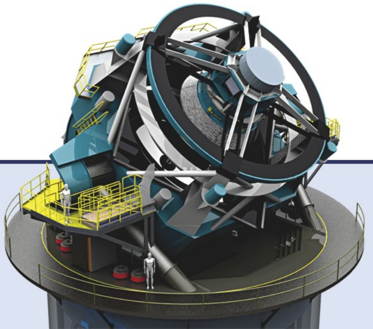
Data Management Status

Jim Bosch
Princeton University



What Data Management has been up to (that's relevant for DESC).

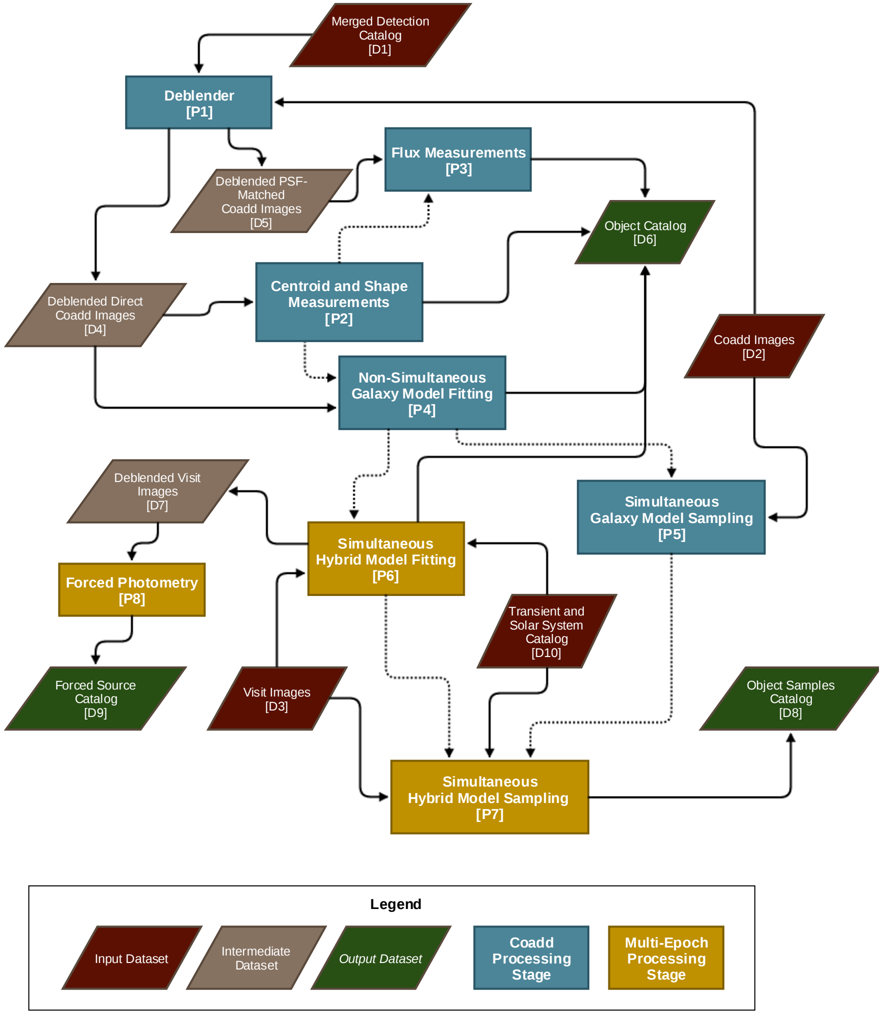
What Data Management and DESC can do for each other.

Planning

1. Update design documents
   - Data Products Definition Document (LSE-163): outputs
   - DM Applications Design (LDM-151): algorithms
2. Identify some intermediate goals (“Define Milestones”)
   - Some of these allow different teams within DM to know when they can expect a dependency to be ready.
   - Some of these allow external users (like DESC) to know when to expect a feature.
3. Figure out how long it will take (“Resource Load”)
   - Do we have enough people?

Flowcharts for the Algorithms

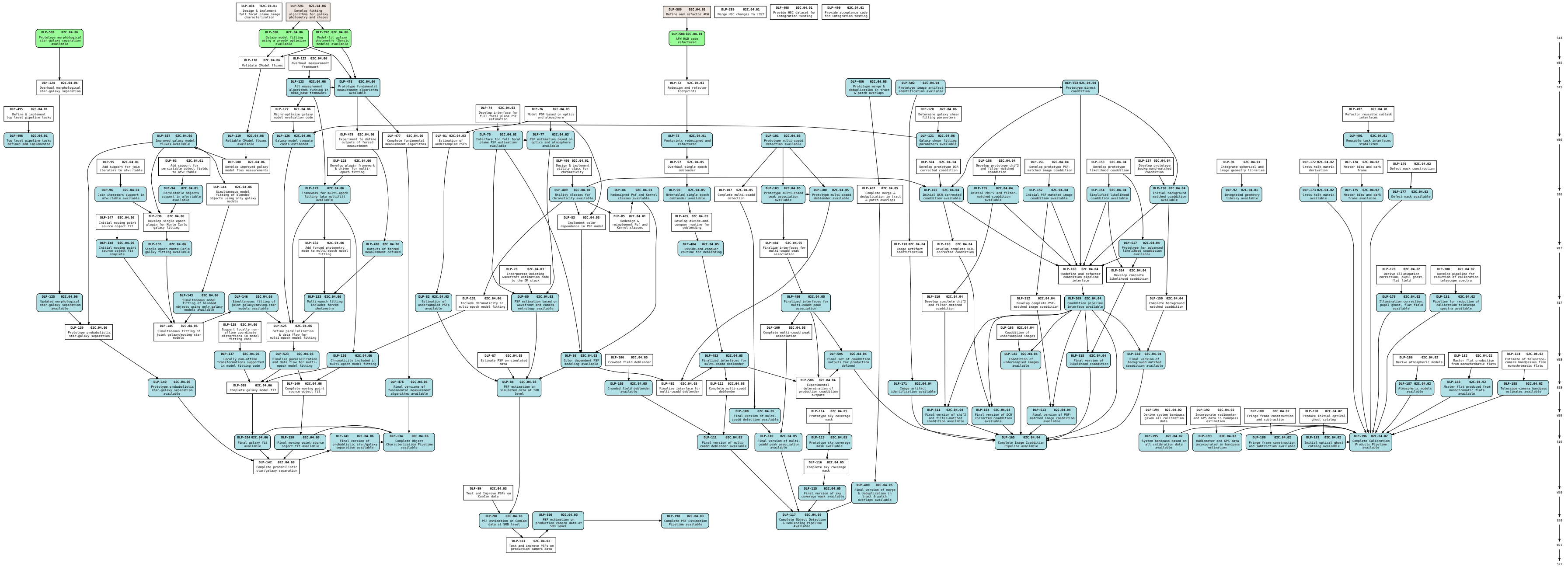
Large Synoptic Survey Telescope


Merged Detection Catalog [D1]
Deblender [P1]
Flux Measurements [P3]
Deblended PSF-Matched Coadd Images [D5]
Object Catalog [D6]
Deblended Direct Coadd Images [D4]
Centroid and Shape Measurements [P2]
Coadd Images [D2]
Non-Simultaneous Galaxy Model Fitting [P4]
Deblended Visit Images [D7]
Simultaneous Galaxy Model Sampling [P5]
Simultaneous Hybrid Model Fitting [P6]
Forced Photometry [P8]
Transient and Solar System Catalog [D10]
Forced Source Catalog [D9]
Visit Images [D3]
Object Samples Catalog [D8]
Simultaneous Hybrid Model Sampling [P7]
Legend
Input Dataset
Intermediate Dataset
Output Dataset
Coadd Processing Stage
Multi-Epoch Processing Stage

Level of detail in design docs: 1-2 paragraphs for each of these boxes.

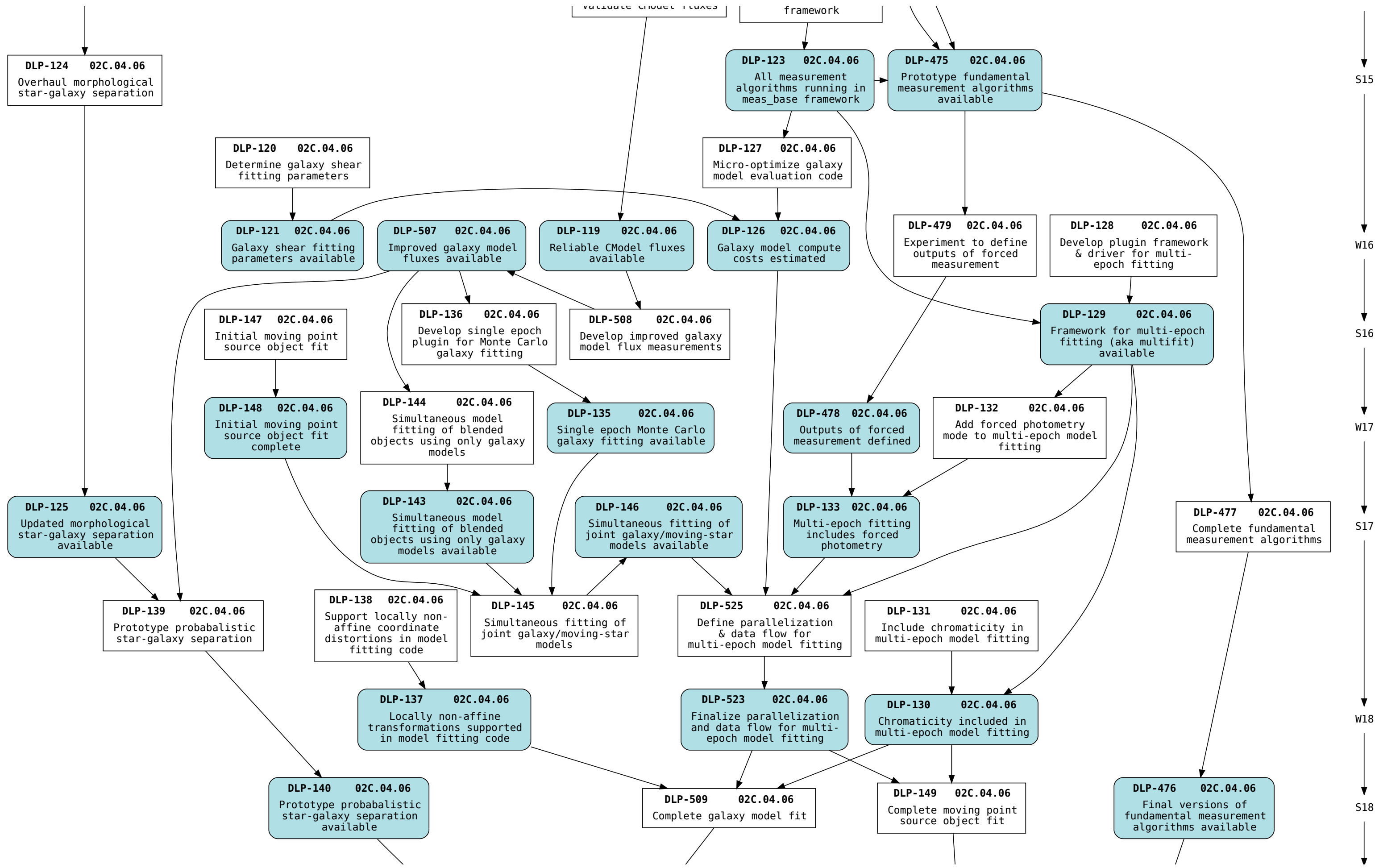
This is only about 1/6 of the data release production pipeline (not including middleware).

Flowcharts for Development


DLP-494 02C.04.01 Design & implement full focal plane image characterization
DLP-591 02C.04.06 Develop fitting algorithms for galaxy photometry and shapes
DLP-593 02C.04.06 Prototype morphological star-galaxy separation available
DLP-590 02C.04.06 Galaxy model fitting using a greedy optimizer available
DLP-592 02C.04.06 Model-fit galaxy photometry (Sersic models) available
DLP-118 02C.04.06 Validate CModel fluxes
DLP-122 02C.04.06 Overhaul measurement framework
DLP-124 02C.04.06 Overhaul morphological star-galaxy separation
DLP-123 02C.04.06 All measurement algorithms running in meas_base framework
DLP-475 02C.04.06 Prototype fundamental measurement algorithms available
DLP-495 02C.04.01 Define & implement top level pipeline tasks
DLP-127 02C.04.06 Micro-optimize galaxy model evaluation code
DLP-496 02C.04.01 Top level pipeline tasks defined and implemented
DLP-507 02C.04.06 Improved galaxy model fluxes available
DLP-119 02C.04.06 Reliable CModel fluxes available
DLP-126 02C.04.06 Galaxy model compute costs estimated
DLP-479 02C.04.06 Experiment to define outputs of forced measurement
DLP-477 02C.04.06 Complete fundamental measurement algorithms
DLP-95 02C.04.01 Add support for join iterators to afw::table
DLP-93 02C.04.01 Add support for persistable object fields to afw::table
DLP-508 02C.04.06 Develop improved galaxy model flux measurements
DLP-128 02C.04.06 Develop plugin framework & driver for multi-epoch fitting
DLP-96 02C.04.01 Join iterators support in afw::table available
DLP-94 02C.04.01 Persistable objects support in afw::table available
DLP-144 02C.04.06 Simultaneous model fitting of blended objects using only galaxy models
DLP-129 02C.04.06 Framework for multi-epoch fitting (aka multifit) available
DLP-147 02C.04.06 Initial moving point source object fit
DLP-136 02C.04.06 Develop single epoch plugin for Monte Carlo galaxy fitting
DLP-148 02C.04.06 Initial moving point source object fit complete
DLP-135 02C.04.06 Single epoch Monte Carlo galaxy fitting available
DLP-132 02C.04.06 Add forced photometry mode to multi-epoch model fitting
DLP-478 02C.04.06 Outputs of forced measurement defined
DLP-125 02C.04.06 Updated morphological star-galaxy separation available
DLP-143 02C.04.06 Simultaneous model fitting of blended objects using only galaxy models available
DLP-146 02C.04.06 Simultaneous fitting of joint galaxy/moving-star models available
DLP-133 02C.04.06 Multi-epoch fitting includes forced photometry
DLP-139 02C.04.06 Prototype probabalistic star-galaxy separation
DLP-145 02C.04.06 Simultaneous fitting of joint galaxy/moving-star models
DLP-138 02C.04.06 Support locally non-affine coordinate distortions in model fitting code
DLP-525 02C.04.06 Define parallelization & data flow for multi-epoch model fitting
DLP-137 02C.04.06 Locally non-affine transformations supported in model fitting code
DLP-523 02C.04.06 Finalize parallelization and data flow for multi-epoch model fitting
DLP-130 02C.04.06 Chromaticity included in multi-epoch model fitting
DLP-140 02C.04.06 Prototype probabalistic star-galaxy separation available
DLP-509 02C.04.06 Complete galaxy model fit
DLP-149 02C.04.06 Complete moving point source object fit
DLP-476 02C.04.06 Final versions of fundamental measurement algorithms available
DLP-524 02C.04.06 Final galaxy fit available
DLP-150 02C.04.06 Final moving point source object fit available
DLP-141 02C.04.06 Final version of probablistic star/galaxy separation available
DLP-134 02C.04.06 Complete Object Characterization Pipeline available
DLP-142 02C.04.06 Complete probabilistic star/galaxy separation
DLP-74 02C.04.03 Develop interface for full focal plane PSF estimation
DLP-76 02C.04.03 Model PSF based on optics and atmosphere
DLP-81 02C.04.03 Estimation of undersampled PSFs
DLP-75 02C.04.03 Interface for full focal plane PSF estimation available
DLP-77 02C.04.03 PSF estimation based on optics and atmosphere available
DLP-490 02C.04.01 Design & implement utility class for chromaticity
DLP-489 02C.04.01 Utility classes for chromaticity available
DLP-84 02C.04.01 Redesigned Psf and Kernel classes available
DLP-83 02C.04.03 Implement color dependence in PSF model
DLP-85 02C.04.01 Redesign & reimplement Psf and Kernel classes
DLP-78 02C.04.03 Incorporate existing wavefront estimation code to the DM stack
DLP-82 02C.04.03 Estimation of undersampled PSFs available
DLP-131 02C.04.06 Include chromaticity in multi-epoch model fitting
DLP-80 02C.04.03 PSF estimation based on wavefront and camera metrology available
DLP-87 02C.04.03 Estimate PSF on simulated data
DLP-86 02C.04.03 Color dependent PSF modeling available
DLP-88 02C.04.03 PSF estimation on simulated data at SRD level
DLP-89 02C.04.03 Test and Improve PSFs on ComCam data
DLP-90 02C.04.03 PSF estimation on ComCam data at SRD level
DLP-500 02C.04.03 PSF estimation on production camera data at SRD level
DLP-198 02C.04.03 Complete PSF Estimation Pipeline available
DLP-501 02C.04.03 Test and improve PSFs on production camera data
DLP-589 02C.04.01 Refine and refactor AFW
DLP-289 02C.04.01 Merge HSC changes to LSST
DLP-498 02C.04.01 Provide HSC dataset for integration testing
DLP-499 02C.04.01 Provide acceptance code for integration testing
DLP-588 02C.04.01 AFW R&D code refactored
DLP-72 02C.04.01 Redesign and refactor Footprints
DLP-73 02C.04.01 Footprints redesigned and refactored
DLP-97 02C.04.05 Overhaul single epoch deblender
DLP-98 02C.04.05 Overhauled single epoch deblender available
DLP-485 02C.04.05 Develop divide-and-conquer routine for deblending
DLP-484 02C.04.05 Divide-and-conquer routine for deblending
DLP-106 02C.04.05 Crowded field deblender
DLP-105 02C.04.05 Crowded field deblender available
DLP-482 02C.04.05 Finalize interface for multi-coadd deblender
DLP-483 02C.04.05 Finalized interfaces for multi-coadd deblender
DLP-112 02C.04.05 Complete multi-coadd deblender
DLP-111 02C.04.05 Final version of multi-coadd deblender available
DLP-486 02C.04.05 Prototype merge & deduplication in tract & patch overlaps
DLP-502 02C.04.04 Prototype image artifact identification available
DLP-120 02C.04.06 Determine galaxy shear fitting parameters
DLP-101 02C.04.05 Prototype multi-coadd detection available
DLP-121 02C.04.06 Galaxy shear fitting parameters available
DLP-107 02C.04.05 Complete multi-coadd detection
DLP-103 02C.04.05 Prototype multi-coadd peak association available
DLP-100 02C.04.05 Prototype multi-coadd deblender available
DLP-487 02C.04.05 Complete merge & deduplication in tract & patch overlaps
DLP-481 02C.04.05 Finalize interfaces for multi-coadd peak association
DLP-480 02C.04.05 Finalized interfaces for multi-coadd peak association
DLP-109 02C.04.05 Complete multi-coadd peak association
DLP-505 02C.04.04 Final set of coaddition outputs for production defined
DLP-506 02C.04.04 Experimental determination of production coaddition outputs
DLP-108 02C.04.05 Final version of multi-coadd detection available
DLP-110 02C.04.05 Final version of multi-coadd peak association available
DLP-114 02C.04.05 Prototype sky coverage mask
DLP-113 02C.04.05 Prototype sky coverage mask available
DLP-116 02C.04.05 Complete sky coverage mask
DLP-115 02C.04.05 Final version of sky coverage mask available
DLP-488 02C.04.05 Final version of merge & deduplication in tract & patch overlaps available
DLP-117 02C.04.05 Complete Object Detection & Deblending Pipeline Available
DLP-503 02C.04.04 Prototype direct coaddition
DLP-504 02C.04.04 Develop prototype DCR corrected coaddition
DLP-156 02C.04.04 Develop prototype chi^2 and filter-matched coaddition
DLP-151 02C.04.04 Develop prototype PSF-matched image coaddition
DLP-153 02C.04.04 Develop prototype likelihood coaddition
DLP-157 02C.04.04 Develop prototype background-matched coaddition
DLP-162 02C.04.04 Initial DCR-corrected coaddition available
DLP-155 02C.04.04 Initial chi^2 and filter-matched coaddition available
DLP-152 02C.04.04 Initial PSF-matched image coaddition available
DLP-154 02C.04.04 Simplified likelihood coaddition available
DLP-158 02C.04.04 Initial background matched coaddition available
DLP-170 02C.04.04 Image artifact identification
DLP-163 02C.04.04 Develop complete DCR-corrected coaddition
DLP-517 02C.04.04 Prototype for advanced likelihood coaddition available
DLP-168 02C.04.04 Redefine and refactor coaddition pipeline interface
DLP-514 02C.04.04 Develop complete likelihood coaddition
DLP-510 02C.04.04 Develop complete chi^2 and filter-matched coaddition
DLP-512 02C.04.04 Develop complete PSF-matched image coaddition
DLP-169 02C.04.04 Coaddition pipeline interface available
DLP-159 02C.04.04 Complete background matched coaddition
DLP-166 02C.04.04 Coaddition of undersampled images
DLP-167 02C.04.04 Coaddition of undersampled images available
DLP-515 02C.04.04 Final version of likelihood coaddition
DLP-160 02C.04.04 Final version of background matched coaddition available
DLP-171 02C.04.04 Image artifact identification available
DLP-511 02C.04.04 Final version of chi^2 and filter-matched coaddition available
DLP-164 02C.04.04 Final version of DCR corrected coaddition available
DLP-513 02C.04.04 Final version of PSF-matched image coaddition available
DLP-165 02C.04.04 Complete Image Coaddition Pipeline available
DLP-91 02C.04.01 Integrate spherical and image geometry libraries
DLP-92 02C.04.01 Integrated geometry library available
DLP-492 02C.04.01 Refactor reusable subtask interfaces
DLP-491 02C.04.01 Reusable task interfaces stabilized
DLP-172 02C.04.02 Cross-talk matrix derivation
DLP-174 02C.04.02 Master bias and dark frame
DLP-176 02C.04.02 Defect mask construction
DLP-173 02C.04.02 Cross-talk matrix available
DLP-175 02C.04.02 Master bias and dark frame available
DLP-177 02C.04.02 Defect mask available
DLP-178 02C.04.02 Derive illumination correction, pupil ghost, flat field
DLP-180 02C.04.02 Develop pipeline for reduction of calibration telescope spectra
DLP-179 02C.04.02 Illumination correction, pupil ghost, flat field available
DLP-181 02C.04.02 Pipeline for reduction of calibration telescope spectra available
DLP-186 02C.04.02 Derive atmospheric models
DLP-182 02C.04.02 Master flat production from monochromatic flats
DLP-184 02C.04.02 Estimate of telescope-camera bandpasses from monochromatic flats
DLP-187 02C.04.02 Atmospheric models available
DLP-183 02C.04.02 Master flat produced from monochromatic flats available
DLP-185 02C.04.02 Telescope-camera bandpass estimates available
DLP-194 02C.04.02 Derive system bandpass given all calibration data
DLP-192 02C.04.02 Incorporate radiometer and GPS data in bandpass estimation
DLP-188 02C.04.02 Fringe frame construction and subtraction
DLP-190 02C.04.02 Produce initial optical ghost catalog
DLP-195 02C.04.02 System bandpass based on all calibration data available
DLP-193 02C.04.02 Radiometer and GPS data incorporated in bandpass estimation
DLP-189 02C.04.02 Fringe frame construction and subtraction available
DLP-191 02C.04.02 Initial optical ghost catalog available
DLP-196 02C.04.02 Complete Calibration Products Pipeline available
S14
W15
S15
W16
S16
W17
S17
W18
S18
W19
S19
W20
S20
W21
S21

Flowcharts for Development

Large Synoptic Survey Telescope



validate CModel fluxes

framework

DLP-124 02C.04.06
Overhaul morphological star-galaxy separation

DLP-123 02C.04.06
All measurement algorithms running in meas_base framework

DLP-475 02C.04.06
Prototype fundamental measurement algorithms available

S15

DLP-120 02C.04.06
Determine galaxy shear fitting parameters

DLP-127 02C.04.06
Micro-optimize galaxy model evaluation code

DLP-121 02C.04.06
Galaxy shear fitting parameters available

DLP-507 02C.04.06
Improved galaxy model fluxes available

DLP-119 02C.04.06
Reliable CModel fluxes available

DLP-126 02C.04.06
Galaxy model compute costs estimated

DLP-479 02C.04.06
Experiment to define outputs of forced measurement

DLP-128 02C.04.06
Develop plugin framework & driver for multi-epoch fitting

W16

DLP-147 02C.04.06
Initial moving point source object fit

DLP-136 02C.04.06
Develop single epoch plugin for Monte Carlo galaxy fitting

DLP-508 02C.04.06
Develop improved galaxy model flux measurements

DLP-129 02C.04.06
Framework for multi-epoch fitting (aka multifit) available

S16

DLP-148 02C.04.06
Initial moving point source object fit complete

DLP-144 02C.04.06
Simultaneous model fitting of blended objects using only galaxy models

DLP-135 02C.04.06
Single epoch Monte Carlo galaxy fitting available

DLP-478 02C.04.06
Outputs of forced measurement defined

DLP-132 02C.04.06
Add forced photometry mode to multi-epoch model fitting

W17

DLP-125 02C.04.06
Updated morphological star-galaxy separation available

DLP-143 02C.04.06
Simultaneous model fitting of blended objects using only galaxy models available

DLP-146 02C.04.06
Simultaneous fitting of joint galaxy/moving-star models available

DLP-133 02C.04.06
Multi-epoch fitting includes forced photometry

DLP-477 02C.04.06
Complete fundamental measurement algorithms

S17

DLP-139 02C.04.06
Prototype probabalistic star-galaxy separation

DLP-138 02C.04.06
Support locally non-affine coordinate distortions in model fitting code

DLP-145 02C.04.06
Simultaneous fitting of joint galaxy/moving-star models

DLP-525 02C.04.06
Define parallelization & data flow for multi-epoch model fitting

DLP-131 02C.04.06
Include chromaticity in multi-epoch model fitting

DLP-137 02C.04.06
Locally non-affine transformations supported in model fitting code

DLP-523 02C.04.06
Finalize parallelization and data flow for multi-epoch model fitting

DLP-130 02C.04.06
Chromaticity included in multi-epoch model fitting

W18

DLP-140 02C.04.06
Prototype probabalistic star-galaxy separation available

DLP-509 02C.04.06
Complete galaxy model fit

DLP-149 02C.04.06
Complete moving point source object fit

DLP-476 02C.04.06
Final versions of fundamental measurement algorithms available

S18

Planning Timeline

A new version of the full plan, with updated design documents and resource-loaded milestones, ready by mid-November

is not going to happen.

We may have a rough plan with a good outline in the design documents by then.

What DM is Up To: Hiring and Onboarding

- We're perhaps a bit more than half of the way to fully staffed.
- We're still missing a few lead positions.
- Still training lots of new developers.

DM Bootcamp a few weeks ago. Materials may be useful
https://community.lsst.org/t/dm-boot-camp-announcement/249

Some new names/roles DESC members should know:

- K-T Lim is now Project Engineer
- Tim Jenness is Deputy Architect
- David Nidever is Science/Data Quality Scientist
- Jonathan Sick is documentation tsar

We're a three-headed dragon now!

Hate our APIs? Tim wants to fix that.

Developing plan for verification datasets and production runs.

Docs are finally *someone's* top priority.

State of the DM Stack: Current Work

About 2 years ago, the HSC pipeline team forked the LSST stack. Since then:

- The HSC stack has become a state-of-the-art end-to-end pipeline with high-quality (for Stage II) algorithms for HSC data reduction – at the expense of a lot of confusing, poorly-documented code.
- The LSST stack has been refactored and improved considerably in terms of code quality, our testing and workflow have been greatly enhanced, and we've made it work on more cameras. But algorithms are largely unchanged, or in some cases have bitrotted.

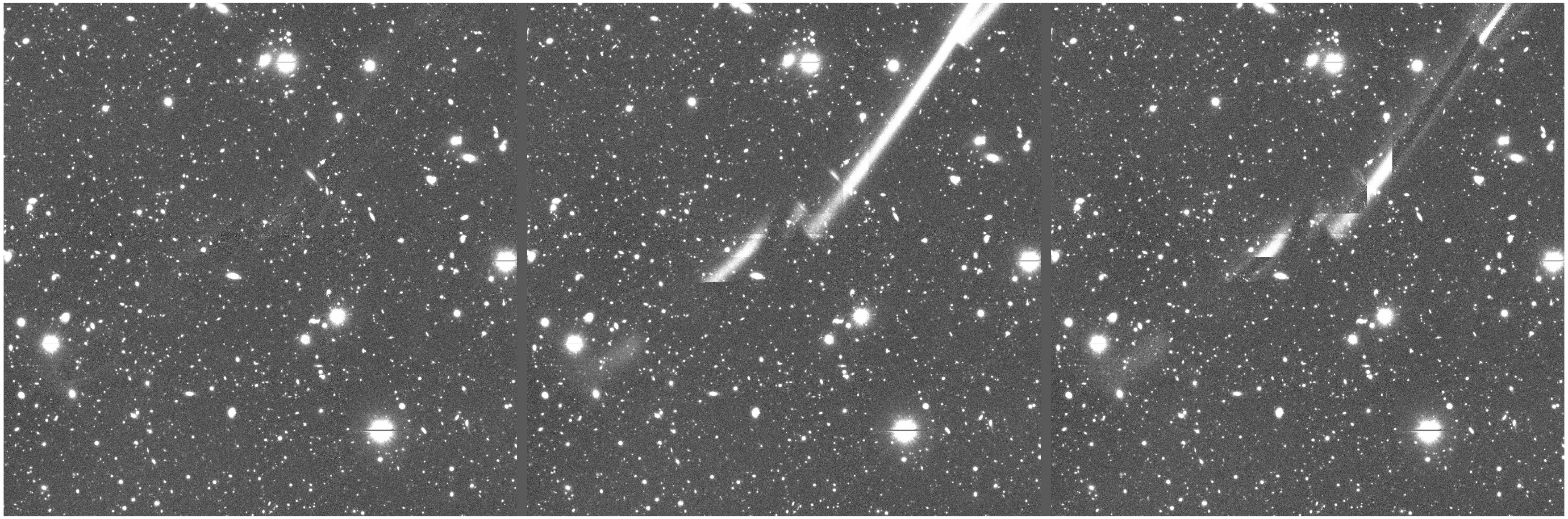
Now we're putting them back together.

And the code release from September is about halfway there.

State of the DM Stack: HSC Fork

- We're making the best coadds in the world.
- We detect and deblend consistently in all bands.
- PSF modeling uses a modified *PSFEx*; already meeting LSST minimum design specs for residual ellipticity correlations.
- We have a basic correction for brighter-fatter (~90% corrected).
- Moderately robust galaxy model fluxes and colors.
- *No multi-epoch fitting.*
- *No PSF-matched coadds.*
- *No modern shear estimation (just Regaussianization).*

“Safe Clipping” in Coadds



safe clipping

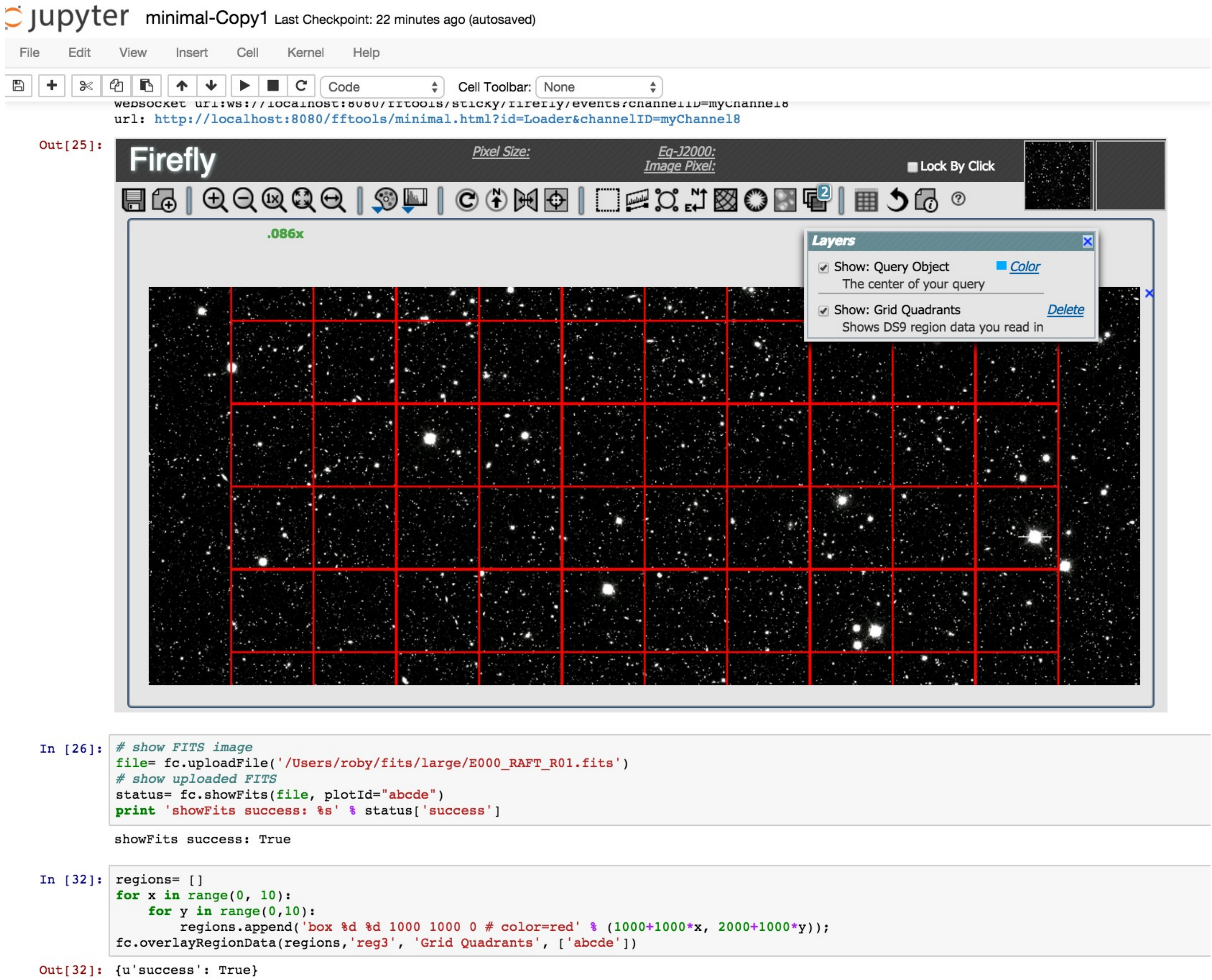
direct mean

3σ clipped mean

- build mean and clipped mean coadds
- difference them, find above-threshold regions
- reject whole regions from individual images

courtesy of Bob Armstrong

State of the DM Stack: SUI



minimal-Copy1 Last Checkpoint: 22 minutes ago (autosaved)

```
websocket url:ws://localhost:8080/fftools/sticky/firefly/events?channelID=myChannel8
url: http://localhost:8080/fftools/minimal.html?id=Loader&channelID=myChannel8
```

Out[25]:

☑ Show: Query Object 🟦 Color
The center of your query

☑ Show: Grid Quadrants Delete
Shows DS9 region data you read in

In [26]:

```
# show FITS image
file= fc.uploadFile('/Users/roby/fits/large/E000_RAFT_R01.fits')
# show uploaded FITS
status= fc.showFits(file, plotId="abcde")
print 'showFits success: %s' % status['success']
```

```
showFits success: True
```

In [32]:

```
regions= []
for x in range(0, 10):
    for y in range(0,10):
        regions.append('box %d %d 1000 1000 0 # color=red' % (1000+1000*x, 2000+1000*y));
fc.overlayRegionData(regions,'reg3', 'Grid Quadrants', ['abcde'])
```

Out[32]:

```
{u'success': True}
```

User interface team at IPAC is experimenting with embedding their `Firefly` image display tool inside IPython notebooks.

State of the DM Stack

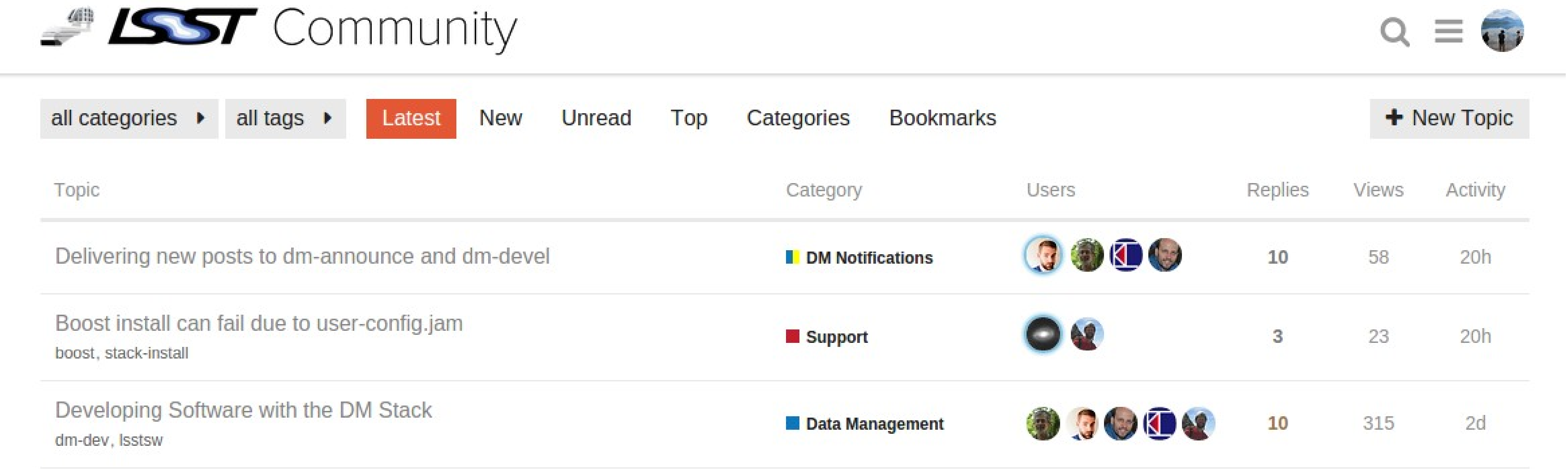
- Pierre Astier and Dominique Boutigny are developing a new relative astrometry and photometry module that's now being integrated into the pipeline; already working on CFHT and HSC data, with DECam coming soon.
- We're putting a lot of work into getting the stack running well on DECam data.
- We're working on the image subtraction code again, focusing on DCR.
- We've greatly improved our continuous integration and automated testing.

Main goals: regular, complete end-to-end tests on real data and improvements that make developers work more efficiently.

What Data Management has been up to (that's relevant for DESC).

What Data Management and DESC can do for each other.

Introducing **community.lsst.org**



Topic	Category	Users	Replies	Views	Activity
Delivering new posts to dm-announce and dm-devel	DM Notifications		10	58	20h
Boost install can fail due to user-config.jam boost, stack-install	Support		3	23	20h
Developing Software with the DM Stack dm-dev, lsstsw	Data Management		10	315	2d

- Preferred over HipChat for long-running, in-depth discussions.
- Most non-announce email lists are being phased out.
- Use/search the Q&A tag for StackOverflow-style FAQs.
- Intended for users, not just DM/Project people (self-signup).

Verification Datasets and Productions

We've started thinking about what real datasets we can use for development and verification before commissioning. No firm plans yet, but we're considering many options:

- DECam – Start with general observer datasets whose owners want to try our code to do their science, *eventually* reprocess significant fraction (all?) of DES.
- HSC – Amount of public data will increase dramatically in the next year.
- CFHTLS – best for early tests of shear estimation, due to publicly available shear catalogs to compare against.
- PanSTARRS – mostly database and user interface testing (probably no reprocessing).

Should you use the DM Stack?

The DM stack is essentially state-of-the-art from an algorithms standpoint – it's about as good as the PanSTARRS, SDSS, and DES pipelines in most respects. It's clearly not as good in some areas (but frequently better in others).

Documentation and ease-of-use are in bad shape compared to AstroPy. They're in pretty good shape compared to any of the above pipelines.

After a few months you'll mostly just be annoyed, and only occasionally frustrated.

Expect to get a lot out of it, but expect to be very frustrated at first.

Feedback and Contributing to the Stack

Please tell us what you don't like. Don't assume you just don't know enough to comment intelligently. This *absolutely* includes changes to bad interfaces.

If you can, create a GitHub PR to fix it (but ask about it before you do a lot of work).

The DM team has *spirited internal debates* about what we think our users want. Please, please just tell us.

The DM/DESC Boundary

If you don't know if DM is responsible for producing something, *please ask*. We may or may not know, but we also might not have even thought about it.

Please don't assume you shouldn't work on a problem just because DM is responsible for it. We want algorithmic input from outside, both in ideas and in code.

Please do talk to us about those algorithms early and often, so we don't duplicate effort or spend time on solved problems.

Please do consider writing code directly within the DM stack. You don't have to be a DM member to contribute to the codebase.

Design Document Development

What DM is planning in detail for algorithms and data products should become more much more clear in the next few months.

If you're particularly interested, you can also follow our design documents on GitHub and comment on the pull requests:

- `https://github.com/lsst/LSE-163` (data products)
- `https://github.com/lsst/LDM-151` (algorithms)

For now, take the content of these documents with a large bucket of salt.

Summary

Planning and design document work will give you all a lot to read (soon).

Verification dataset runs will enable a ton of new DESC investigations (soon).

We've turned the corner on hiring – contributions from new people now outweigh the hiring/onboarding effort.

DM stack is steadily improving: still a big learning curve, but it's really a complete pipeline.

Join community.lsst.org! Ask questions! Contribute!