** survey start + 36 months: ~ Dec 2020

- **Survey start:** ~ late 2017

- **Core deliverables for the above:**
  - epochal science images + catalogs + ancillary products (metadata)
  - co-add images + catalogs + ancillary products (metadata)
  - above products searchable through user-inteface according to spatial constraints/survey parameters
  - lightcurve access through user-interface (refined source-matching from epochal PSF-fit catalogs)

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 ~ 1 TB

Total per night: ~ 5.65 TB

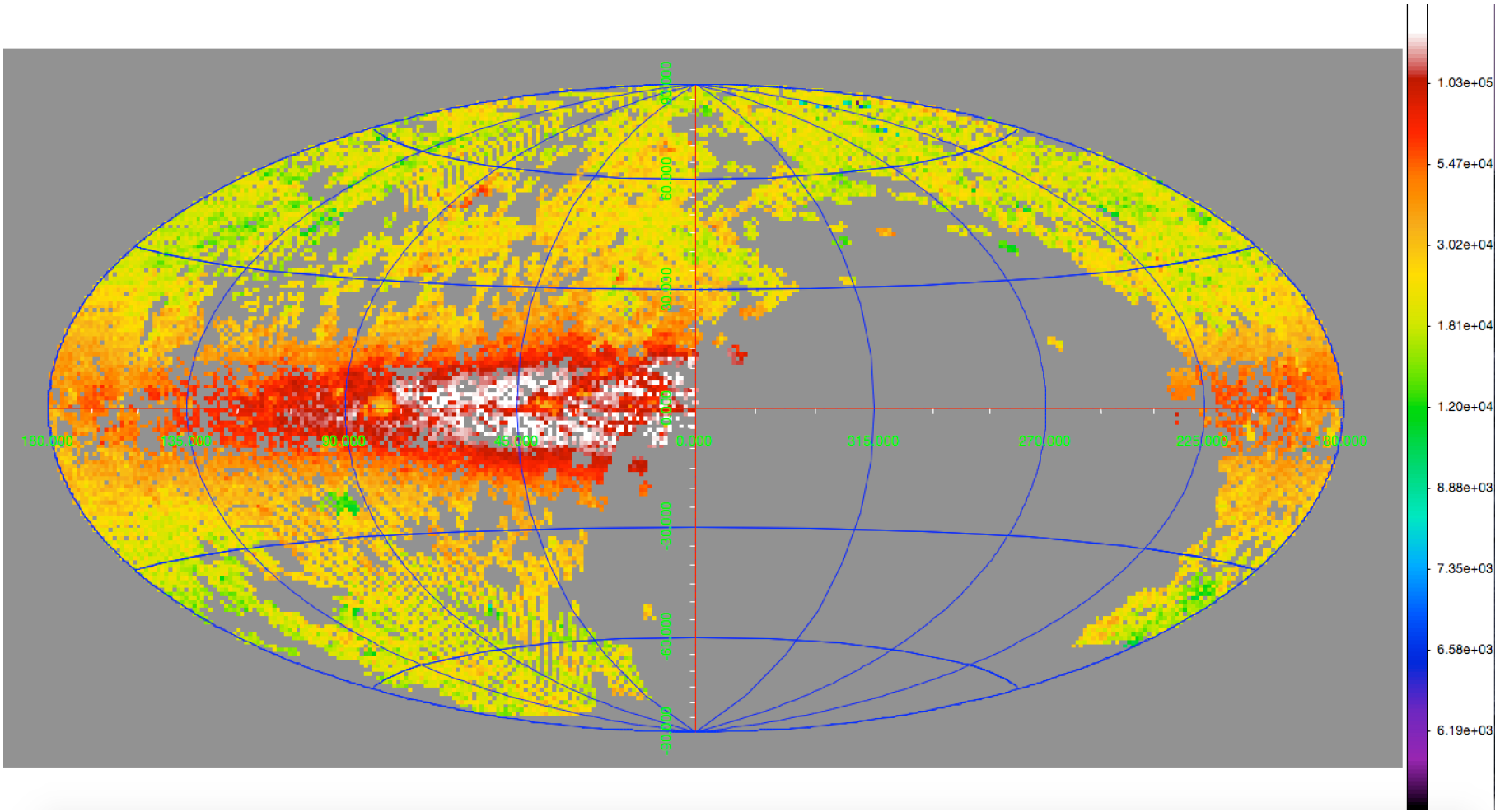
For three-year survey:

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

Total image/catalog file products: ~ 3 PB

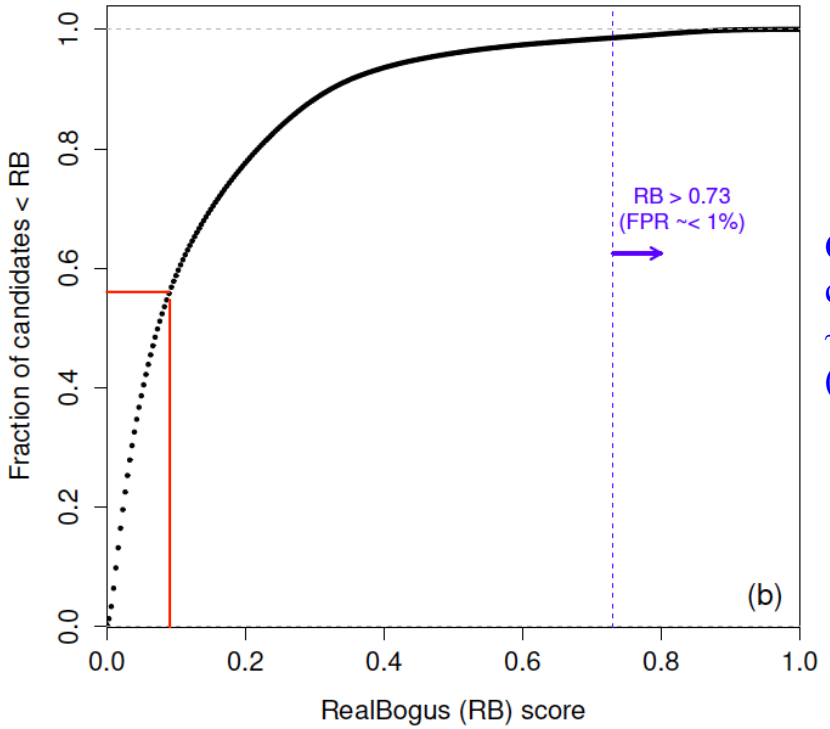
Density of PSF-fit extractions from PTF CCDs

Extrapolating from PTF, we expect of order 900 million PSF-fit extractions **per night** for ZTF(!)



Benefit of Machine Learning

- 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).

Expected number of transients/cutouts for ZTF

- Assume (following internal real-bogus filtering) ~ 20 point-source candidates per readout-channel image.
- With ~ 700 exposures / night resulting in 44,800 positive subtraction images, this leads to ~ 1 million *target* point-source candidates / night.
- These exposures (and candidates) are ~ equally apportioned across multiple filters and two epochs (coverages) / night. This point is not relevant to the overall sizing estimate.
- For each *target* point-source candidate, there will be three image cutouts (*sci, ref, and sci-minus-ref*), yielding ~ 3 million *target* candidate image cutouts / night.
- We also expect ~ 200,000 *streak*-candidates / night, from which only *sci-minus-ref* image cutouts will be generated.
- From the last two bullets, that's ~ 3.2 million image-cutouts per night (or ~ 5,300 cutouts / minute) *for target candidates only.*
- At the time of writing, each JPG cutout is ~ 1.3kB in size. This amounts to ~ 4.2 GB in cutouts alone per night (or ~ 7 Mbytes / minute) *for target candidates only.*
- Assuming **Ndhist** = 3 days when generating *forced* historical image cutouts on each *point-source target* candidate only, the above numbers are multiplied by three, to yield potentially ~ 9 million cutouts per night (or ~ 15,000 / minute).