DARK ENERGY
SURVEY

The Dark Energy Survey Data Management System

Robert Gruendl
NCSA/University of Illinois

DARK ENERGY
SURVEY

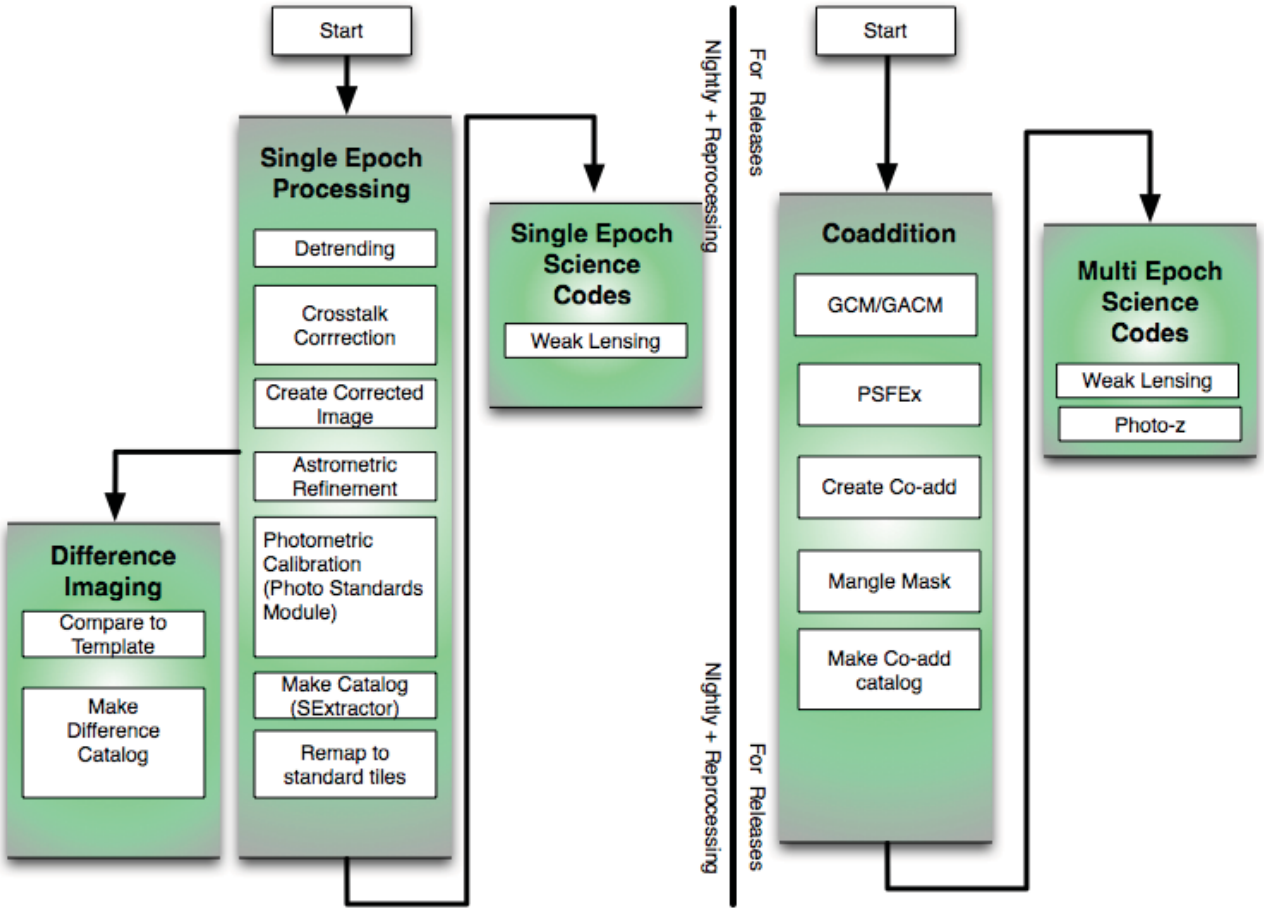
Outline

- A Brief History of Pipelines
- Overview of the Main Pipeline Modules
- Some Lessons Learned and Challenges that LSST might still have time to avoid

Data Challenge 6B (2011)
12 months before DECam first light

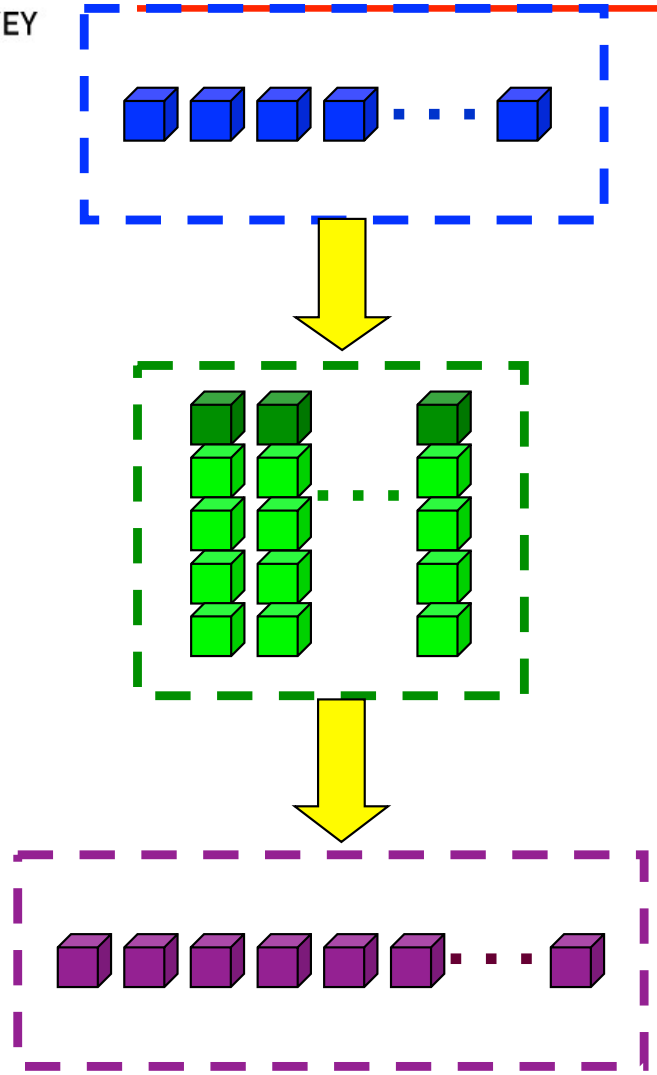
DARK ENERGY SURVEY

- Reduced 10 simulated nights (**relatively ideal**)
- SE-pipeline (**a monolith**) took ~36-48 hrs to process one night (**required success for all exposures in a night**)
- Calibrations depended on good nightly data (e.g. **nightly illumination/fringe correction**)
- SNDiff pipeline needed flat with **~5-10x fidelity** achievable with a single night's calibrations
- Many "science codes" (e.g. WL and photo-z) were undergoing heavy, continuous development by WGs (**poor turn around of tests**)



Pipeline Parallelism

DARK ENERGY SURVEY



Crosstalk Block: X pipeline jobs

Modules in pipeline: Crosstalk

CreateCor Block: Y pipeline jobs

Modules in pipeline: mkbiascor, mkflatcor

Note: mkflatcor is repeated for each band

Detrend Block: Z pipeline jobs (Z >> Y)

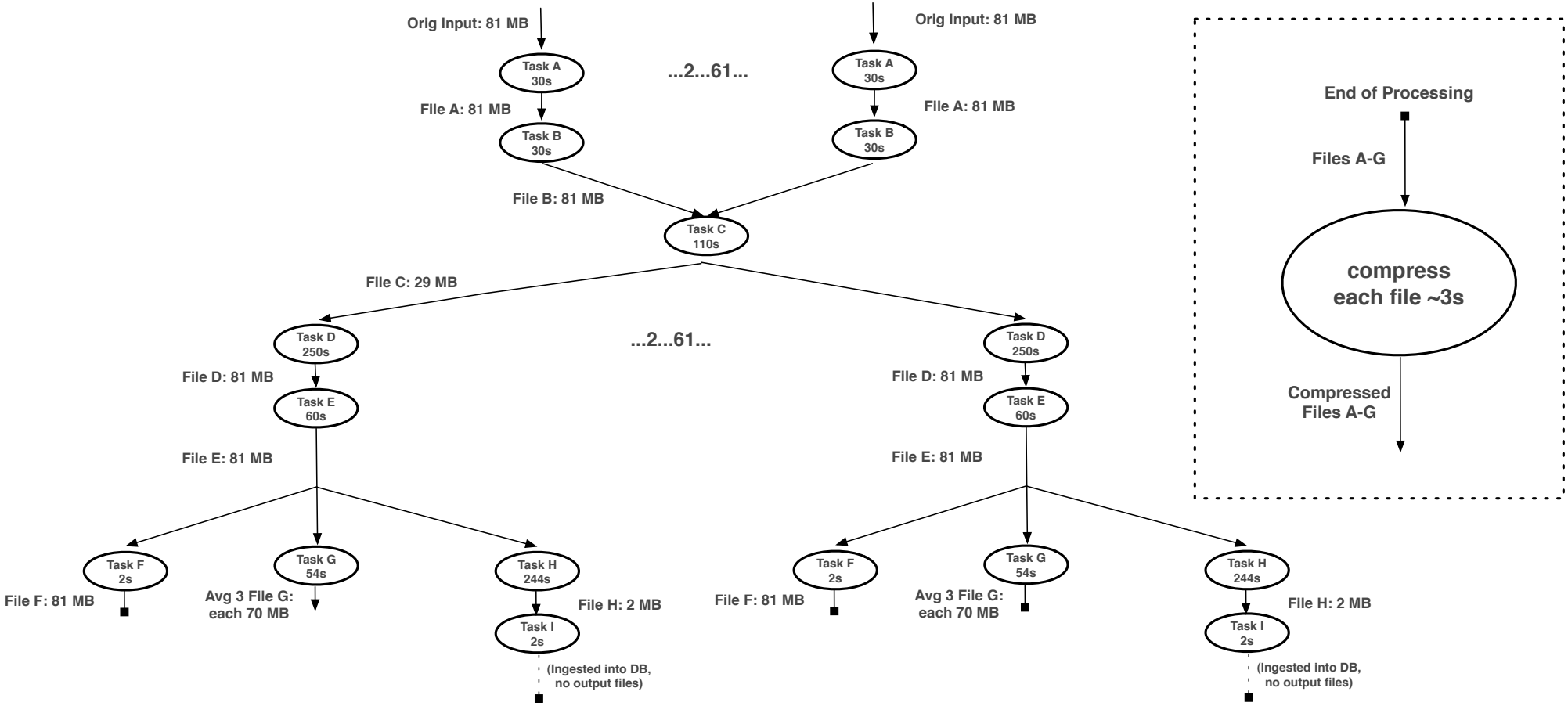
Modules in pipeline: imcorrect

Example for Execution Paths

DARK ENERGY

FinalSE Pipeline

All numbers are estimates with some more accurate than others
Codes use less than 2G memory (Most less than 1G)
Extra calibration files not illustrated ~18G total
Misc smaller output files not illustrated
62 pieces almost independent trivially-parallel pipelines (exception task C)



DARK ENERGY SURVEY

The Intermediate Solution

Bring the computation closer to the data preferably on resources with a well provisioned file system

- The original DESDM proposal was that it would use HPC resources available through TeraGrid/XSEDE centers. In reality these were not ideal for most of the processing needs
  - Always required heavy file transport
  - Typically had LUSTRE file systems (not good for heavy i/o)
- Arranged for use of iForge (part of the PSP program at NCSA)
- NCSA also stood up a GPFS storage condominium (to better serve both DES and other facility projects)

Revamp DB infrastructure

- Eliminate live replica versions at secondary sites.
  - Helped removes latency/bottlenecks when ingesting large amounts of data
- Replace web-portal DB access with light-weight tools
  - Open access for the DES Collaboration to all data

DARK ENERGY SURVEY

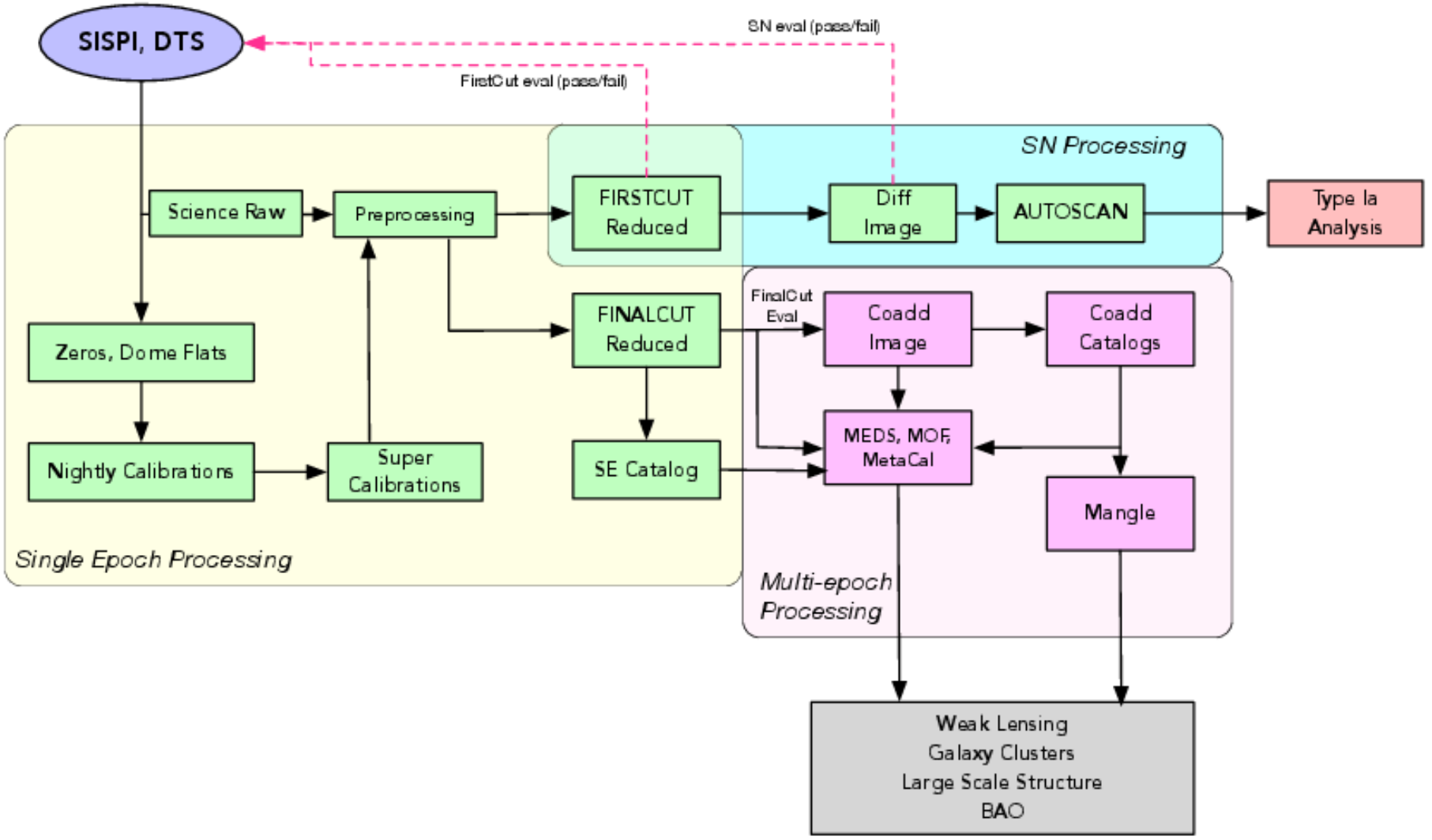
The Intermediate Solution

Break up the pipeline computation into native chunks with realizable goals

- SE processing was altered to work on individual exposures
  - one problematic exposure no longer halts all processing
- Pipelines responsible for building calibrations decoupled from science pipeline processing
  - Long term boon because DECam is an incredibly stable instrument
  - Removed the one-and-done mentality for calibration
  - Allows knowledge about the changing state of the instrument to filter back and inform calibration process
- FIRSTCUT → rapid feedback/QA
- FINALCUT → best reduction possible (after calibration is understood)
- Pipelines with heavy development were temporarily off-loaded to WGs
  - Most have been re-united with DESDM processing as afterburners as of the Y3 production campaign

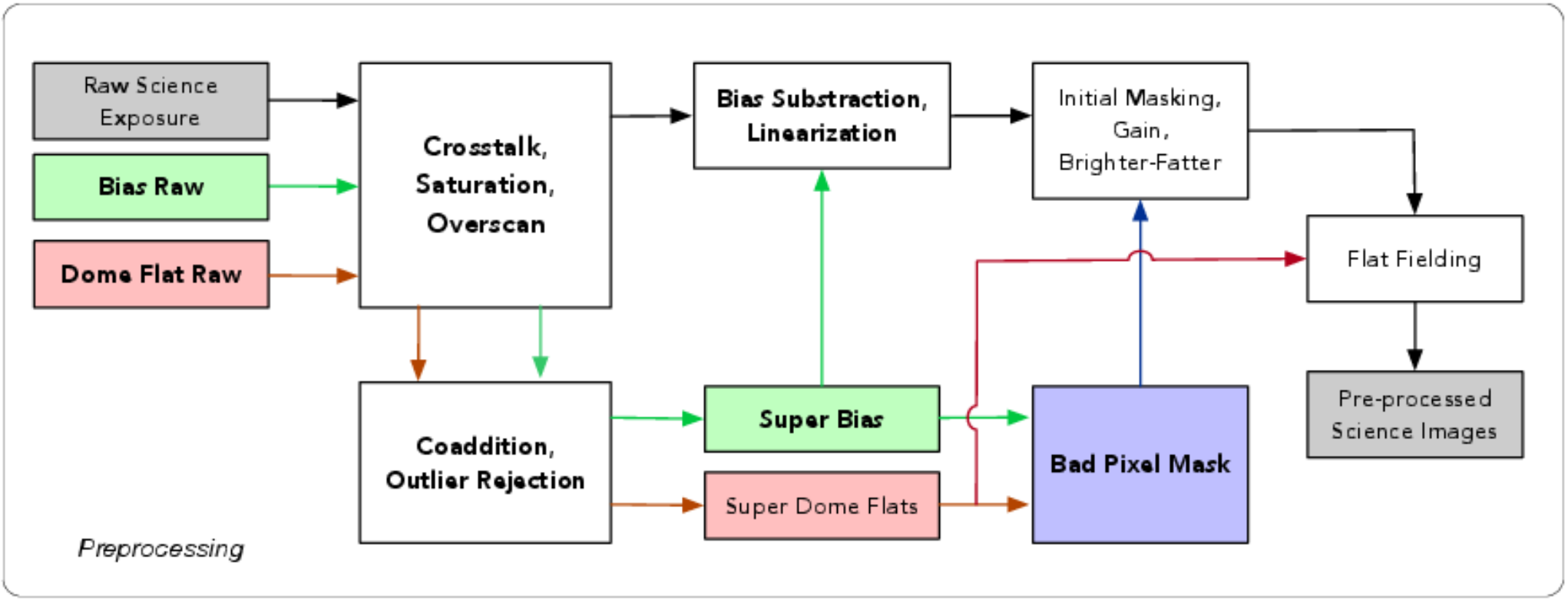
Current Dataflow and Pipelines

DARK ENERGY SURVEY



Detrending in the DES SE pipelines

DARK ENERGY SURVEY



DARK ENERGY SURVEY

The Long-Term Solution

Revamp Middleware

- PERL → Python
- Allow for use of Open Science Grid resources (e.g. FermiGrid)
  - \+ More locations to run
  - \+ Use of a local filesytem where only local job(s) are competing for i/o.
  - − More data transfer necessary (all calibrations to each node)
- Require unique filenames for all products
- Allow for provenance tracking (using the Open Provenance Model) so that the interdependence of products can be traced

Revamp Science Codes so that Collaboration more able to contribute

- Move from C/C++ → Python (w/ wrapped C where needed)
- Bring DES Collaboration expertise to bear
  - More eyes, more minds (spot and understand problems/"features")
  - Provide solutions (up to and including actual code)
- EUPS → to provide more regimented production environment
  - Also provides for stand-alone development (and even stand-alone production)

DARK ENERGY SURVEY

The Long-Term Solution (hardware)

Replace iForge with dedicated DES production nodes

- Illinois Campus Cluster (with 40 Gb/s, 3-hop to main file storage)\
- Well provisioned nodes 28 cores, 8GB/core, 12 TB local disk
- Use FermiGrid and BlueWaters for peaking

Second Generation DB Hardware

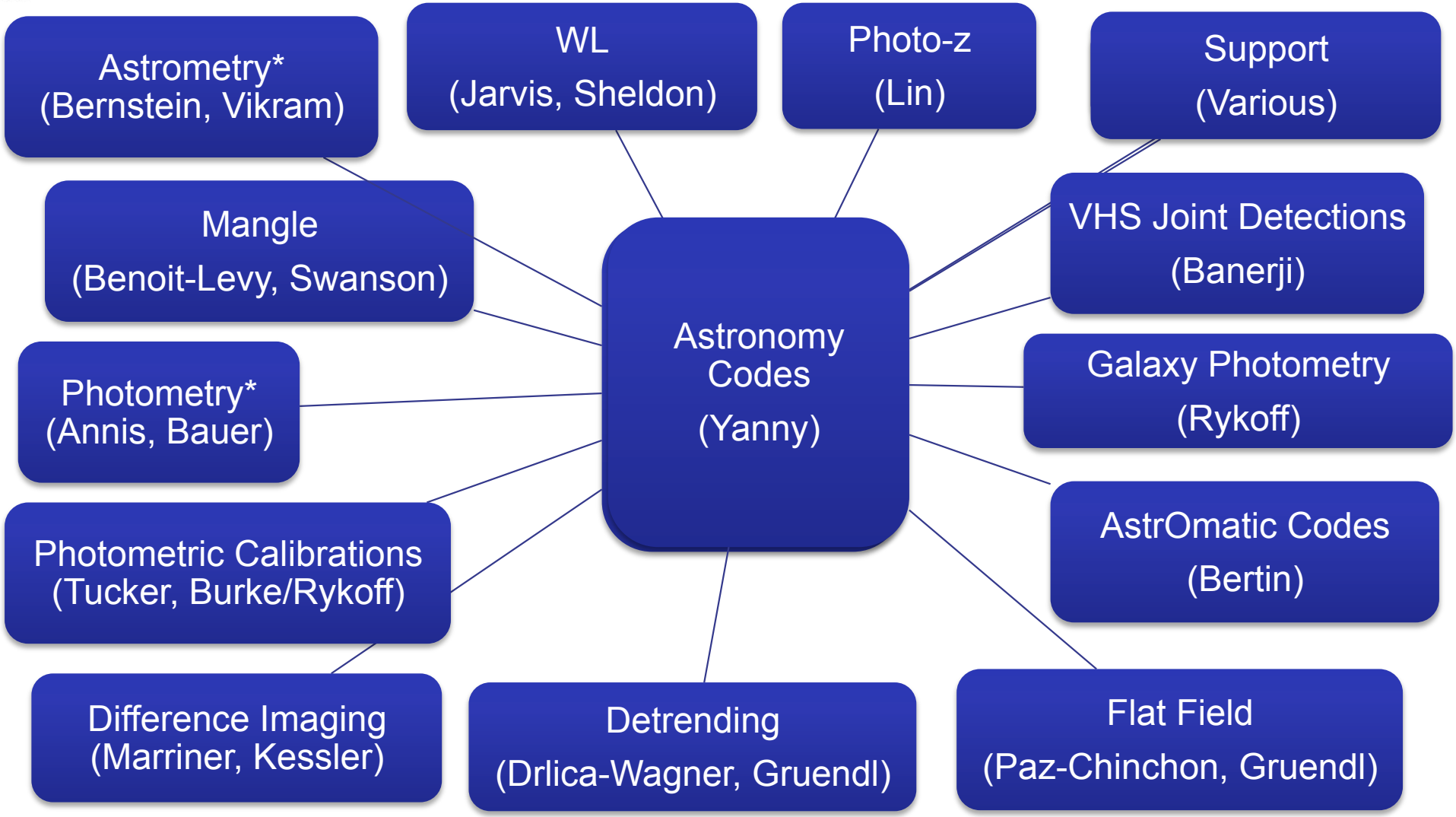
- Replaces a cobbled together system (stone-soup DB cluster)
- Adds SSDs for user and temporary table spaces

More nimble Disaster Recovery

- Spinning Backup (using JBODs composed of shingle media)
- Provides nimble than recovering from tape

DARK ENERGY SURVEY

Distributed Development



DARK ENERGY SURVEY

Pipelines: FirstCut vs. FinalCut

Minimal processing needed to provide data quality assessment:

- Detrending:
  - MIN: demonstrates that artifacts are properly understood
  - +++: early science is possible
- Astrometry:
  - MIN: exposure location verified, distortion is stable
  - +++: moving and variable object discovery
- Cataloging
  - MIN: rough PSF analysis (FWHM, ellipticity, 2nd moments)\
  - MIN: depth/sensitivity are adequate
  - MIN: preliminary photometric solution
  - +++: preliminary science, moving and variable object discovery

DARK ENERGY SURVEY

Pipelines: FirstCut vs. FinalCut

FirstCut: uses a preliminary set of calibrations and produces a result suitable to for Data Quality Assessment and feedback to the observing team.

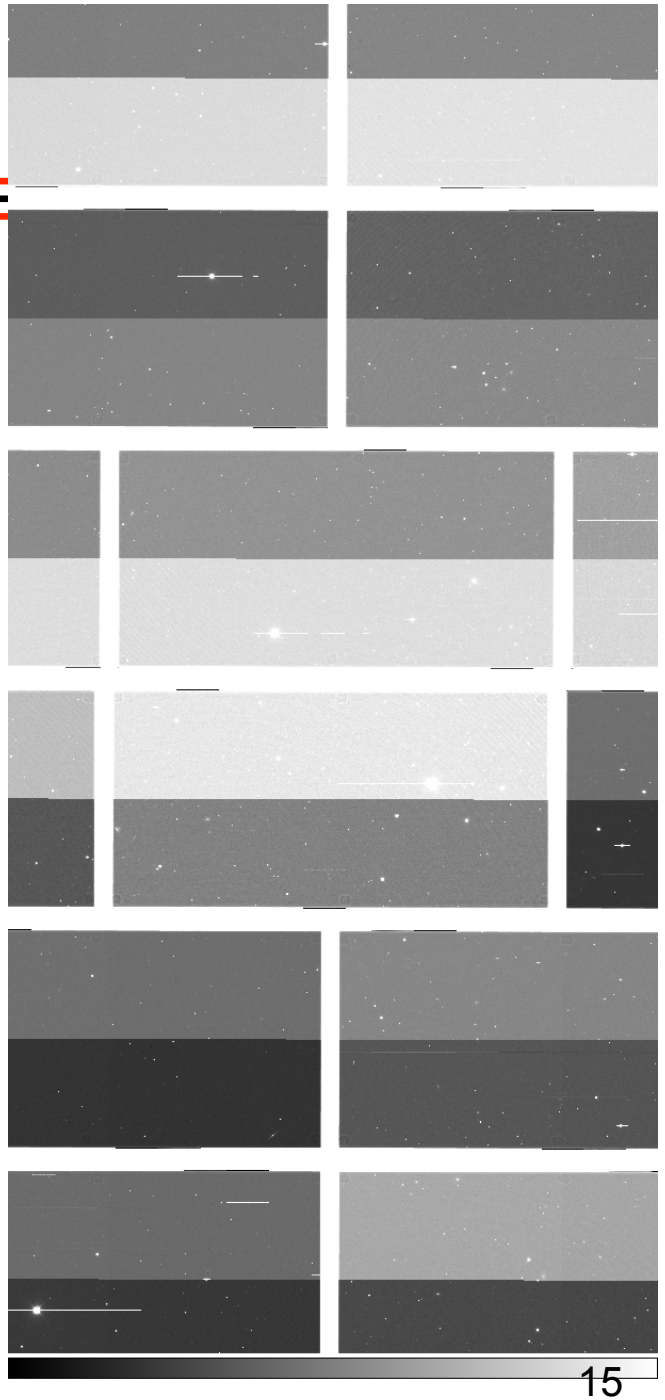
FinalCut: uses a best set of calibrations (draws from experience and QA gained running FirstCut) and produce results that are release to DES Collaboration (and eventually the astronomical community).

By Year 1, the pipelines were streamlined to the point that software changes incorporated in FinalCut were routinely made available for FirstCut processing.

DARK ENERGY
SURVEY

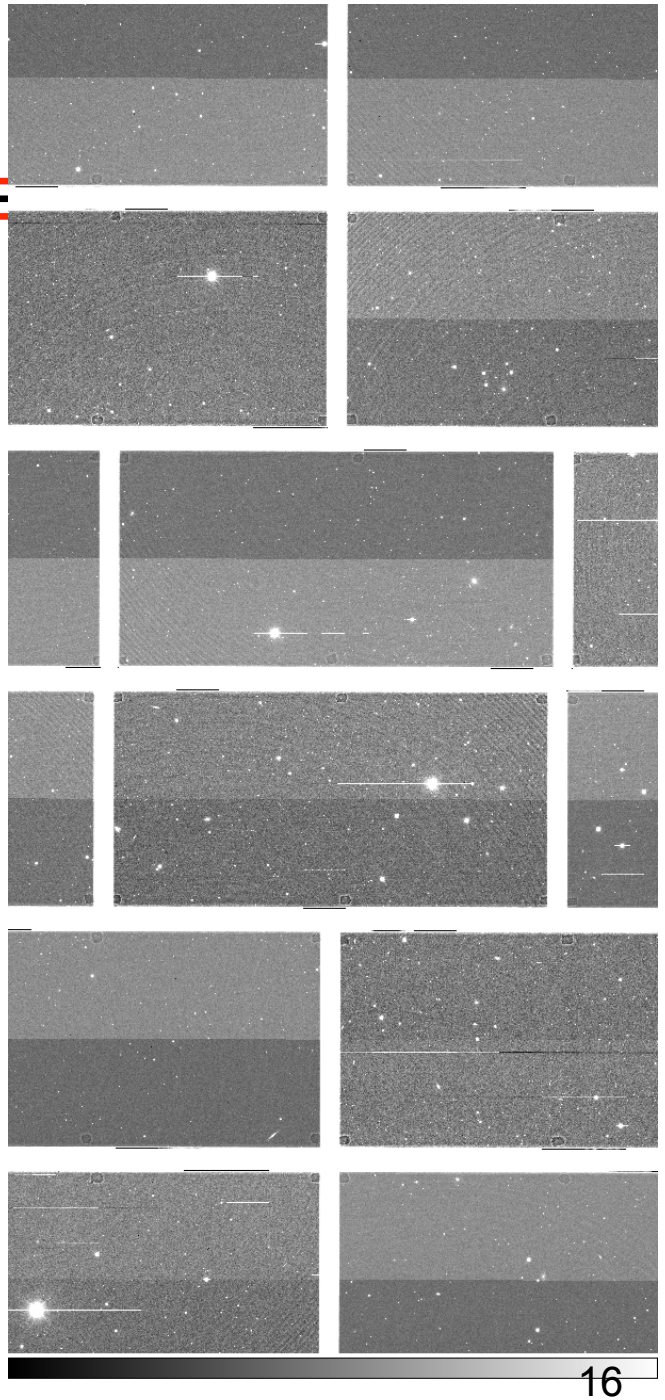
DECam (raw)

Raw Exposure from the telescope



DARK ENERGY SURVEY

Crosstalk & Overscan



Remove overscan:

- Currently uses line-by-line average with outlier rejection
- Functional fitting and splines would require knowledge of bias jumps for backplanes containing focus chips

Crosstalk:

- Crosstalk removal (mostly inter-ccd) but has a non-linear behavior and super-saturated sources cannot be corrected.

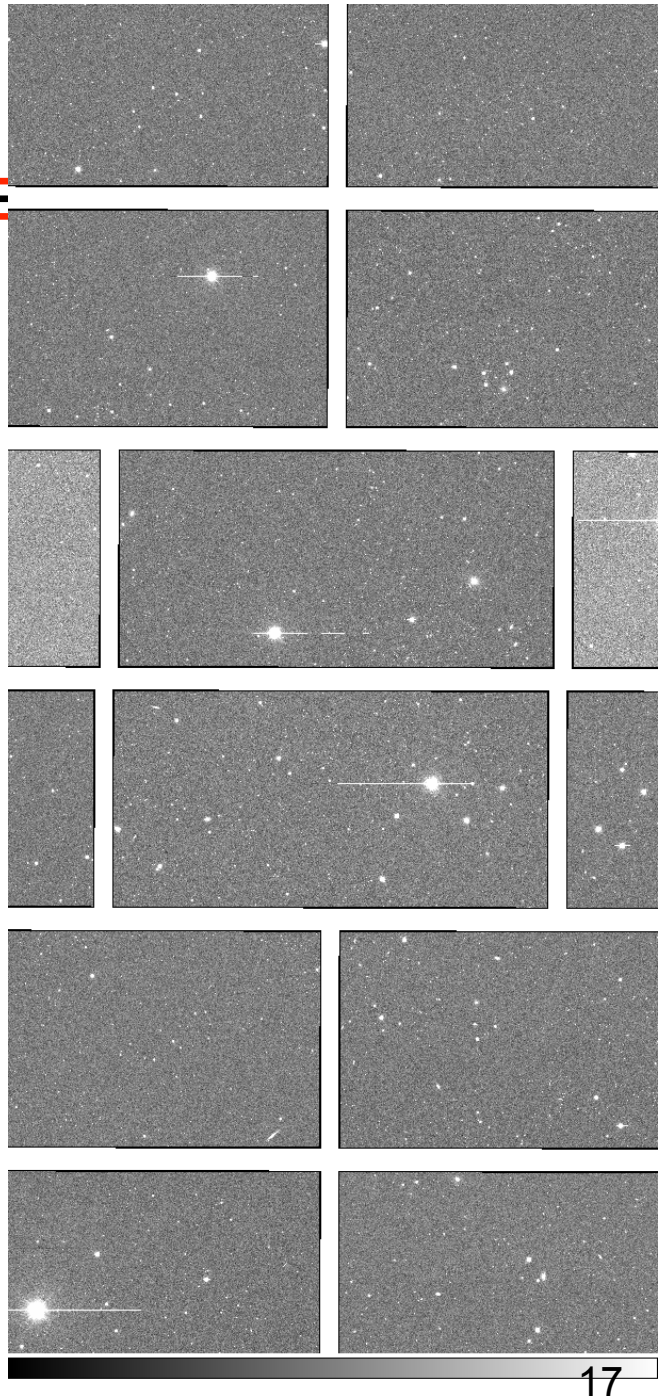
DARK ENERGY SURVEY

Detrend

- Bias (either nightly or super-bias)
- Linearity Correction
- Gain Correction (**added for Y3**)
- Brighter Fatter (**Y3A1 and later**)
- Flat (super-flat)
  - Y3A1 switched to normalization across focal plane to enable full focal plane sky subtraction (not shown in figure)
- ~~Pupil/Illumcor derived from starflats~~
- ~~Fringe (zY-bands only)~~

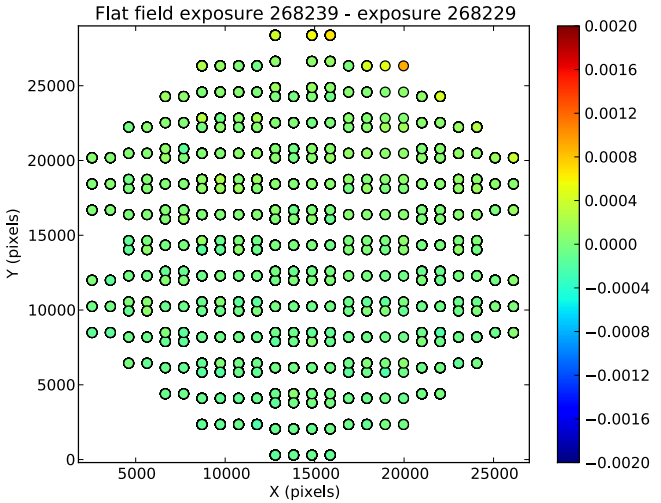
Replaced in Y3

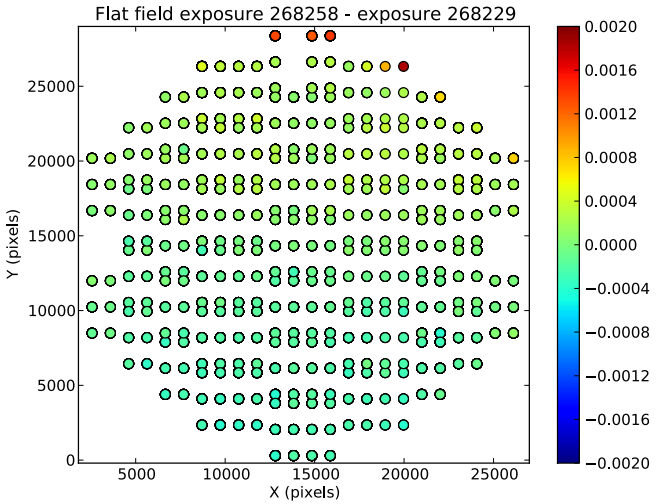
- Y3: replaced by full focal plane sky subtraction using a PCA fit including fringe
- Y3: Illumumination Correction → Star Flat



Intentional Dome Misalignment: ~2 mmag effect

DARK ENERGY SURVEY



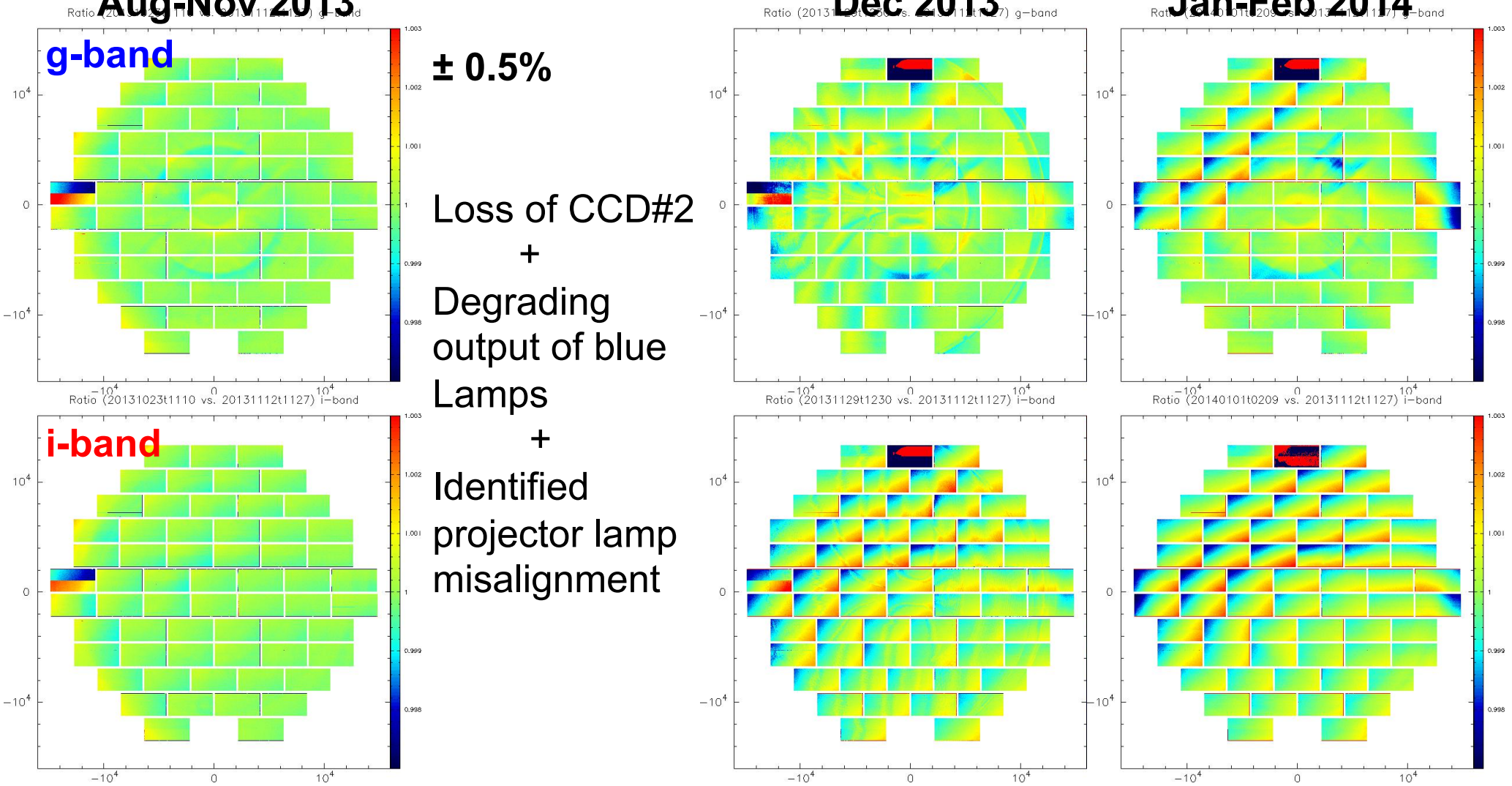


Experiment by
G. Bernstein and
D. James

Flat Field Monitor

DARK ENERGY SURVEY

Aug-Nov 2013 | **Dec 2013** | **Jan-Feb 2014**



± 0.5%

Loss of CCD#2
+
Degrading output of blue Lamps
+
Identified projector lamp misalignment

DARK ENERGY SURVEY

Astrometry

Up through Y3A1:

Used SCAMP and UCAC-4 along with a single predetermined distortion correction → 250 mas RMS

Y4:

Switched to GAIA-DR2 for a reference → 70 mas RMS

Add per epoch distortion estimates (from star-flats) → 25-50 mas RMS

Astrometric Distortion

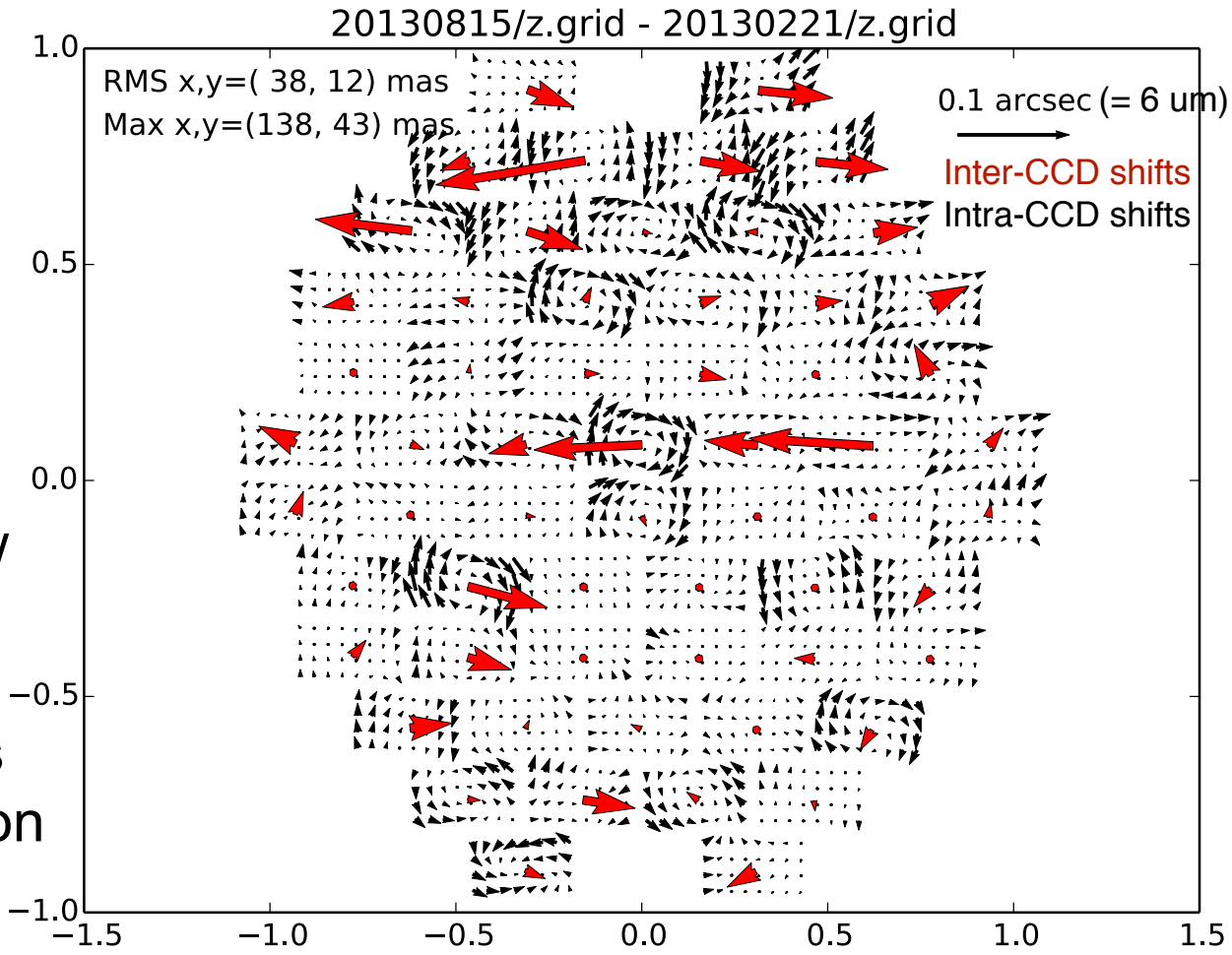
DARK ENERGY SURVEY

Y1:

- update addressed band dependence issues but also identified temporal changes in astrometric distortion pattern

Y4:

- Temporal variation is now accomodated within the pipelines.
- Traced to actual changes in CCD location/orientation when camera warm-up occurs

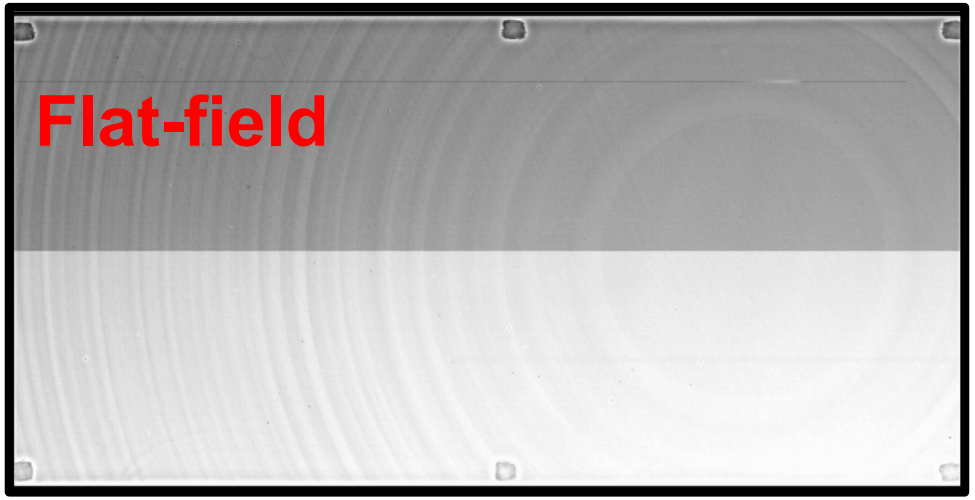


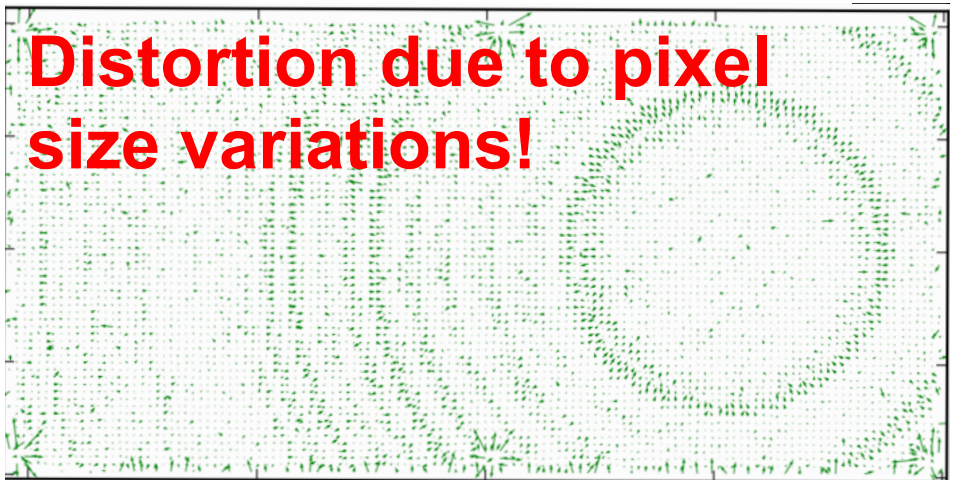
(from G. Bernstein)

Astrometry (longer term)

DARK ENERGY SURVEY

Current (Y5) effort will attempt to refine relative astrometry still further/ Tests show systematic residuals of 50-80 mas can be modeled and removed.

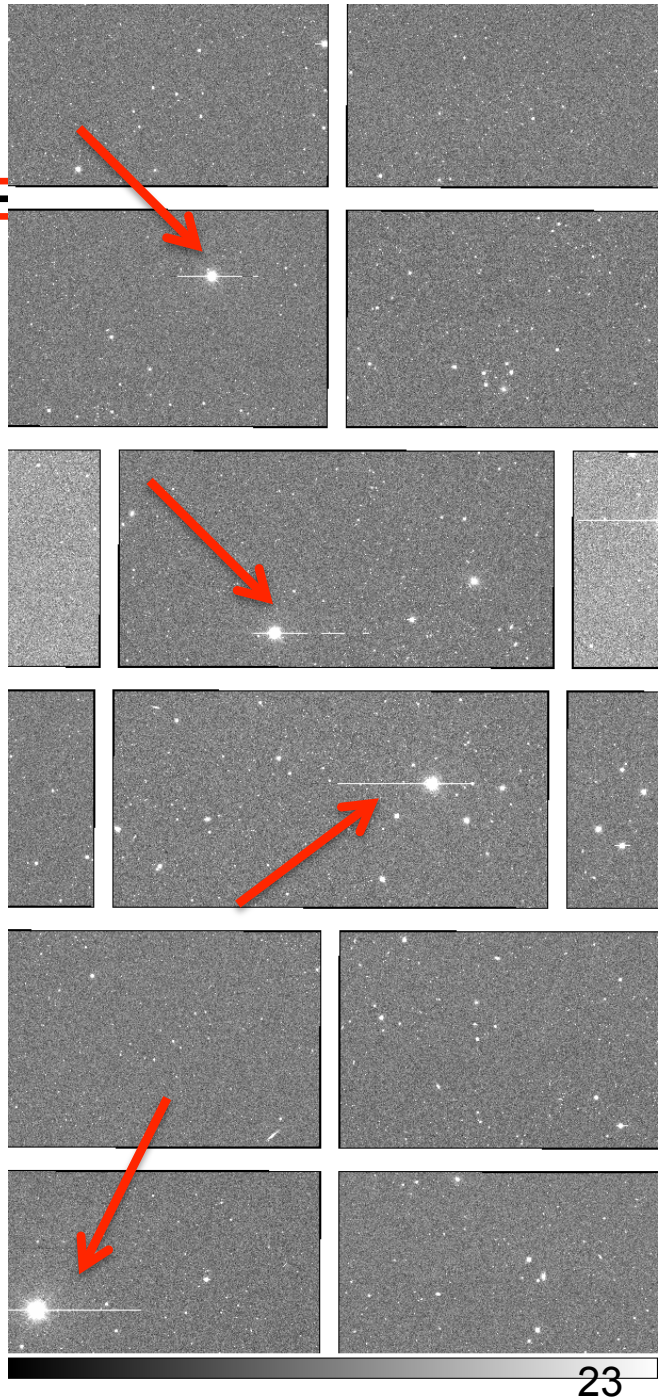




(From G. Bernstein)

DARK ENERGY SURVEY

Bleed



Bleed trail identification by searching for extended structures stemming form saturated island.

Y1 additions:

- Mask dilation in the cross-trail direction to better remove strong bleeds.
- Search for edge-bleed conditions for trails that intercept the read registers.

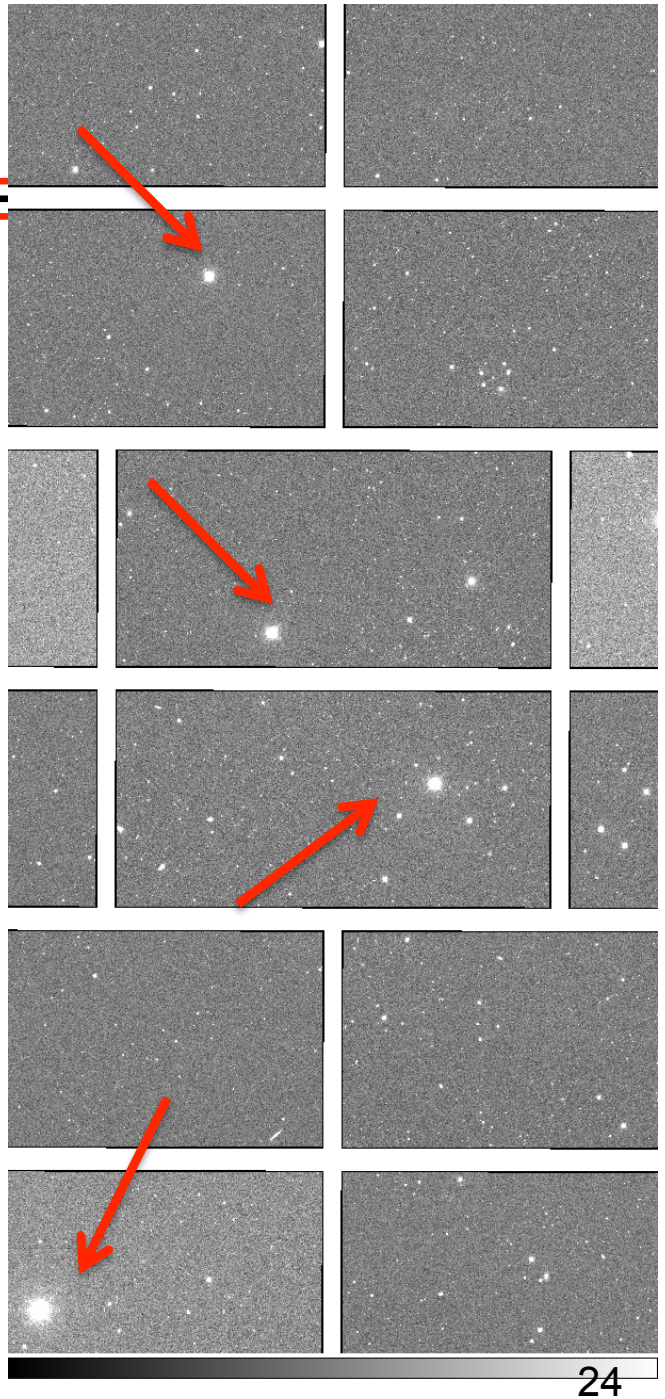
DARK ENERGY SURVEY

Bleed

Bleed trail identification by searching for extended structures stemming form saturated island.

Y1 additions:

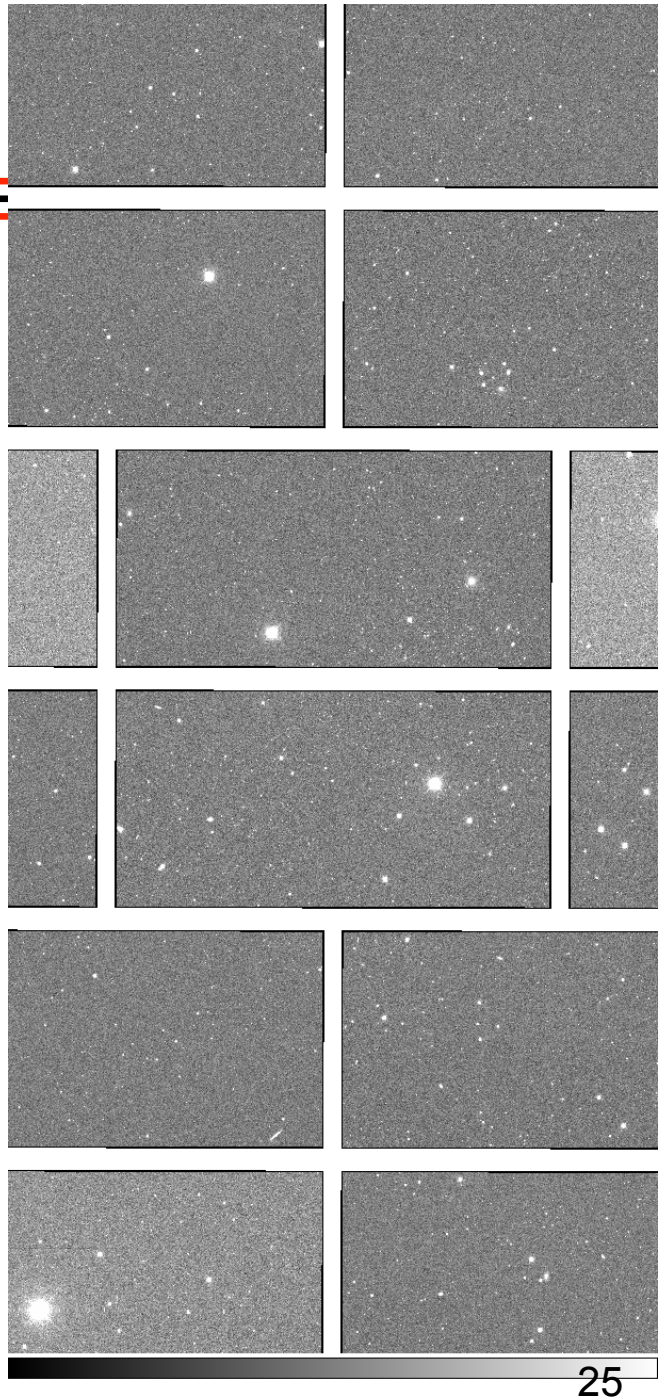
- Mask dilation in the cross-trail direction to better remove strong bleeds.
- Search for edge-bleed conditions for trails that intercept the read registers.



CR-reject & streak finder

DARK ENERGY SURVEY

- Early CR-rejection used neural net identification (only partially effective).
- SV: Single-Image CR-rejection was via gradient (better)
- **Y1: Implemented LSST-stack CR-rejection algorithm within DESDM pipelines.**
- Streak finder deployed in Y1 uses identification via Hough transform
- Y4: adding a truth-table/testbed to investigate further improvements for Y5



DARK ENERGY SURVEY

Exposure Based Assessment

Evaluate each reduced exposure based on SE products. Primary goal is to determine whether each observation meets basic survey requirements/ standards.

Primary decision based on the effective exposure time:

$$T_{eff} = (0.9\ k / FWHM)^2\ (\ Bkgd_{dark} / Bkgd)\ (10^{-2\ cloud\ /\ 2.5})$$

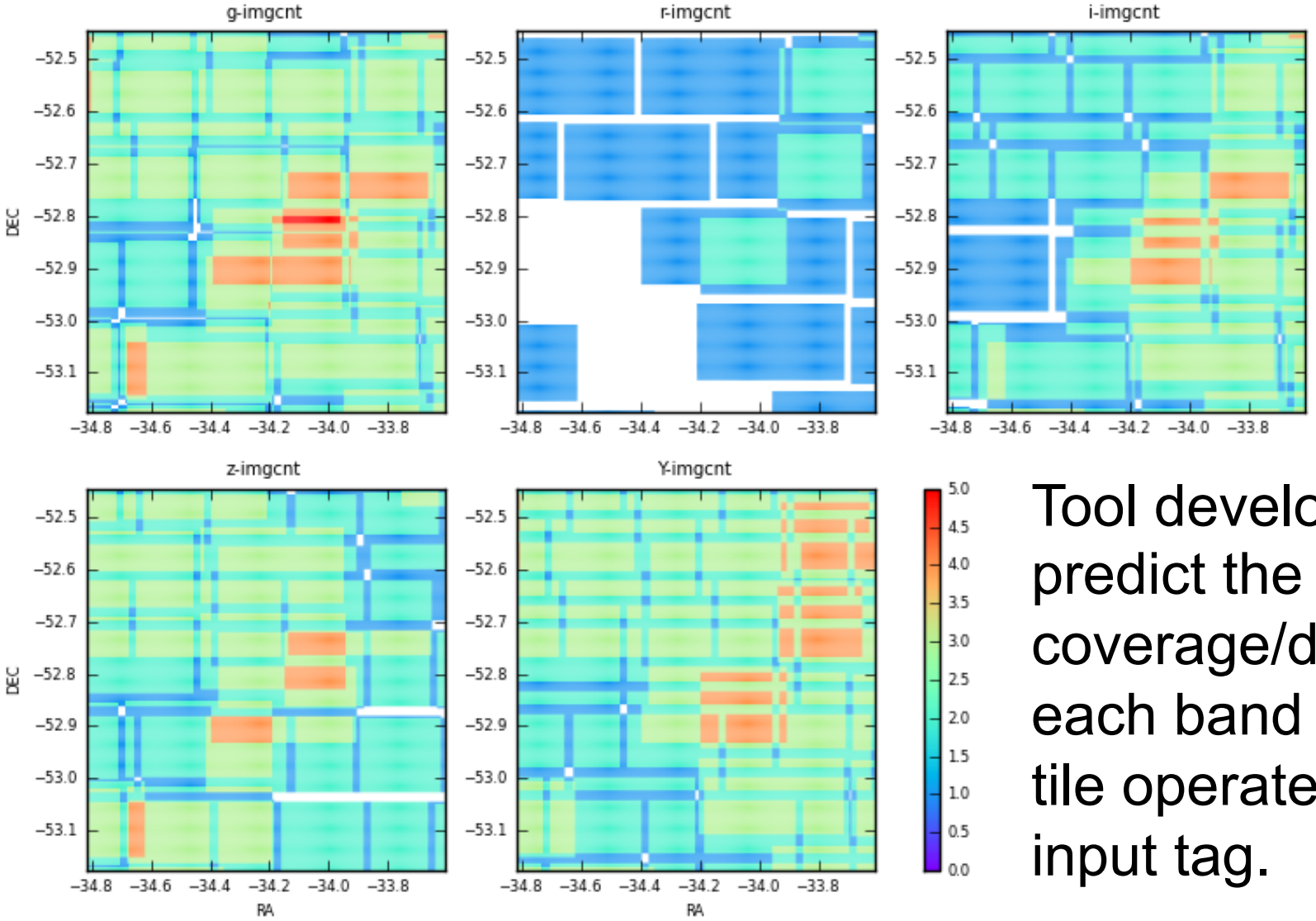
$$= F_{eff}\ B_{eff}\ C_{eff}$$

Current cutoffs are $T_{eff} > 0.3$ (riz-bands), > 0.2 (gY-band)

Further cuts can be placed based on individual components or other QA (e.g. astrometry, PSF) to form input TAG for COADD or other analysis.

Coverage Check Prior to COADD

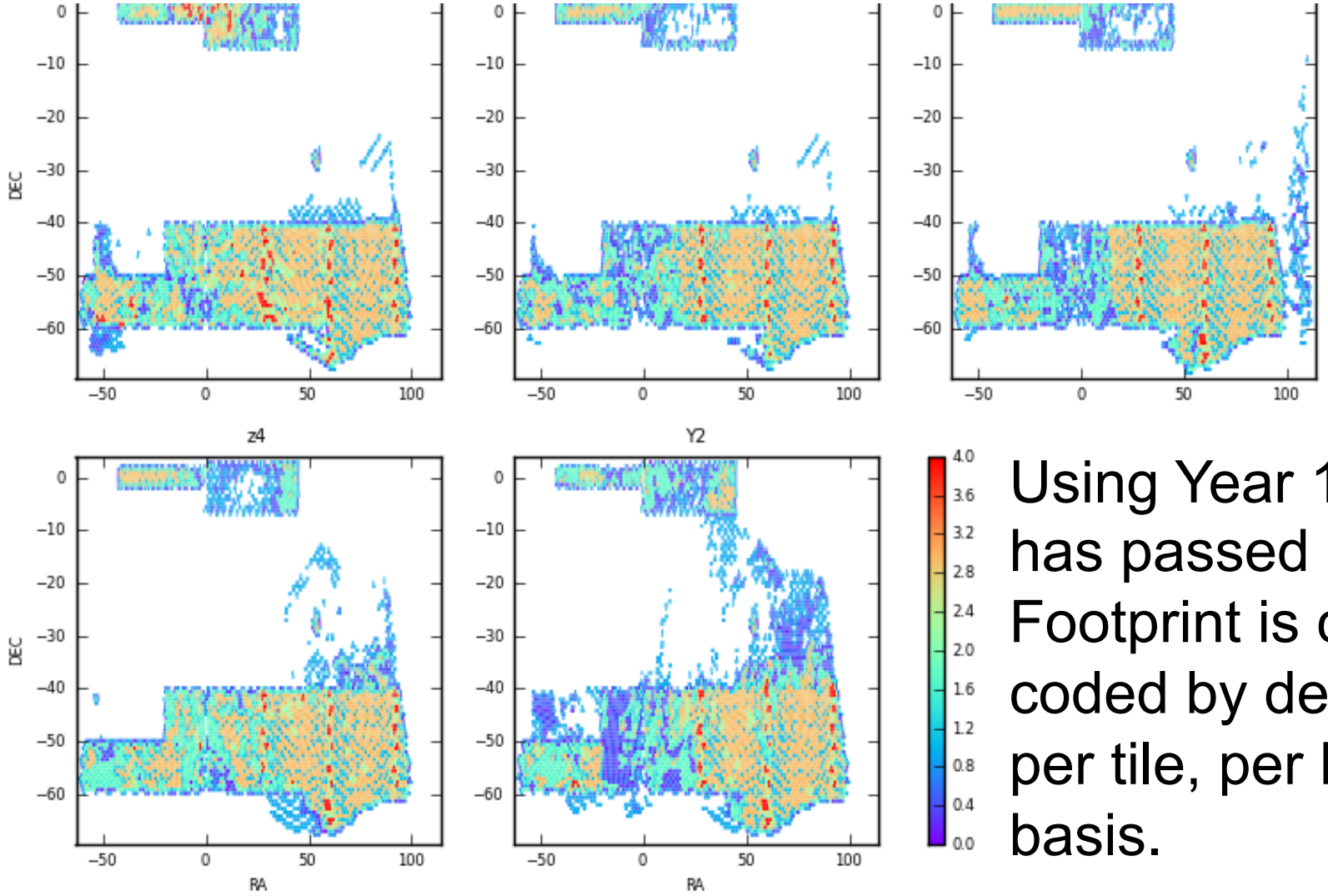
DARK ENERGY SURVEY



Tool developed to predict the detailed coverage/depth at each band for each tile operates on an input tag.

Projected Y1A1 Footprint

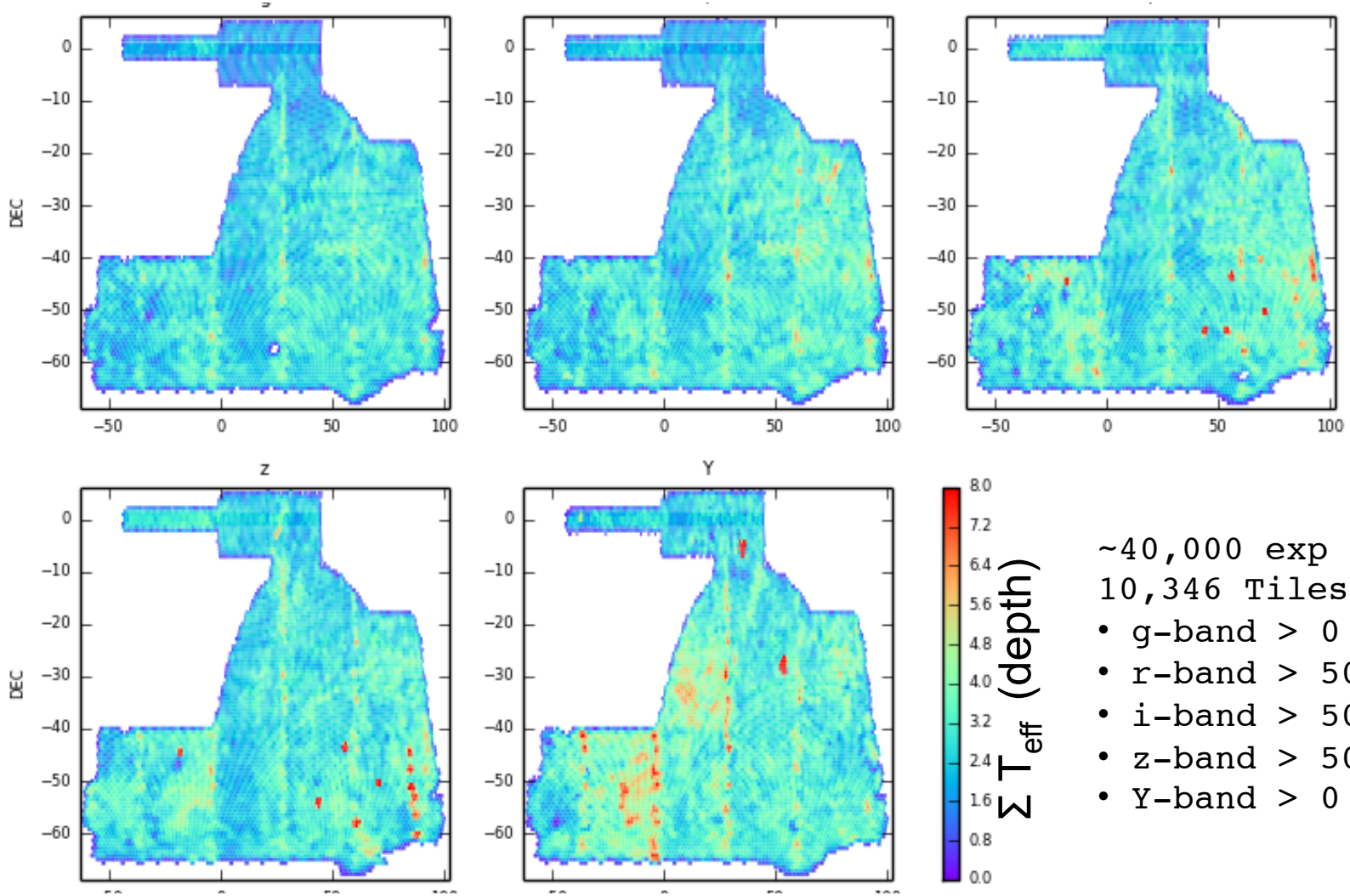
DARK ENERGY SURVEY



Using Year 1 data that has passed FirstCut, Footprint is color coded by depth on a per tile, per band basis.

Y3A1 COADD Footprint (Y1+Y2+Y3)

DARK ENERGY SURVEY

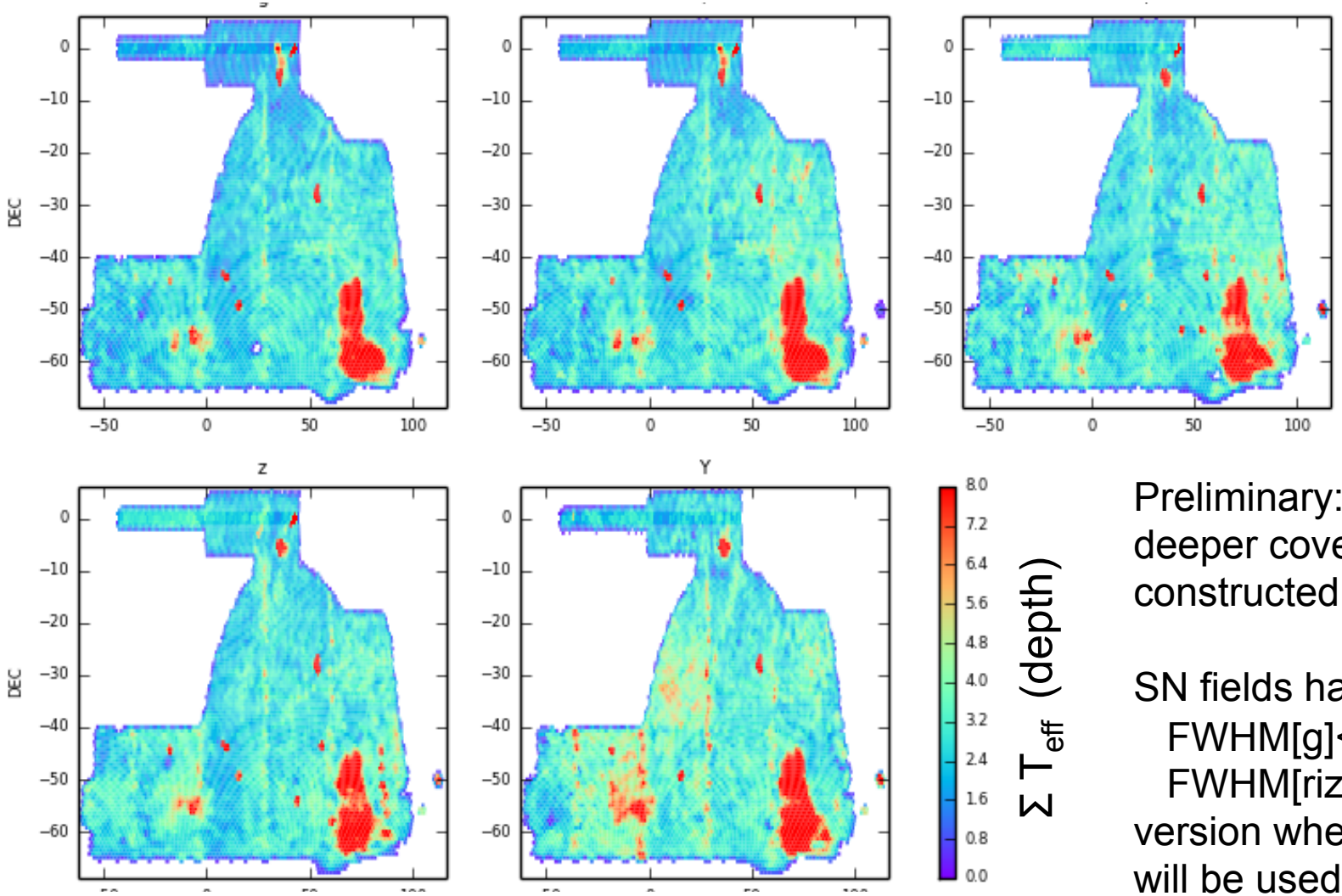


~40,000 exp contribute to 10,346 Tiles with:

- g-band > 0
- r-band > 50% (@ depth 2)
- i-band > 50% (@ depth 2)
- z-band > 50% (@ depth 2)
- Y-band > 0

Y3A1 COADD Footprint (DEEP)

DARK ENERGY SURVEY



Preliminary: O(1000) tiles w/ deeper coverage to be constructed in a 2nd campaign.

SN fields have special cuts:
FWHM[g]<1.1”
FWHM[riz]<1.0”
version where all DES data will be used in their generation.

COADD

DARK ENERGY SURVEY

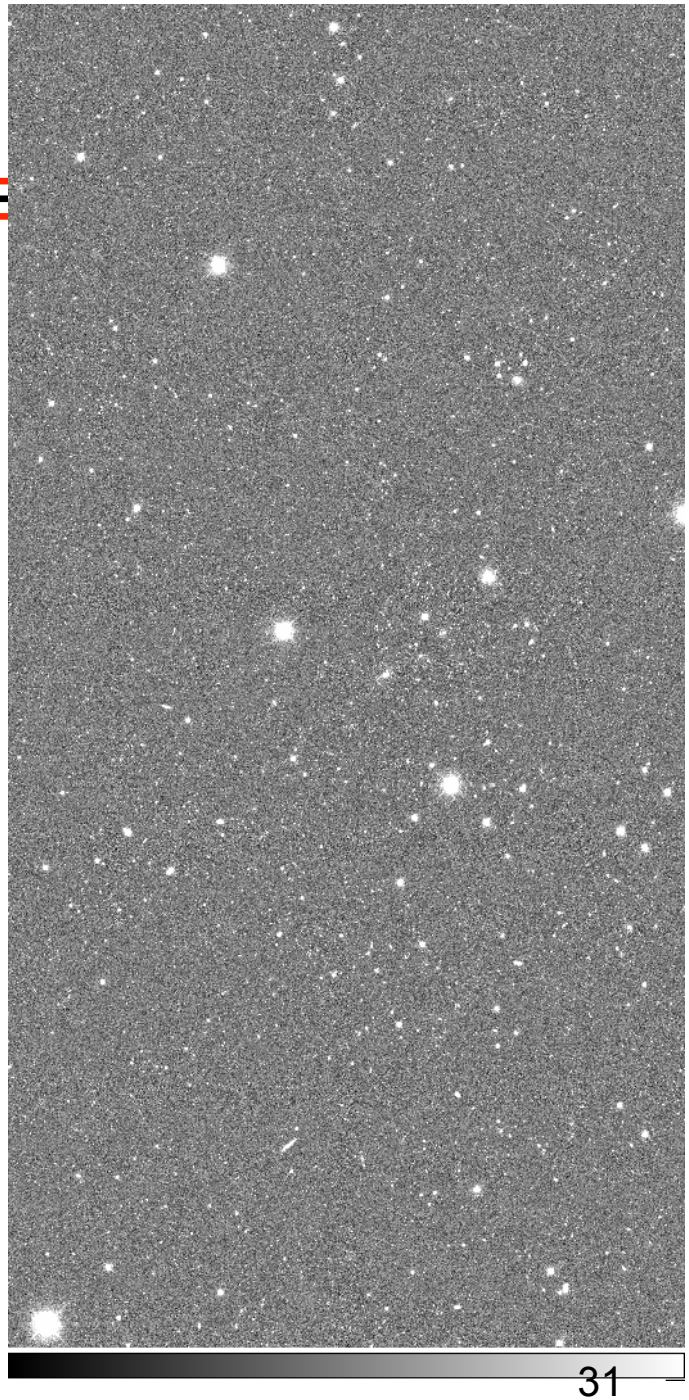
COADDITION of single-epoch images requires a global calibration based on single epoch photometry.

- Y1: GCM → 2-3 percent
- Y3: FGCM → sub-percent accuracy

Y1 added astrometric refinement reduce the relative astrometric residuals to ~50 mas.

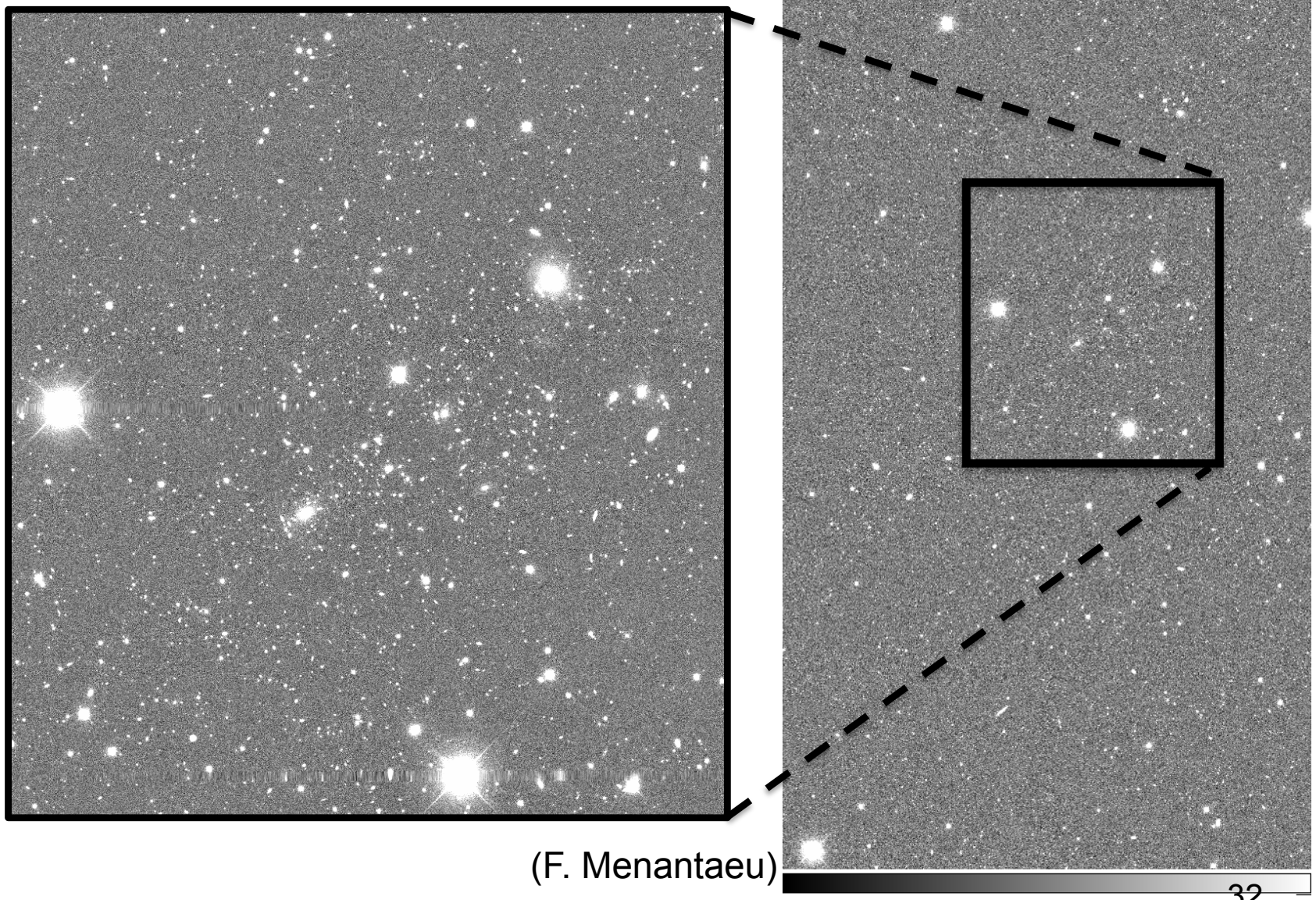
Y3 added:

- pipeline generation of MEDs products (ties COADD detection to single-epoch pixels)
- Afterburner pipelines use MEDs for:
  - MOF: multi-epoch, -object,-band photometry
  - Ngmix: shear estimates for WL



DARK ENERGY SURVEY

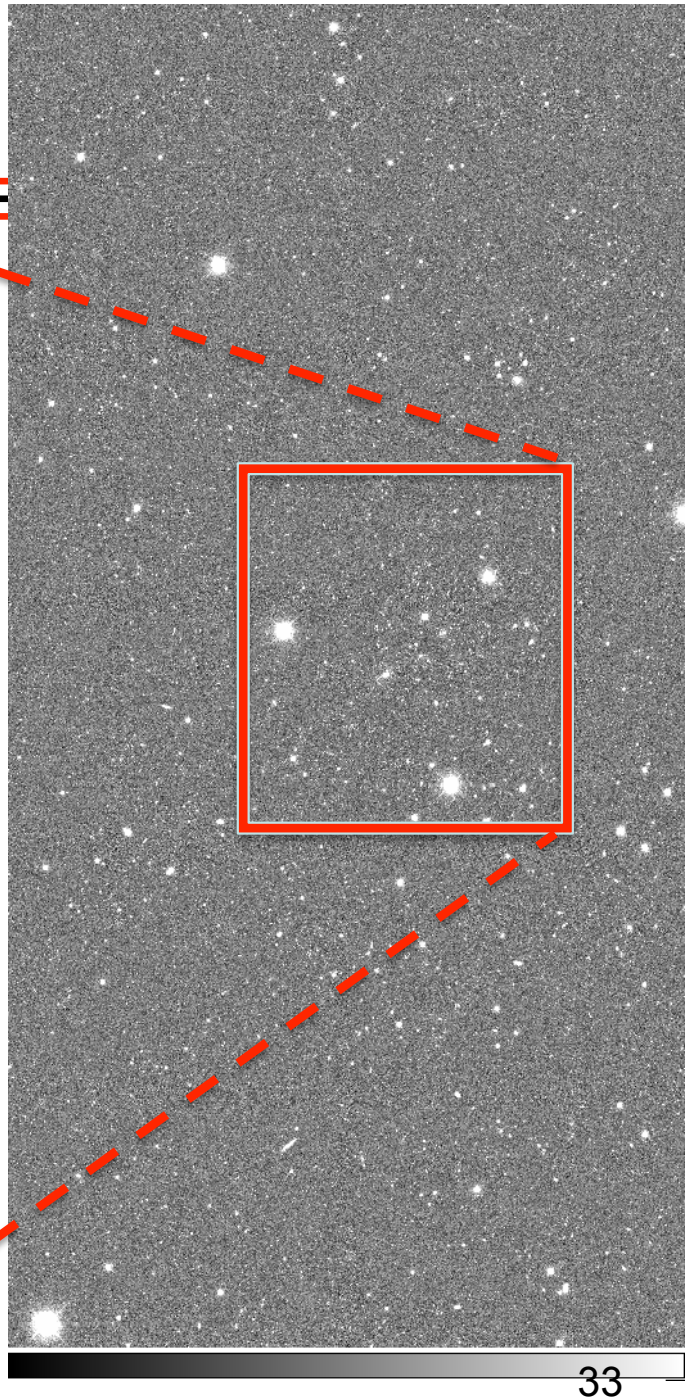
COADD



(F. Menantaeu)

DARK ENERGY SURVEY

COADD



(F. Menantaeu)

DARK ENERGY SURVEY

Some Lessons Learned

DES has been extremely successful but there was a vast amount of improvement that was needed.

1. Do not assume that you can have the final word in calibration at the time an observation occurs
2. Simulations are good but real data are better
3. Build QA metrics that can identify observations taken in poor conditions so that pipelines have a means to exit gracefully
   - Record the occurrence so that you can easily re-identify problematic data
   - FAILURE should be an option
4. Build pipelines so that lessons learned in DRP can influence subsequent nightly processing
5. Do not completely eliminate the ability to use human judgment to determine what data will go through DRP
   - DES had to add both TAGs and BLACKLISTs to identify/remove problematic observations above and beyond data flagged by QA metrics

DARK ENERGY SURVEY

Some More Lessons Learned

6. Feedback from users may eventually help but it cannot be relied on in the short-term.
   - DES's next DRP campaign generally begins just as the collaboration is beginning to dig into the data.
7. Do not underestimate the value of carrying provenance with the production
   - Allows a means to trace the the downstream impact of a problematic (exposure/ cal/code)
8. The perfect can be the enemy of the good
   - Expect that automation and QC assessment will mature over time
   - Afterburners can be your friend
9. "Data management" should not be an oxymoron

DARK ENERGY
SURVEY

Unspoken Thoughts