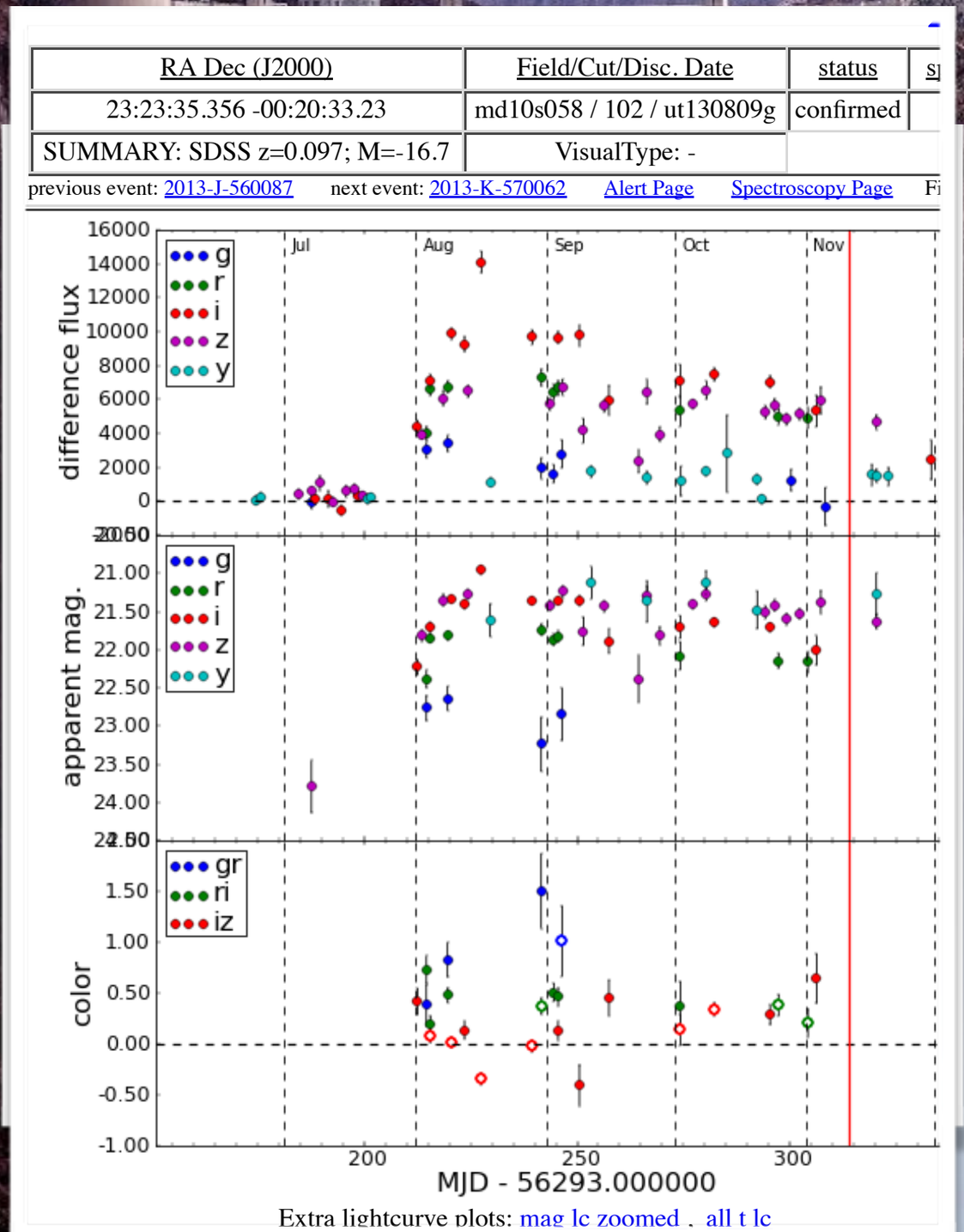
ANTARES: PROGRESS UPDATE!

GAUTHAM NARAYAN
NOAO

UA: Rick Snodgrass, John Kececioglu, Carlos Scheidegger
NOAO: Abi Saha, Tom Matheson, Monika Soraisam
CSS: Rob Seaman
LSST: Tim Jenness
Grads: Zhe Wang, Clark Taylor, Eric Welch, Shuo Yang
Jackson Toeniskotter, Jen Dempsey
NOAO REU + Honors: Tayeb Zaidi (Macalester)

MOTIVATION: THE PETAFLOOD

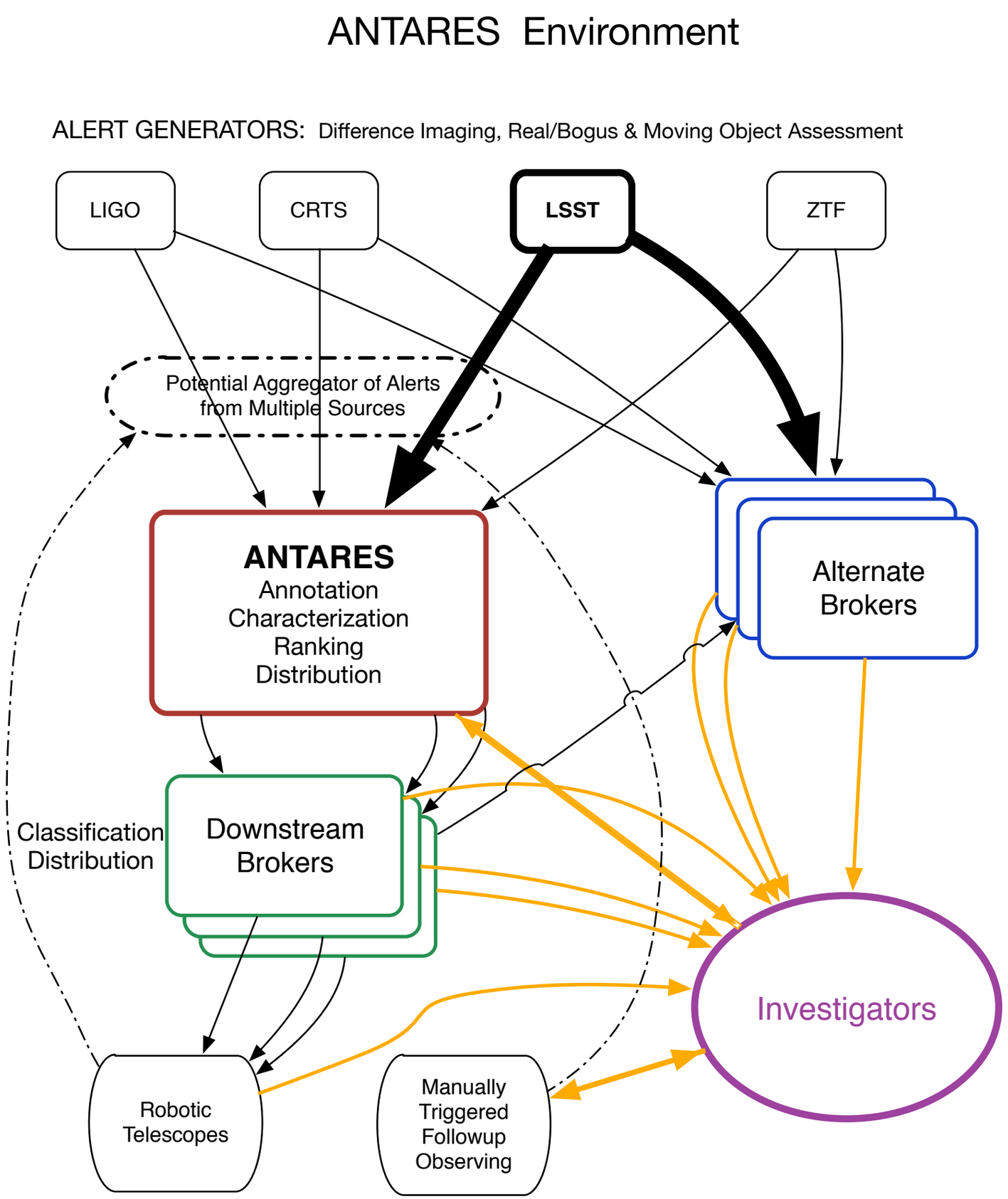
- Transient searches have relied on eyeballs for alerts
- LSST will produce several PBs of images, TB of catalogs, GBs every night
- ~ 1-10 million alerts per night
- Problem is **rate** more than scale
  - Roughly 37 seconds to process each image worth of alerts

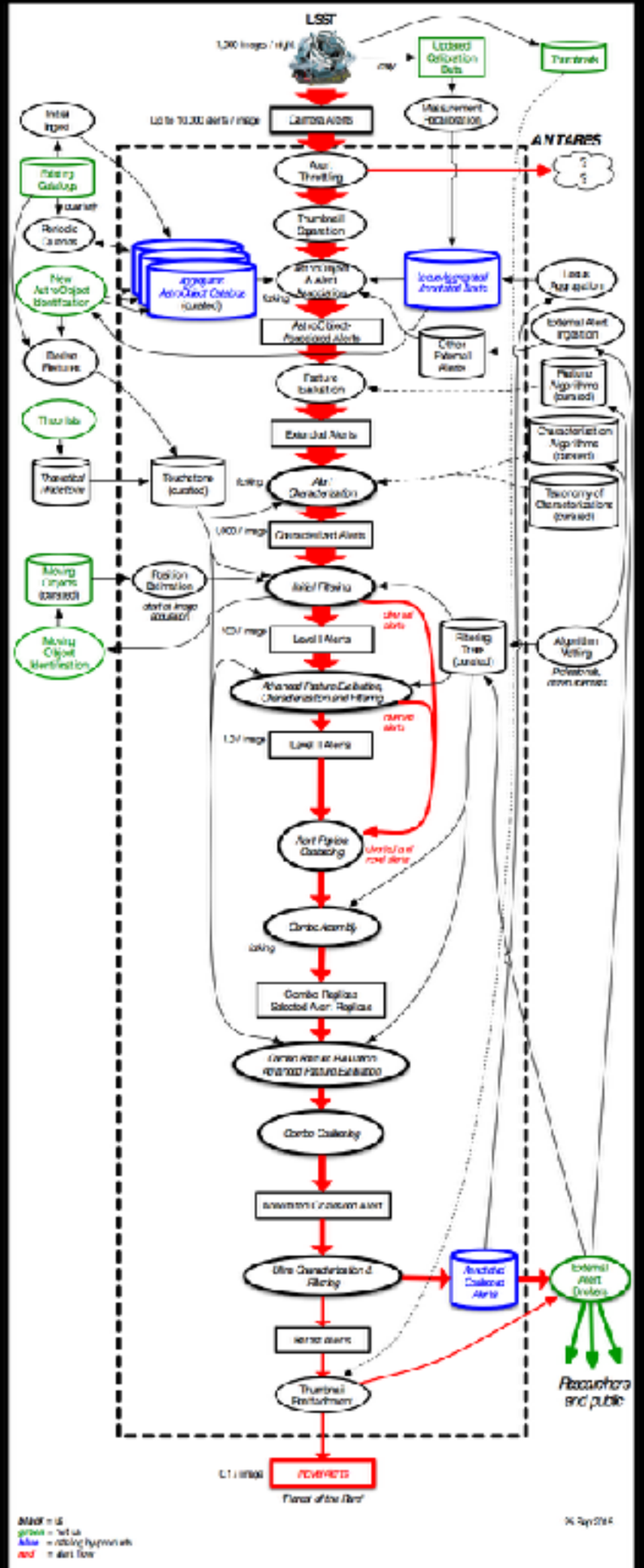


OVERVIEW

ANTARES

- Arizona-NOAO Temporal Analysis and Response to Events System
- Multi-disciplinary collaboration between astronomy, stats and computer science
- #NSFfunded! NSF INSPIRE proposal (IIA-1344024)
- System to characterize alerts automatically - not an LSST data product!
- Will winnow down the number of LSST alerts and let you pick the most interesting
- Archived alerts with contextual information, open source, open access in near real-time





ANTARES Architecture

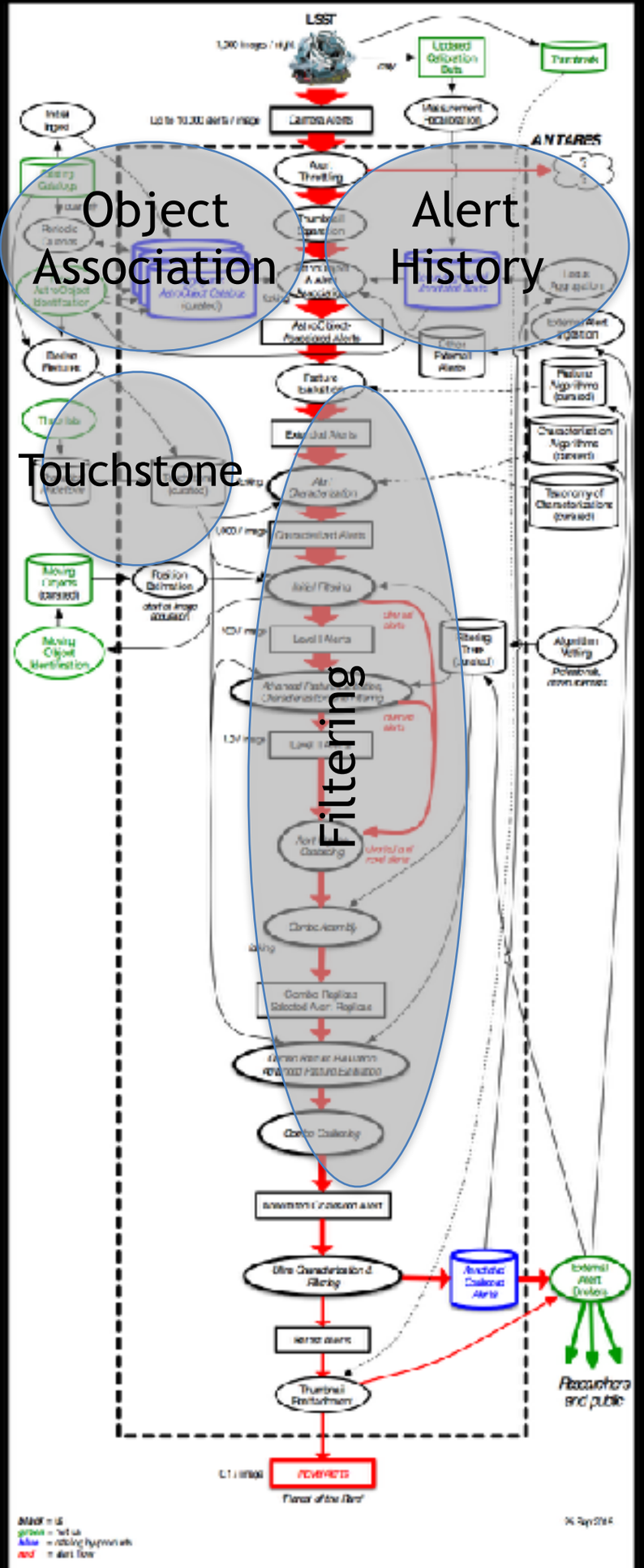
Alerts get annotated with external information and past history

Filtered on features

‘Rarest’ selected (protoype goal)

Flexible/modular system

What can it do?



ANTARES Architecture

Alerts get annotated with external information and past history

Filtered on features

‘Rarest’ selected (protoype goal)

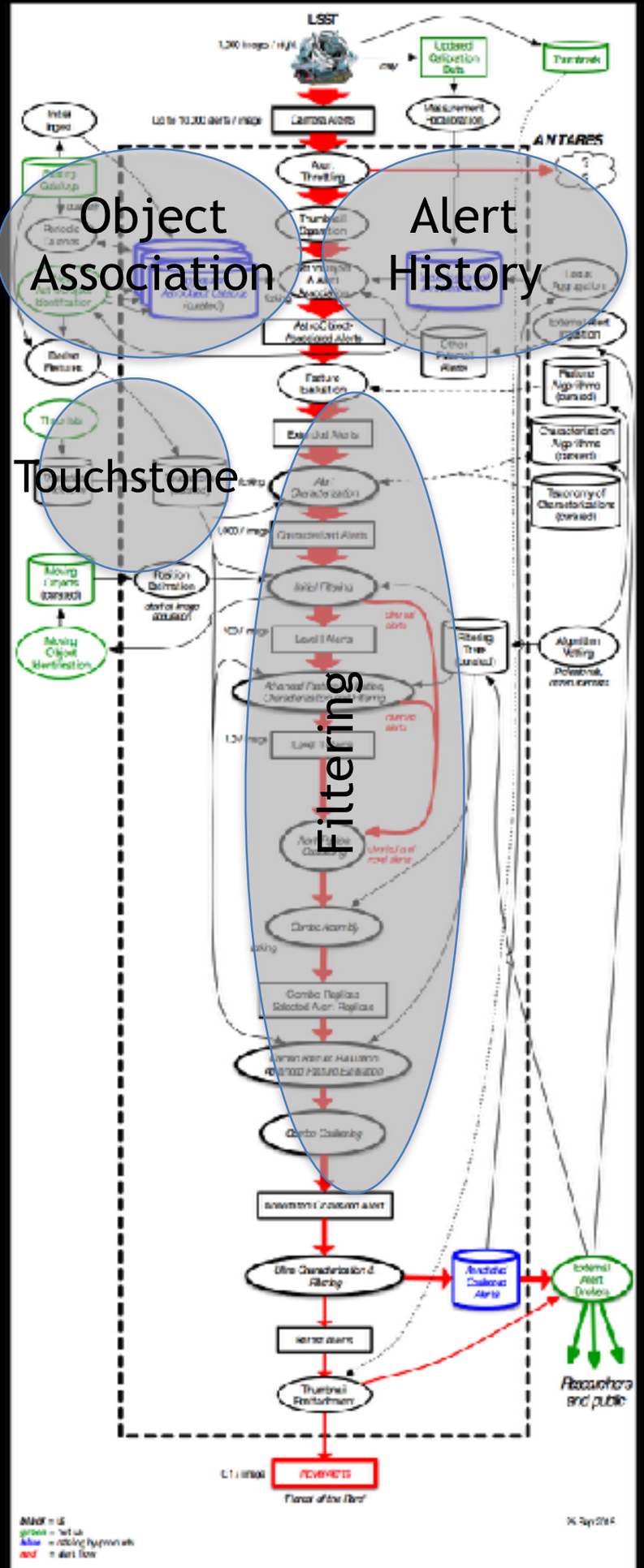
Flexible/modular system

What can it do?

The Touchstone :
A knowledge repository of properties and features of known or predicted kinds of variable astronomical sources.

A locus aggregated annotated alert can be compared to the objects or models in the touchstone for filtering and decision making.

Modus Operandi :
Alerts are ‘diverted’ when deemed ‘commonplace’ as they go through a succession of filtering stages. As stream thins, more complex algorithms can be applied.



Value is added by associating incoming alert with known source(s) at that location, and appending known properties of that associated source.

The Locus Aggregated Annotated Alerts :
Value added history of ***all*** past alerts. Repeat alerts at same location provides time sequence that can be used in conjunction with an incoming alert.

Also a database that can be queried for non-immediate follow-up use cases, with queries on characteristics and features.

OUR VIEW OF THE COMMUNITY'S FOLLOW UP INTERESTS

i.e. what needs we are trying to serve

- Things that need immediate follow up
  - Unknown things because they are rare and need rapid follow up before they disappear
  - Known kinds of transients that need rapid characterization for prioritized detailed follow-up
  - Predicted kinds of variables, but never seen (e.g. neutron star mergers)
- Things that can be studied at "leisure"
  - Variable stars & compilations thereof
- "Normal" variables or transients doing unusual (rare) things

(AKA ZOMBIE VERSION)

ANTARES IN MAY 2015 AT HOTWIRED IV

SO WHAT'S NEW SINCE THAT???

- Real-time cross-matching, feature derivation (even for "expensive" features) and classification.

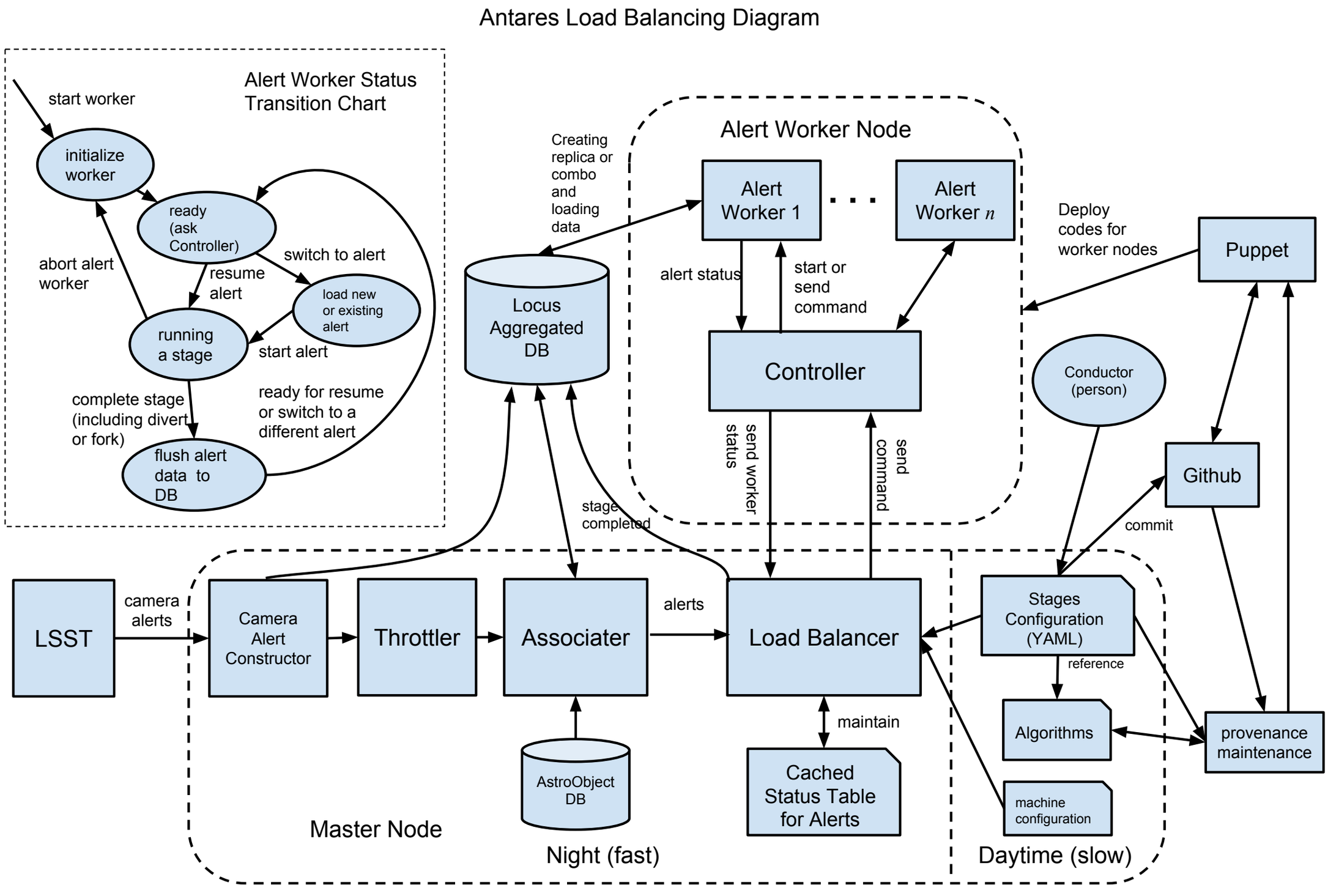
- Every alert is real, noisy, multi-band data, as is the Touchstone, and there are no common objects. We don't have classifications for our alerts, or features - i.e. we're doing science.

- We're running on a cluster (not VMs!) with a much more developed API

- Lots of fault-tolerance and failure resiliency has been built into the pipeline

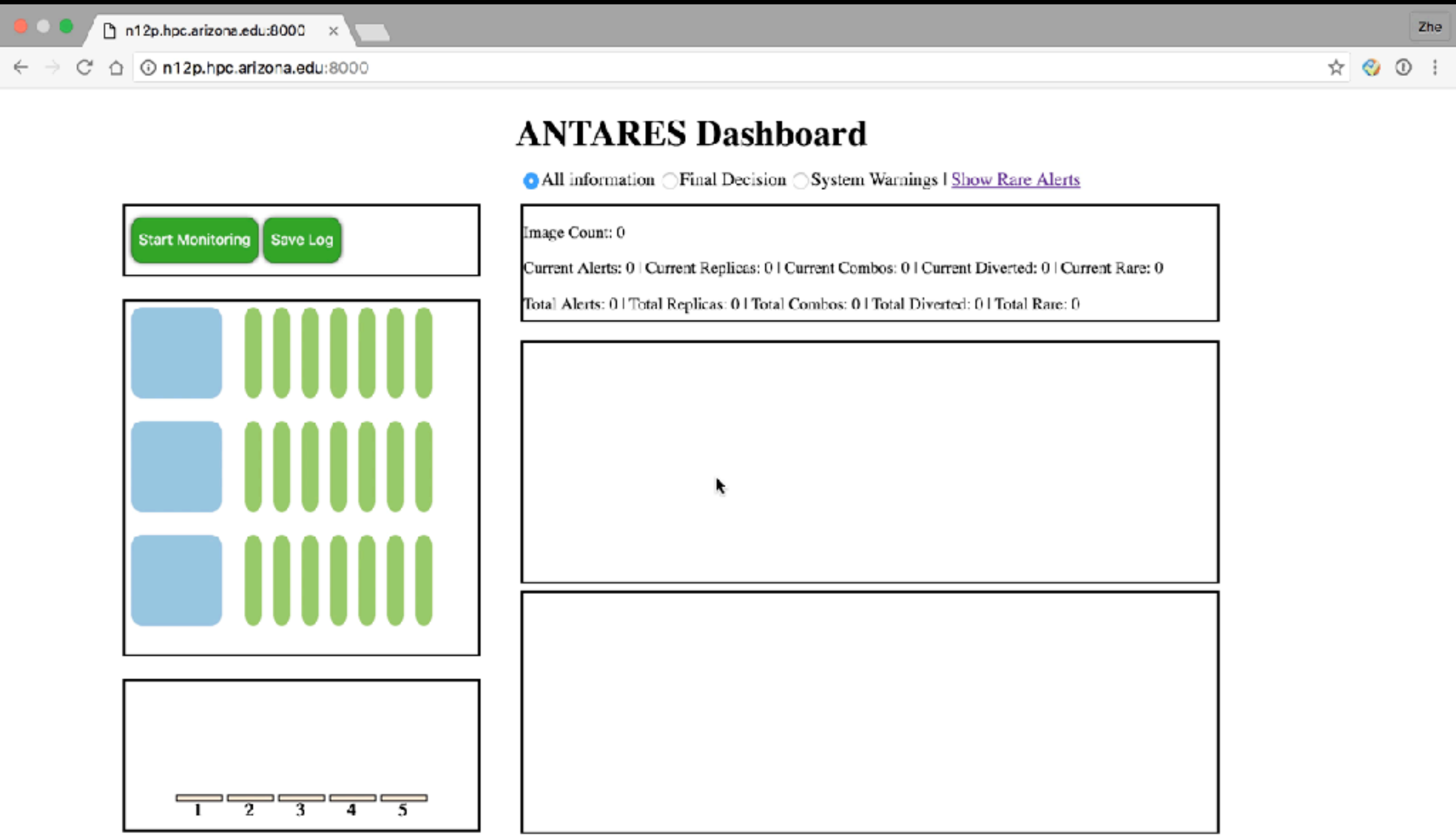
CHALLENGE: WE HAVE ~40 SECONDS TO PROCESS EACH IMAGE

FULLY PARALLEL LOAD BALANCING ARCHITECTURE WITH PROVENANCE



PROCESSING MONITOR + LOGGER

DASHBOARD



CROSS-MATCHING AND ASSOCIATION

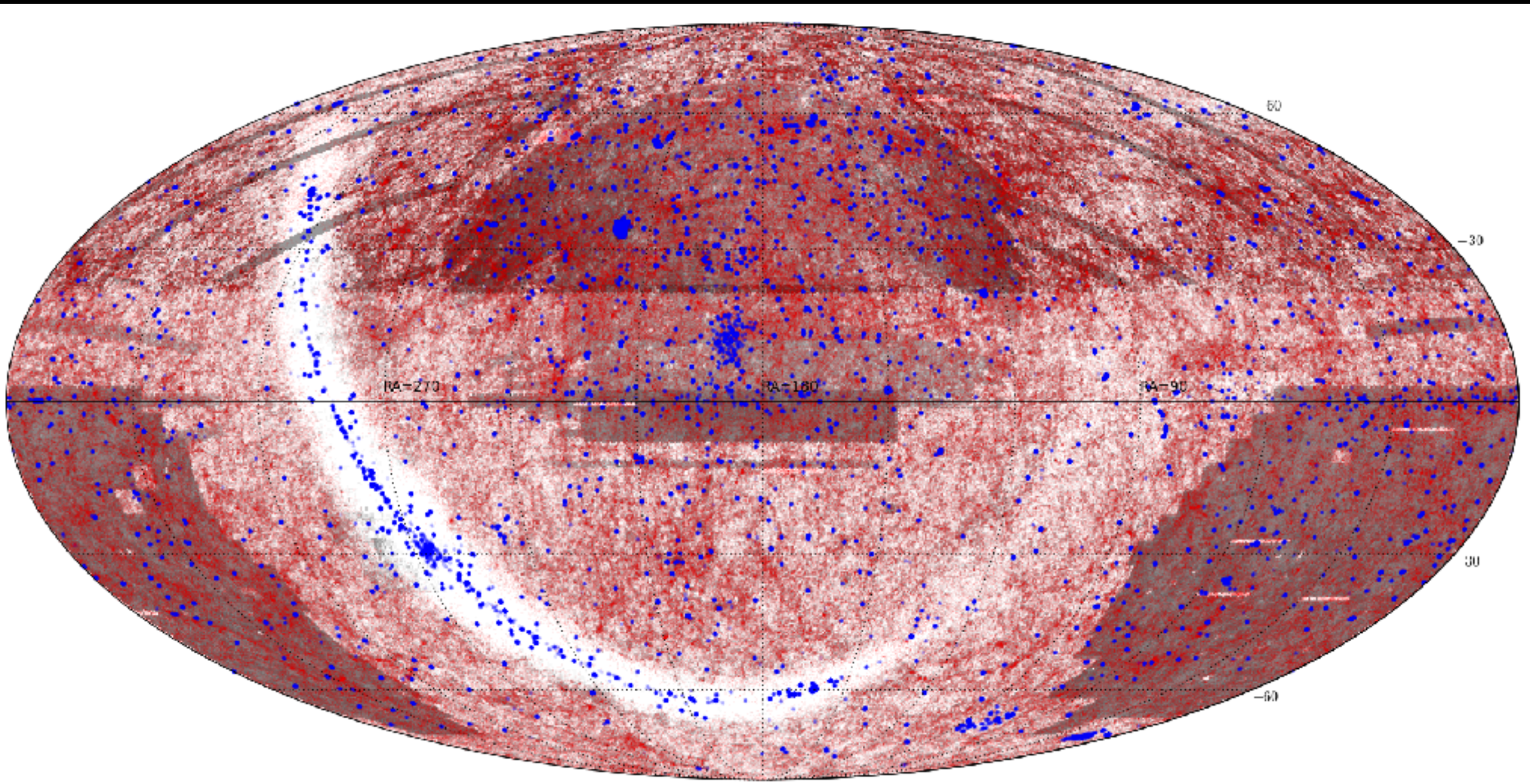
CROSS MATCHING AND ANNOTATING ALERTS: ADDING VALUE

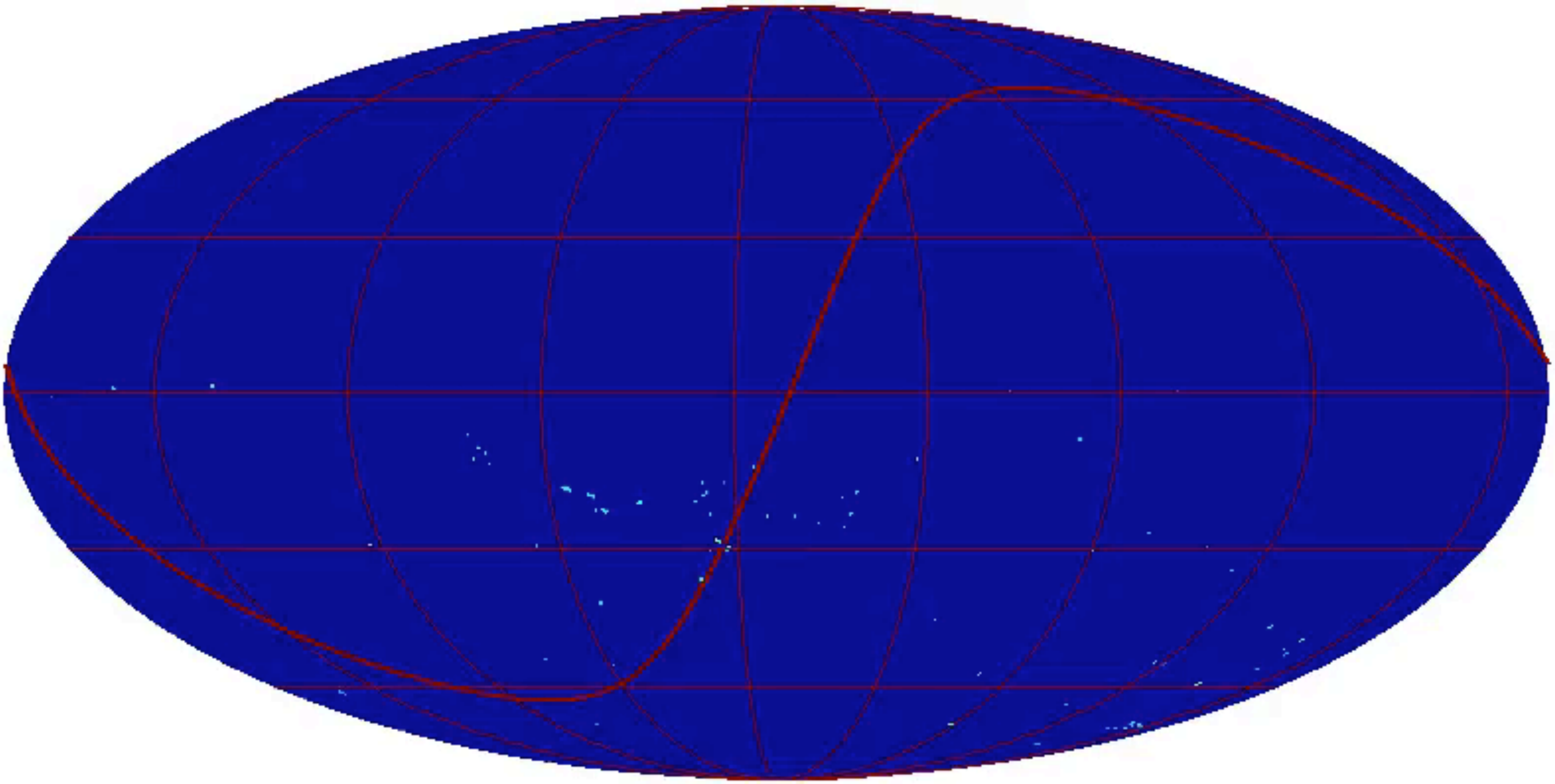
- External information is heterogeneous; LSST rate is too high for web-based queries
- Collecting and cross-indexing hundreds of astronomical catalogs (*SDSS, 2MASS, NED, IRSA, OGLE, Chandra, WISE, GAIA, PS1**)
- Cross-matching using cone search on HTM using SciSQL with adaptation for galactic plane
- Will be available to the wider astronomical community through the NOAO Community Science and Data Center (CSDC)

COVERING THE E-M SPECTRUM IS ABIG DATA PROBLEM

- SDSS DR7 + NED* (Grey density maps) + 2MASS XSC (Red, large-scale structure) + Chandra (Blue, high energy)

- This is an actual density map, not just a plot of SDSS stripes + NED fields - just under 190 million objects were parsed for this map. Expect few billion before LSST.





- CSDC is not just data access, but tools + support for data intensive projects, including time-domain

- Exploiting on-going southern surveys will drastically improve the time-baseline we can probe with LSST

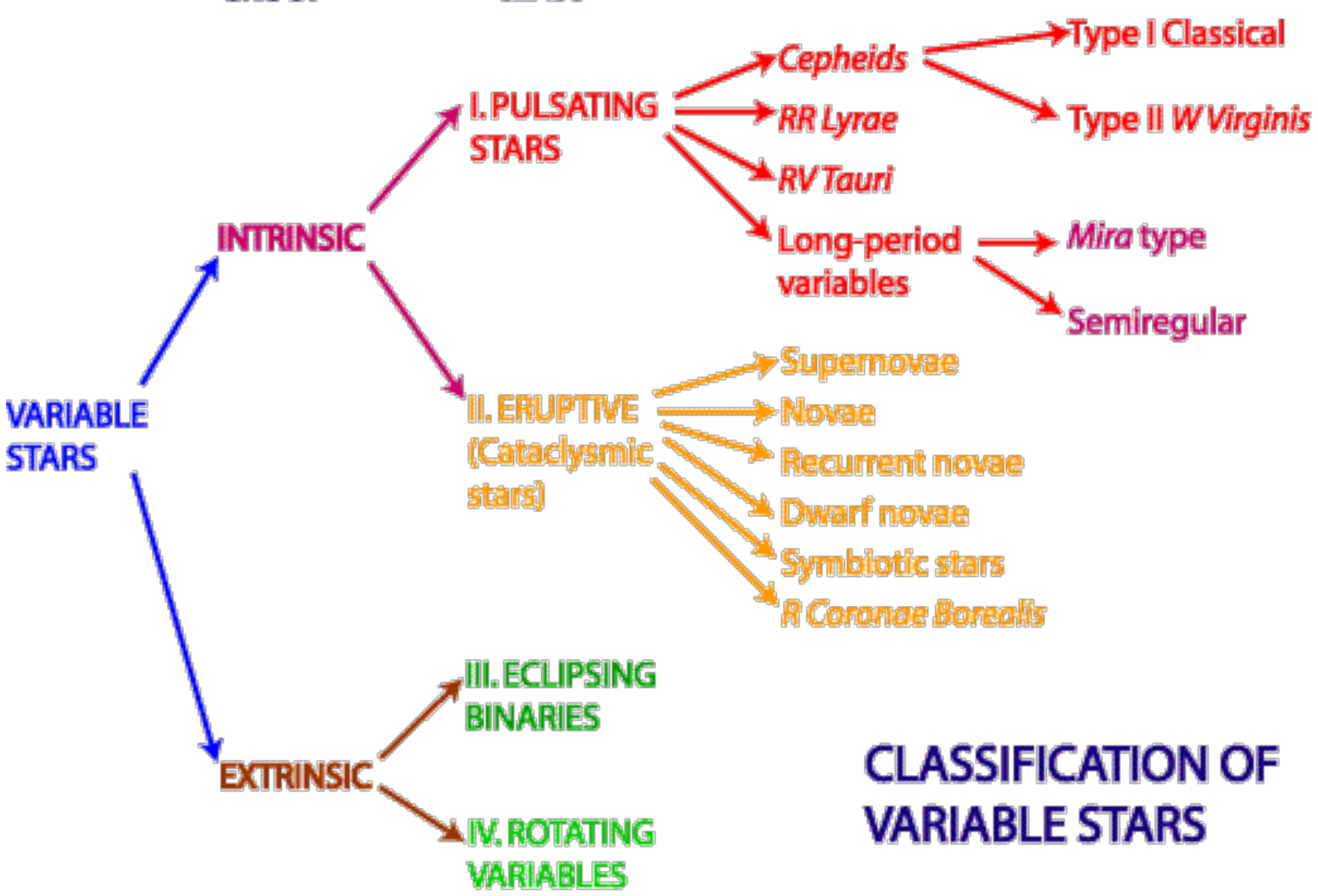
- Proposals to get time-domain observations for early science + library for training ANTARES, and calibrating LSST DD fields

Nifty view of CSDC holdings by DECALS/MzLS

http://legacysurvey.org/viewer

http://datalab.noao.edu/interact.php

THE TOUCHSTONE REFERENCE LIBRARY



WHAT ASTROPHYSICS IS IN THE TOUCHSTONE

(half the equation)

DATA SOURCES: OGLE + LINEAR + SNE.SPACE

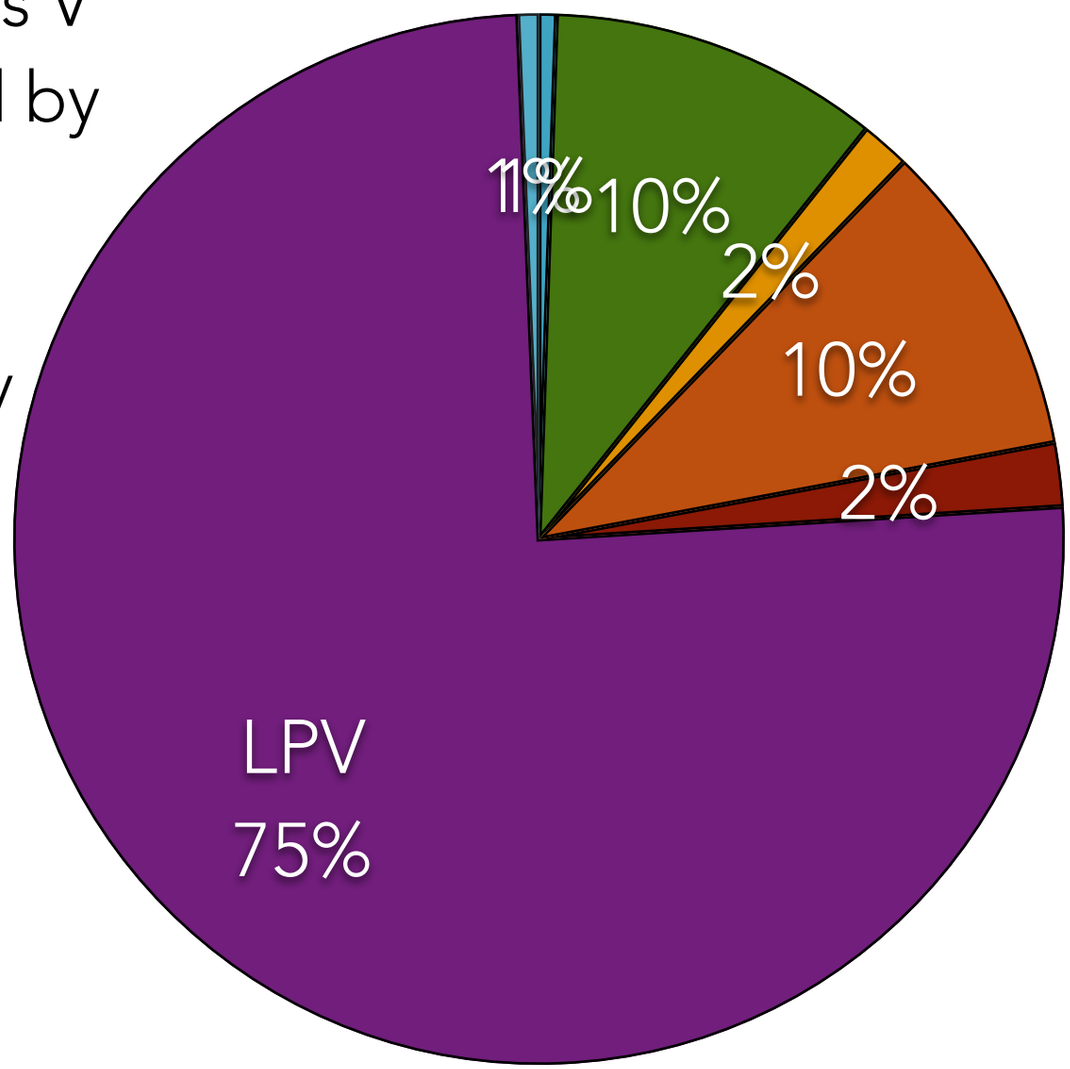
Misc Ecl SNe RR Lyr Cepheids LPV Delta Scuti

- The bulk of the data is V or V,I, and dominated by bright sources

- The SN data is mostly SNIa

- Some of the data is almost certainly mislabelled

- Many classes of variability and transient behavior are completely missing



FEATURE DERIVATION AND CHARACTERIZATION

FEATURES WE NEED IN THE ALERT OR WE'LL DERIVE FROM IT
(the other half of the equation)

MOST OF THESE LSST WILL PROVIDE*

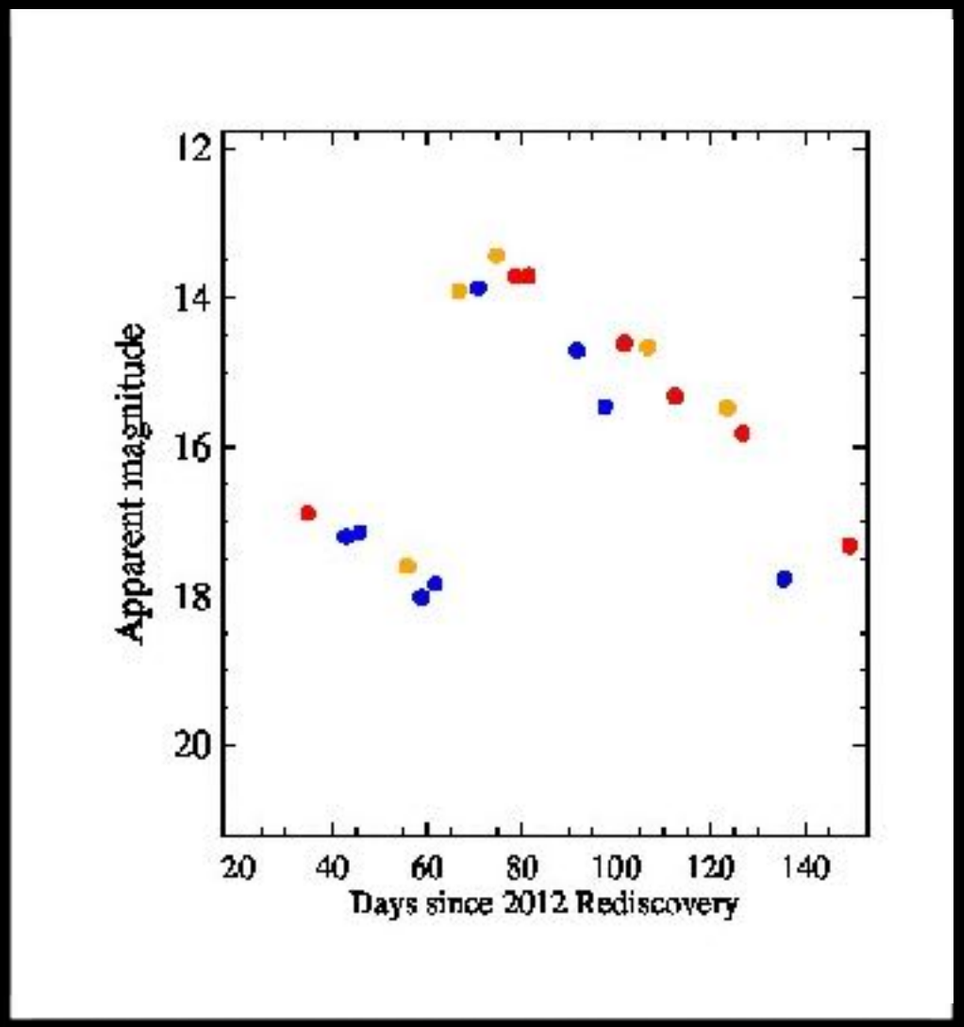
*MANY WILL NOT BE RELIABLE UNTIL AFTER A YEAR OF OPERATIONS

Description	Data location[a]	Quantity	Data Type	Why
Data quality information	A/I	x,y on array	float(s)	near edge/bad pixel?
Data quality, cosmic ray	A/I	x,y on array	boolean	If a c.r. is rejected, variance higher
Data quality information	I	Seeing (arcsec)	float	point source or not
Data quality information	I	Point Spread Function	float [0,1]	point source or not
Equatorial Coords. (RA/DEC)	A	Basic coordinates	floats	position on sky, AstroObject assoc.
Galactic (Milky Way) Coords.	D	Galactic lat/long	floats [0,1]	probability of Galactic origin
Ecliptic Coords.	D	Ecliptic lat/long	floats [0,1]	probability of Solar System origin
Measured brightness	A	mag	float	objects will have different potential brightness ranges
Change in brightness	A	Δmag	float	objects will have different potential Δmags
Prior amplitude range	D	min max range/filter	floats	Test deviation from known variation
Time scale	A/L/D	Δt	float	Different phenomena will have different time scaless
Known source (outside LSST)	C	$\text{Flux}_{\text{external filter}}$	various[b]	Different phenomena emit over different EM ranges
Nearest object on sky	C	distance in arcsec	various	Different phenomena associate with different objects
If galaxy	C	Type, redshift, pos. in gal.	various	type and distance will guide expected phenomena
If star	C	Type, mag, π	various	type and distance will guide expected phenomena
Periodic	D/C	P	float	sort known periodic phenomena
Fourier components [TBD]	D	First n components	floats	characterize variability
Color	D	$m_x - m_y$	foat	narrow range of possible objects
Light curve	D	mag vs. time	floats	different phenomena have different light curves
Light curves (multiband)	D	$[\text{mag vs. time}]_{\text{filter}}$	floats	correlate different bands ($\rho_{f_1 f_2}$)
Moving object	M	Prob. moving object, μ, π, PSF shape	floats	eliminate moving objects

[a] (I)mage-level data, (A)lert-level data, (L)ocus-aggregated alert information, (C)atalog information, (D)erived quantity, (M)oving object (from LSST/MOPS)
[b] In this context, various means that the associated object will have a name (string), and various measurements of flux, position, etc. (floats).

WHY SHOULD YOU CARE ABOUT CADENCE?

- LSST has only simulated a few cadences
- Metrics Analysis Framework (MAF) to assess different cadences
- **You** write metrics to optimize for the science you care about, LSST runs simulations



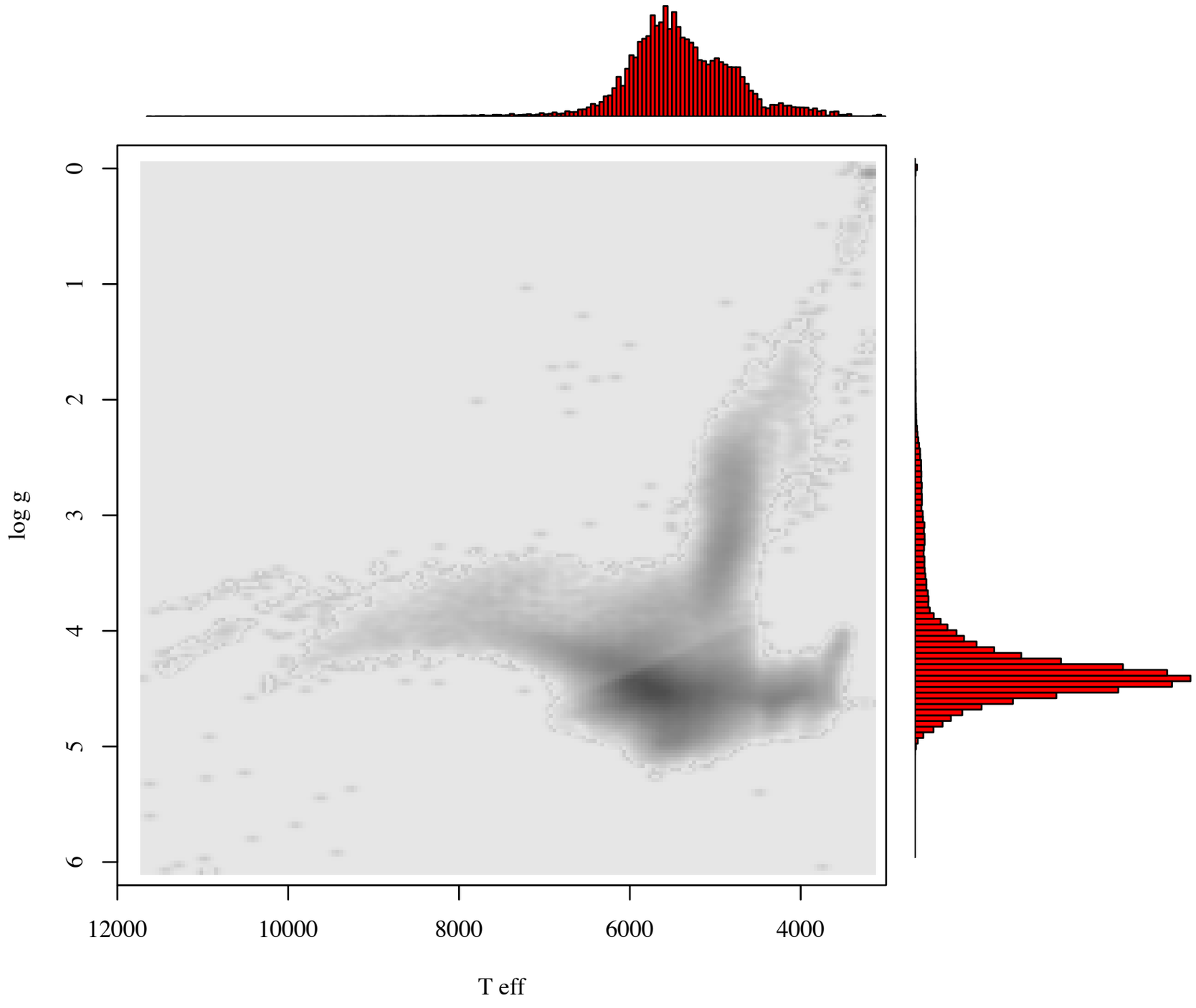
https://confluence.lsstcorp.org/display/SIM/Metric+Analysis+Framework,+Design+Review

So if we want to address these drawbacks, we need to build

- Something with a cheap computation
- That does not use the full lightcurve
- Doesn’t need the redshift
- Doesn’t use anything that can introduce a cosmological bias

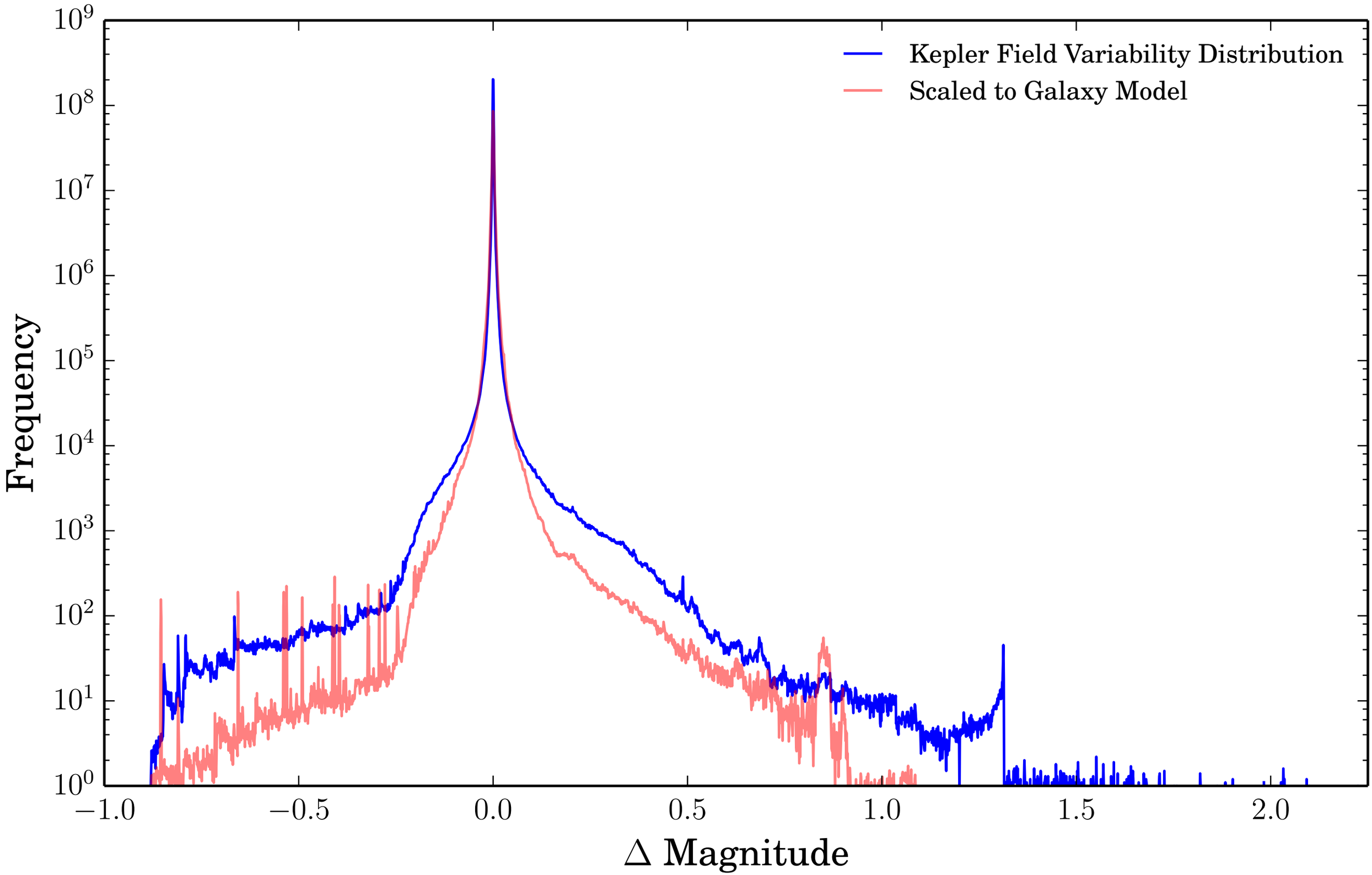
ASSESSING THE LIKELIHOOD OF VARIABILITY

- Kepler provides an empirical model for variability of Galactic stars
- 155k+ stars over wide range of spectral types
- Scale Kepler variability distribution to match Besançon model of spectral types



VARIABILITY PROBABILITY DISTRIBUTION

Can predict variability given galaxy model and galactic coordinates of alerts using Kepler Q13 data (*Ridgway+ arXiv:1409.3265*)



IF WE WANT TO EXTEND THIS TO TRANSIENTS:

- Compute all combinations of rates of change of observed light curves*
- Any sequence of alerts from LSST should lie in this space
- Compute rates of change between consecutive alerts at the same locus
    - Things are fall within the mapped space are known
    - Things that fall outside are potentially interesting
    - Additionally, Can use kNN to attempt to classify

* Ideally, the spectroscopically labelled sample would reflect the underlying rates, but they don’t.

SEPARATION, CHARACTERIZATION, MODELING

- We can quickly determine if something is behaving like a variable from VPDF + NN2.

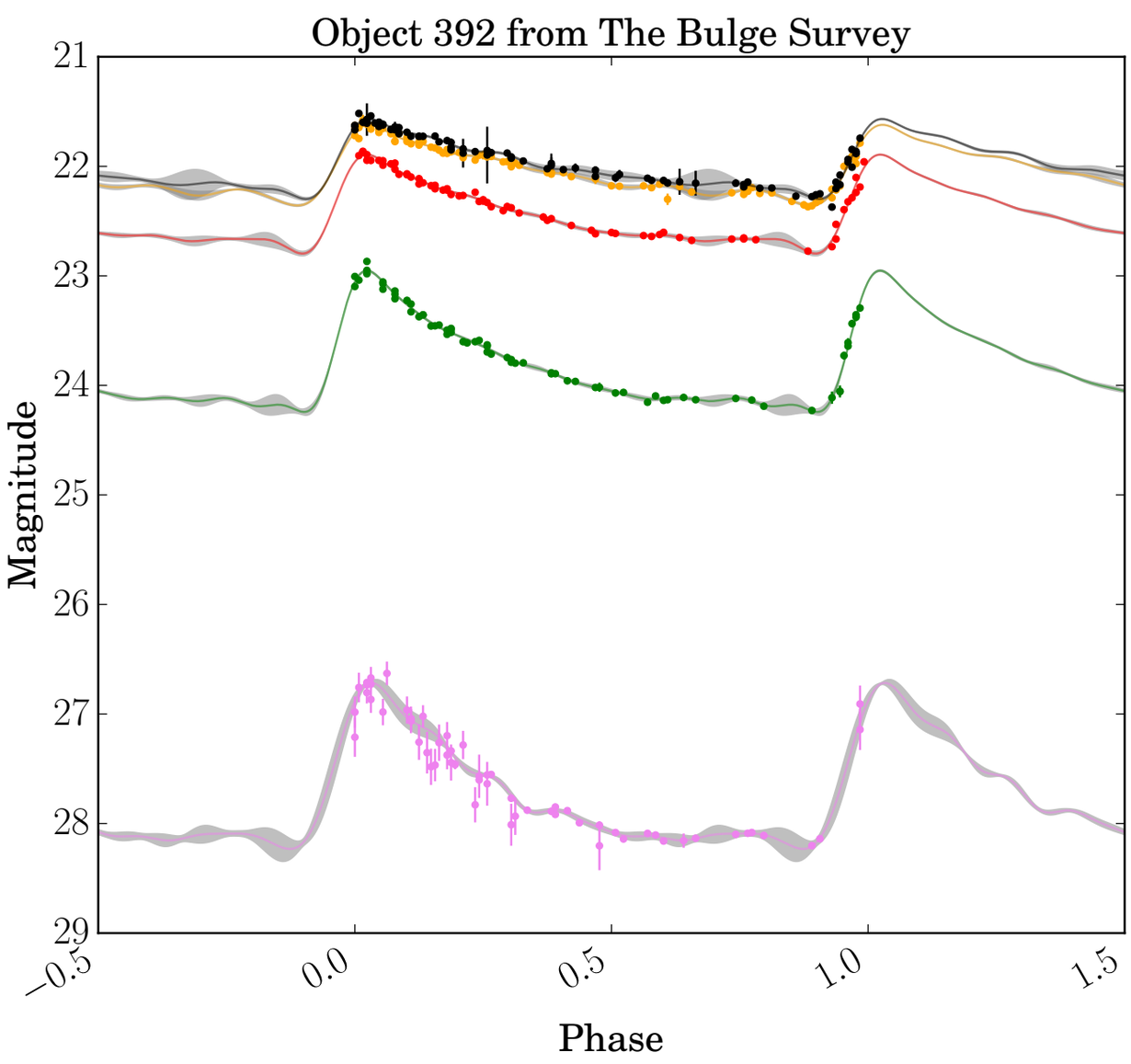
- We use realtime Lomb-Scargle to identify likely periodic variable candidates (gatspy)

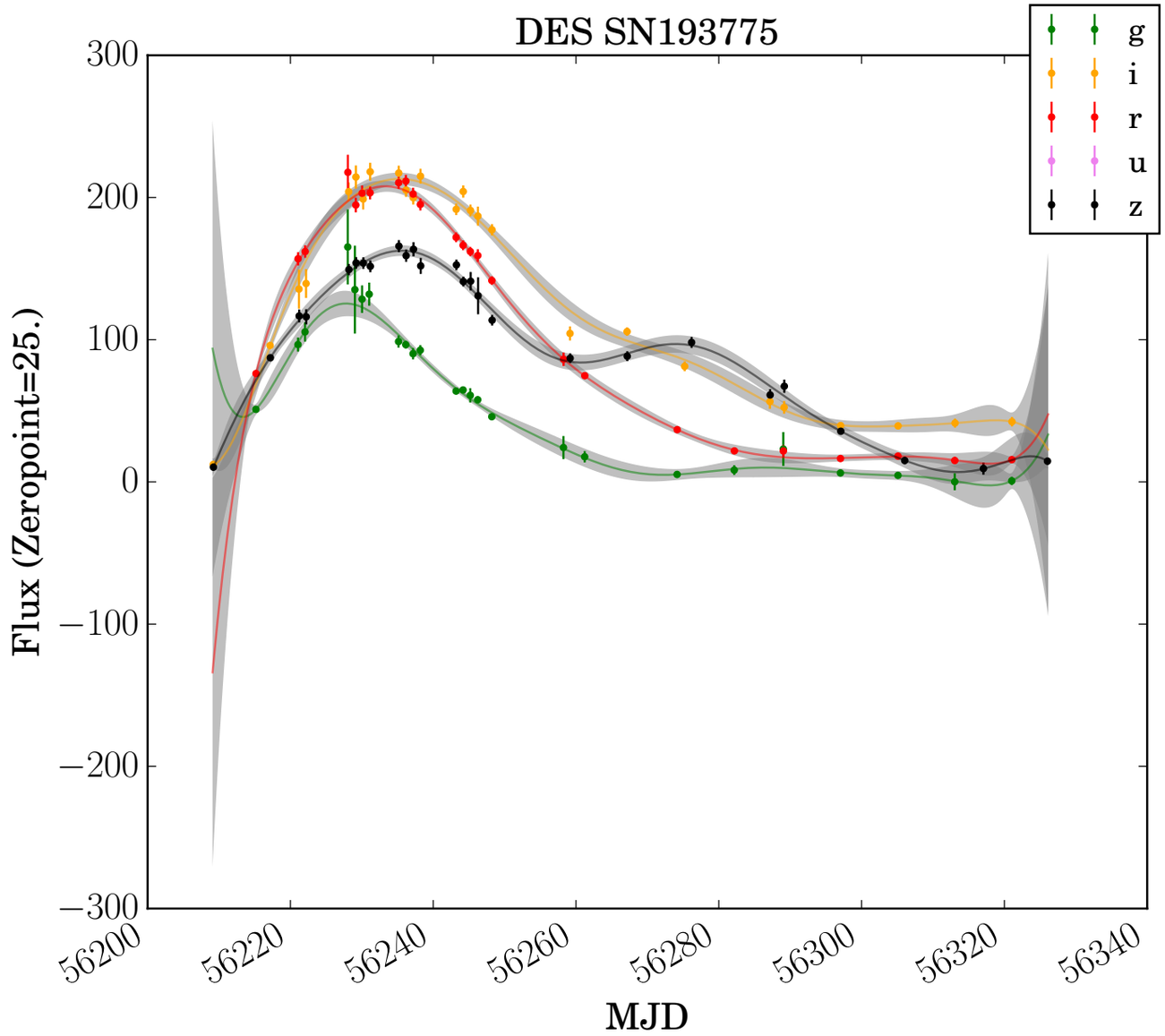
  - With enough data we (re-)calculate the period

- With timescales characterized, we can model the data if we have enough observations

WORKS FOR DIFFERENT CLASSES/TIMESCALES/FLUXSCALES

GAUSSIAN PROCESSES TO MODEL LIGHT CURVES

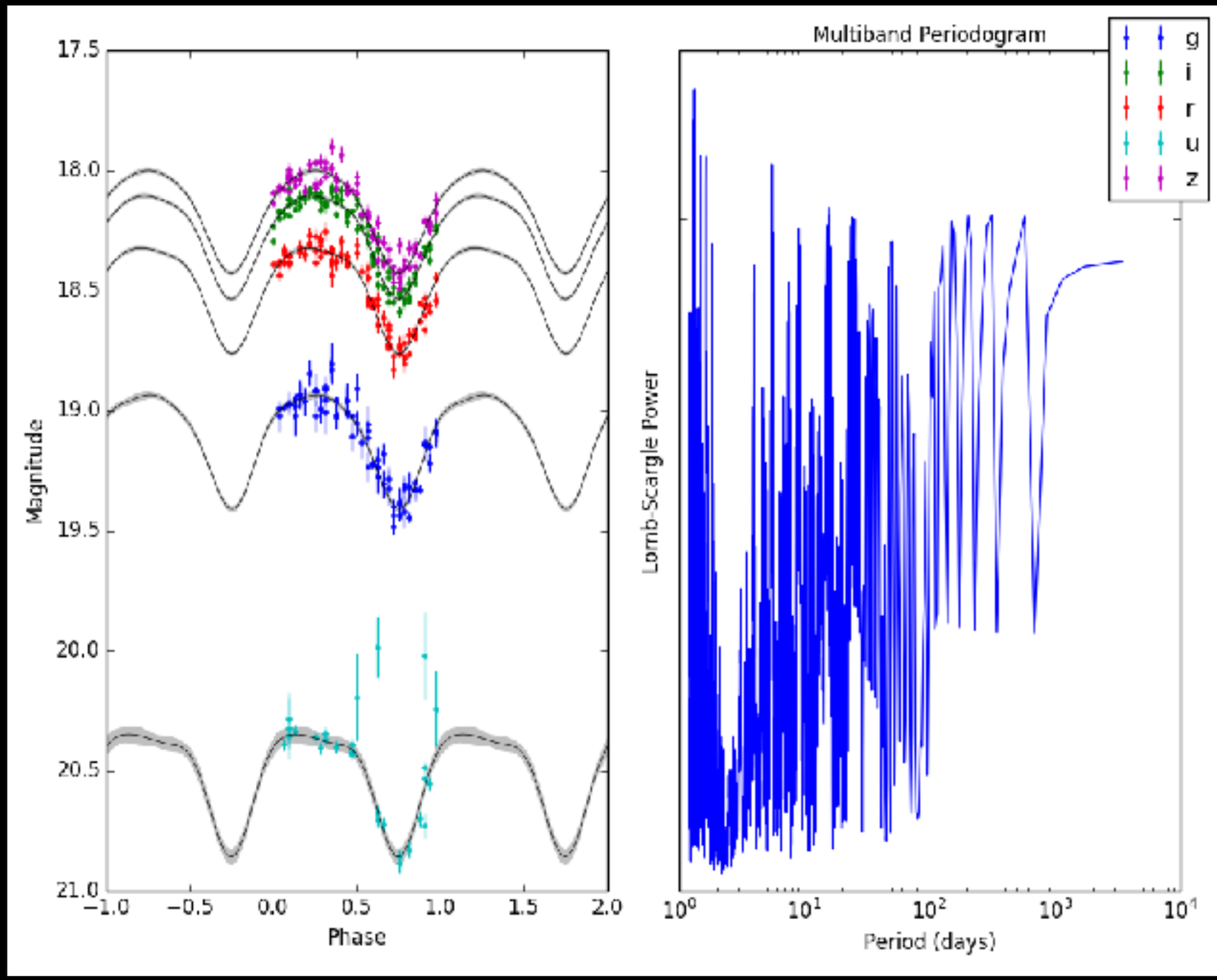


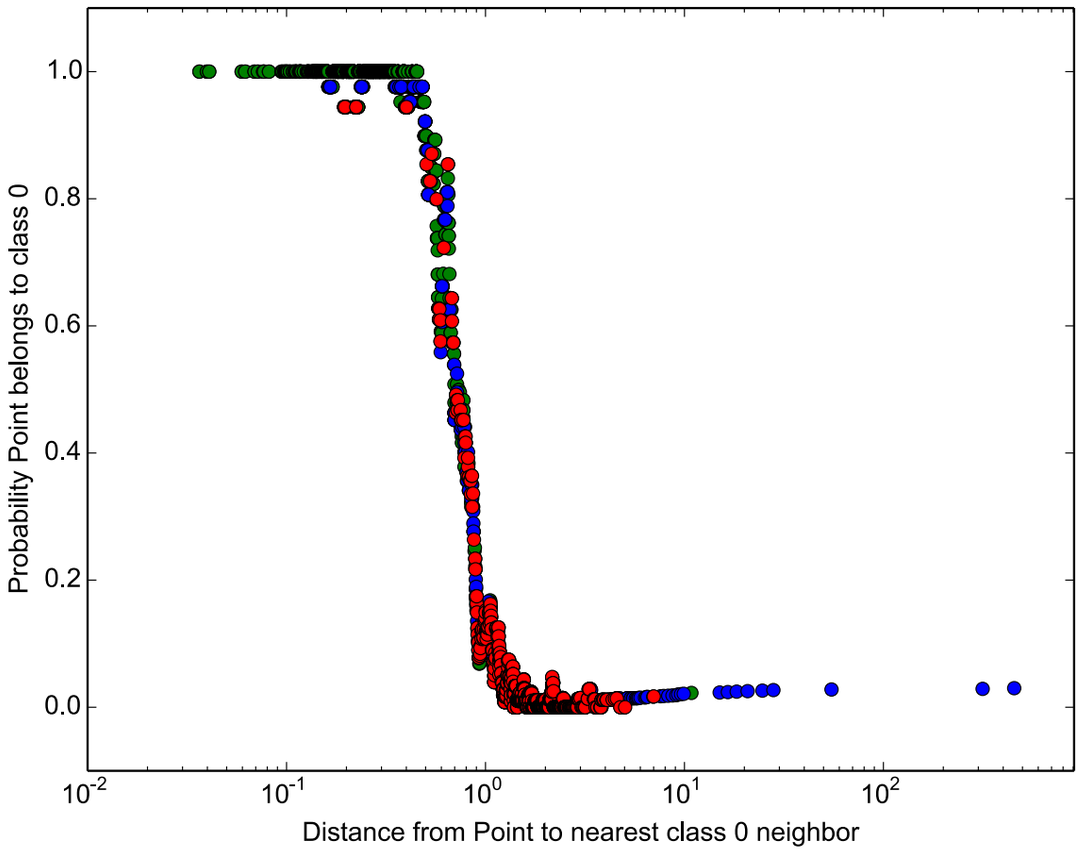


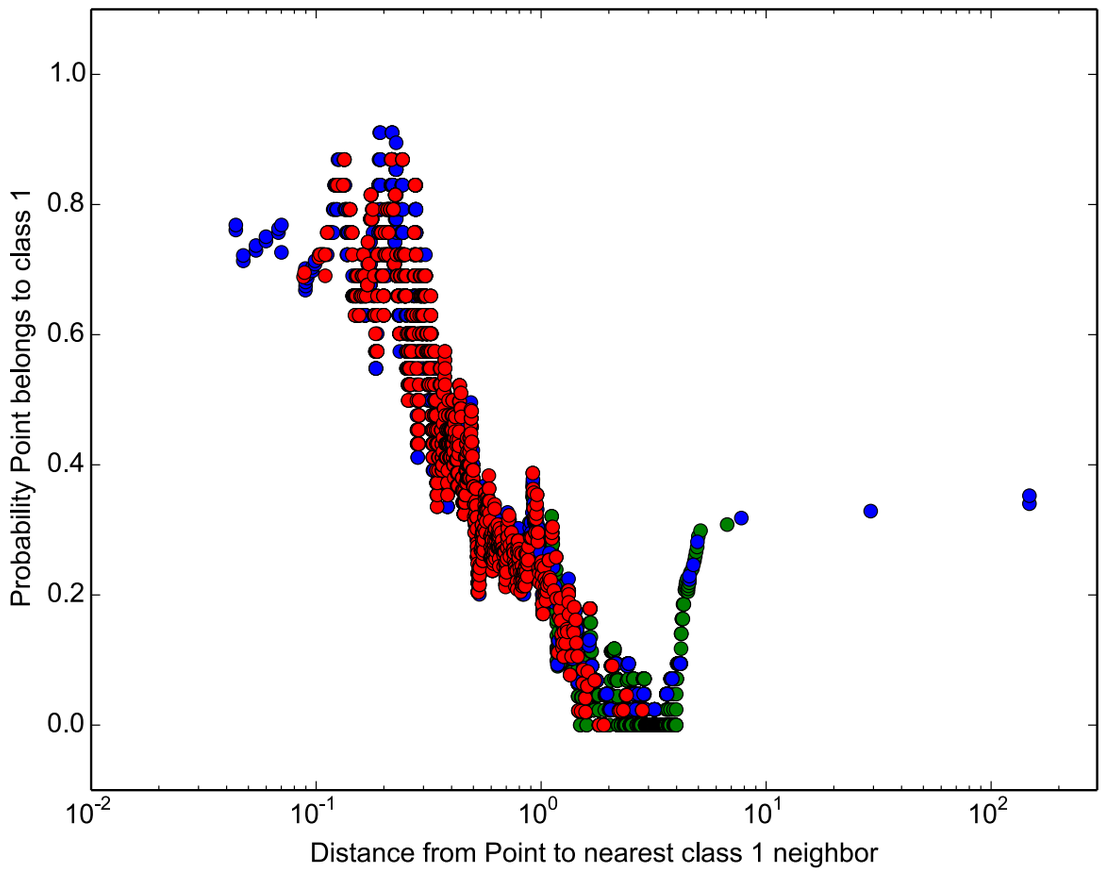
PROJECT TOUCHSTONE GP ONTO DATA IF POORLY SAMPLED, OR MODEL OBSERVATIONS WITH GP, AND COMPARE HYPERPARAMETERS

CAVEATS/FAILURE MODES

- Calculating reliable periods
  - takes lots of data
  - is too slow (alerts will get diverted for taking too long to process)
    - done during daytime processing
- Ideally, we'd train the Gaussian processes on the Touchstone, which would give us priors for the hyperparameters of each kernel
  - If sampling is very different than Touchstone, then we're comparing data with unrepresentative model
  - Still better in practice than "splines gone wild" but need alert simulator to understand failure modes - IN PROGRESS
- Results only going to be as good as data-quality from LSST





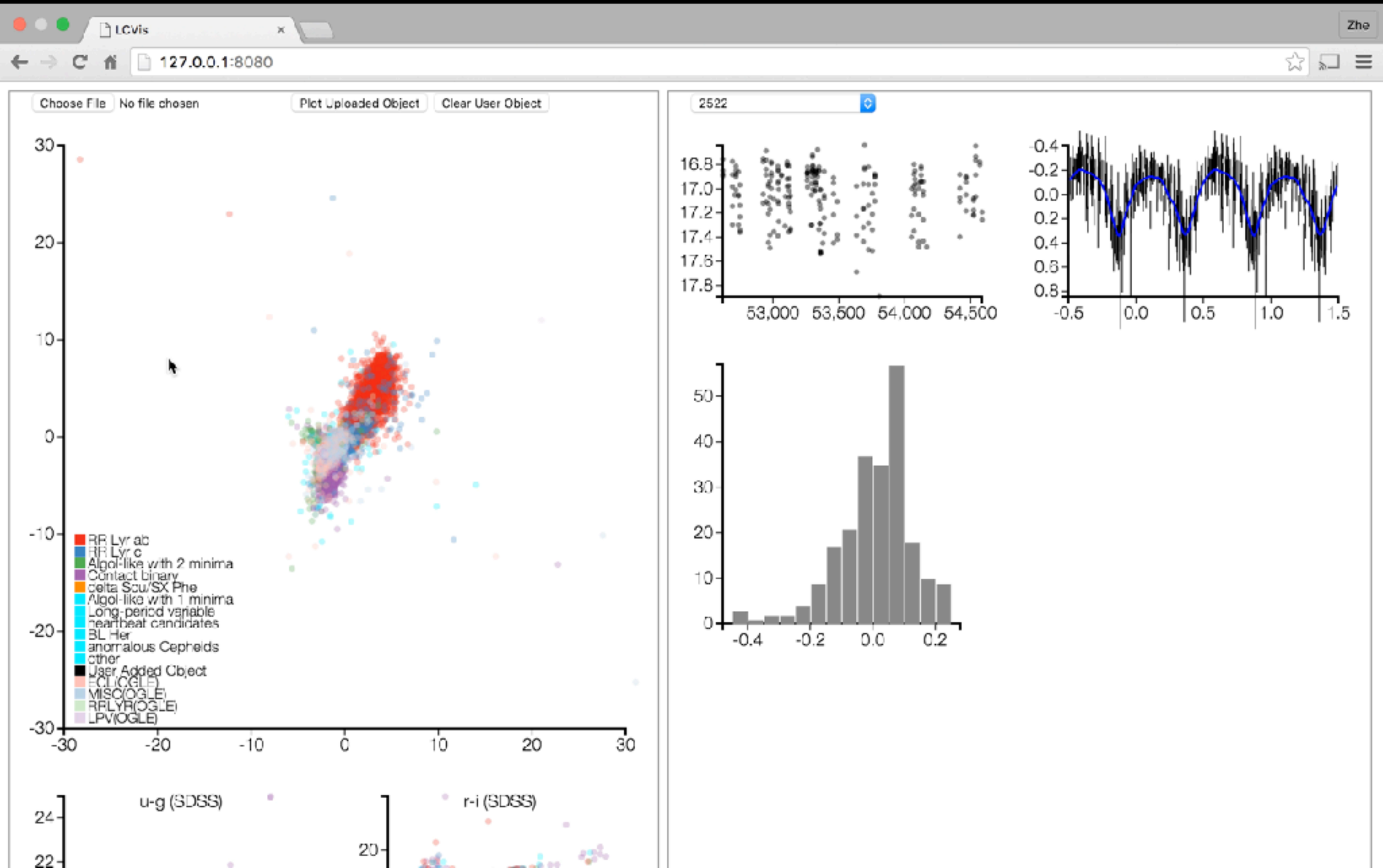


CHARACTERIZATION

- Given high dimensional feature space + contextual info of an alert, what can we say about it?
- Assemble large touchstone with representations of astrophysical events, real & predicted (OGLE + LINEAR variables, all available SNe light curves, etc)
- No single algorithm works best - use ensemble - kNN, random forests, your favorite algorithm…
- kNN is fast, insensitive to population size, can be boosted; sparse PCA to reduce dimensionality, but needs labelled data
- kNN augmented with learned distances gets us to ~90%* accuracy with just one passband, five features

EXAMINING HIGH DIMENSIONAL DATASETS & PLACING ALERTS IN CONTEXT

PCA + VISUALIZATION DEMO



ROADMAP

DEVELOPMENT ROADMAP Q1/Q2 2016

- Astronomy
  - ~~Touchstone database design~~
  - ~~Stellar variability characterization~~
  - ~~Expanded Feature Sets~~
  - Synthetic streams (Current work)
  - Moving object probability (this may not actually be informative)
- Architecture
  - ~~Single-point failure resiliency~~
  - ~~Database design~~
  - ~~Full API implementation/documentation~~
- Algorithms
  - ~~Relocate distance-learning in architecture~~
  - ~~Alternative algorithms~~
- Visualization
  - Speed
  - ~~Operations visualization~~

DEVELOPMENT ROADMAP Q3/Q4 2016

- Astronomy
  - Empirical feature set
  - Light curve anatomy
  - Heterogeneous alerts
  - Community engagement
  - External review (Dec 2016)
- Architecture
  - Database design
  - Scaling
  - Multi-point failure resiliency
- Algorithms
  - Ensemble comparisons
  - Adapt and generalize algorithms
- Visualization
  - Enhanced feature space visualization

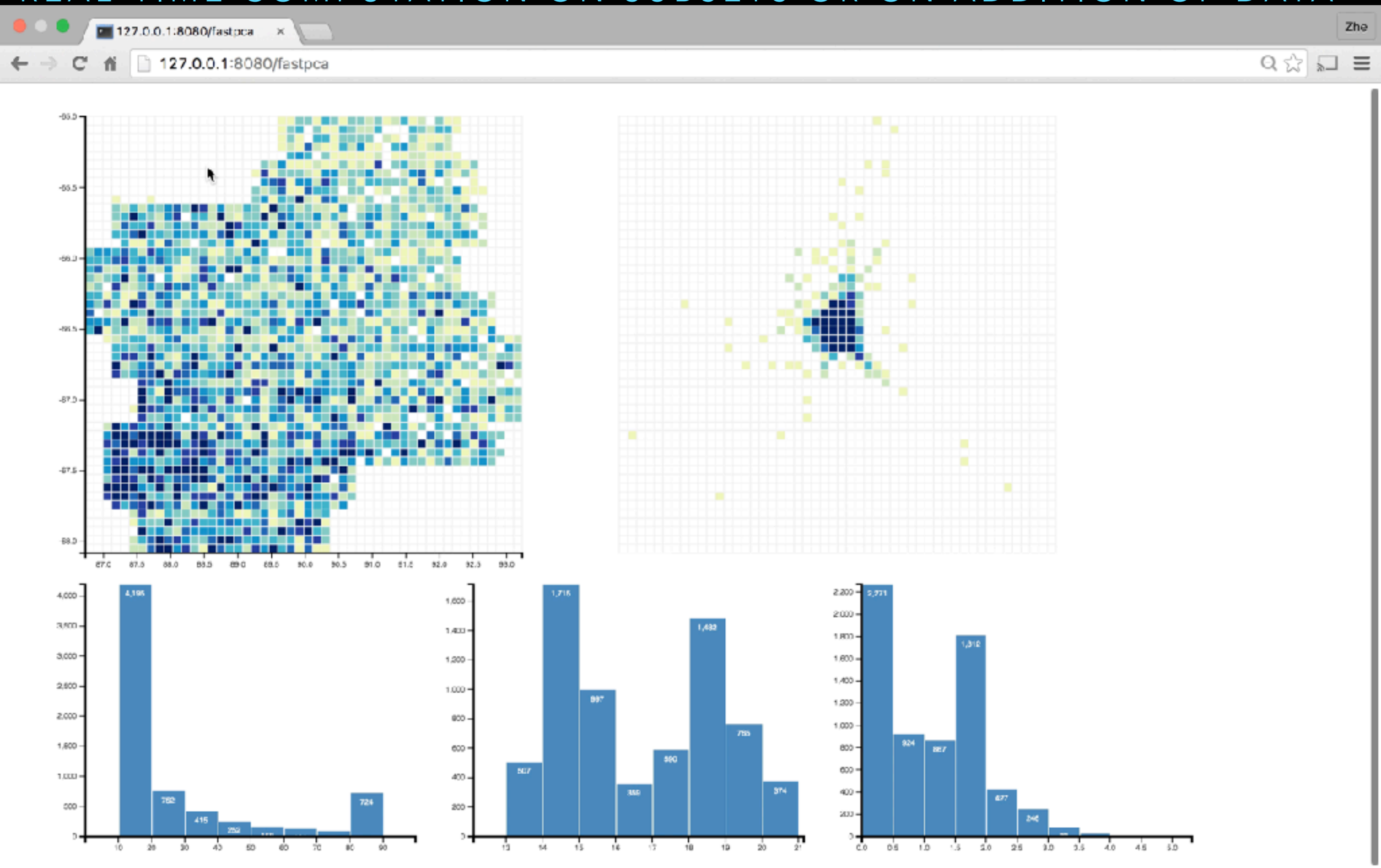
DEVELOPMENT ROADMAP 2017

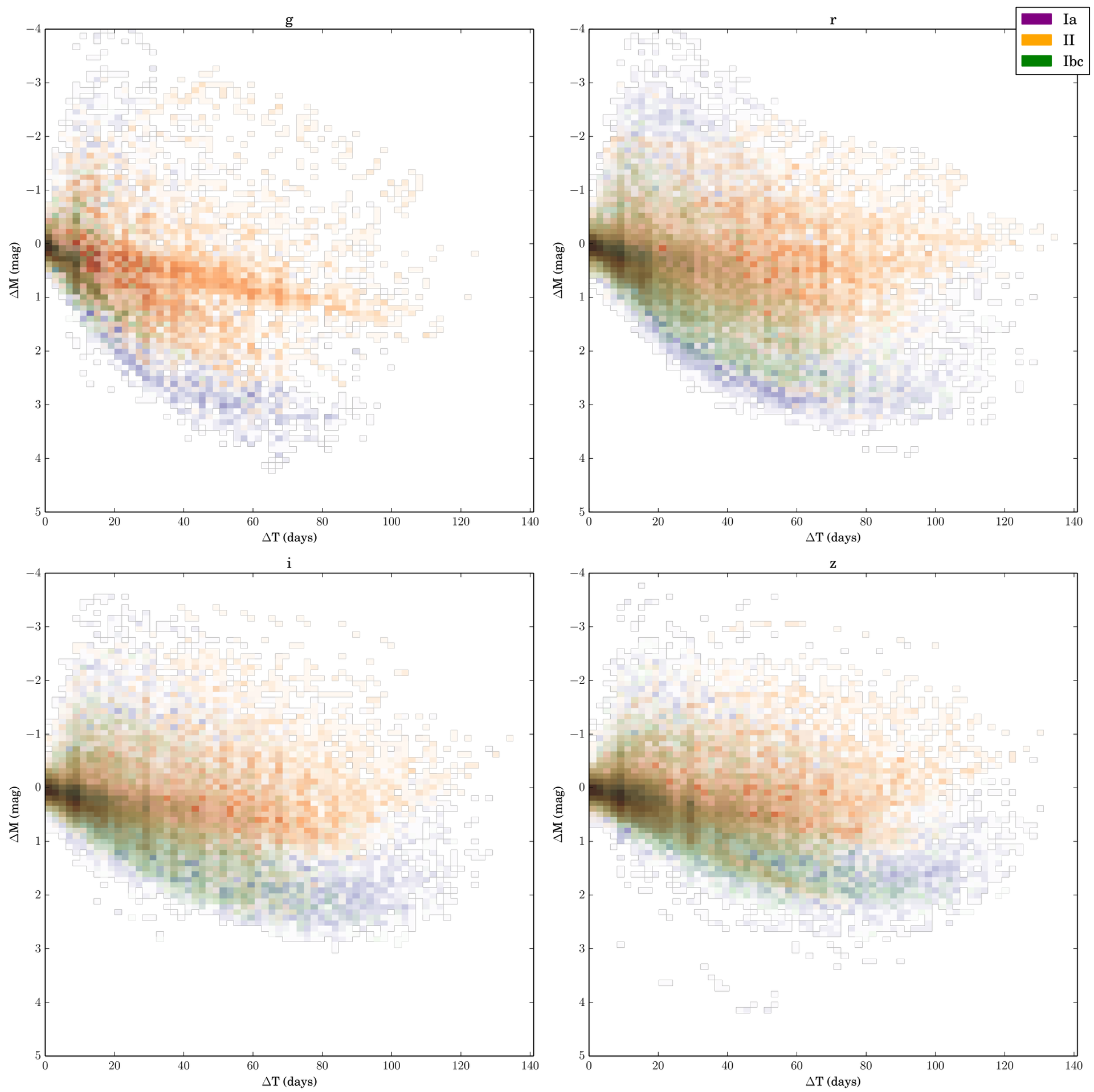
- Astronomy
  - User algorithms
  - Feature sets from theoretical predictions
  - Live alert streams
  - Direct community engagement (workshop)
  - Multiple infrastructure instances (process by category)
- Architecture
  - Scaling/speed
- Algorithms
  - Ensemble algorithms, scaling/speed
- Visualization
  - Operations assessment tool

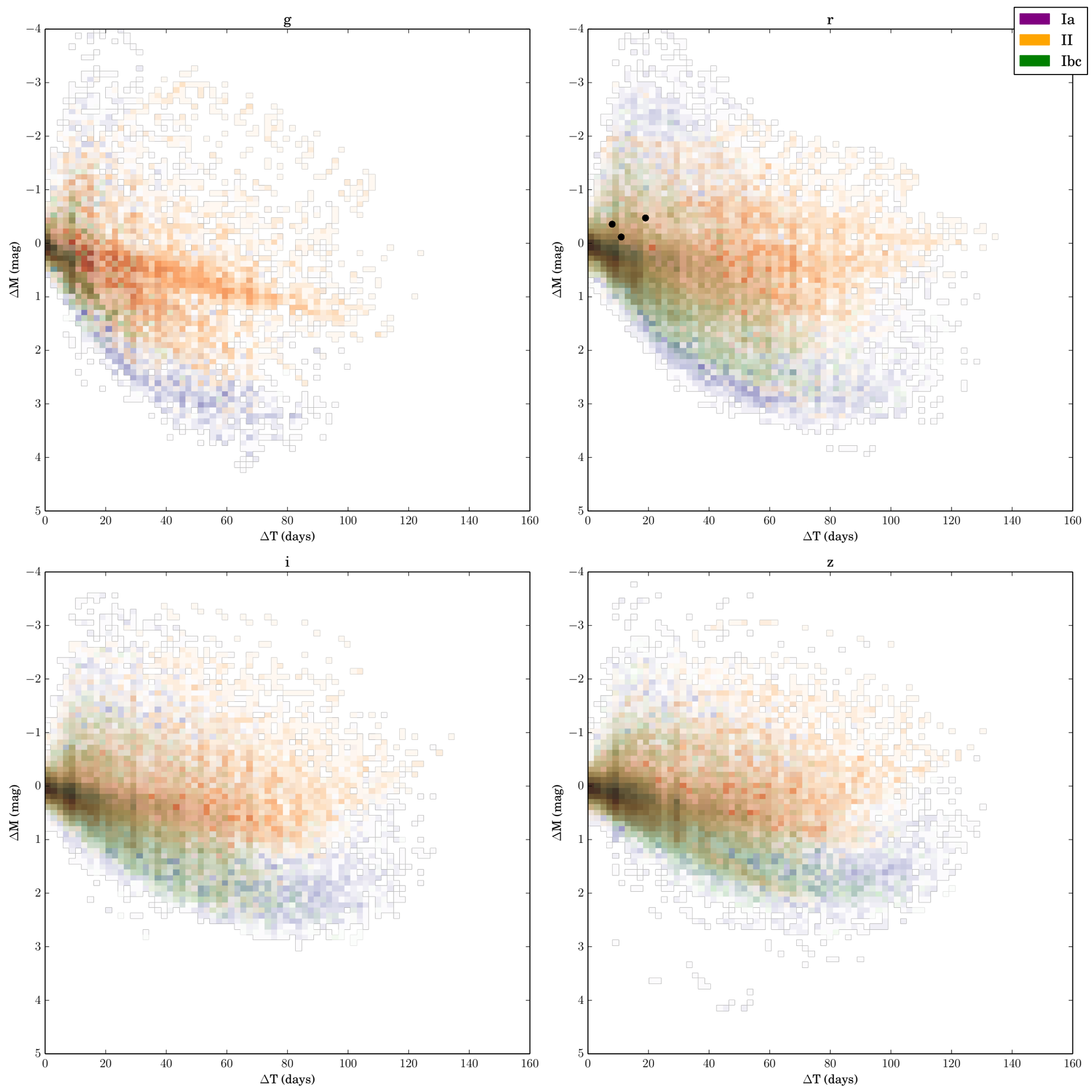
EXTRA THINGS

INTERARCIVE PCA DEMO

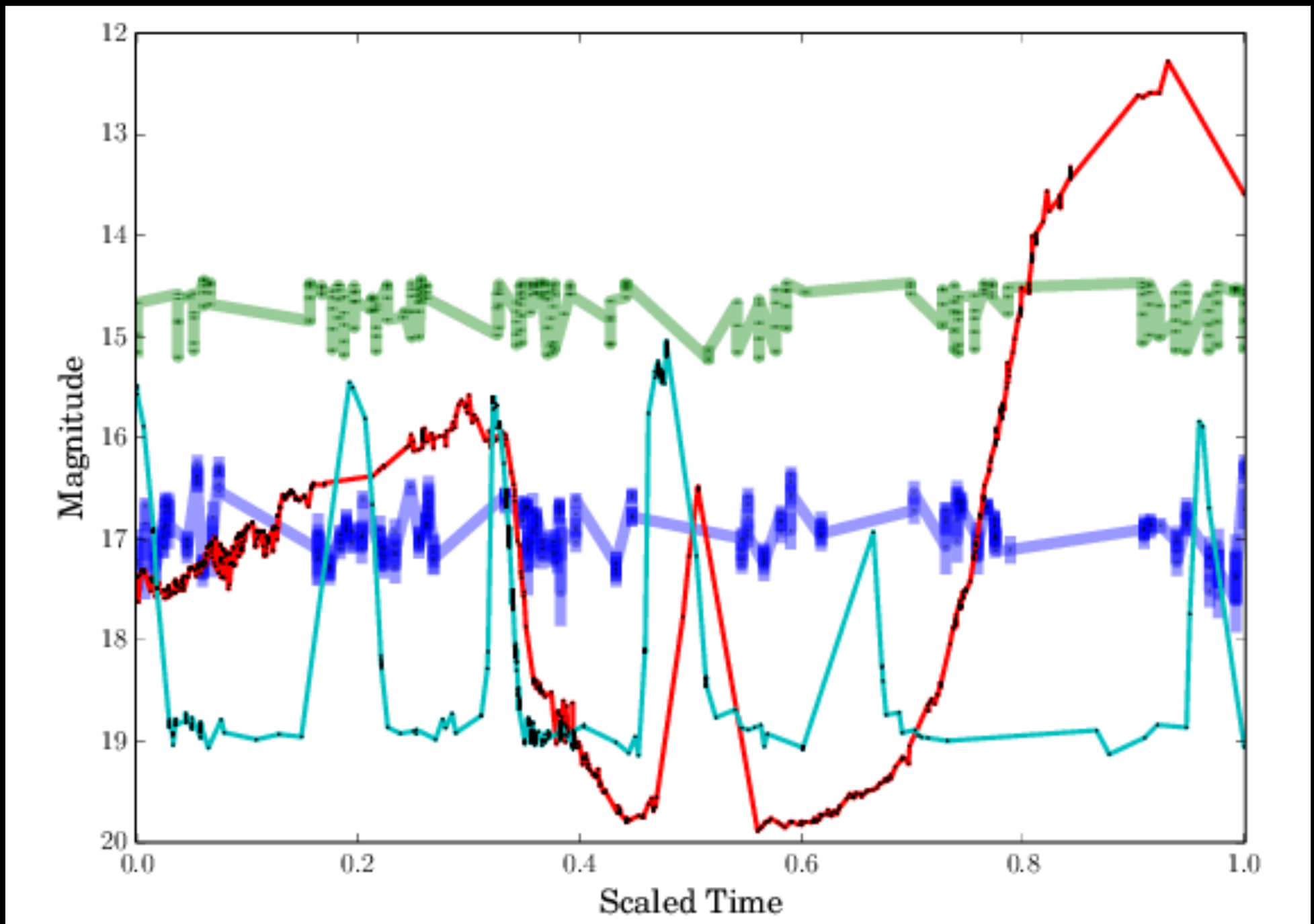
REAL-TIME COMPUTATION ON SUBSETS OR ON ADDITION OF DATA




g
r
i
z
Ia
II
Ibc
ΔM (mag)
ΔT (days)
−4
−3
−2
−1
0
1
2
3
4
5
0
20
40
60
80
100
120
140


g
r
i
z
Ia
II
Ibc
ΔM (mag)
ΔT (days)
−4
−3
−2
−1
0
1
2
3
4
5
0
20
40
60
80
100
120
140
160

TESTING WITH REAL ASTROPHYSICAL DATA

AN ILLUSTRATIVE EXAMPLE

LIGHTCURVES OF "RARE" OBJECTS USED IN PCA VIDEO
DERIVED FROM TT ARI, R CRB