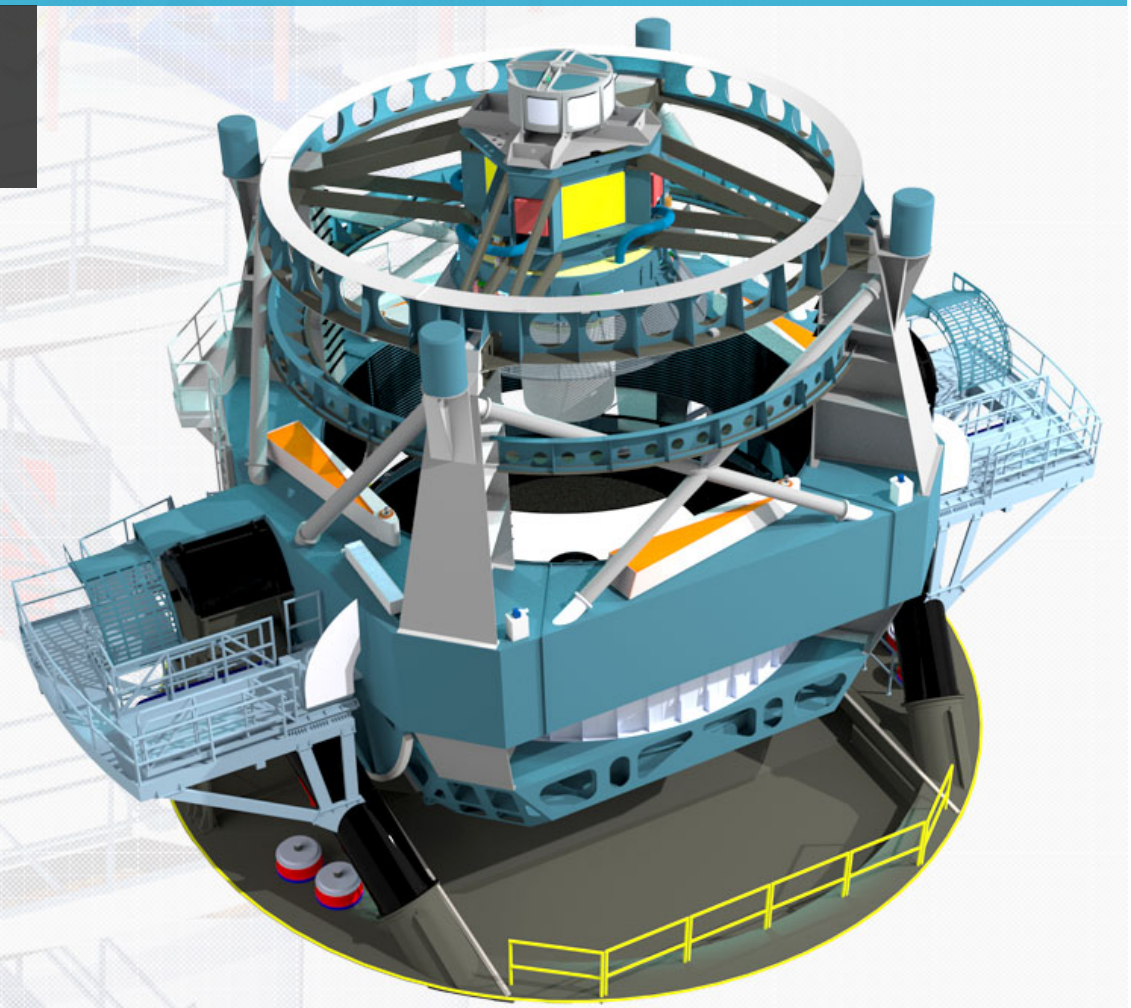
Searching for High Redshift Quasars in the LSST-Reprocessed Stripe 82 Imaging

Yusra AlSayyad
University of Washington

Ian McGreer,
Xiaohui Fan,

Andrew Connolly,
Željko Ivezić,
Mario Jurić,
Andrew Becker
Simon Krughoff
Russell Owen

Greg Daues
and the LSST Data
Management Team



Lessons Learned converting 20TB images into catalogs

Since this meeting is about **tools…**

LSST

Data Management Stack

XSEDE

Extreme Science and Engineering Discovery Environment

globus toolkit

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

MySQL™

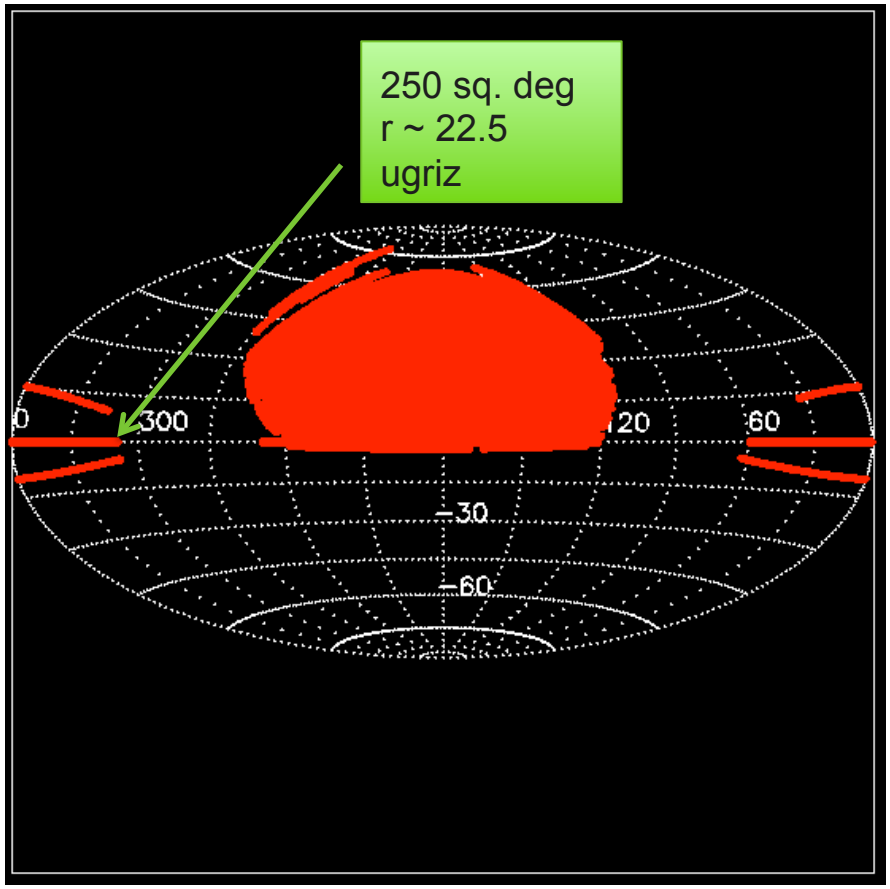
scikit learn

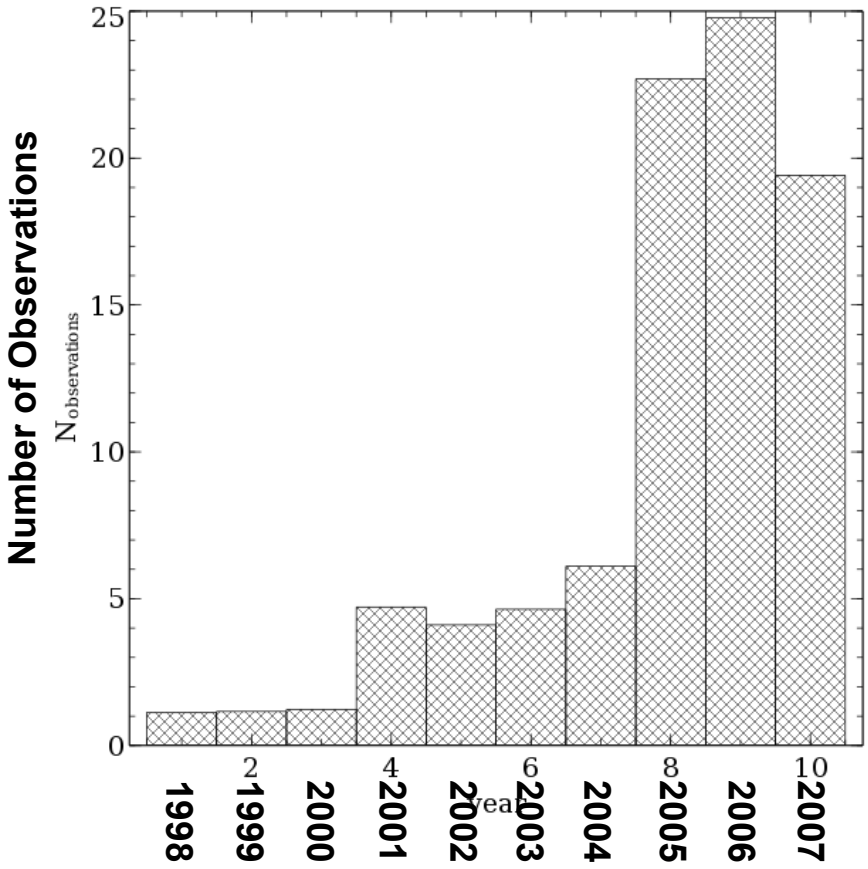
Spark

LSST Data Processing pixels → catalogs

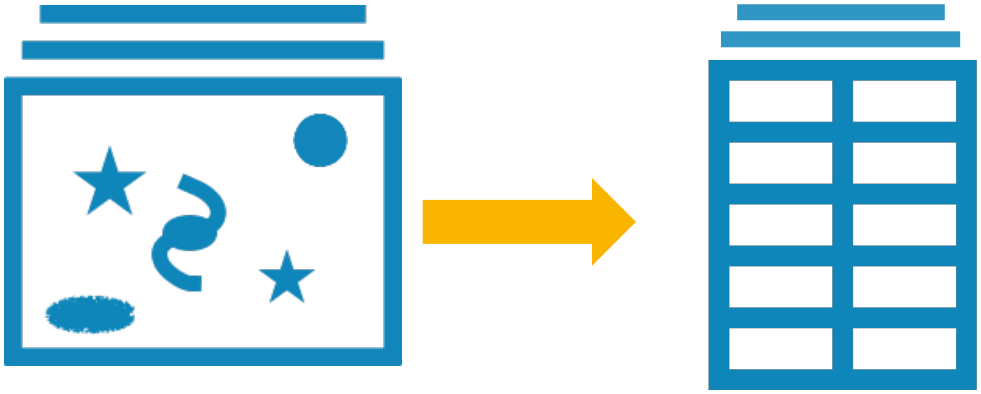
- LSST Data Management processes data two ways:
  1. **Real-time alert stream**: Support science cases requiring rapid identification and follow-up (transients, fast-moving NEOs)
  2. **Annual Data Release Productions**: Support deep static-sky science and statistical studies of variability.
- Run prototype code on simulated images and real data to test:
  - **Accuracy**
  - **Scalability**

SDSS Repeatedly scanned Stripe 82





20 TB Images to lightcurves and colors



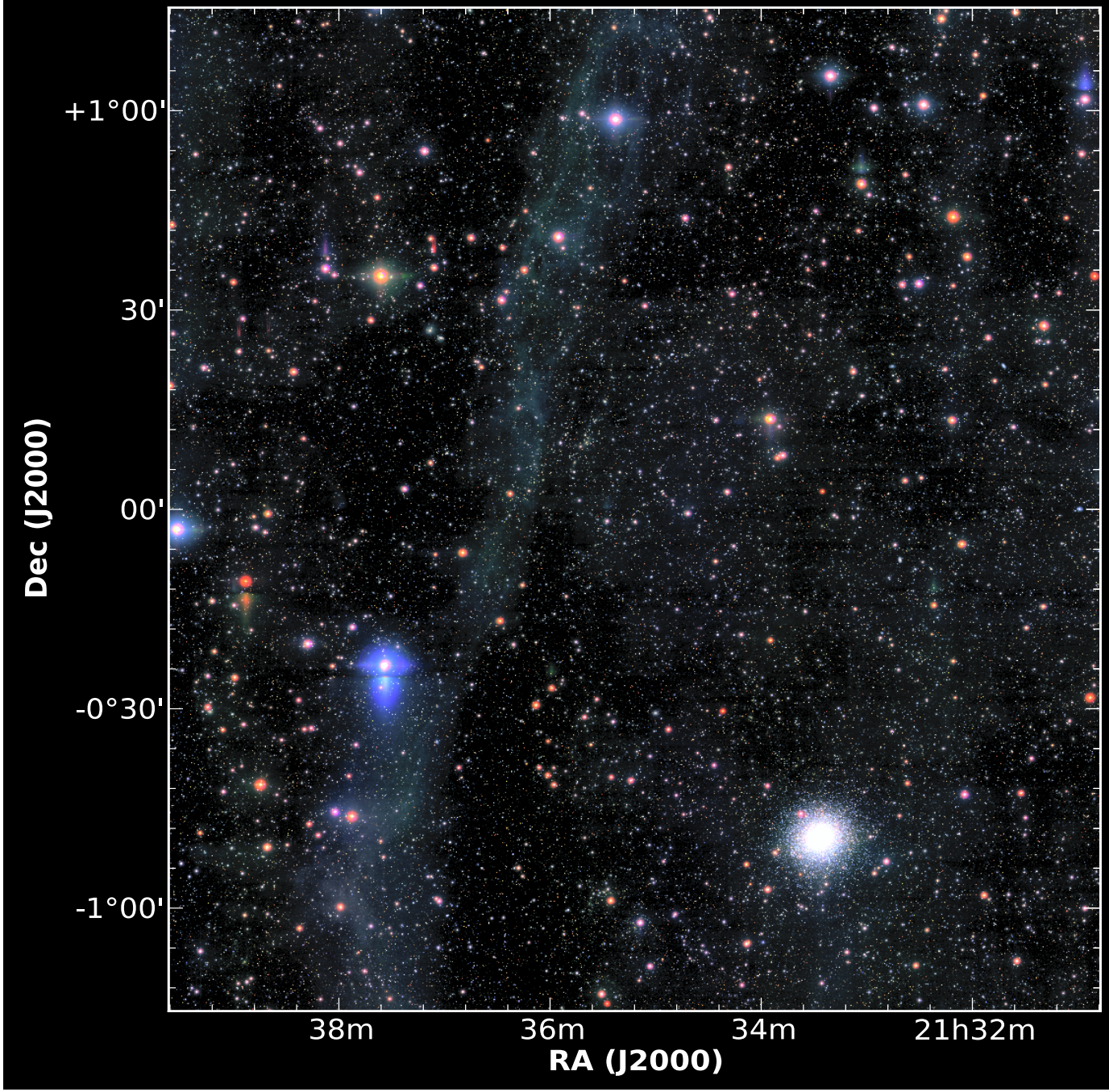
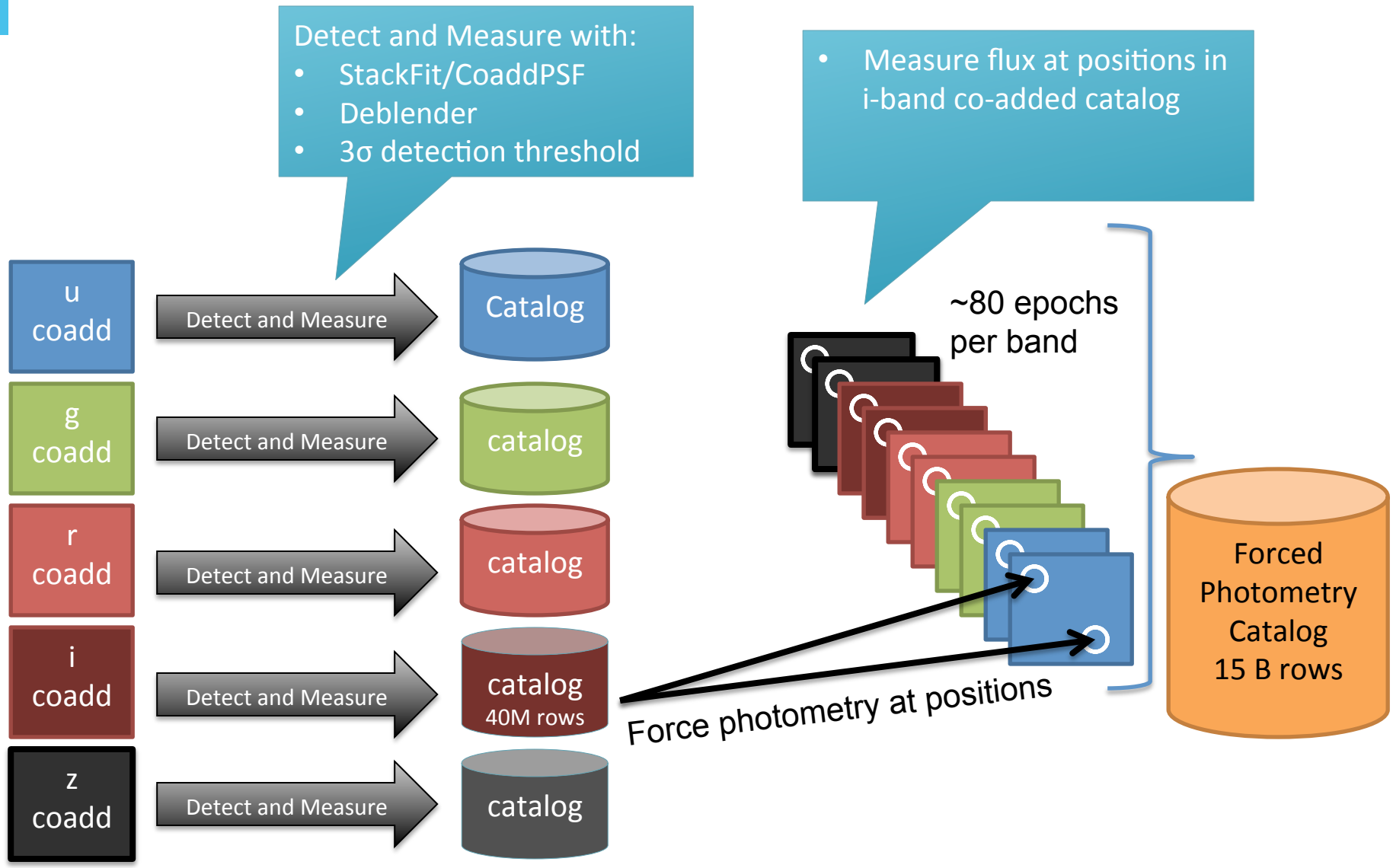


Figure:
5 sq. deg.
background-matched
coadd composite

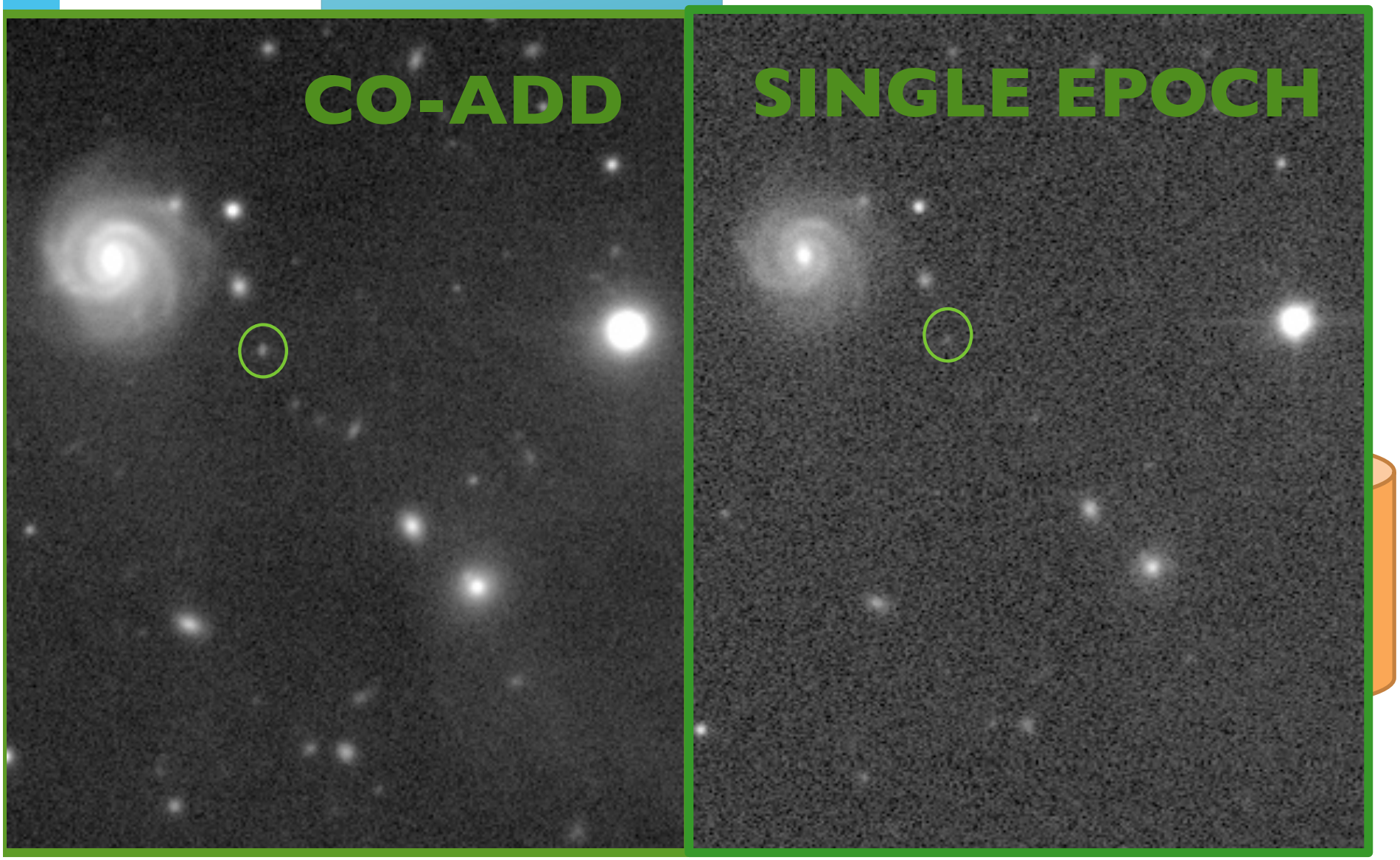
(g,r,i)
~55 epochs

Region: Aqr
Galactic lat = -35.0

Algorithmic pipeline converts images to flux measurements



Algorithmic pipeline converts images to flux measurements



XSEDE is not just for High Performance Computing

XSEDE

Extreme Science and Engineering Discovery Environment

Requested:
200,000 CPU-hours on Lonestar
118TB on Shared Disk

COMPONENT	TECHNOLOGY	PERFORMANCE/SIZE
Nodes(blades)	2 Hex-core Xeon 5680 processors	1,888 Nodes / 22,656 Cores
Memory	Distributed	45TB (Aggregate)
Shared Disk	Lustre, parallel File System	1 PB
Local Disk	SATA (146GB)	276TB (Aggregate)
Interconnect	InfiniBand Mellanox Switch	QDR 40 Gbit/s

Takeaways from running pipeline on XSEDE

- Very simple allocation request process.
- Service was fantastic. Tickets responded to within hours.

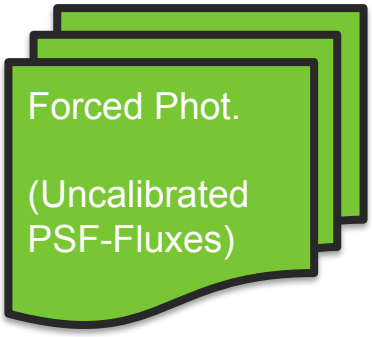
Lessons Learned:

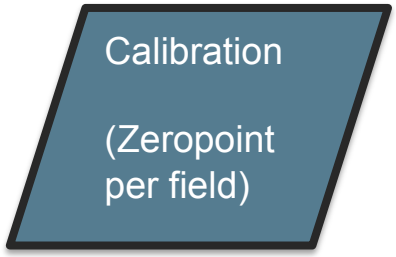
- Remember to copy the code to the individual nodes
- Expect to tweak configuration of any centralized resources
- Expect to be surprised

Centralized database for example:

```
max_connections=500 → 1000
open-files-limit=4096 → 16364
```

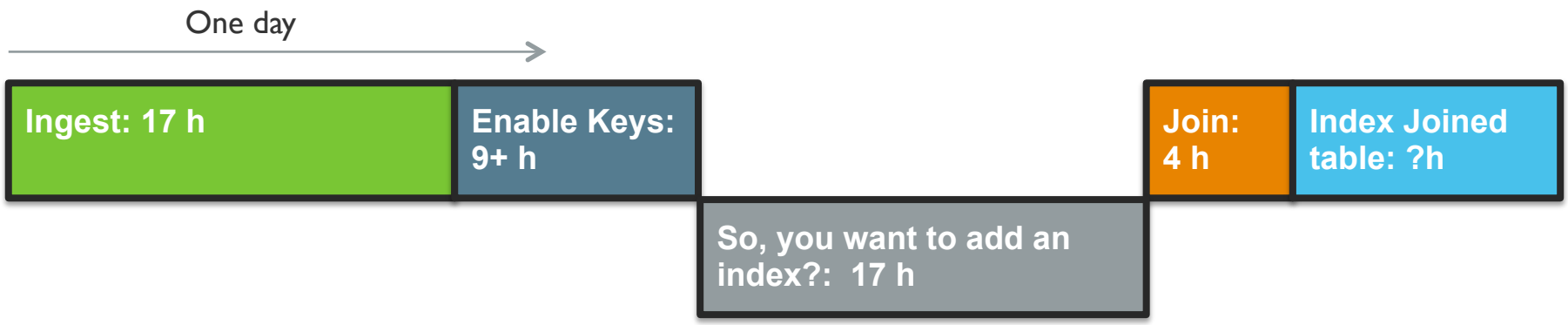
Transform 8 billion measurements to colors and lightcurves





MySQL (mysiam) on 8 billion rows (3.5TB)

- Partitioned by filter (ugriz) each with 1.7B rows



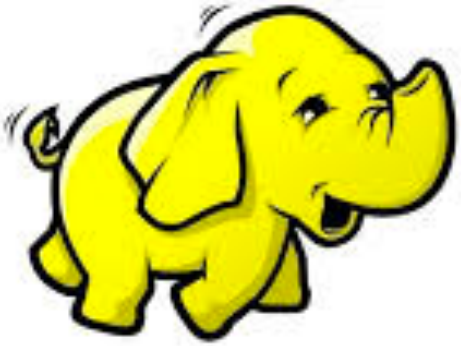
Lessons Learned:

- Intuition from one tool does not necessarily apply to another
- RDMS still appropriate for this task, but use smaller partitions

Industrial Big Data Tools?

The Apache Software Foundation

hadoop



cassandra

Spark

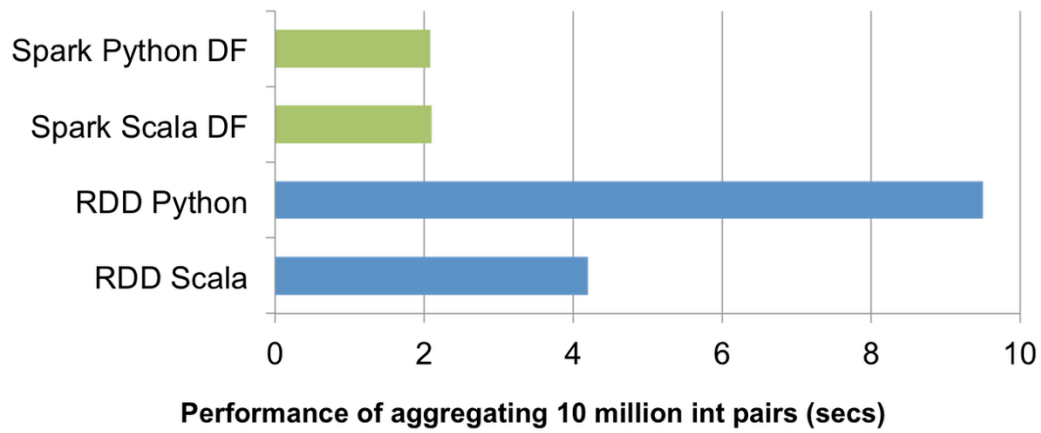
Keep an eye out for new interfaces

Spark

DataFrame

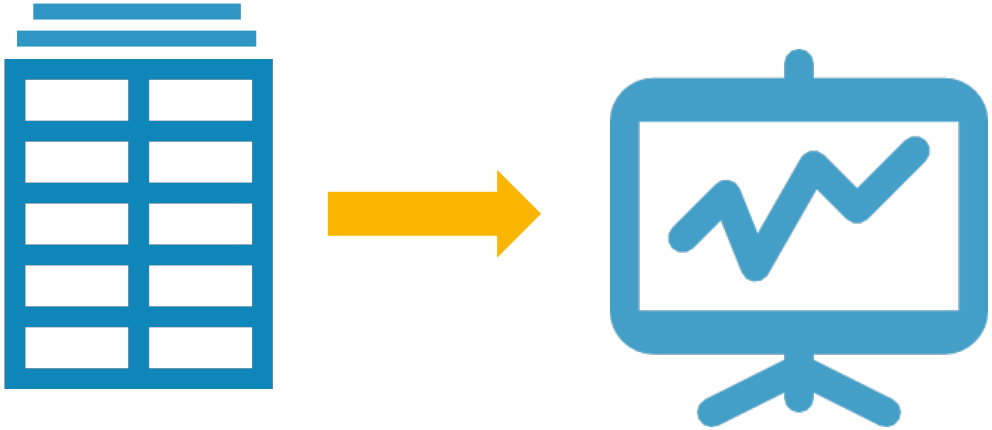
1.3.0 release "early March"

Credit: databricks.com

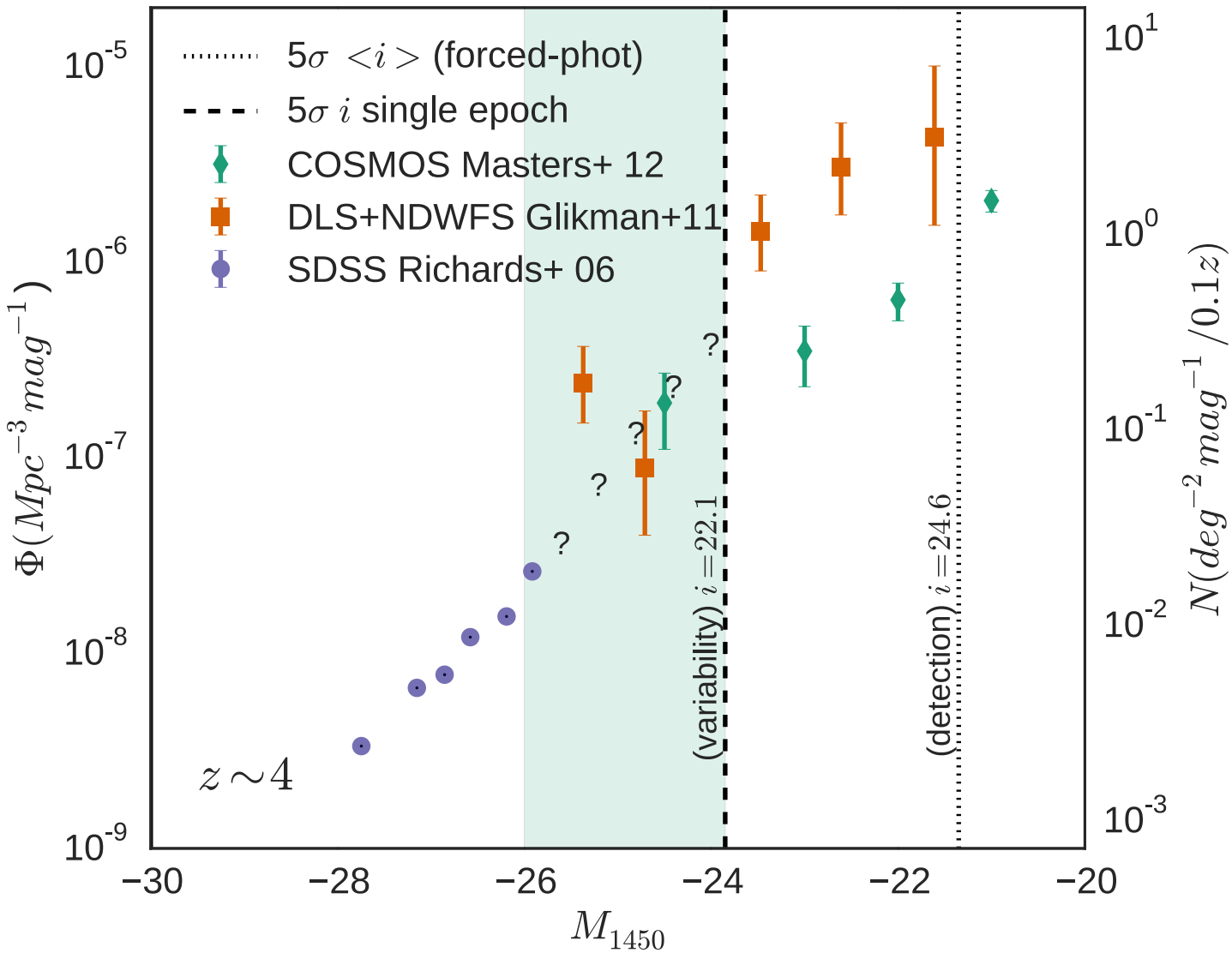


SparklingPandas

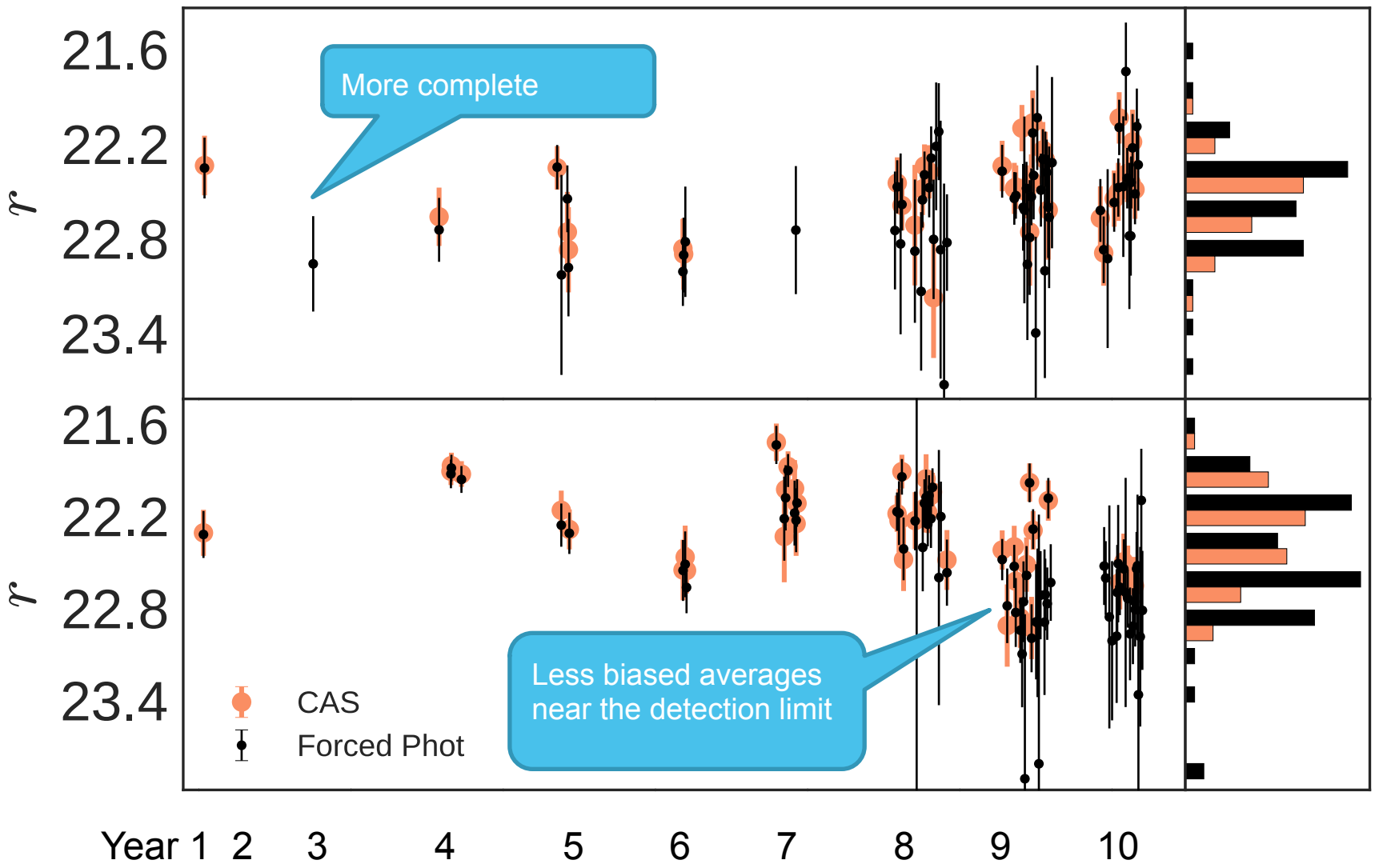
Lightcurves/colors to Science



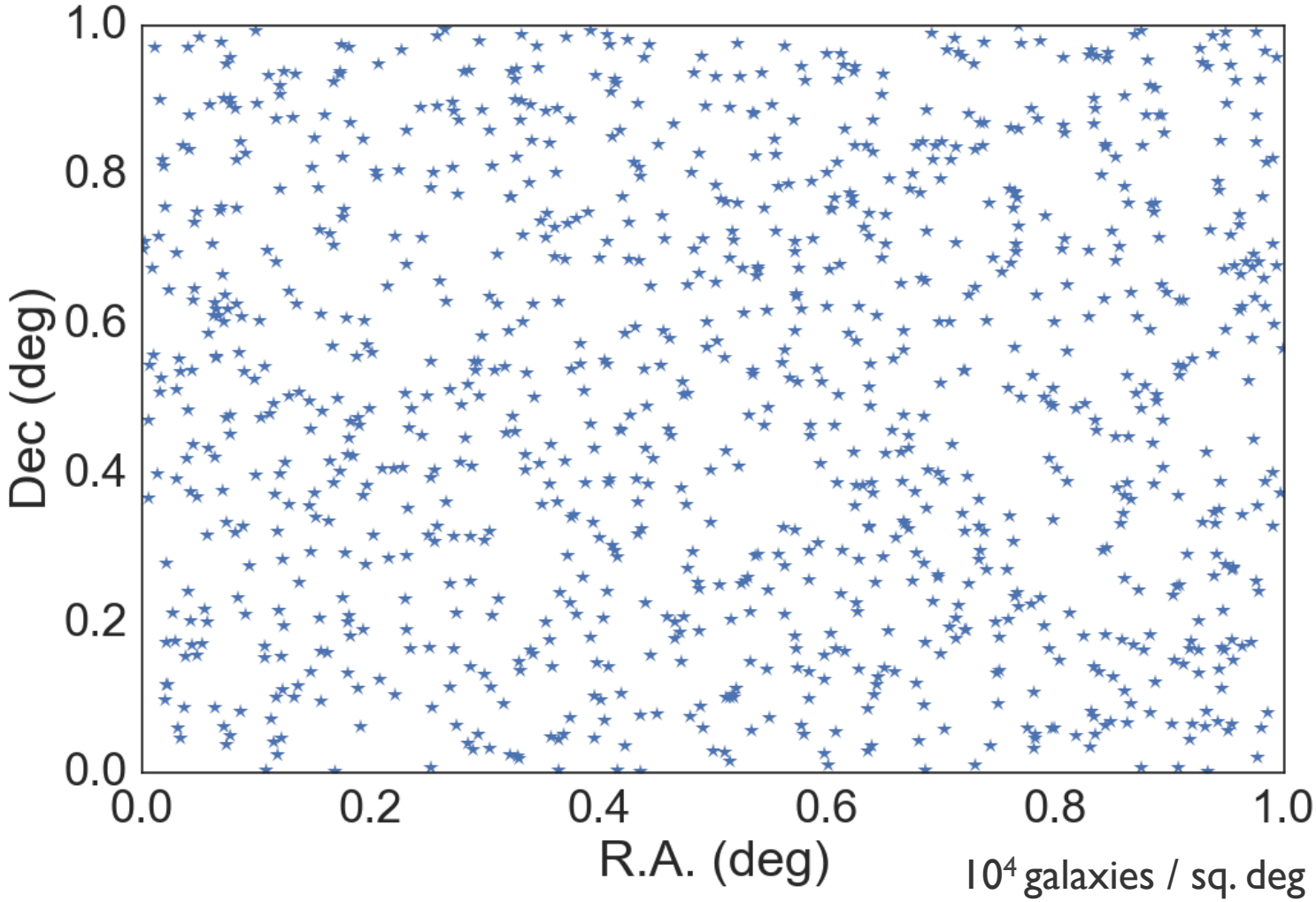
Goal: Measure z~4 quasar luminosity function 20 < i < 22.5



Forced photometry enables variability studies down to detection limit

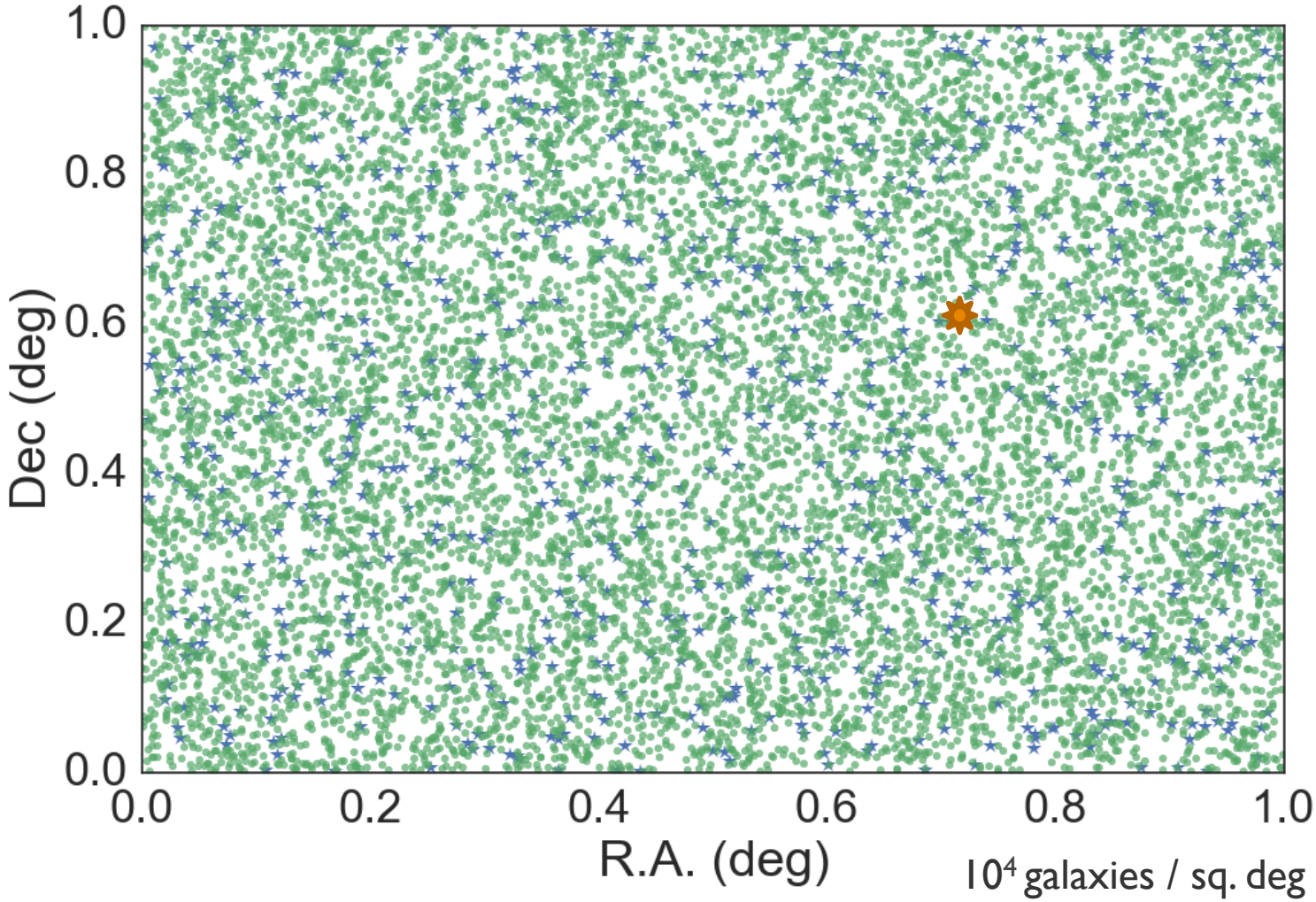


One Square Degree at *i* ~22 has…



10^4 galaxies / sq. deg / 0.5mag
10^3 stars
1 z~4 quasar

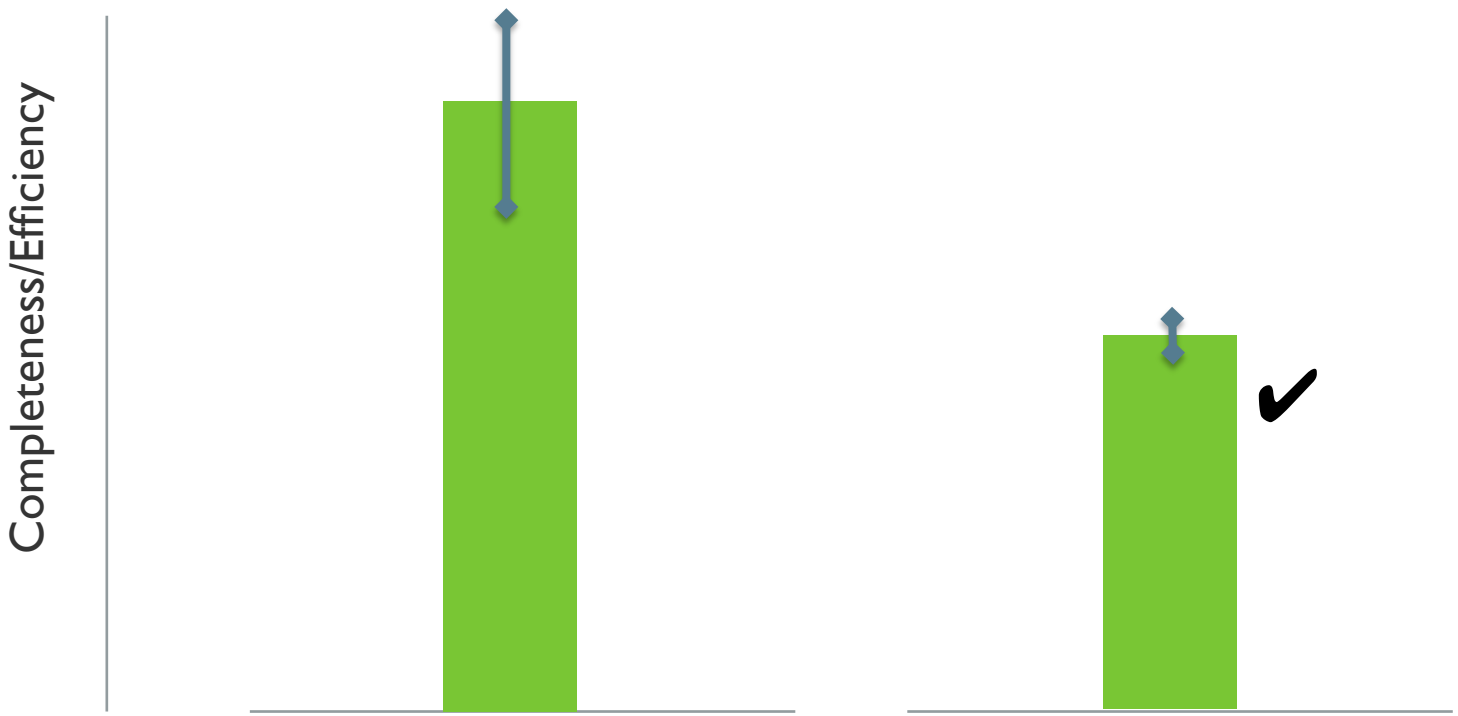
One Square Degree at *i* ~22 has…



10^4 galaxies / sq. deg / 0.5mag
10^3 stars
1 z~4 quasar

To make this QLF measurement we need:

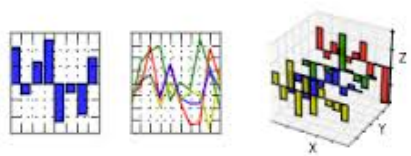
1) Automated selection from 10 million objects
2) Estimation of the selection function
   - (disconnect between astronomy and machine learning community)



Feature Engineering: variables to characterize lightcurves

- Damped Random Walk CAR(1) model:
  - MacLeod+ 2011;
  - Butler & Bloom 2011;
  - Choi+ 2014 (extended AGN)
- Structure Function Slope
  - Schmidt+ 2010;
  - Palanque-Delabrouille+ 2011;
  - Peters+ in prep (see poster!)

Tool to featurize lightcurves: pandas




```
>>> df = pandas.DataFrame.from_csv('toy.csv')
>>> anotherDf = pandas.DataFrame.from_csv('toyCategory.csv')

>>> df.join(anotherDf, on='category')

   category  value     Name
id
1         A    100   Apples
2         A     10   Apples
3         A     10   Apples
4         B     20  Bananas
5         B     20  Bananas

>>> grouped = df.groupby('category')
>>> grouped.mean()
          value
category
A            40
B            20

>>> grouped.apply(np.max)
         category  value
category
A               A    100
B               B     20

>>> grouped.apply(yourOwnSuperFancyFunction)
```

Meant to be more R-like and bring DataFrames to python:

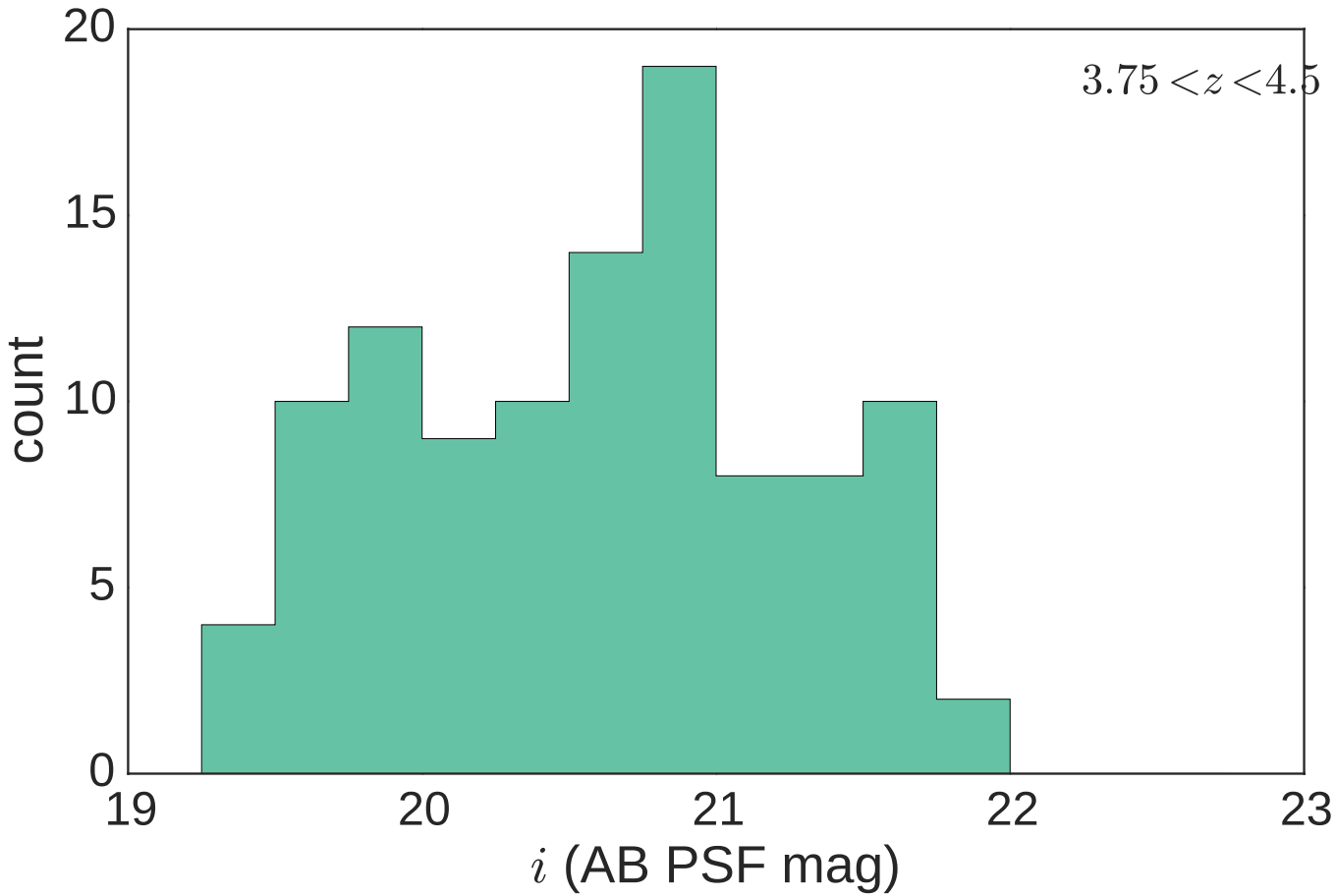
Joins

User defined aggregate functions!

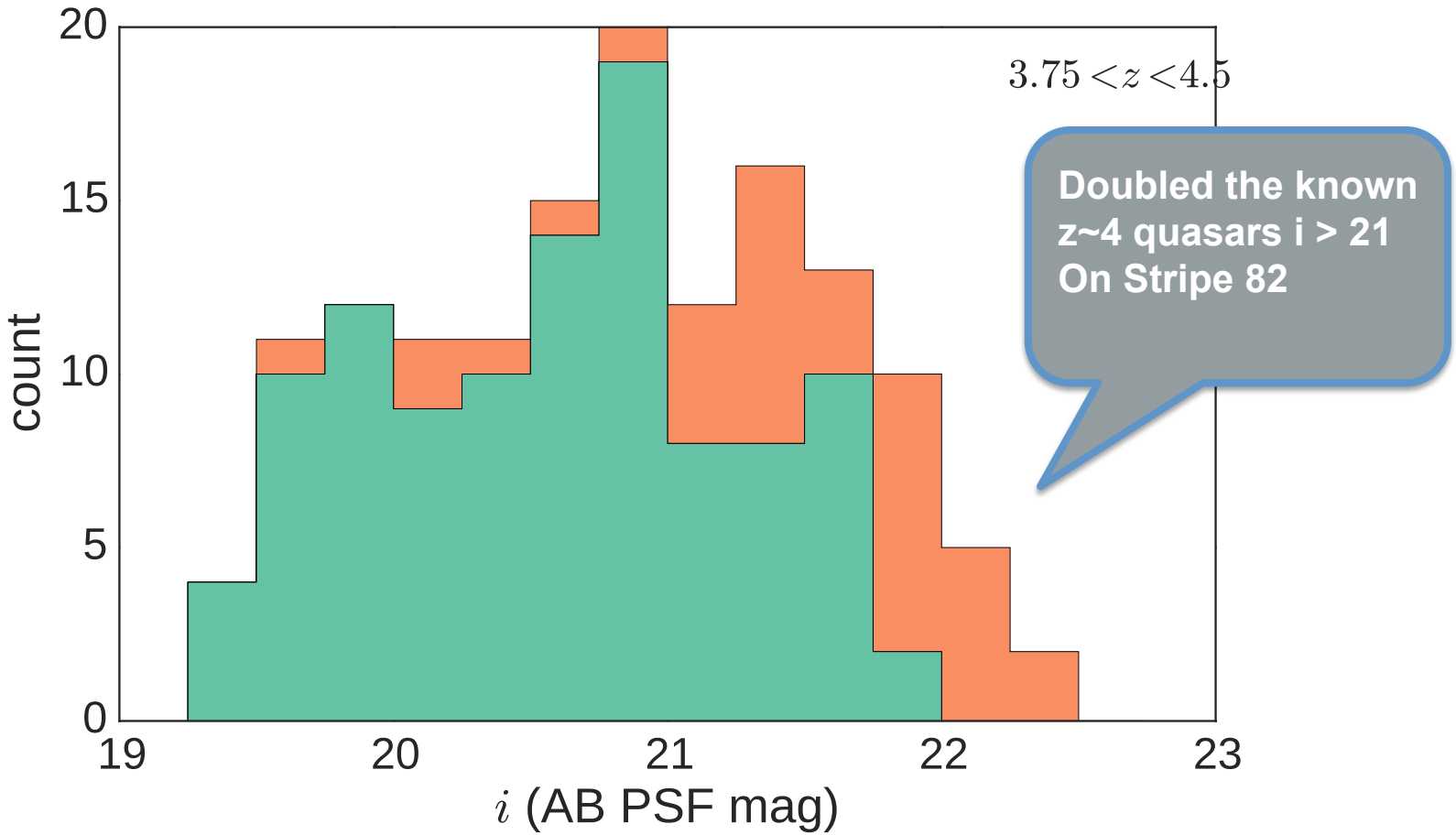
Build the Training Set

- To build the training set we cross-matched the point source catalog with spectroscopically confirmed quasars from:
  - BOSS DR12 Quasar Catalog (in prep.)
  - Jiang et al. (2006, 2008)
  - McGreer et al. (2013)
  - 2dF-SDSS LRG and QSO Survey (2SLAQ; Croom et al. 2008)
  - Palanque-Delabrouille et al. (2012)

Important that your training set looks like data



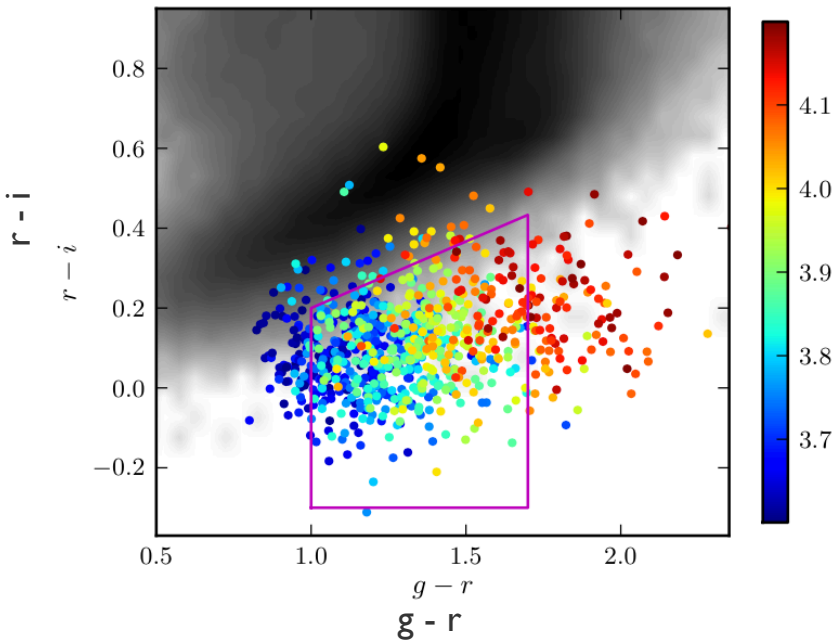
…so we took spectra of fainter z~4 quasars



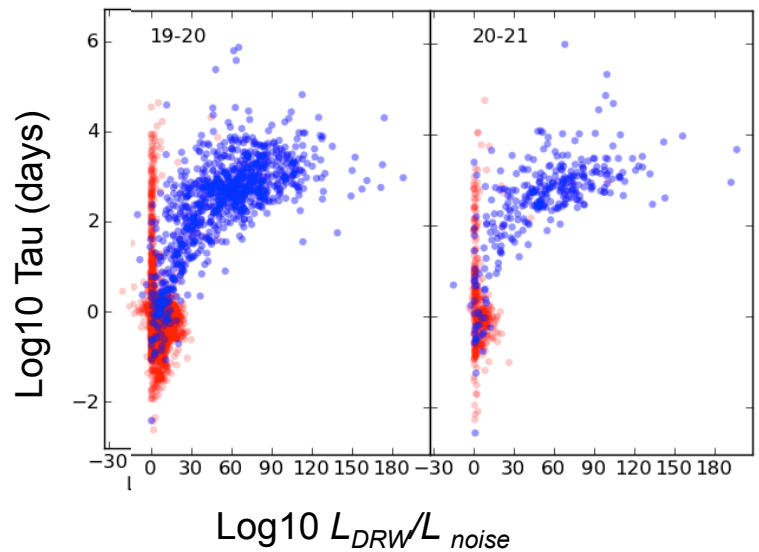
Preliminary Classifier Performance

- Colors AND variability
- For illustration:

Simulated quasar spectra from McGreer (in prep)

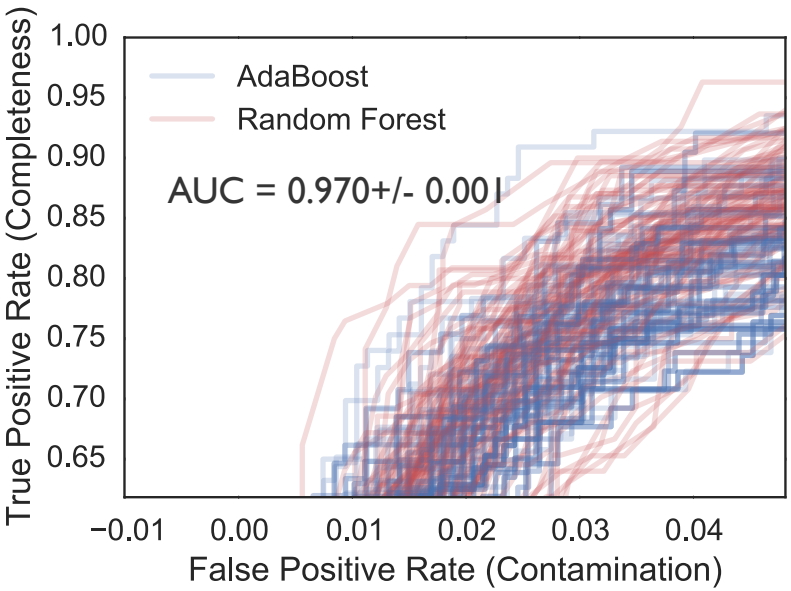


Time-scales for brighter quasars (blue)

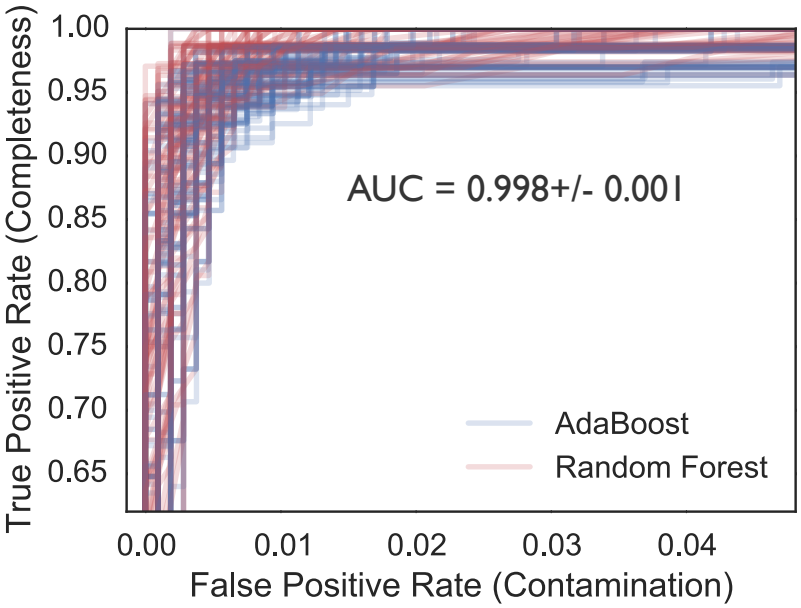


Preliminary Classifier Performance

Variability and Colors Separately



Variability and Colors Simultaneously



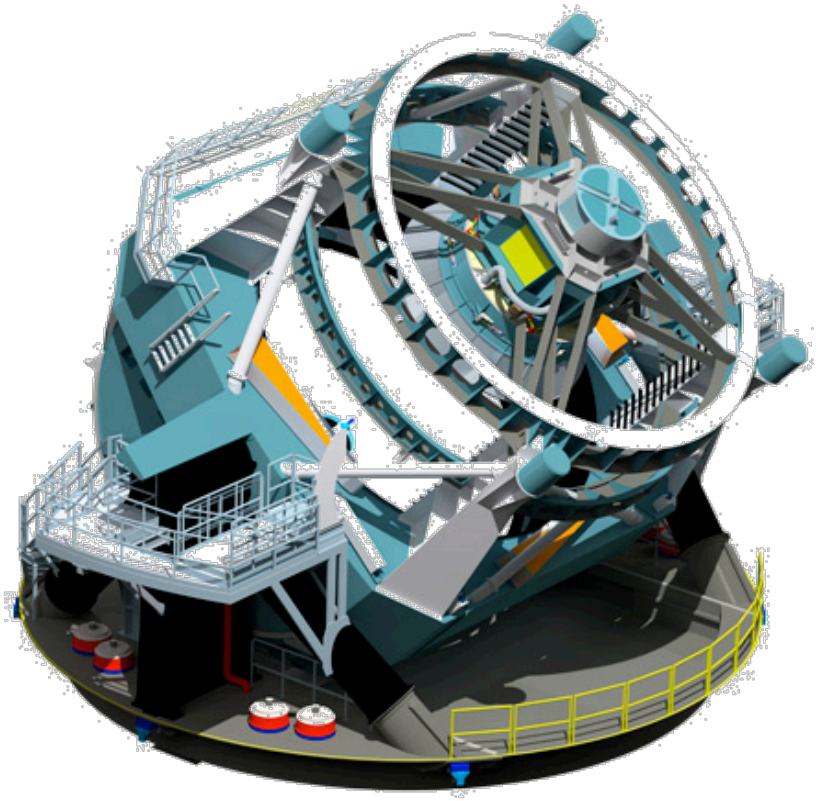
Variability-only classifier for color-selected objects.

Work in Progress and Questions

- Expecting the photometric sample to be large enough for clustering measurement at z~4
- Estimating the selection function w/ model spectra and lightcurves.
- Are median colors and variability parameters independent?
  - Need to model varying spectra?
- Can we train classifier with the model/simulated colors and lightcurves?

Finally

- LSST DM Stack is open source: http://dm.lsst.org
- Thank you



References

- Butler, N. R. & Bloom, J. S. 2011, AJ, 141, 93
- Croom, S. M., et al. 2008, arXiv, astro-ph, 19–44
- Glikman, E., et al. 2010, ApJ, 710, 1498–1514
- Jiang, L., et al. 2006, AJ, 131, 2788
- Jiang, L., et al. 2008, AJ, 135, 1057
- Kozłowski, S., et al. 2010, ApJ, 708, 927–945
- MacLeod, C. L., et al. 2010, ApJ, 721, 1014
- MacLeod, C. L., et al. 2011, ApJ, 728, 26
- Masters, D., et al. 2012, ApJ, 755, 169
- McGreer, I. D., et al. 2013, ApJ, 768, 105
- Palanque-Delabrouille, N., et al. 2011, A&A, 530, A122
- Richards, G. T., et al. 2006, AJ, 131, 2766–2787
- Schmidt, K. B., et al. 2012, ApJ, 744, 147