COPING WITH THE DATA STREAM

GAUTHAM NARAYAN
NOAO

GSNARAYAN

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)

OR PUTTING ALL OF THE THINGS TOGETHER

GAUTHAM NARAYAN
NOAO

GSNARAYAN

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)

THINGS I WORK ON:

- Understanding dark energy, particularly SNIa cosmology
- Bayesian inference with generative models (DA WDs, SNIa)
- Machine learning to identify and characterize transients, Monte Carlo simulations
- Supernova impostors, misbehaving supernovae
- Image reduction pipelines & calibration for wide-field surveys (ESSENCE + Pan-STARRS is ~ 900TB of data)
- Feel free to come by NOAO #118, or send me an email

OVERVIEW

OVERVIEW

- In DSFP, you've covered:

OVERVIEW

- In DSFP, you've covered:
  - <u>time-series analysis</u>

OVERVIEW

- In DSFP, you've covered:
    - time-series analysis
    - machine learning

OVERVIEW

- In DSFP, you've covered:
  - time-series analysis
  - machine learning
  - object-oriented programming

OVERVIEW

- In DSFP, you've covered:
  - time-series analysis
  - machine learning
  - object-oriented programming
  - databases

OVERVIEW

- In DSFP, you've covered:
  - time-series analysis
  - machine learning
  - object-oriented programming
  - databases
  - profiling code

OVERVIEW

- In DSFP, you've covered:
  - time-series analysis
  - machine learning
  - object-oriented programming
  - databases
  - profiling code
  - visualization…

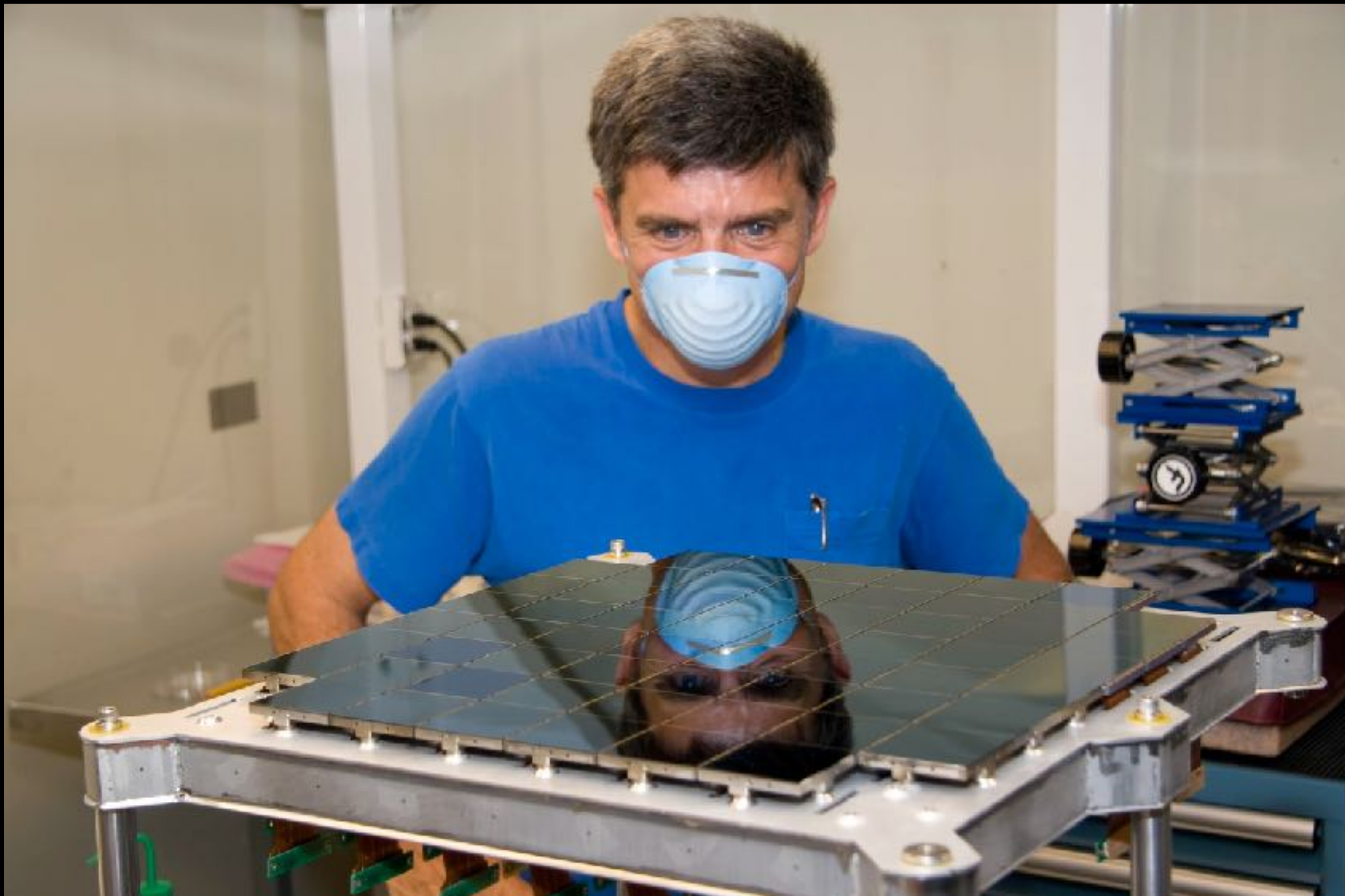
OVERVIEW

- In DSFP, you've covered:
  - time-series analysis
  - machine learning
  - object-oriented programming
  - databases
  - profiling code
  - visualization…
- Combine these elements (+1) in the context of a single LSST project

LSST

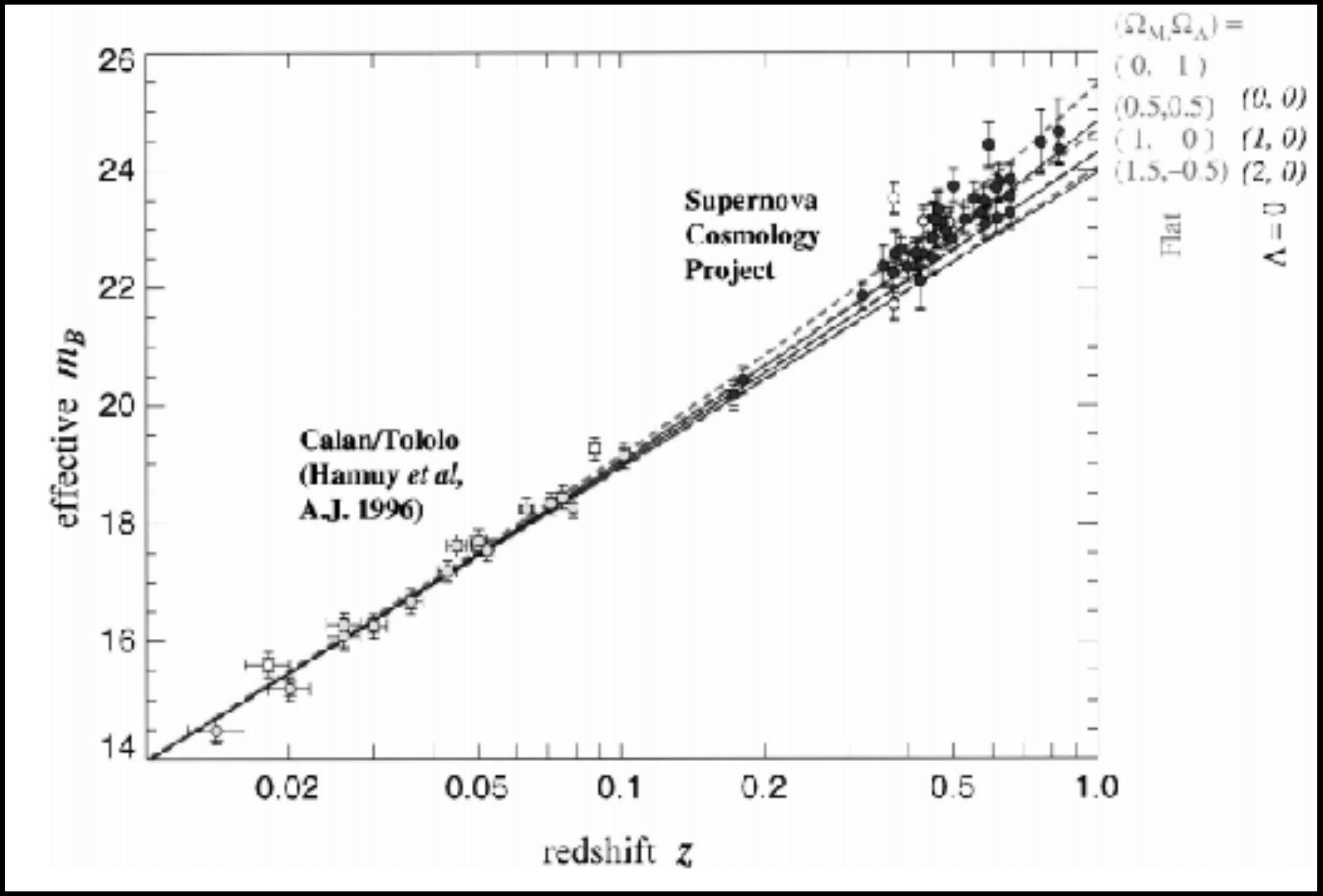
This picture is both exciting and terrifying

J. B. Oke holding a 0.24 megapixel CCD detector circa 1990

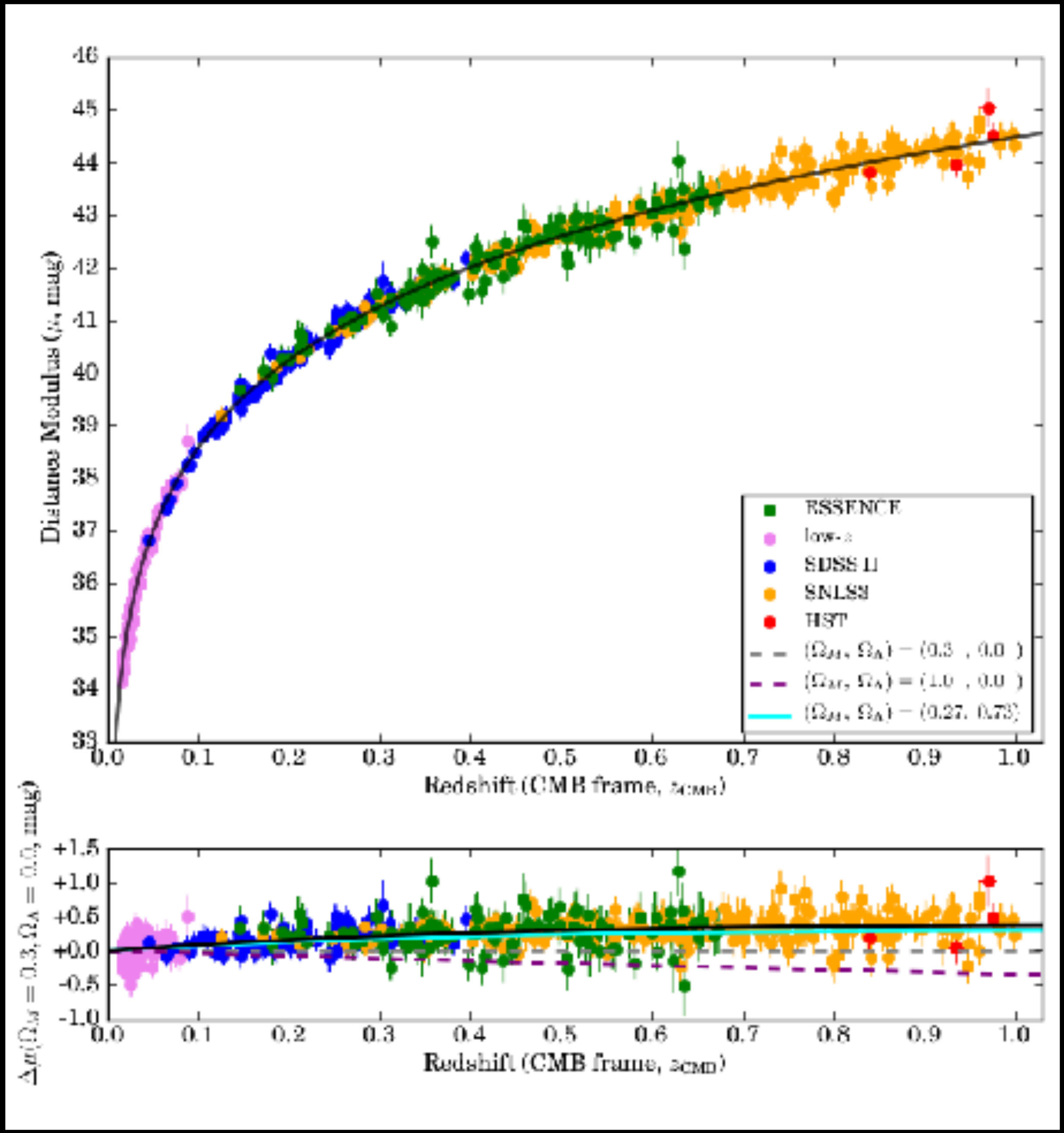


John Tonry looking at the 1.4 Gigapixel GPC1 circa 2010

DATA RATES HAVE RAPIDLY INCREASED



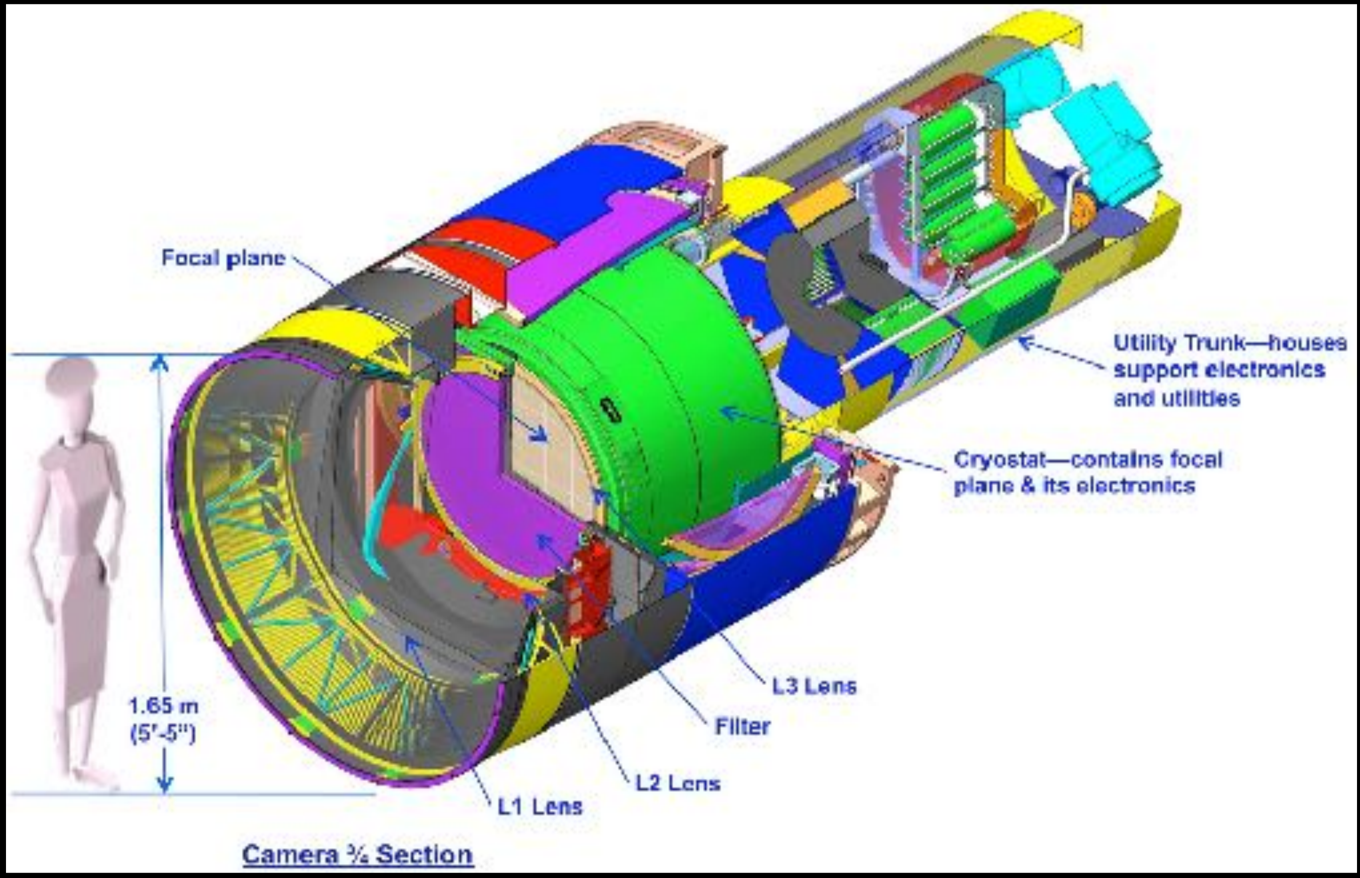
Perlmutter+ 1999



Narayan+ 2016

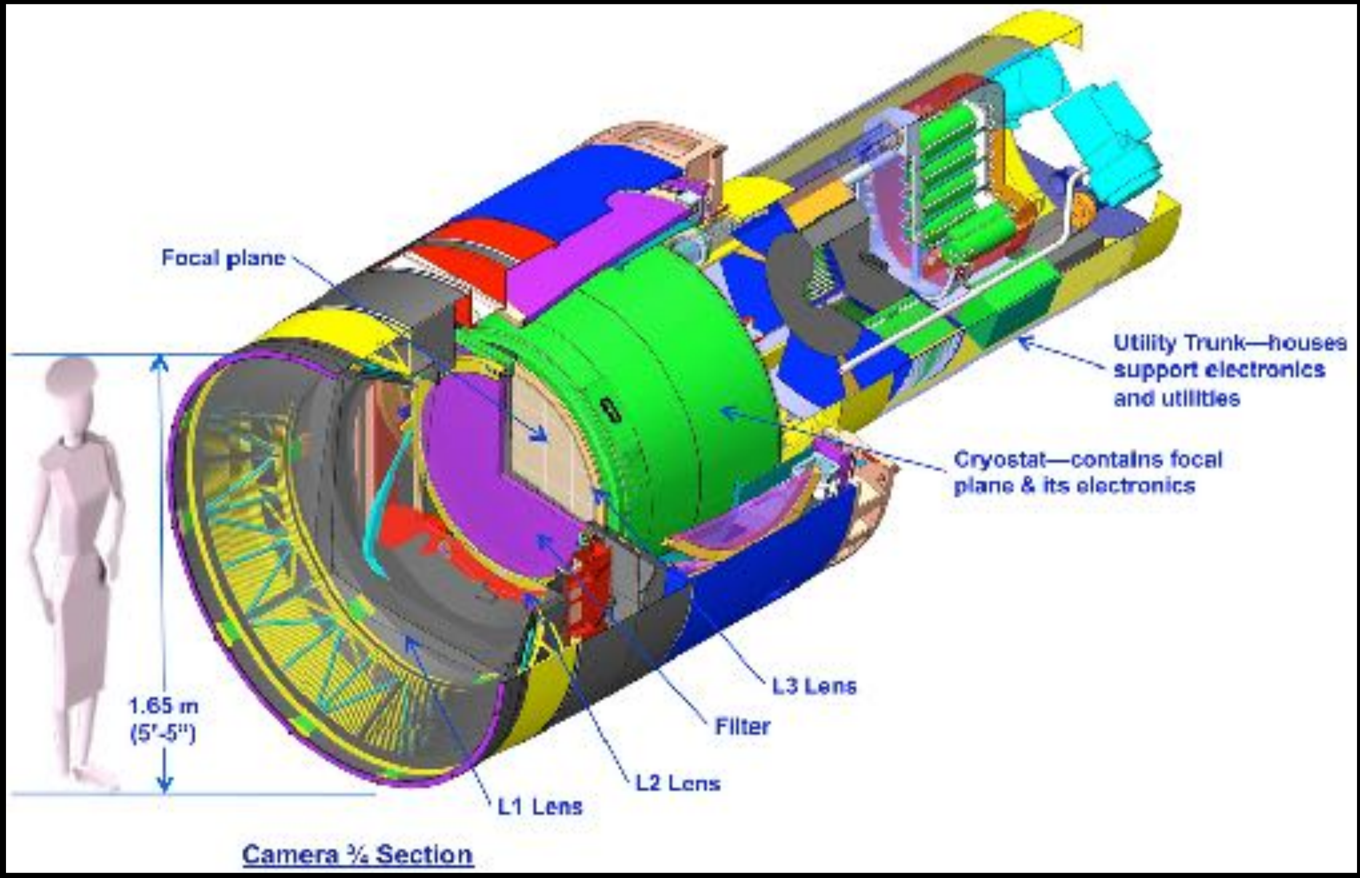
The evolution of the Hubble diagram over ~15 years

MOTIVATION 1: THE 'L' IN LSST



MOTIVATION 1: THE 'L' IN LSST

- Transient searches have relied on eyeballs for alerts



MOTIVATION 1: THE 'L' IN LSST

- Transient searches have relied on eyeballs for alerts
- LSST will produce several PBs of images/yr i.e. TB of catalogs, GBs every night: PETAFLOOD

MOTIVATION 1: THE 'L' IN LSST

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

POP QUIZ:

WHAT DOES SYNOPTIC MEAN?

POP QUIZ:

WHAT DOES SYNOPTIC MEAN?

"observations that give a broad view of a subject"

POP QUIZ:

WHAT DOES SYNOPTIC MEAN?

"observations that give a broad view of a subject"

i.e. many different scientists and science goals that are not always mutually compatible

OUR VIEW OF THE COMMUNITY'S FOLLOW UP INTERESTS

i.e. what needs we are trying to serve

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/rare things that need rapid follow up
  - Known things that need rapid characterization & detailed follow up
  - Known unknowns - predicted but never seen

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/rare things that need rapid follow up
  - Known things that need rapid characterization & detailed follow up
  - Known unknowns - predicted but never seen

- Large pure samples of photometrically selected things
  - Things that can be studied at "leisure"

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/rare things that need rapid follow up
  - Known things that need rapid characterization & detailed follow up
  - Known unknowns - predicted but never seen
- Large pure samples of photometrically selected things
  - Things that can be studied at "leisure"
- "Normal" things being wonky

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/rare things that need rapid follow up: **Strongly lensed SNe**
  - Known things that need rapid characterization & detailed follow up: **Microlensing**
  - Known unknowns - predicted but never seen: **GW counterparts**

- Large pure samples of photometrically selected things: **SNe**
  - Things that can be studied at "leisure": **Variable Stars**

- "Normal" things being wonky: **KIC 8462852 aka Tabby's Star**

OUR VIEW OF THE COMMUNITY'S FOLLOW UP INTERESTS

i.e. what needs we are trying to serve

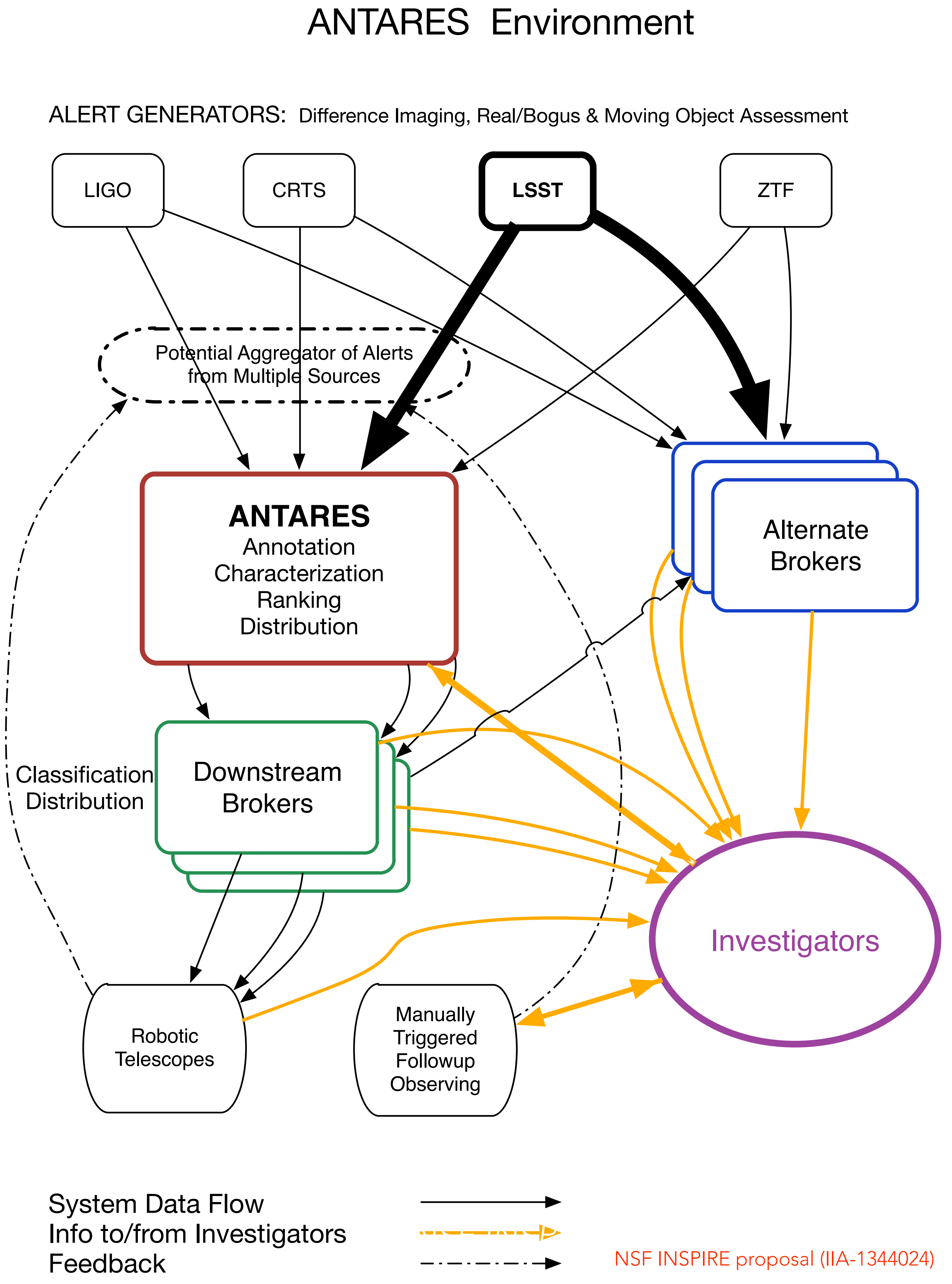
- Things that need immediate follow up
  - Unknown/rare things that need rapid follow up: **Strongly lensed SNe**
  - Known things that need rapid characterization & detailed follow up: **Microlensing**
  - Known unknowns - predicted but never seen: **GW counterparts**

- Large pure samples of photometrically selected things: **SNe**
  - Things that can be studied at "leisure": **Variable Stars**

- "Normal" things being wonky: **KIC 8462852 aka Tabby's Star**

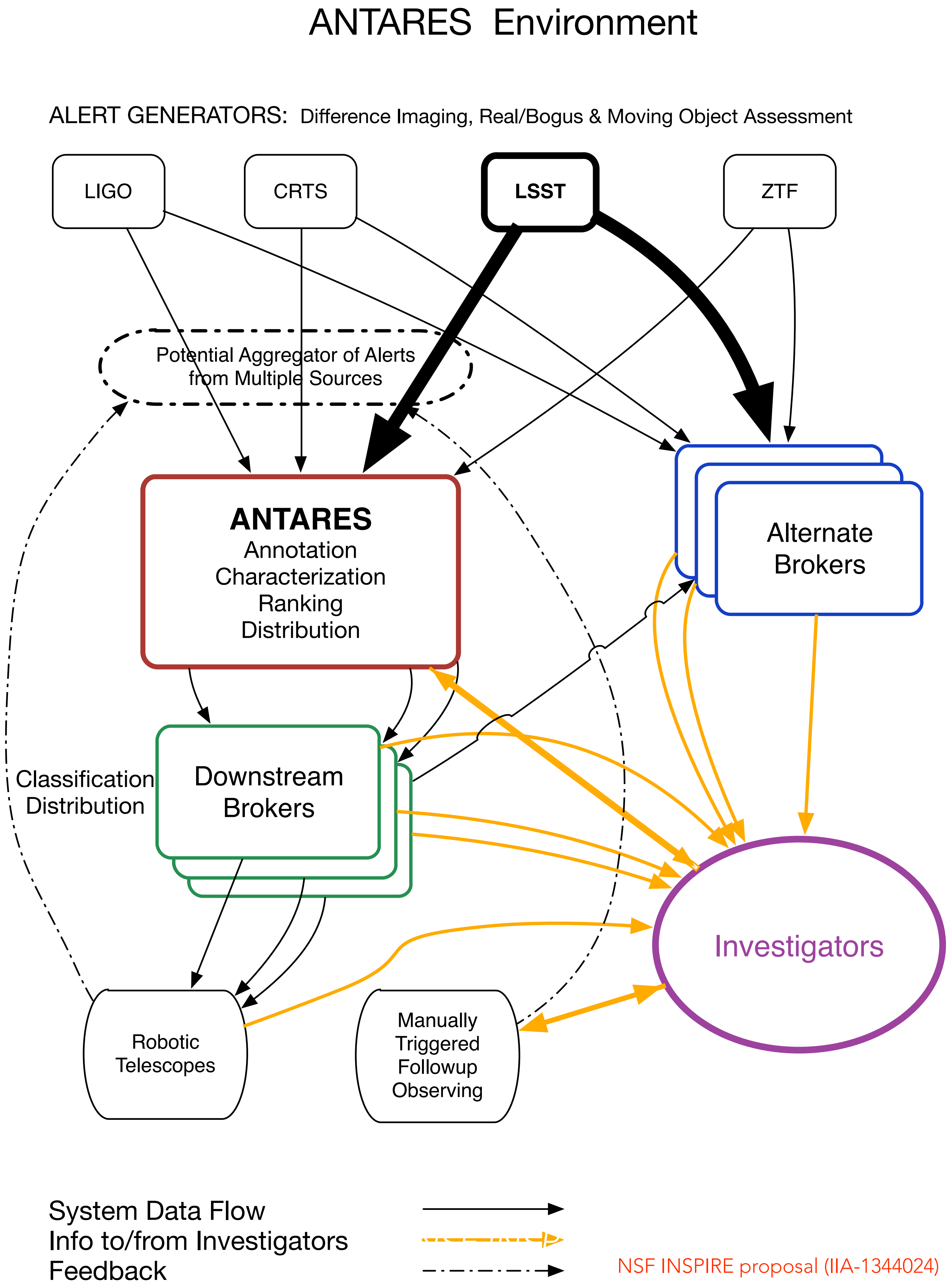
There is no single feature set & ML algorithm that is optimal for all these use cases.

ANTARES


ANTARES Environment
ALERT GENERATORS: Difference Imaging, Real/Bogus & Moving Object Assessment
LIGO
CRTS
LSST
ZTF
Potential Aggregator of Alerts from Multiple Sources
ANTARES
Annotation
Characterization
Ranking
Distribution
Alternate Brokers
Classification Distribution
Downstream Brokers
Investigators
Robotic Telescopes
Manually Triggered Followup Observing
System Data Flow
Info to/from Investigators
Feedback
NSF INSPIRE proposal (IIA-1344024)

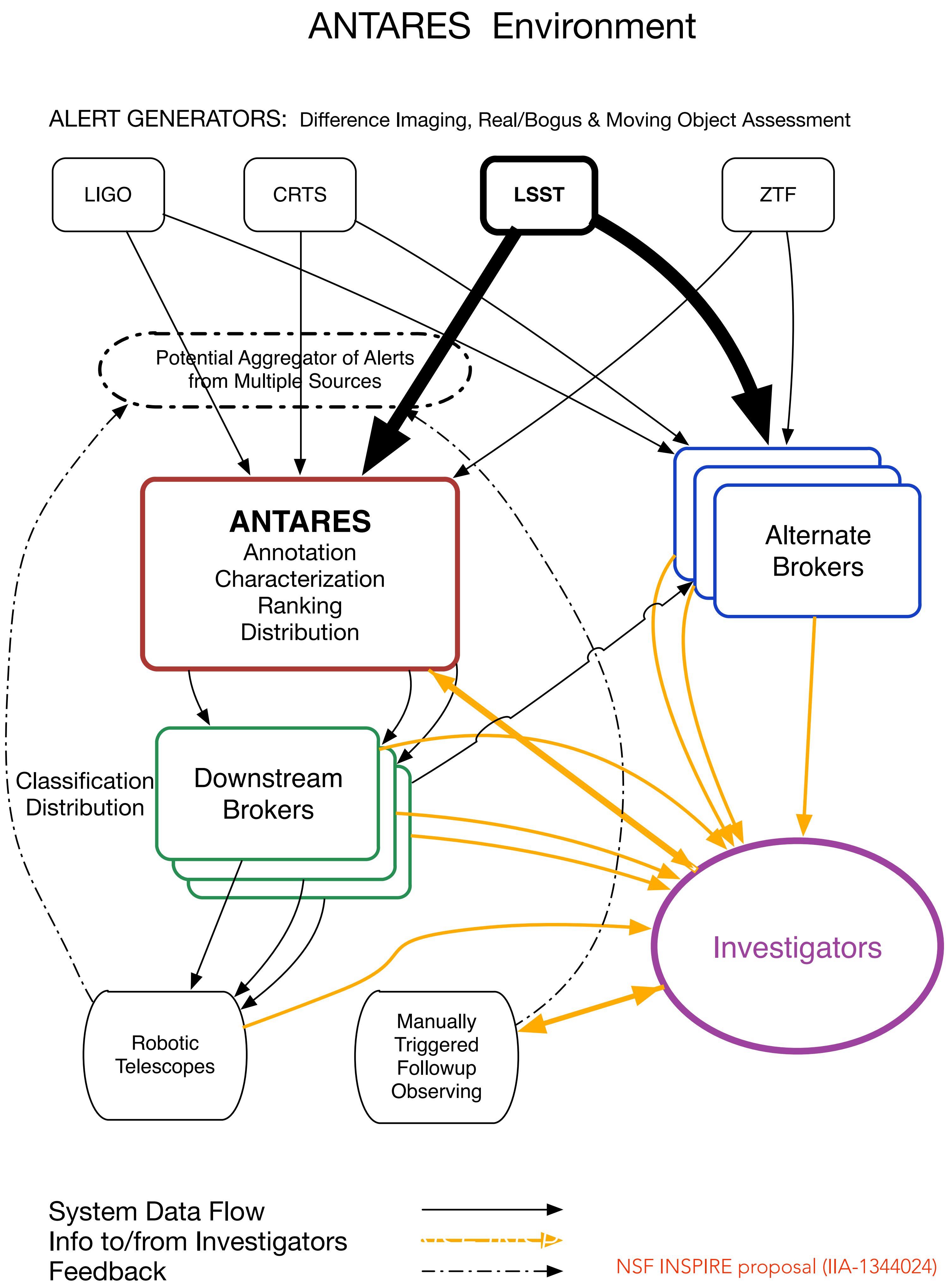
ANTARES

- Arizona-NOAO Temporal Analysis and Response to Events System



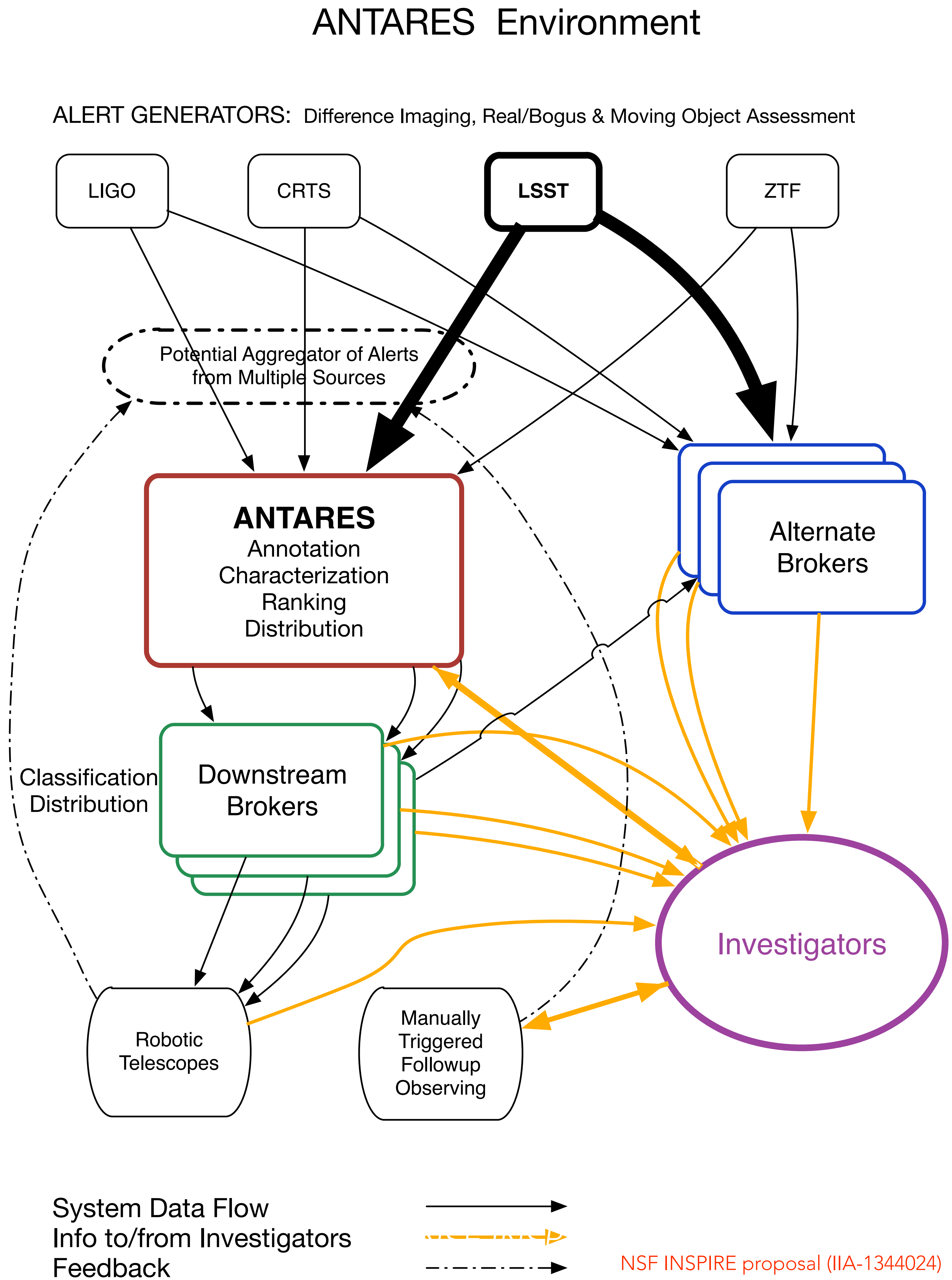
ANTARES

- Arizona-NOAO Temporal Analysis and Response to Events System
- Multi-disciplinary collaboration between astronomy, stats and computer science



ANTARES

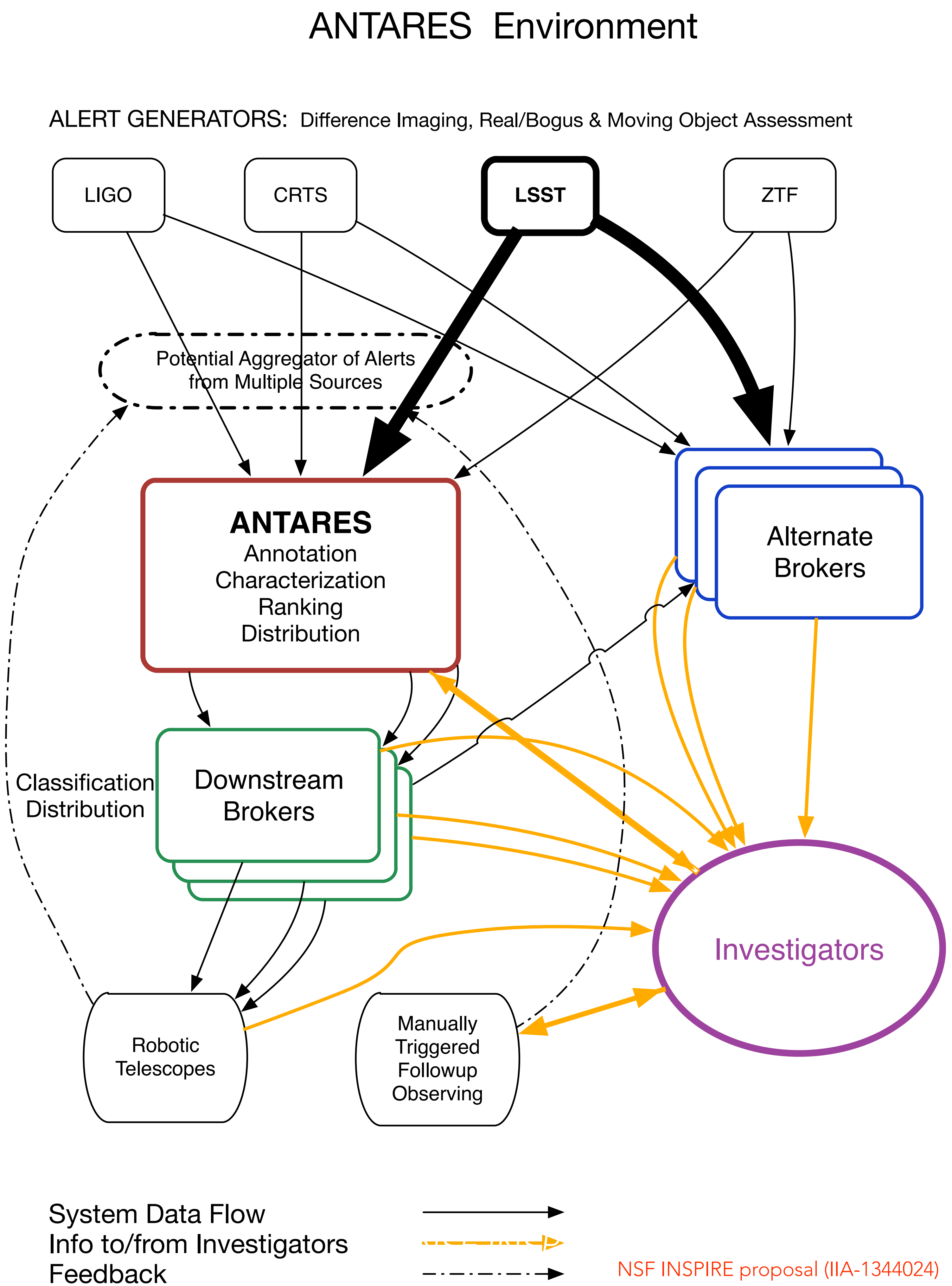
- Arizona-NOAO Temporal Analysis and Response to Events System
- Multi-disciplinary collaboration between astronomy, stats and computer science
- System to characterize alerts automatically - not an LSST data product!



NSF INSPIRE proposal (IIA-1344024)

ANTARES

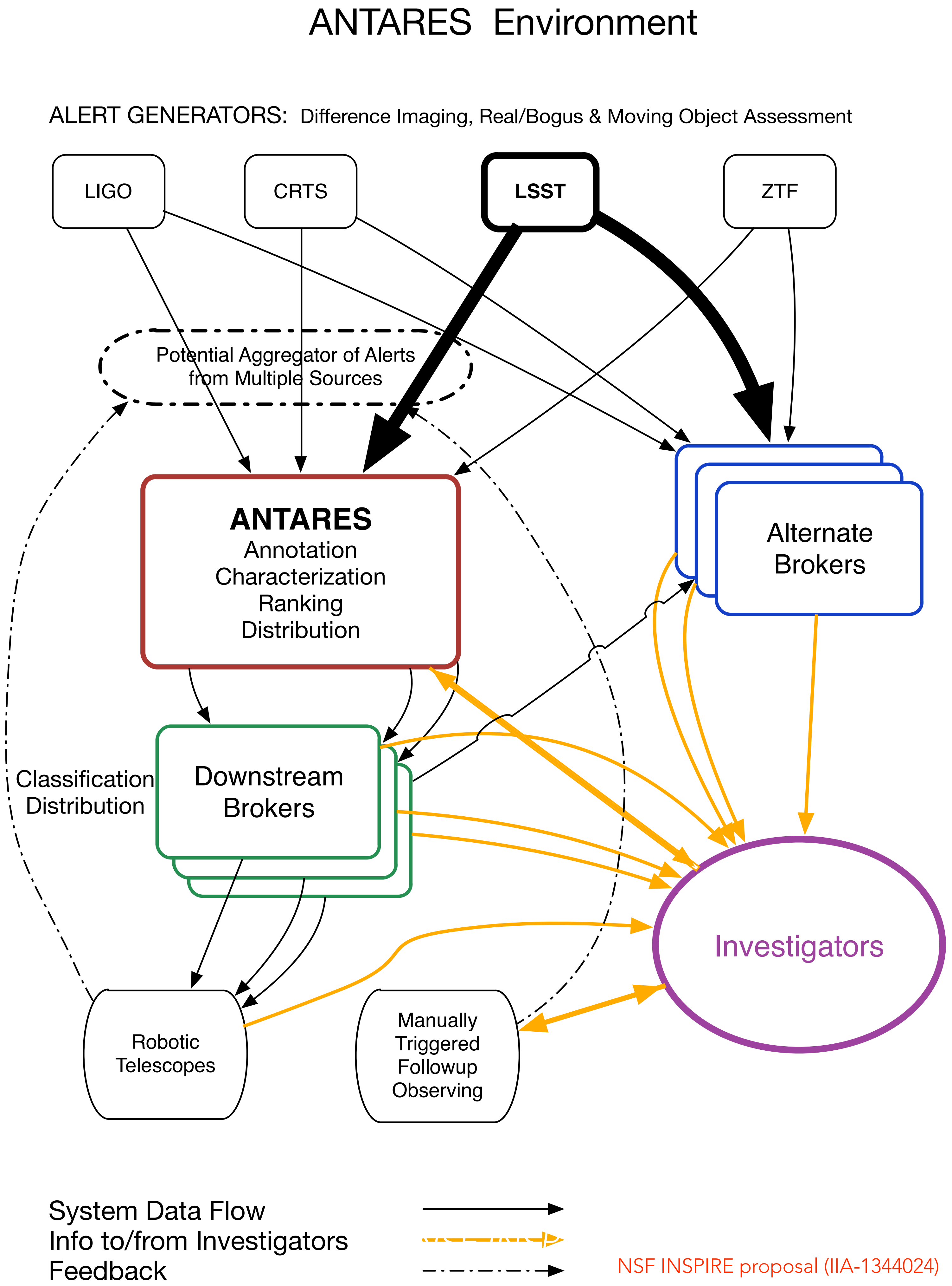
- Arizona-NOAO Temporal Analysis and Response to Events System
- Multi-disciplinary collaboration between astronomy, stats and computer science
- System to characterize alerts automatically - not an LSST data product!
- Will winnow down the number of LSST alerts and let you pick the most interesting



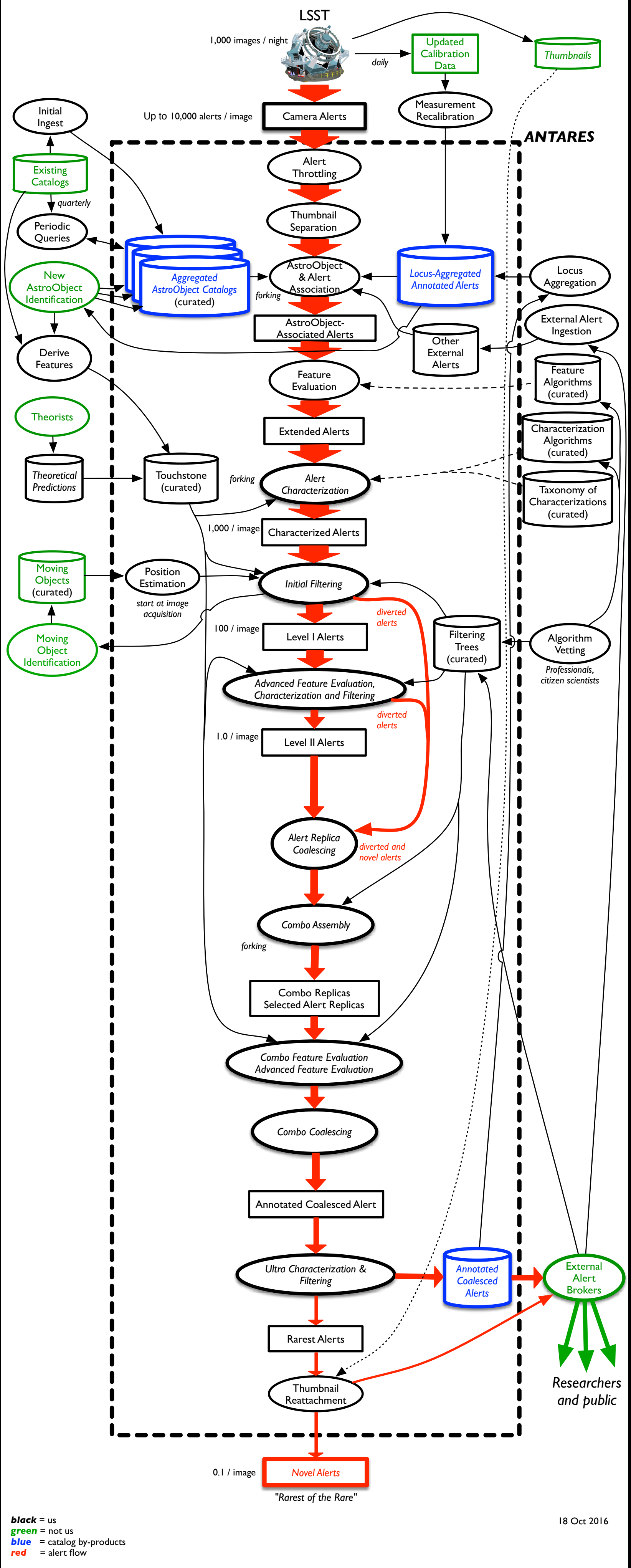
NSF INSPIRE proposal (IIA-1344024)

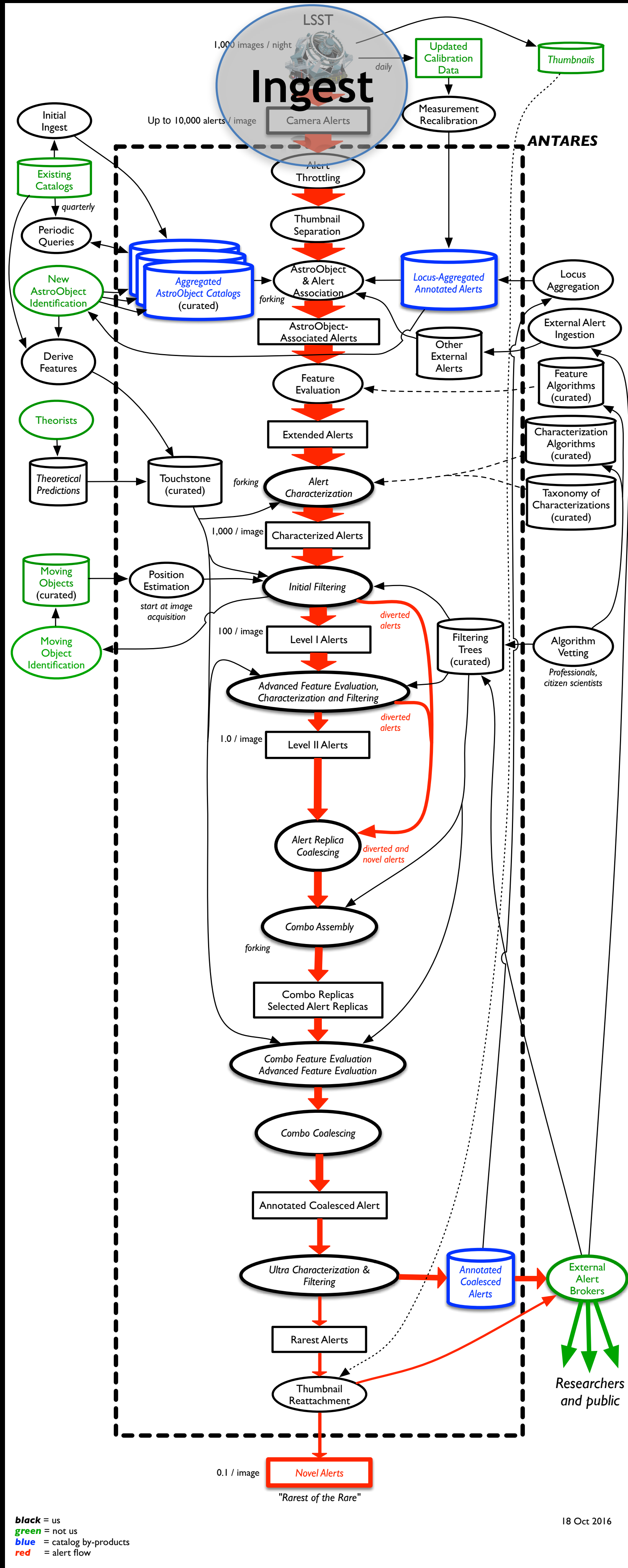
ANTARES

- Arizona-NOAO Temporal Analysis and Response to Events System
- Multi-disciplinary collaboration between astronomy, stats and computer science
- 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



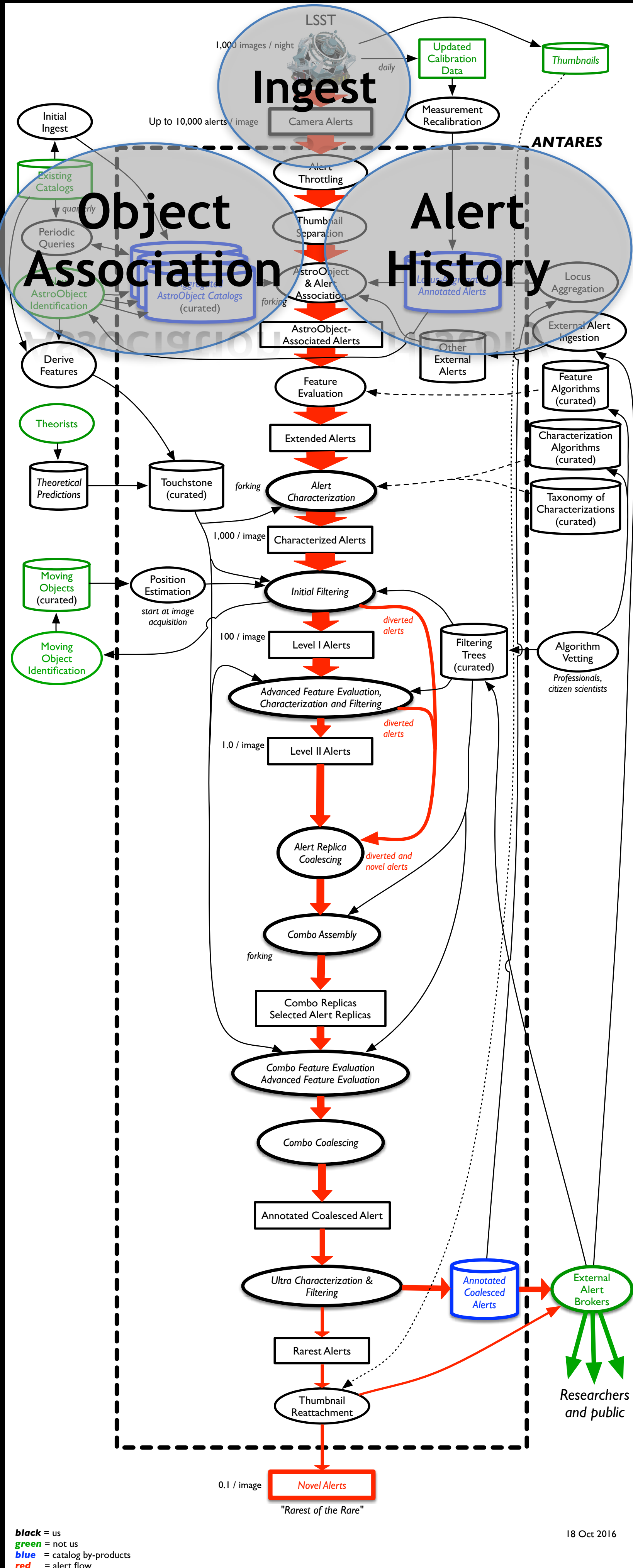
ANTARES Architecture





ANTARES Architecture

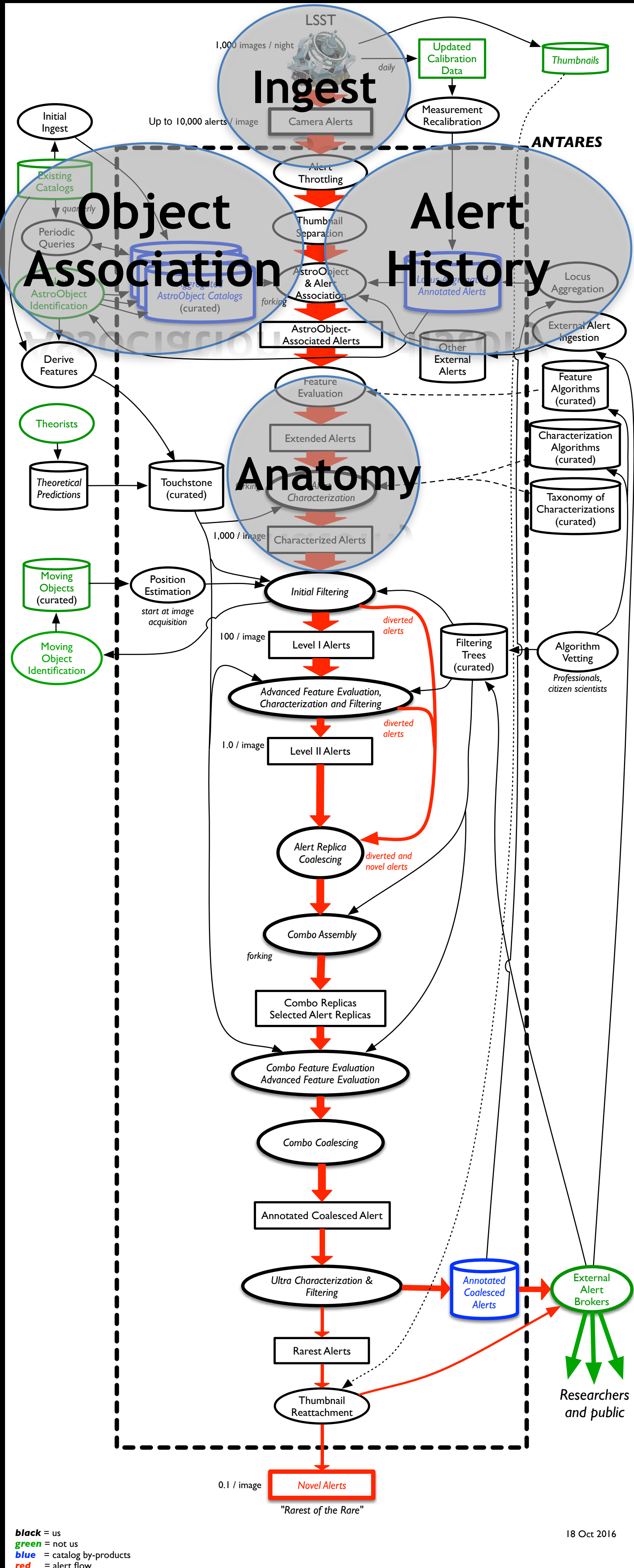
Alerts are generated from real light curves, outside the system, and then ingested into: **streaming data**



ANTARES Architecture

Alerts are generated from real light curves, outside the system, and then ingested into: **streaming data**

Alerts get annotated with external information and past history: **databases**

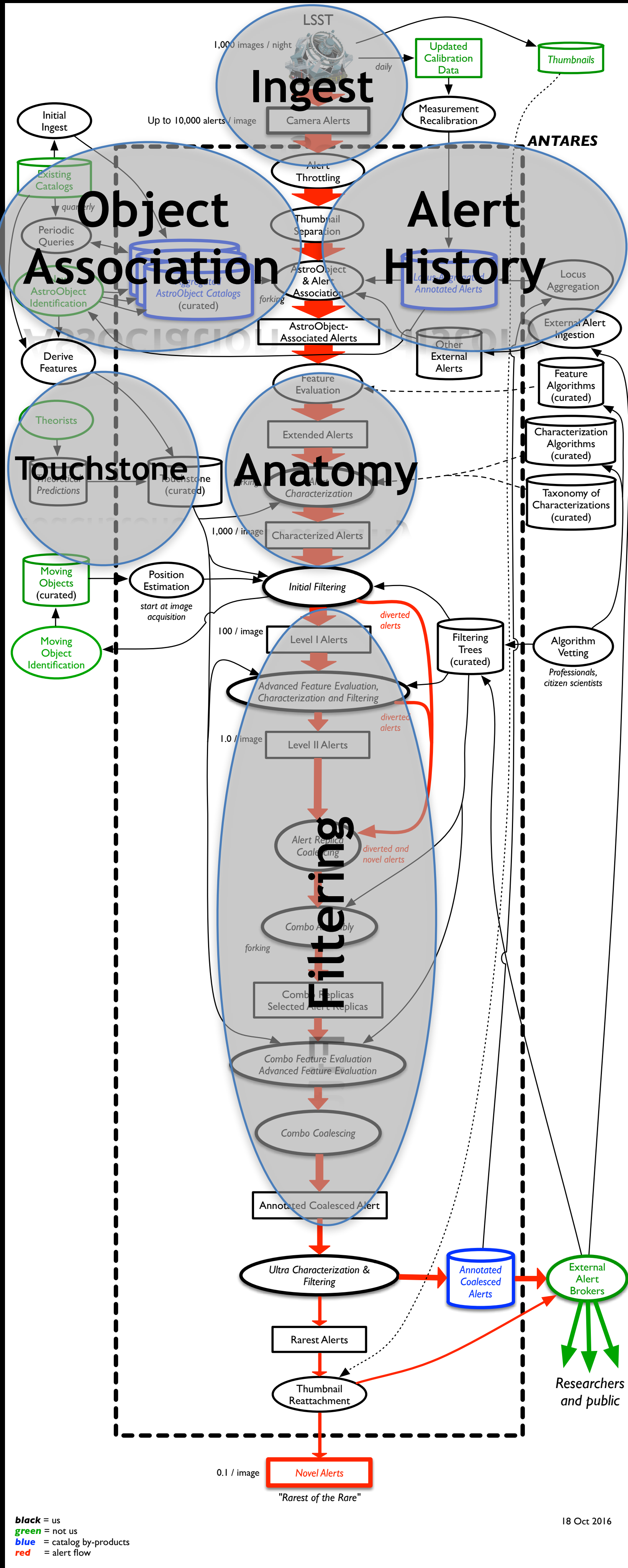


ANTARES Architecture

Alerts are generated from real light curves, outside the system, and then ingested into: **streaming data**

Alerts get annotated with external information and past history: **databases**

Several base features are derived for all the alerts: **OOP, profiling**



ANTARES Architecture

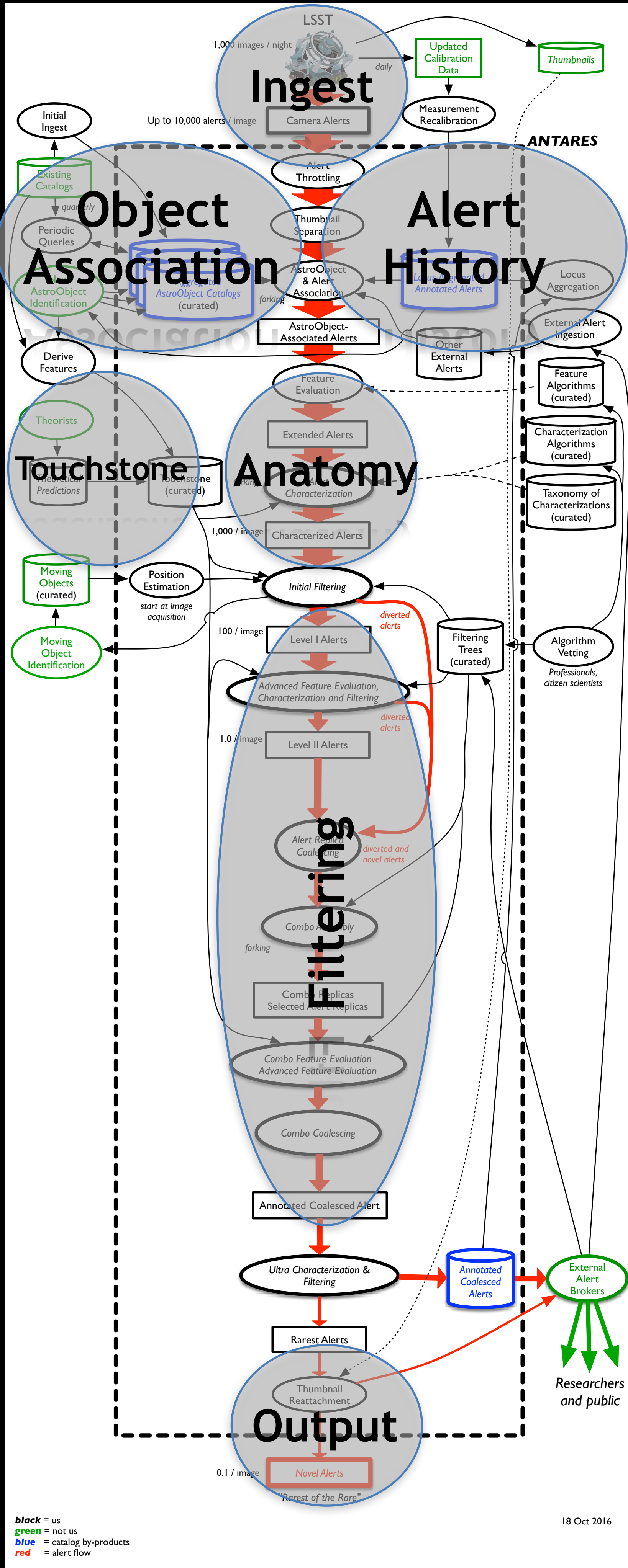
Alerts are generated from real light curves, outside the system, and then ingested into: **streaming data**

Alerts get annotated with external information and past history: **databases**

Several base features are derived for all the alerts: **OOP, profiling**

Alerts are characterized and compared to the Touchstone: **Time Series Analysis** and filtered: **Machine Learning**

ANTARES Architecture



Alerts are generated from real light curves, outside the system, and then ingested into: **streaming data**

Alerts get annotated with external information and past history: **databases**

Several base features are derived for all the alerts: **OOP, profiling**

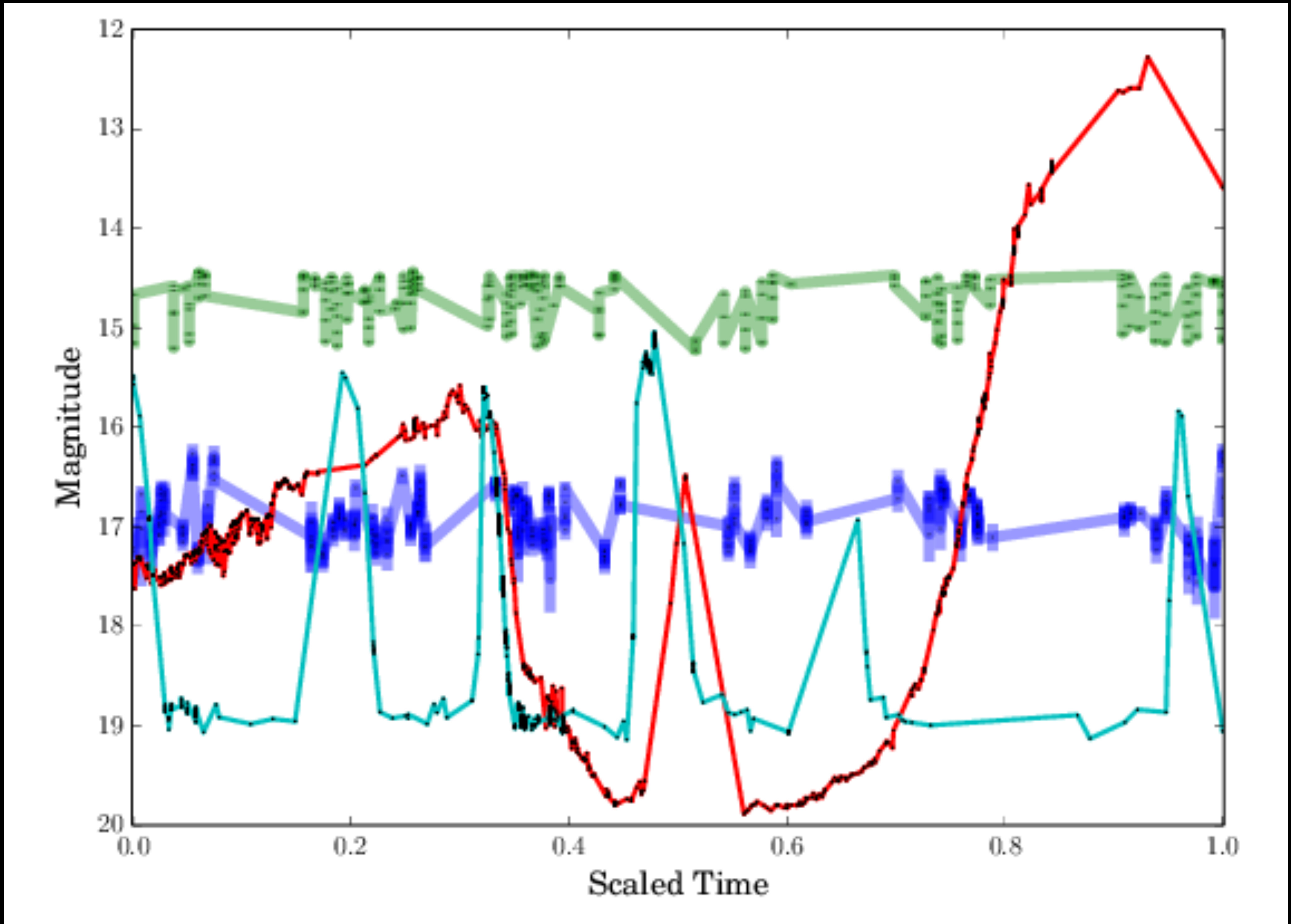
Alerts are characterized and compared to the Touchstone: **Time Series Analysis** and filtered: **Machine Learning**

'Rarest' selected (protoype goal)

TESTING WITH REAL ASTROPHYSICAL DATA

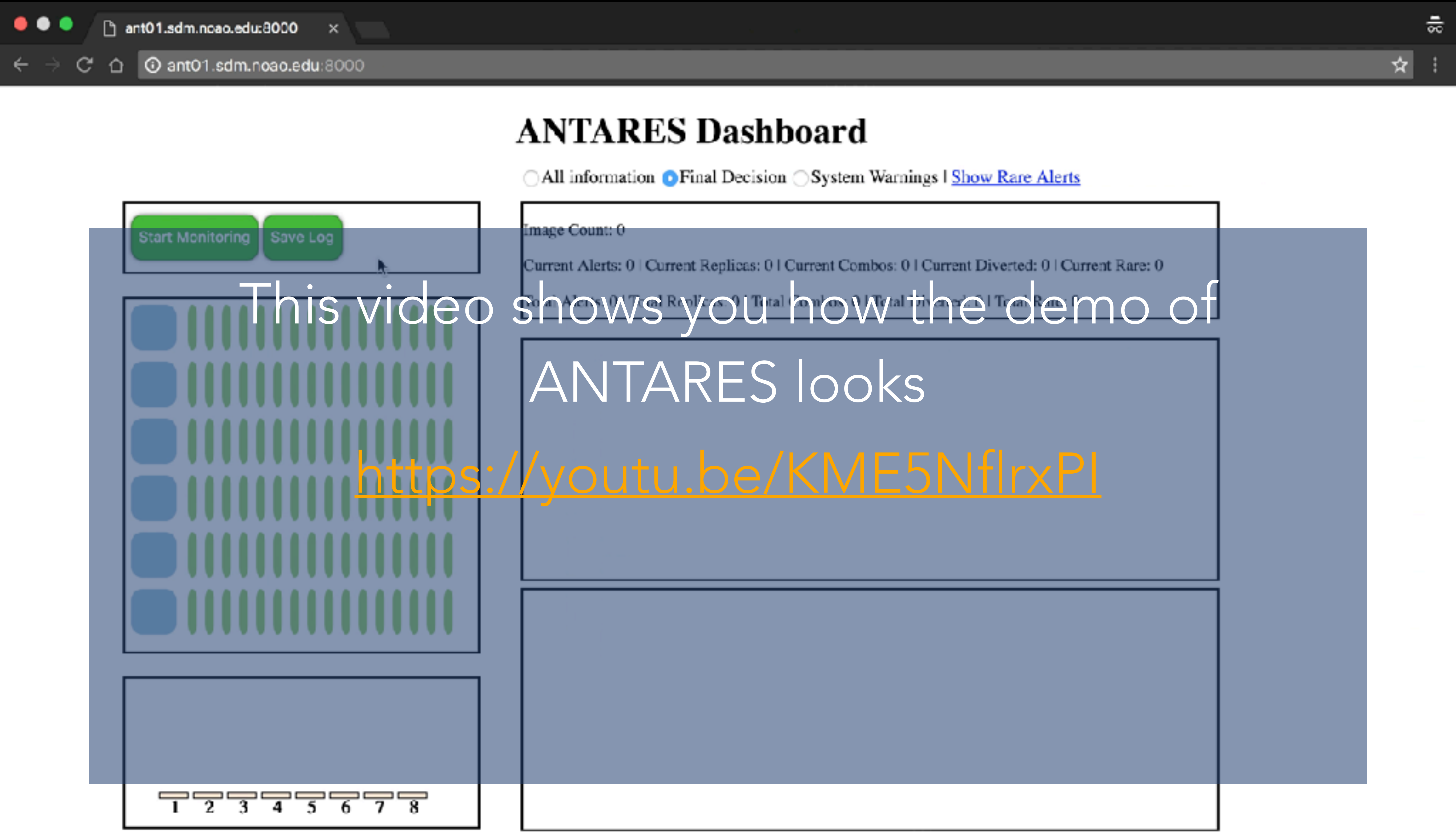
AN ILLUSTRATIVE EXAMPLE

LIGHTCURVES OF SOME OF THE "RARE" OBJECTS USED IN THIS DEMO
DERIVED FROM TT ARI, R CRB



PROCESSING MONITOR + LOGGER

DASHBOARD

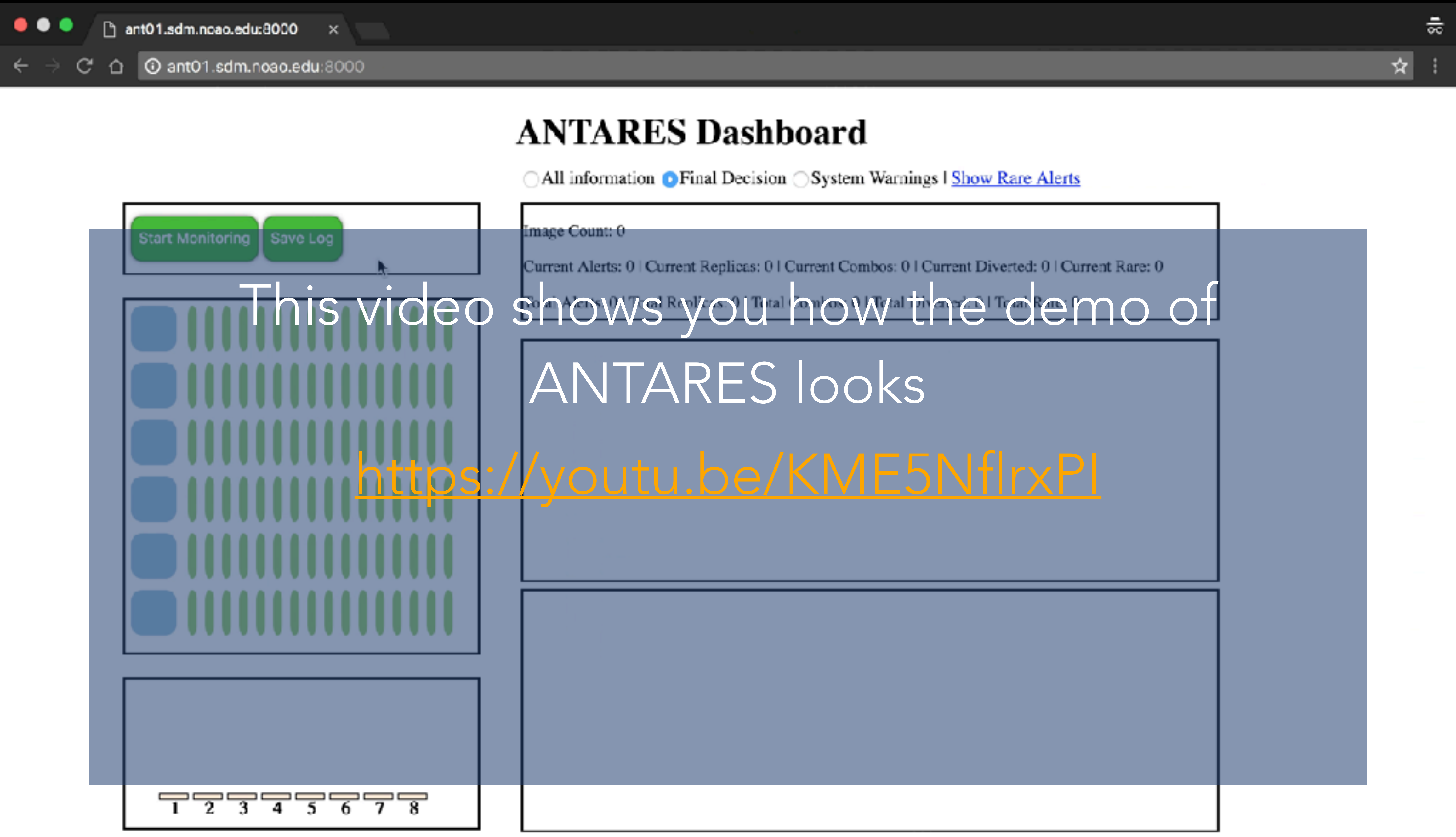


This video shows you how the demo of ANTARES looks

https://youtu.be/KME5NflrxPI

PROCESSING MONITOR + LOGGER

DASHBOARD



This video shows you how the demo of ANTARES looks

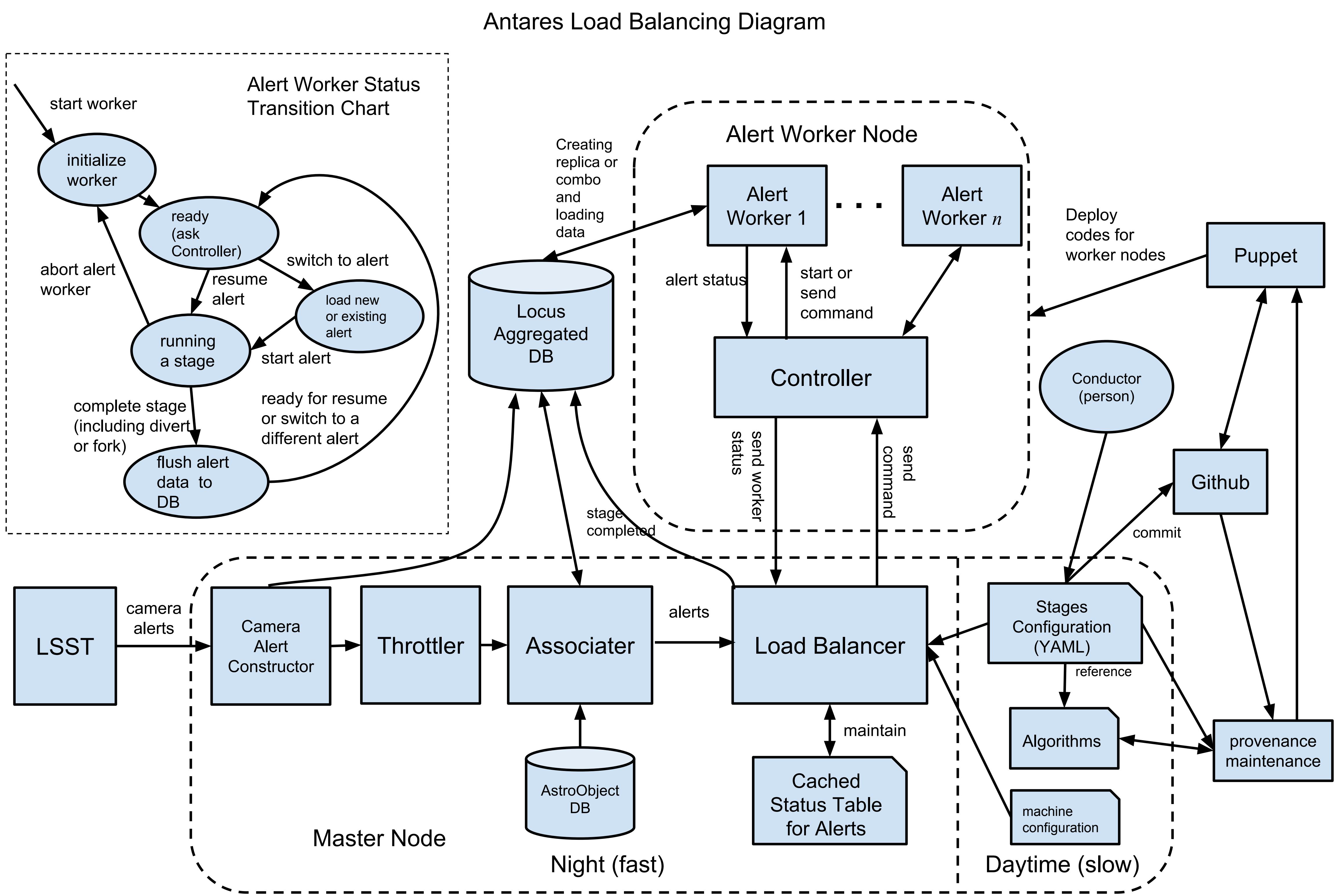
https://youtu.be/KME5NflrxPI

HOW IS IT HAPPENING?

STARTUP

CHALLENGE: WE HAVE ~40 SECONDS TO PROCESS EACH IMAGE

FULLY PARALLEL LOAD BALANCING ARCHITECTURE WITH PROVENANCE



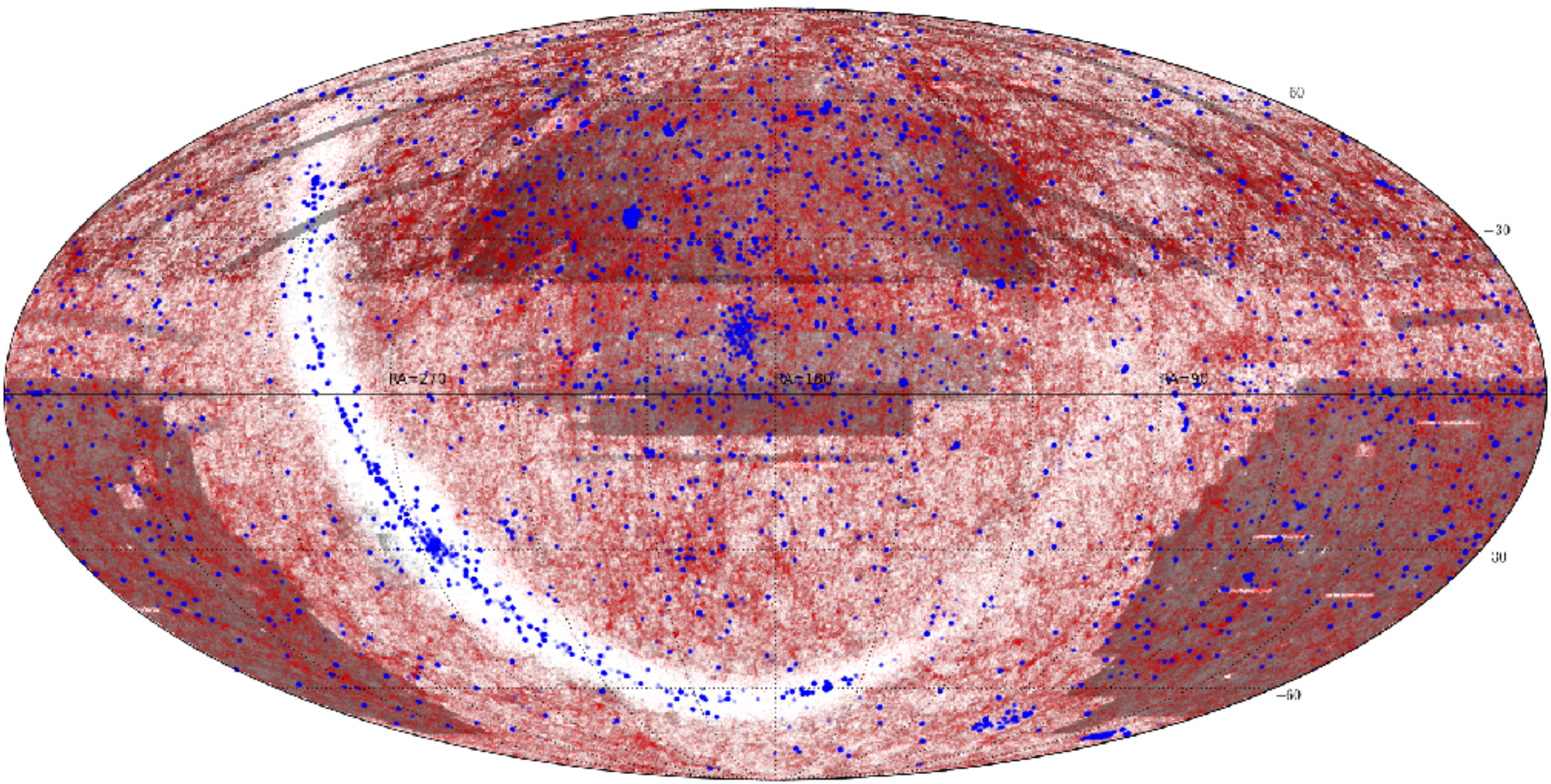
CROSS-MATCHING AND ASSOCIATION

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, 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.



This video shows you how the holdings at the Data Lab have built up - i.e. our time-domain history

http://datalab.noao.edu/files/may16Movie.mp4

- Data Lab 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

This video shows you how the holdings at the Data Lab have built up - i.e. our time-domain history

http://datalab.noao.edu/files/may16Movie.mp4

- Data Lab 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

FEATURE DERIVATION AND CHARACTERIZATION

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

MOST OF THESE LSST WILL PROVIDE*

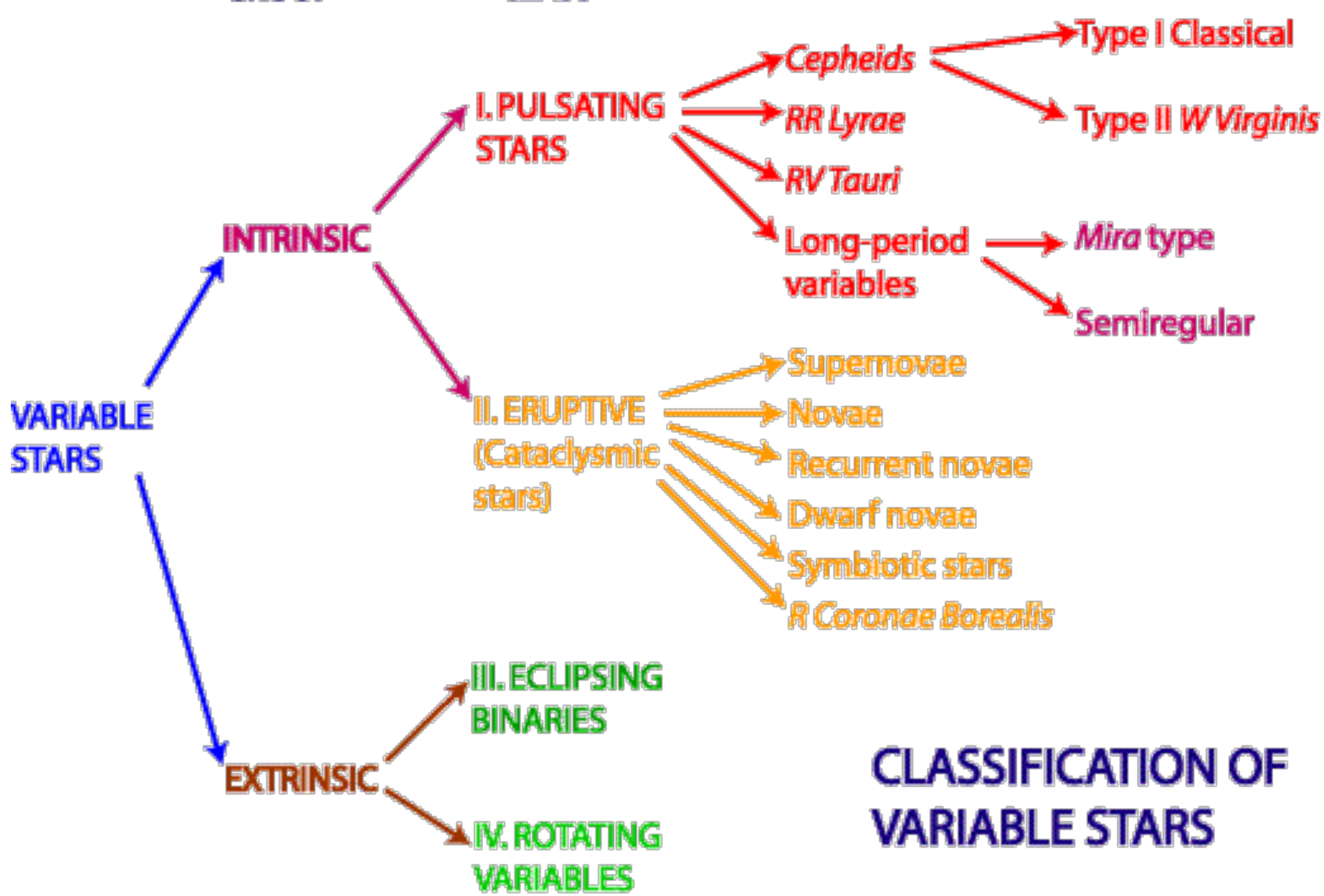
*MANY WILL NOT BE RELIABLE UNTIL AFTER A YEAR OF OPERATIONS

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).

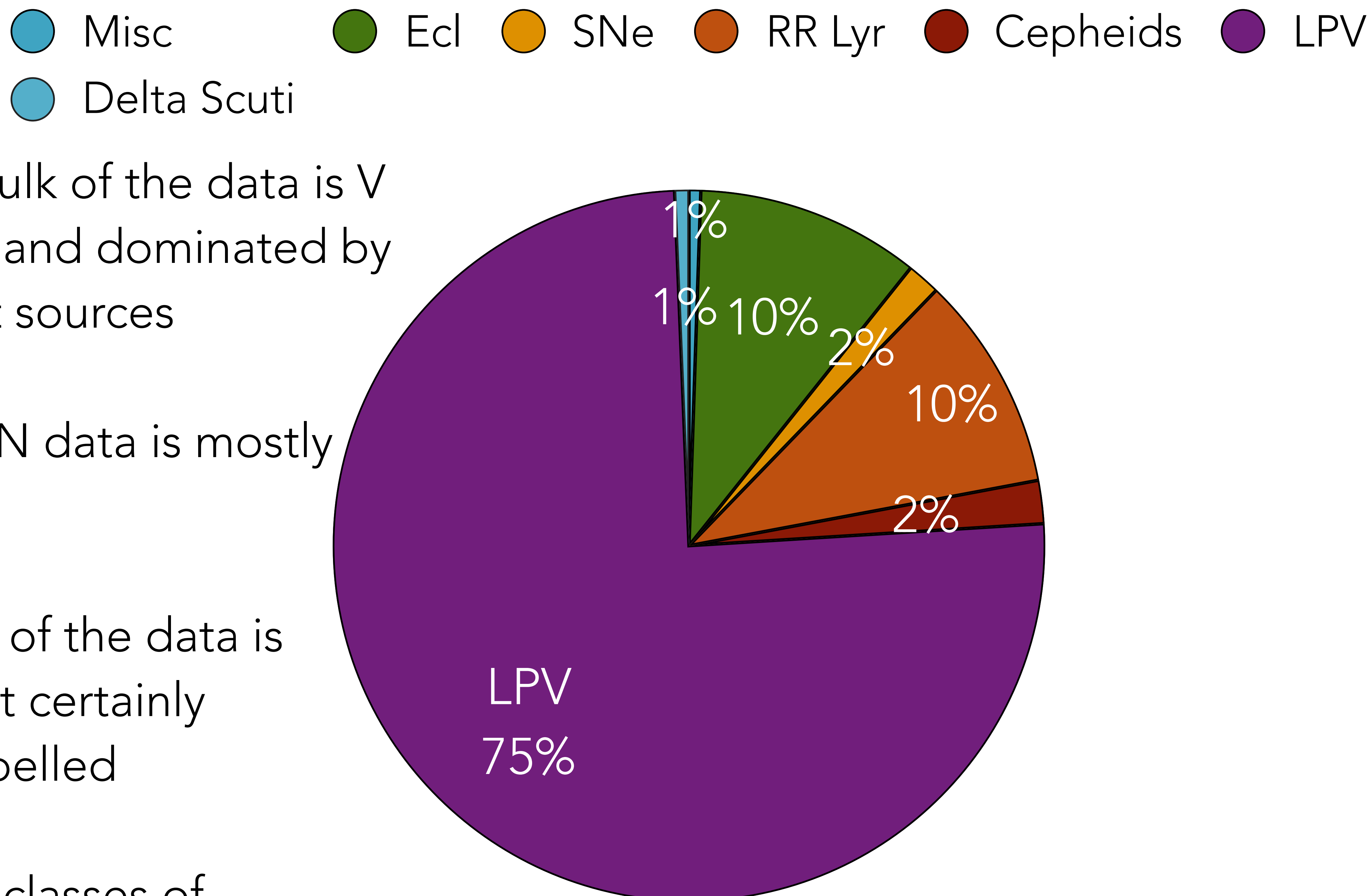
THE TOUCHSTONE REFERENCE LIBRARY



WHAT ASTROPHYSICS IS IN THE TOUCHSTONE

(the other half of the equation)

DATA SOURCES: OGLE + LINEAR + SNE.SPACE



- 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

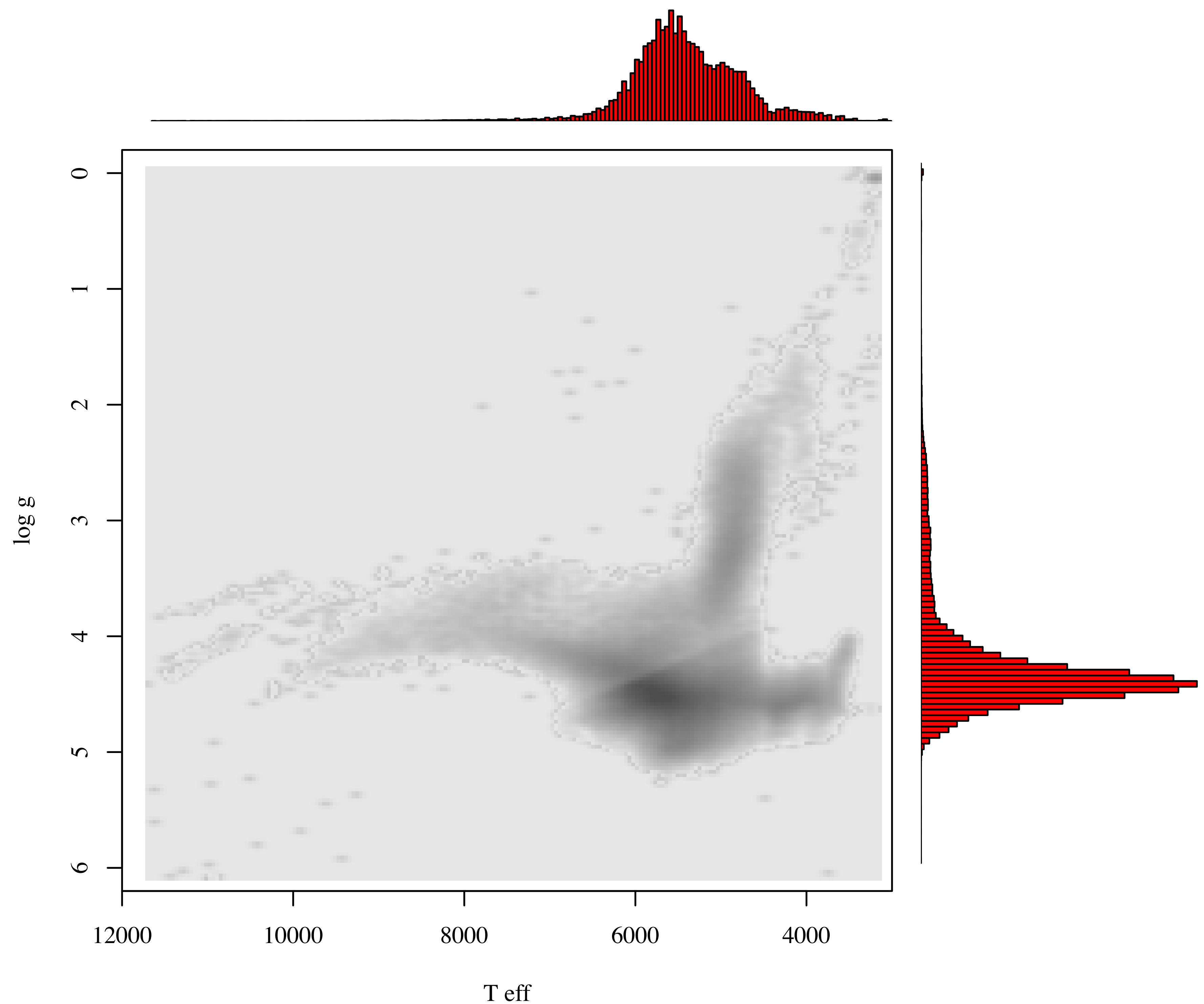
FILTERING

ASSESSING THE LIKELIHOOD OF VARIABILITY

- Kepler provides an empirical model for variability of Galactic stars

- 155k+ stars over wide range of spectral types

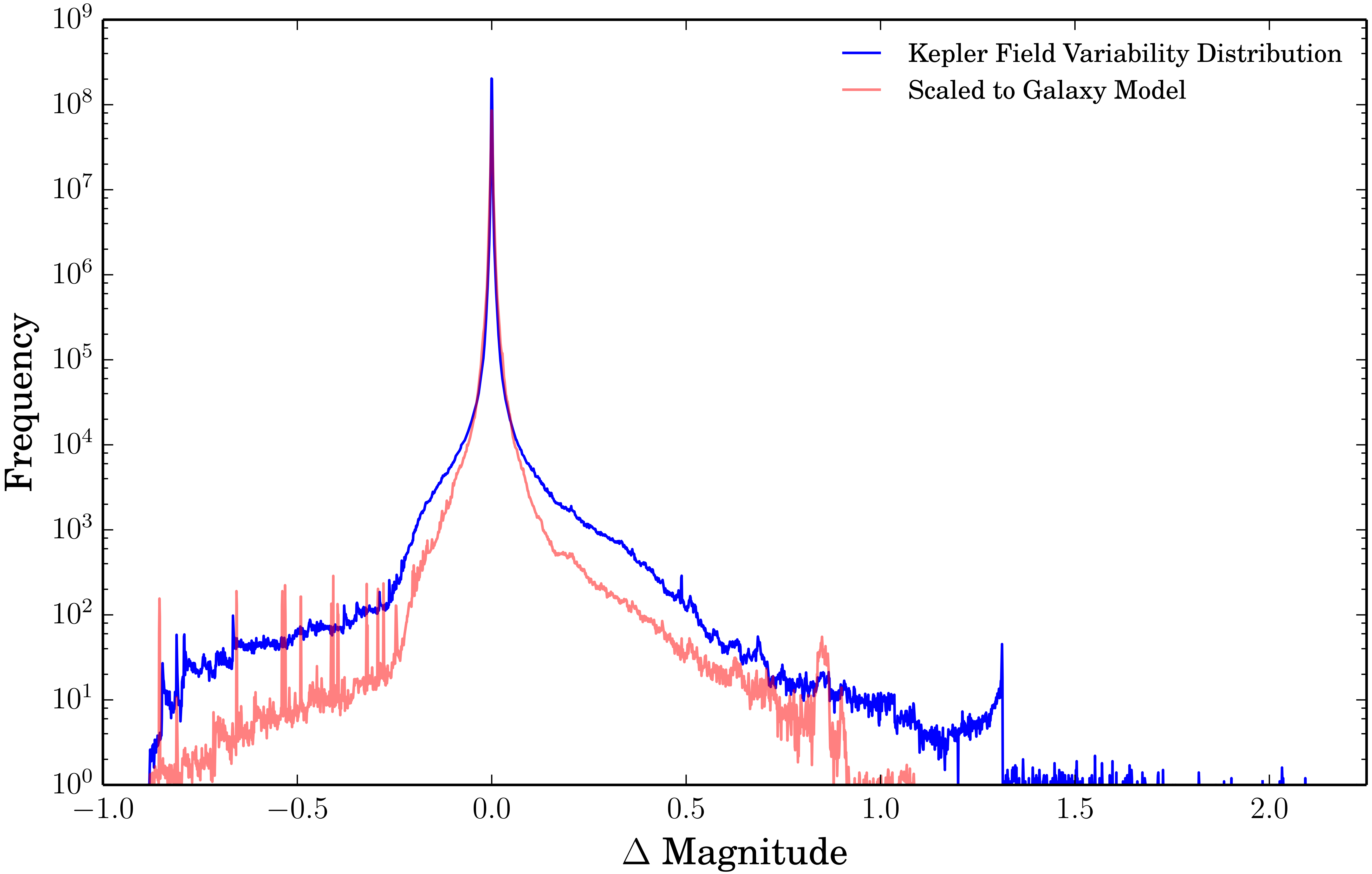
- Scale Kepler variability distribution to match Besançon model of spectral types



VARIABILITY PROBABILITY DISTRIBUTION

Can predict variability given galaxy model and galactic coordinates of alerts using Kepler Q13 data

(*Ridgway+ arXiv:1409.3265*)

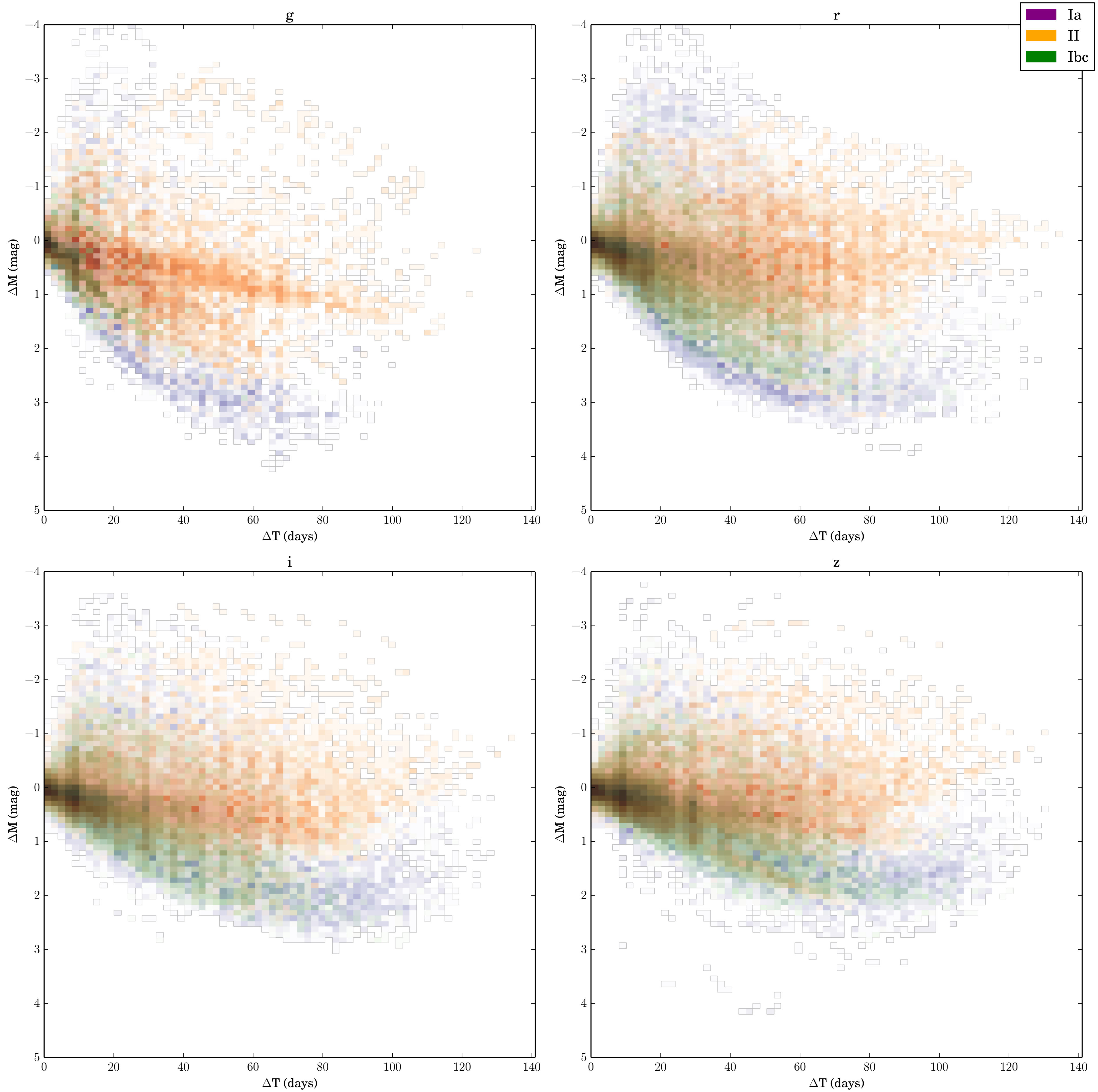


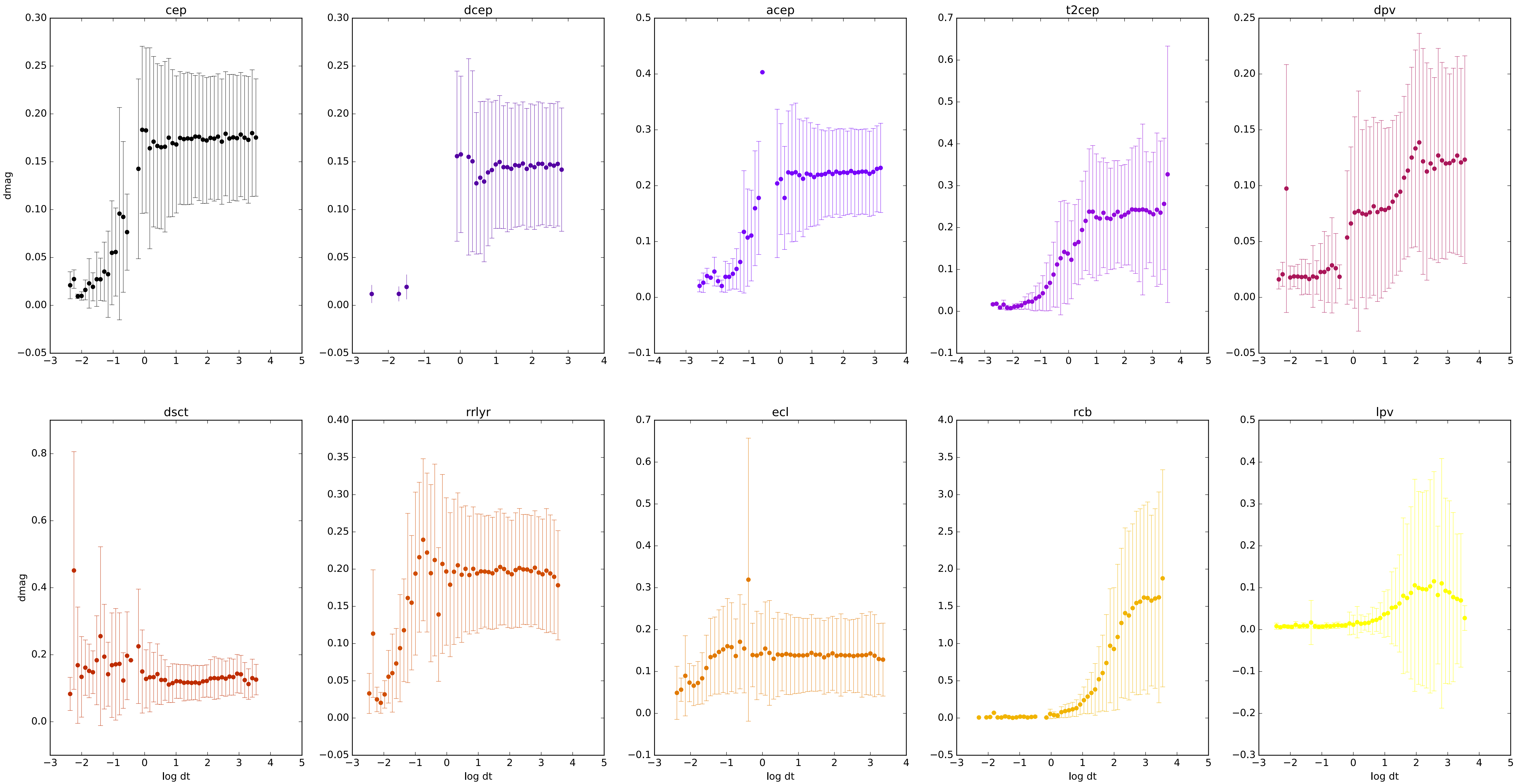
IF WE WANT TO EXTEND THIS TO TRANSIENTS:

- 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
  - Can use kNN to attempt to classify

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

- Even with just the structure function of observed objects, you can distinguish type I vs II SNe, given data in multiple bands over a long enough baseline
- Most often, we'll have some host association and therefore at least a photo-z as well
- Irregularly sampled, and biased, but can convert to KDE





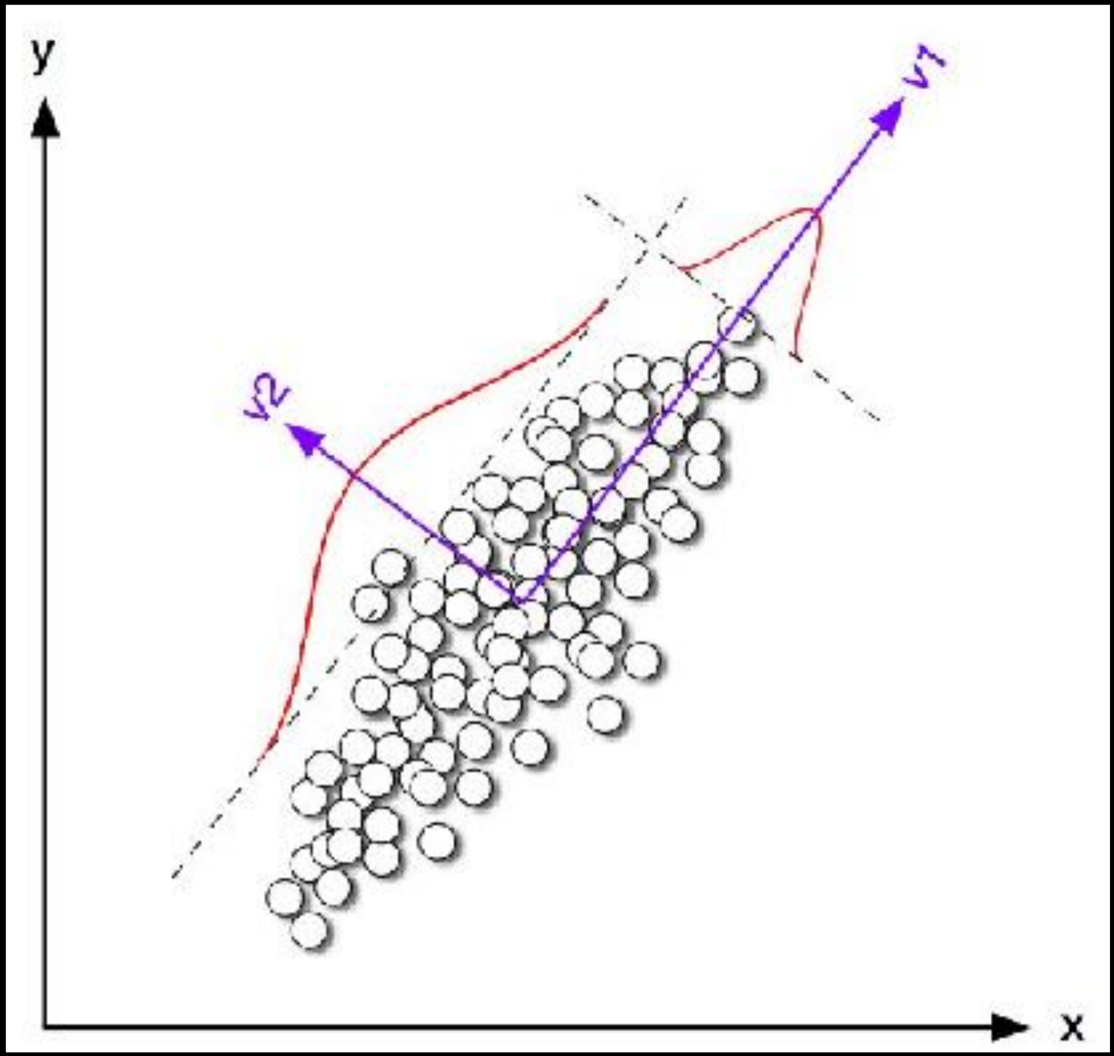
- We can also construct these distributions for different classes of variable stars (as long as we have a sufficient number of light curves to build a distribution)

- The advantage of doing this s that you can (potentially) label something as variable WITHOUT finding a period: CHEAP/FAST

PRINCIPAL COMPONENT ANALYSIS

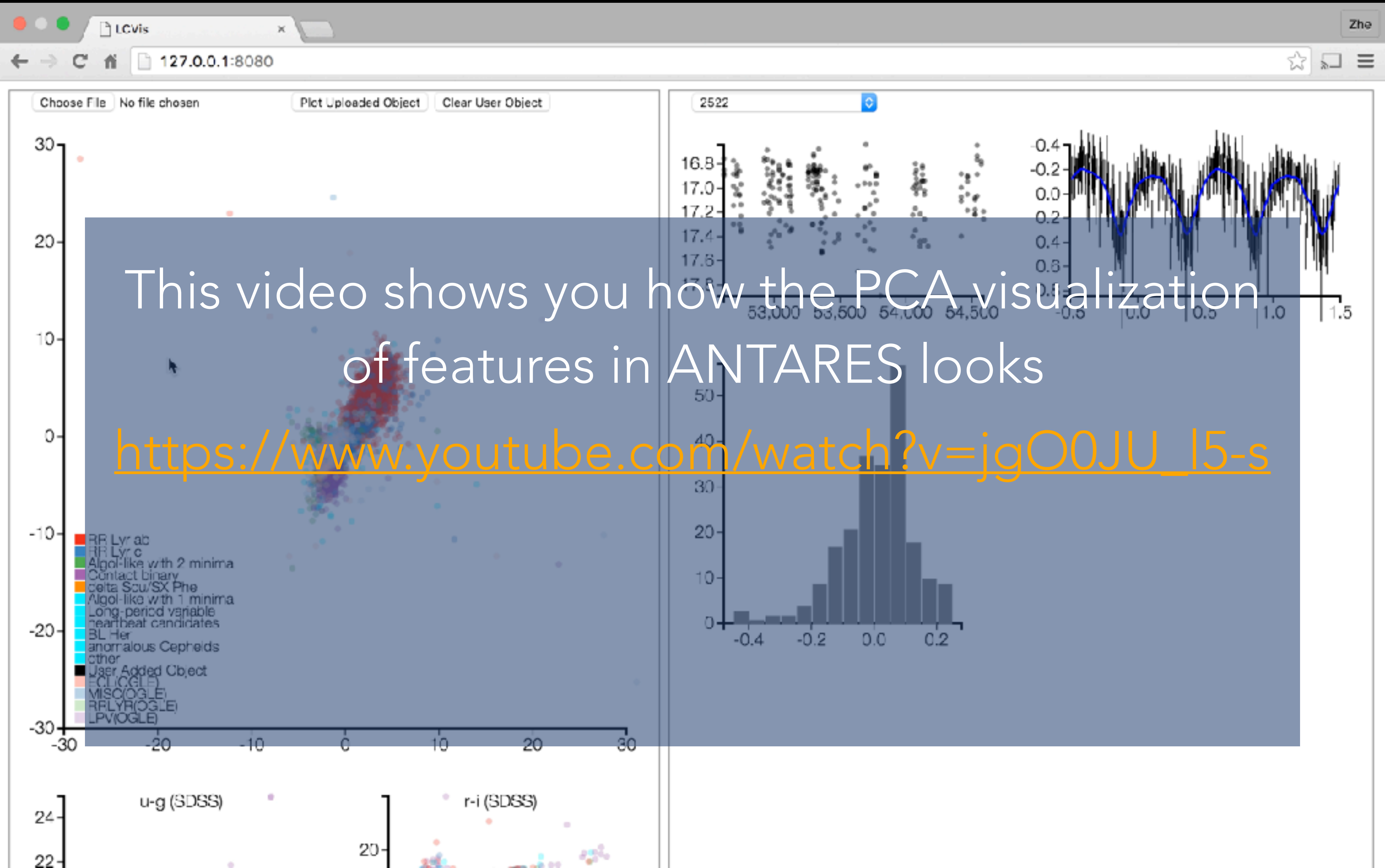
- Orthogonal transformation to map correlated values into linearly uncorrelated basis vectors (i.e. Principal Components)
  - Reveals internal structure of data in a way that best explains variance

- Probably the most common dimensionality reduction method (i.e. lots of stackexchange posts for when it doesn't work like you think it should)
  - Prevents overfitting
  - Only use first few principal components for feature extraction



EXAMINING HIGH DIMENSIONAL DATASETS & PLACING ALERTS IN CONTEXT

PCA + VISUALIZATION DEMO

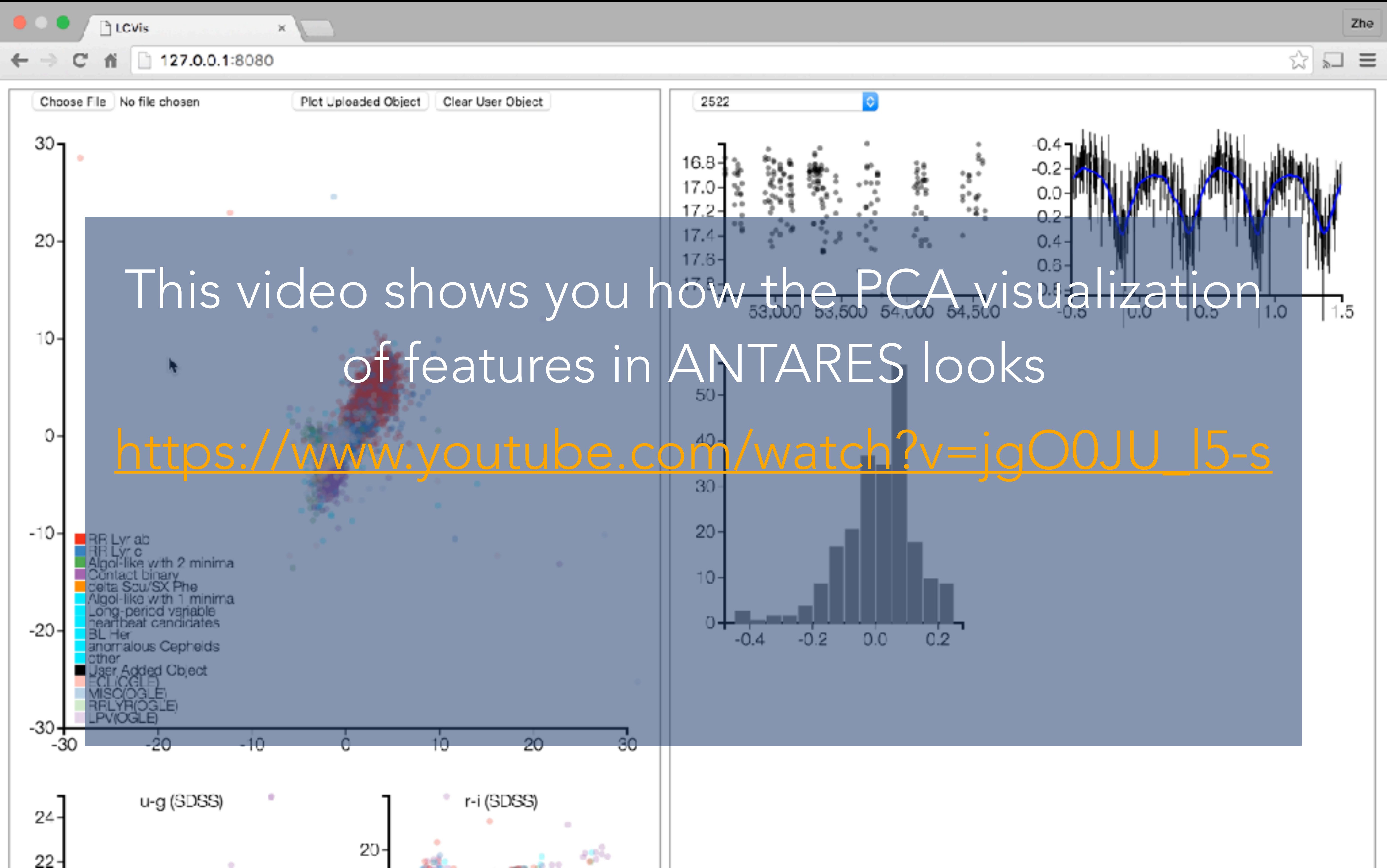


This video shows you how the PCA visualization of features in ANTARES looks

https://www.youtube.com/watch?v=jgO0JU_l5-s

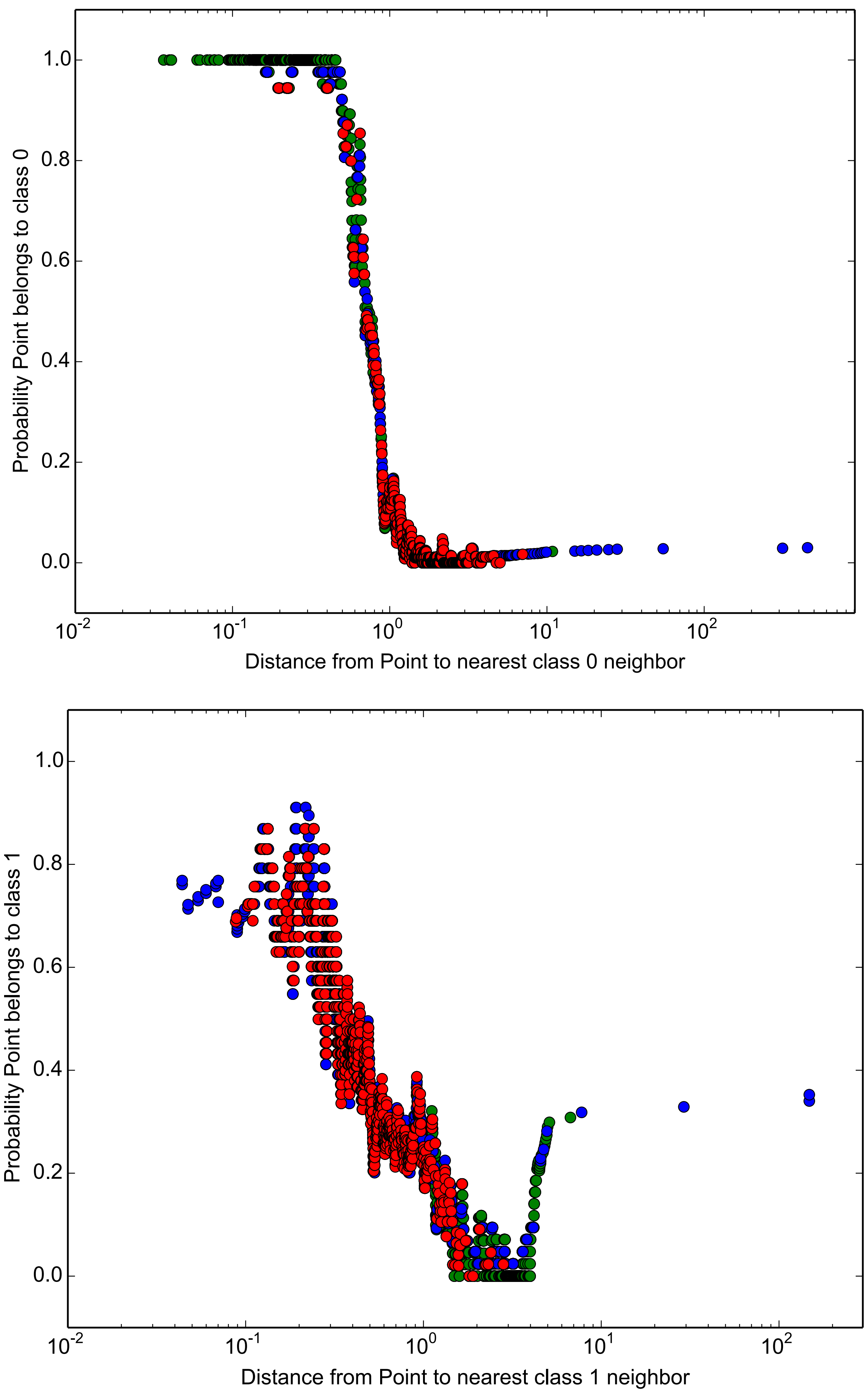
EXAMINING HIGH DIMENSIONAL DATASETS & PLACING ALERTS IN CONTEXT

PCA + VISUALIZATION DEMO



This video shows you how the PCA visualization of features in ANTARES looks

https://www.youtube.com/watch?v=jgO0JU_I5-s



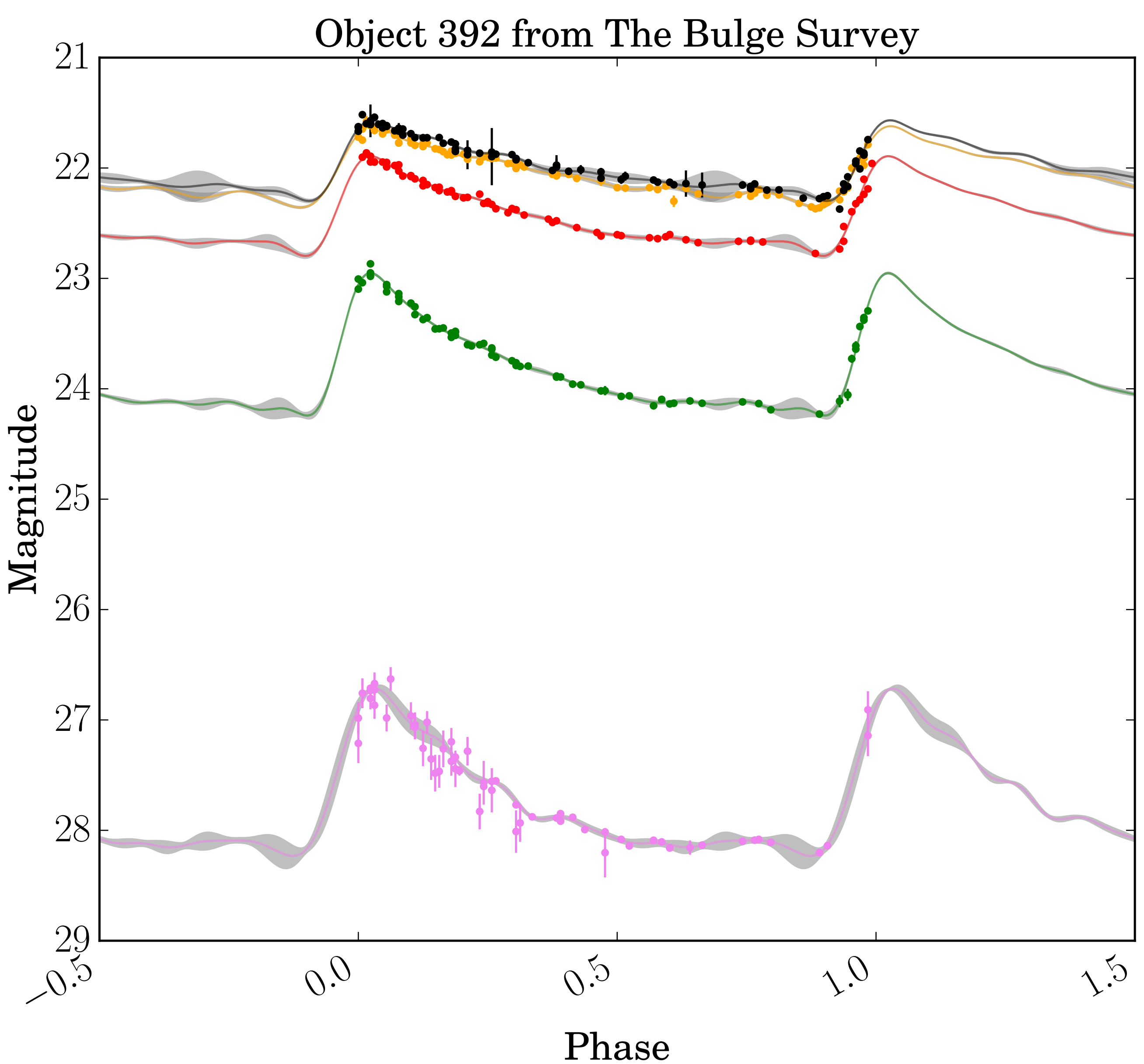
CHARACTERIZATION

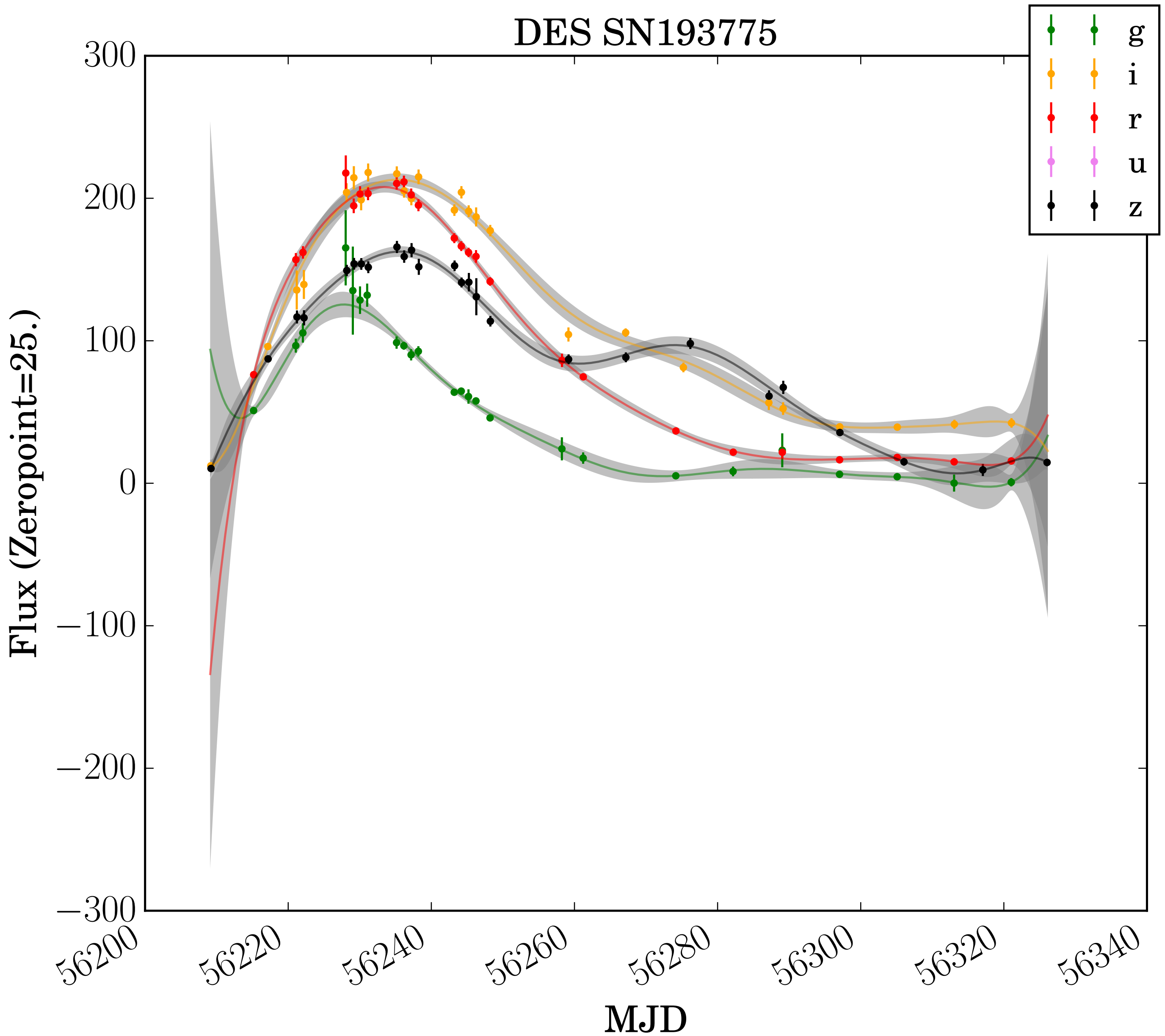
- Given high dimensional feature space + contextual info of an alert, what can we say about it?
- Assemble large touchstone with representations of astrophysical events, real & predicted (OGLE + LINEAR variables, all available SNe light curves, etc)
- No single algorithm works best - use ensemble - kNN, random forests, SVM, neural nets
- 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

SEPARATION, CHARACTERIZATION, MODELING

- We can do a coarse characterization with simple point estimates - amplitude/skewness/kurtosis
- We want to characterize better as we get more observations
- We can start to do time-series analysis
  - Lomb-Scargle/Phase Dispersion Minimization/Fourier analysis/conditional entropy
- Can model the data if we have enough observations: Gaussian Processes, Wavelet decomposition

GAUSSIAN PROCESSES TO MODEL LIGHT CURVES

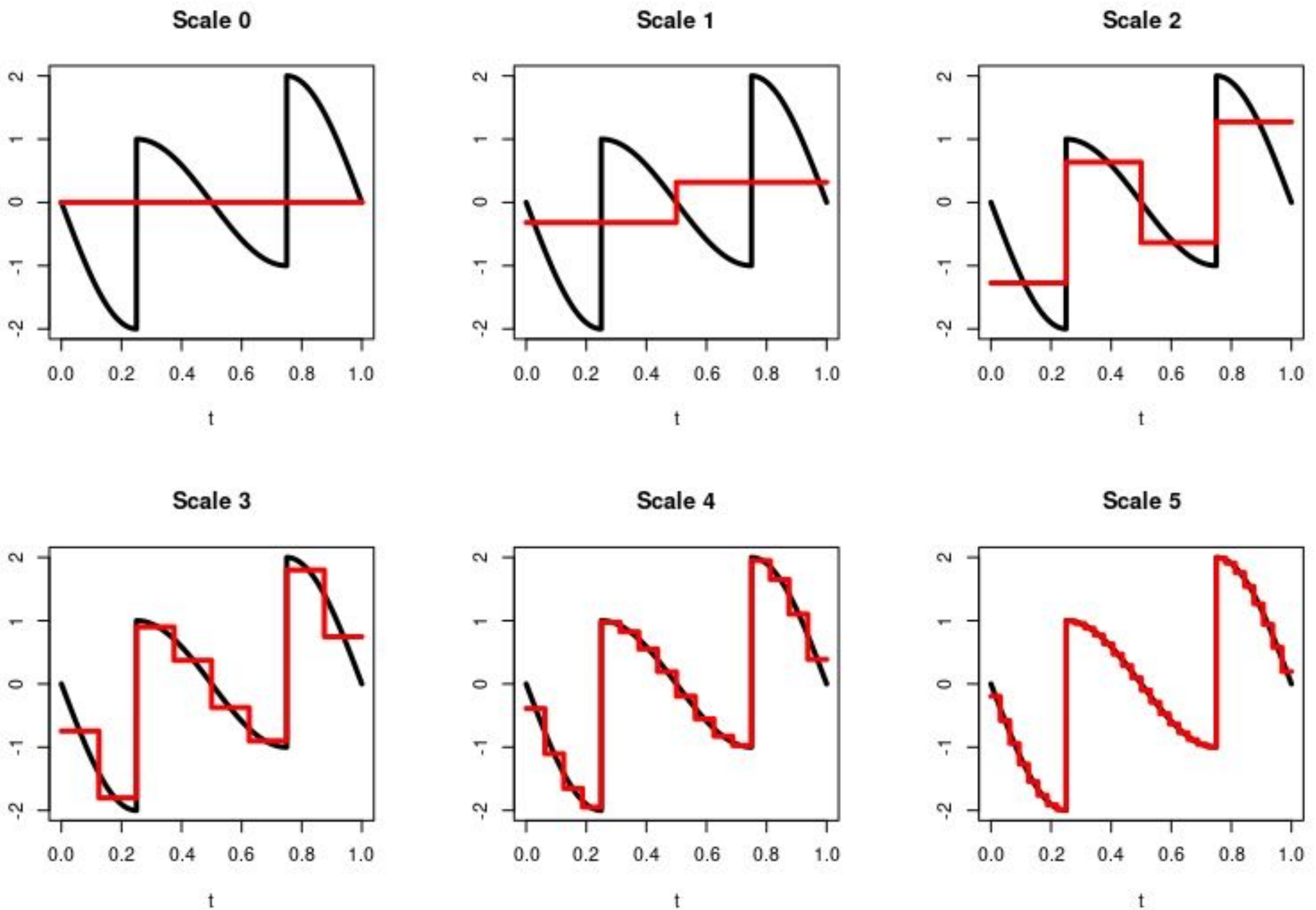




PROJECT TOUCHSTONE GP ONTO DATA IF POORLY SAMPLED, OR MODEL OBSERVATIONS WITH GP, AND COMPARE HYPERPARAMETERS

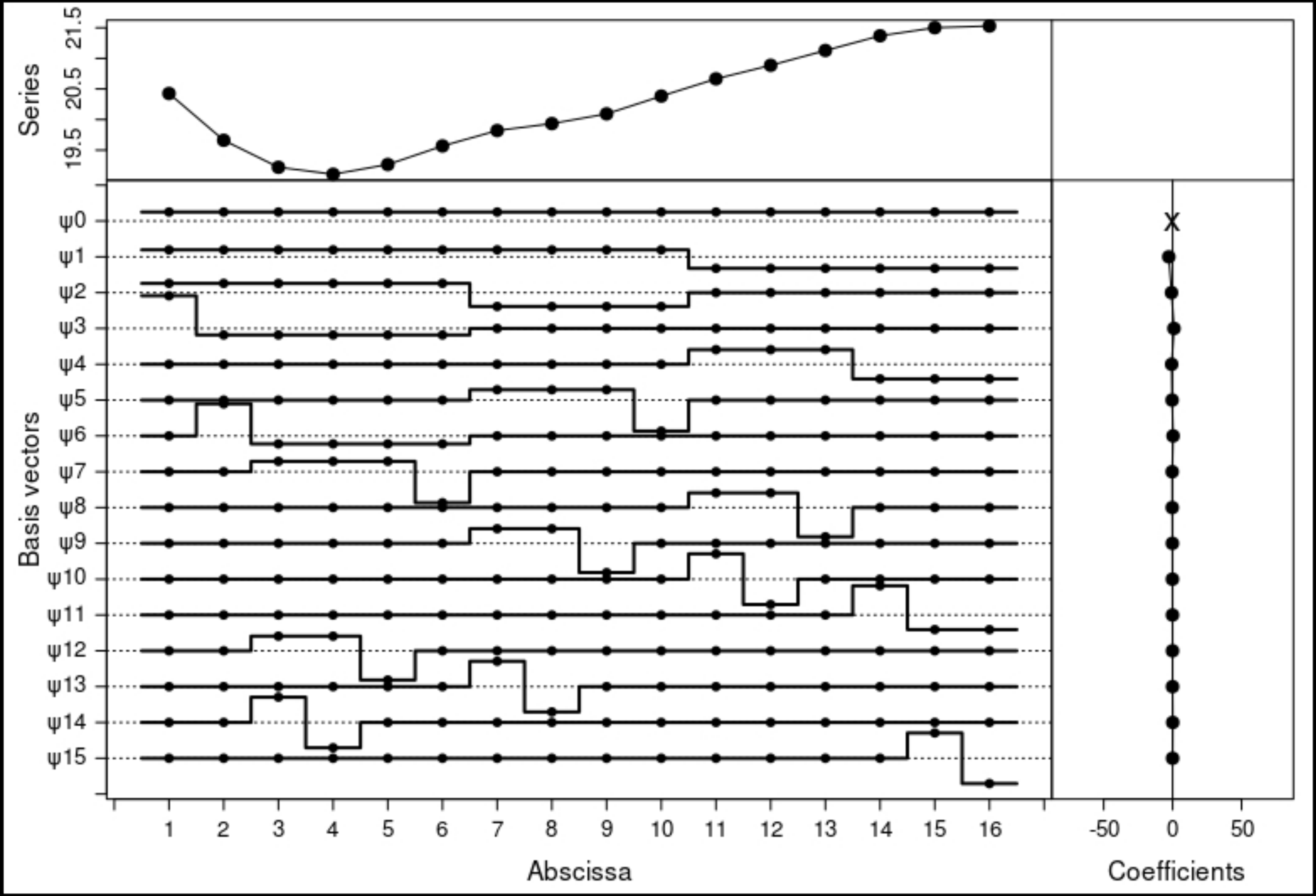
WAVELET DECOMPOSITION TO DESCRIBE LIGHTCURVE SHAPE

Wavelet decompositions can be translationally invariant



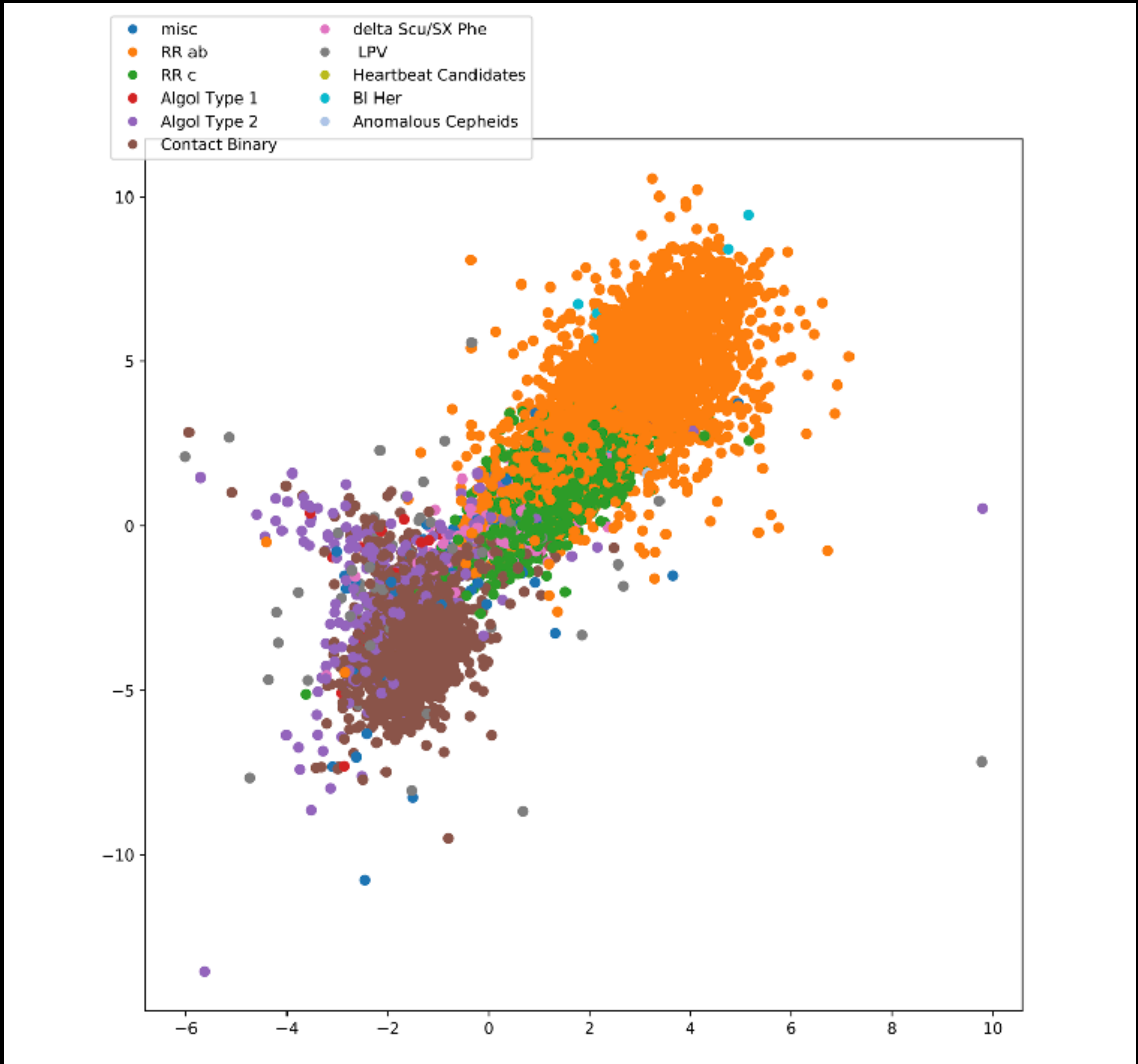
BAGIDIS: BASIS GIVING DISTANCES USING AN UWHT

DECOMPOSE SMOOTH LIGHTCURVES INTO FEW BASIS VECTORS

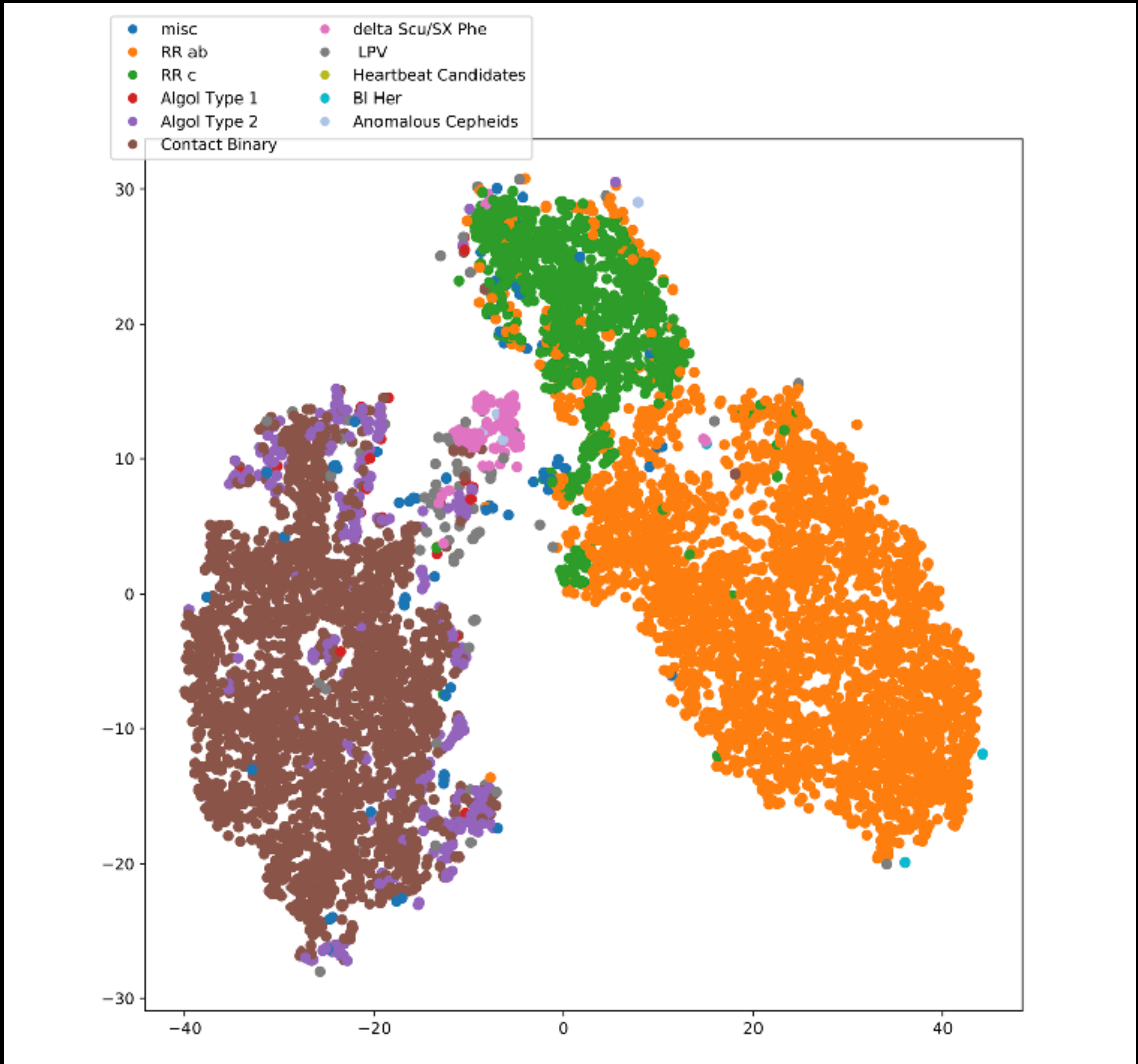


Work by my undergrad, Tayeb Zaidi and in collaboration with Michelle Lochner

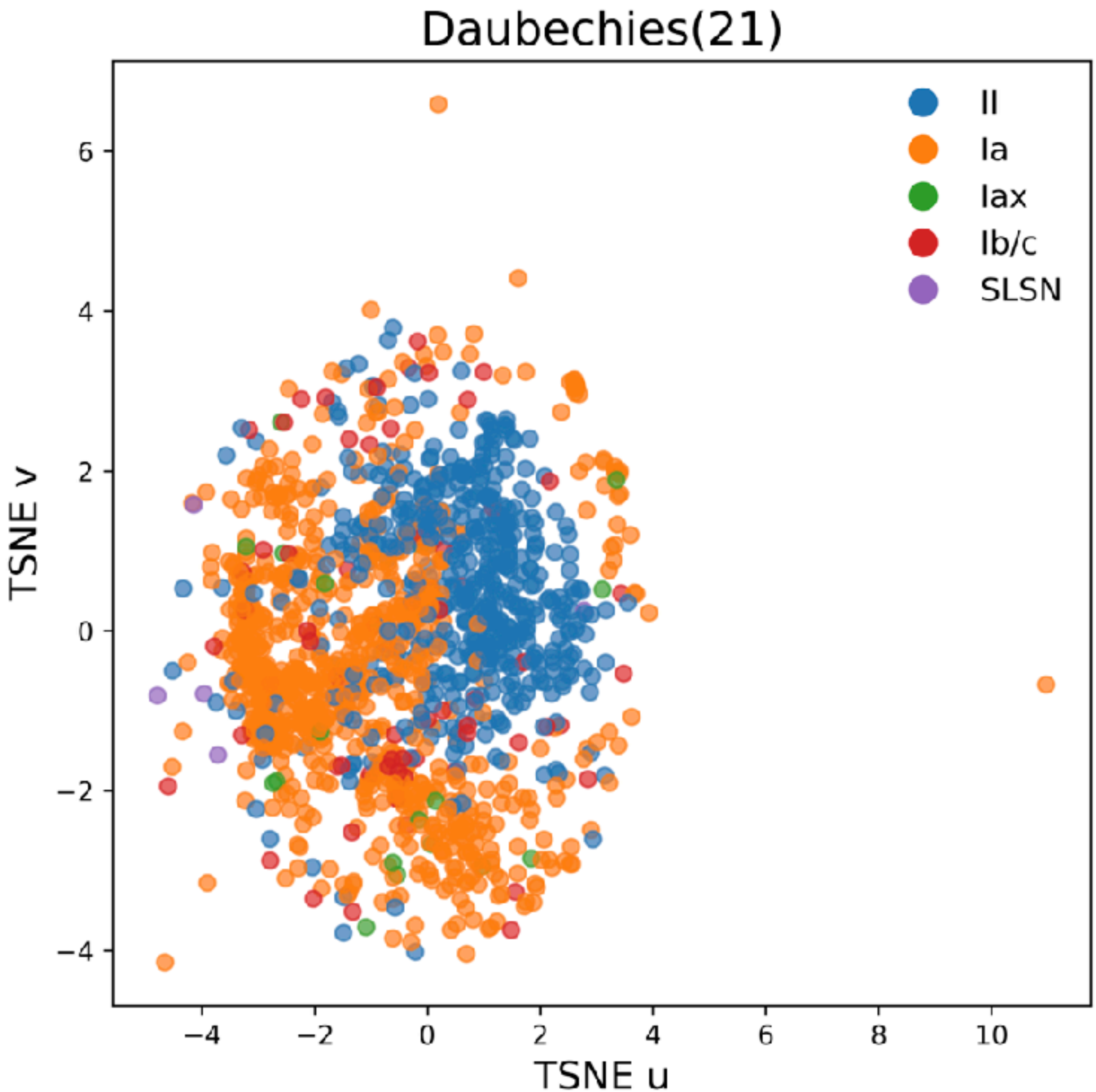
THIS WORKS ASTONISHINGLY WELL FOR BOTH VARIABLES AND TRANSIENTS



THIS WORKS ASTONISHINGLY WELL FOR BOTH VARIABLES AND TRANSIENTS



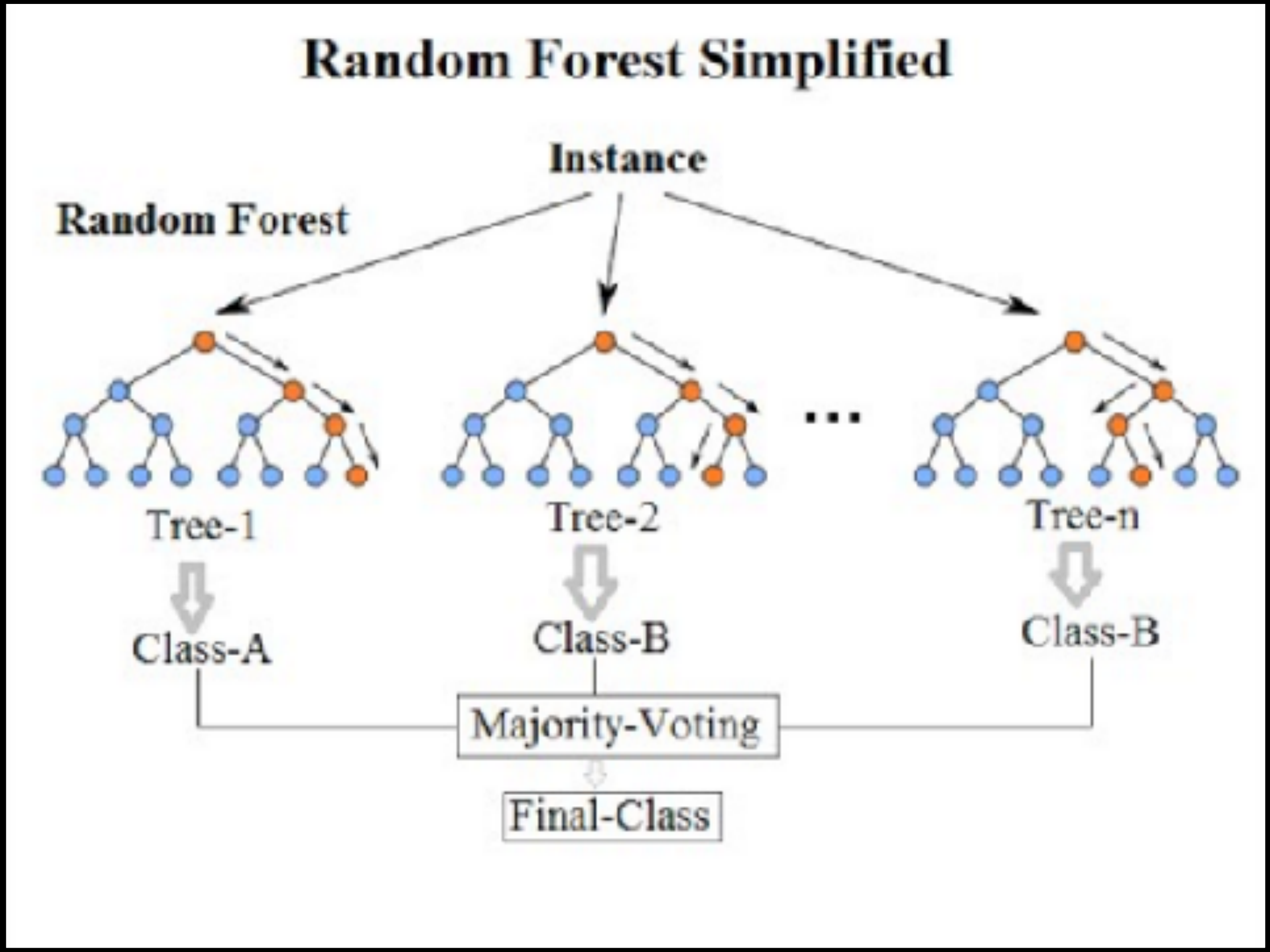
This is an example of the t-SNE you saw in Session 2



Daubechies(21)
II
Ia
Iax
Ib/c
SLSN
TSNE v
TSNE u
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10

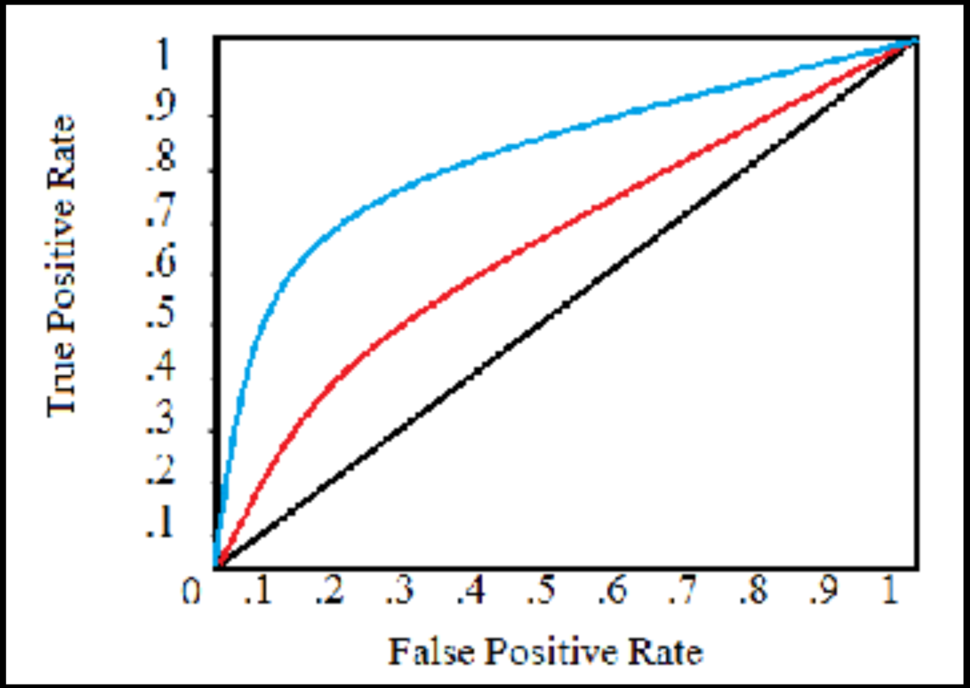
CLASSIFICATION ALGORITHMS

- Mapping from input features to output classes
  - Many require labeled data (i.e. the Touchstone) - supervised learning algorithms
  - Unsupervised learning algorithms generally work well if your data doesn't have degeneracies
  - **Only as good as your feature selection**
- Decision Trees
  - Sequence of decision rules (*if* statements) generated recursively on the training data
- Random Forests
  - Simple, robust classifier consisting of multiple individual decision trees
  - All trees vote on the class, and the mode of the votes is selected



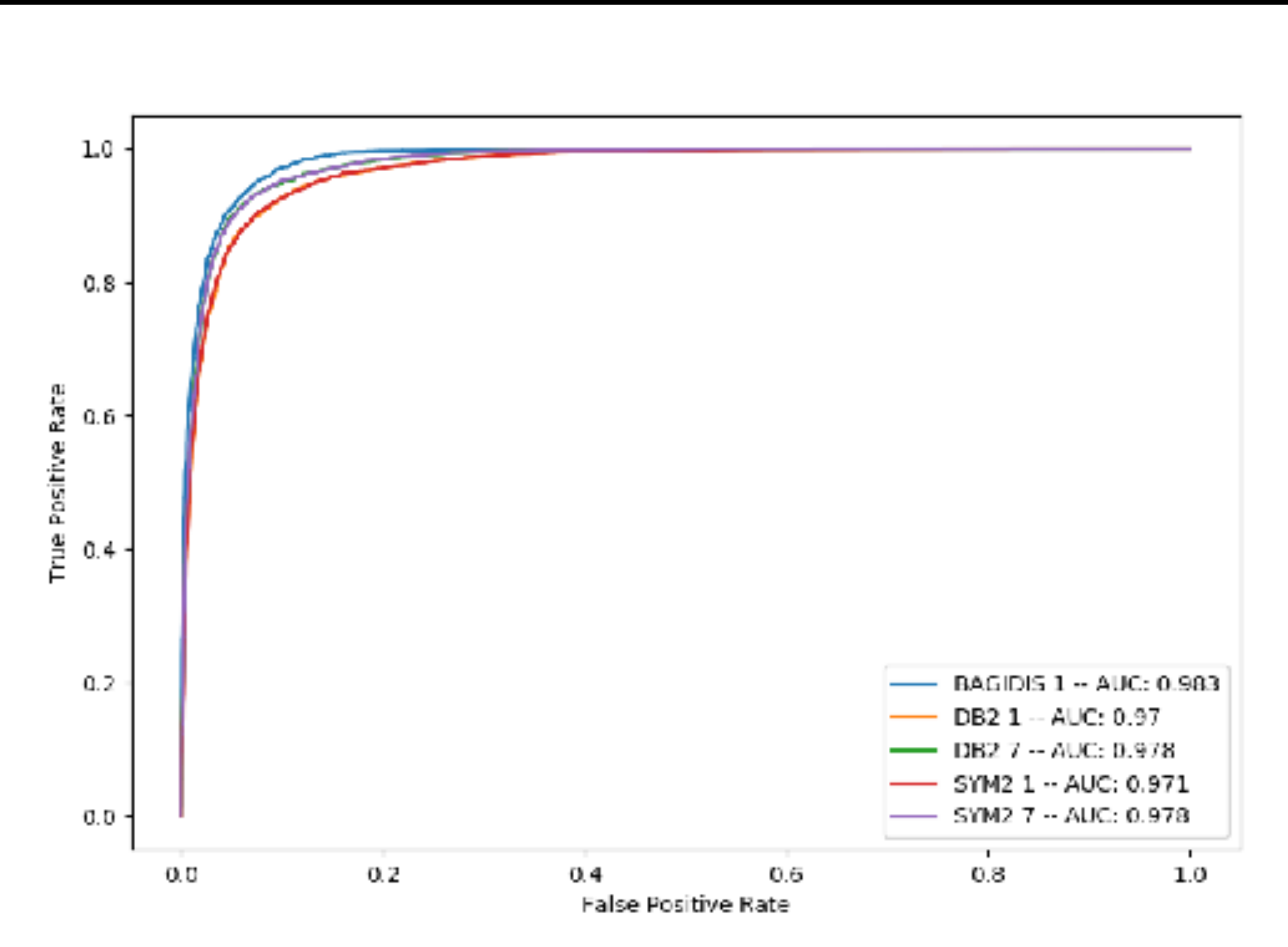
ROC CURVES

- Receiver Operating Characteristic Curves
  - Plot the True Positive Rate vs False Positive Rate *at each threshold*
    - True Positive Rate (TPR) = TP / (TP + FN)
    - False Positive Rate (FPR) = FP / (FP + TN)
  - Resilient to imbalances in class representation
  - Allows careful selection of threshold for different astrophysical uses
- AUC measures the Area Under the Curve
  - 0.5 for random guessing, 1 for perfect classification

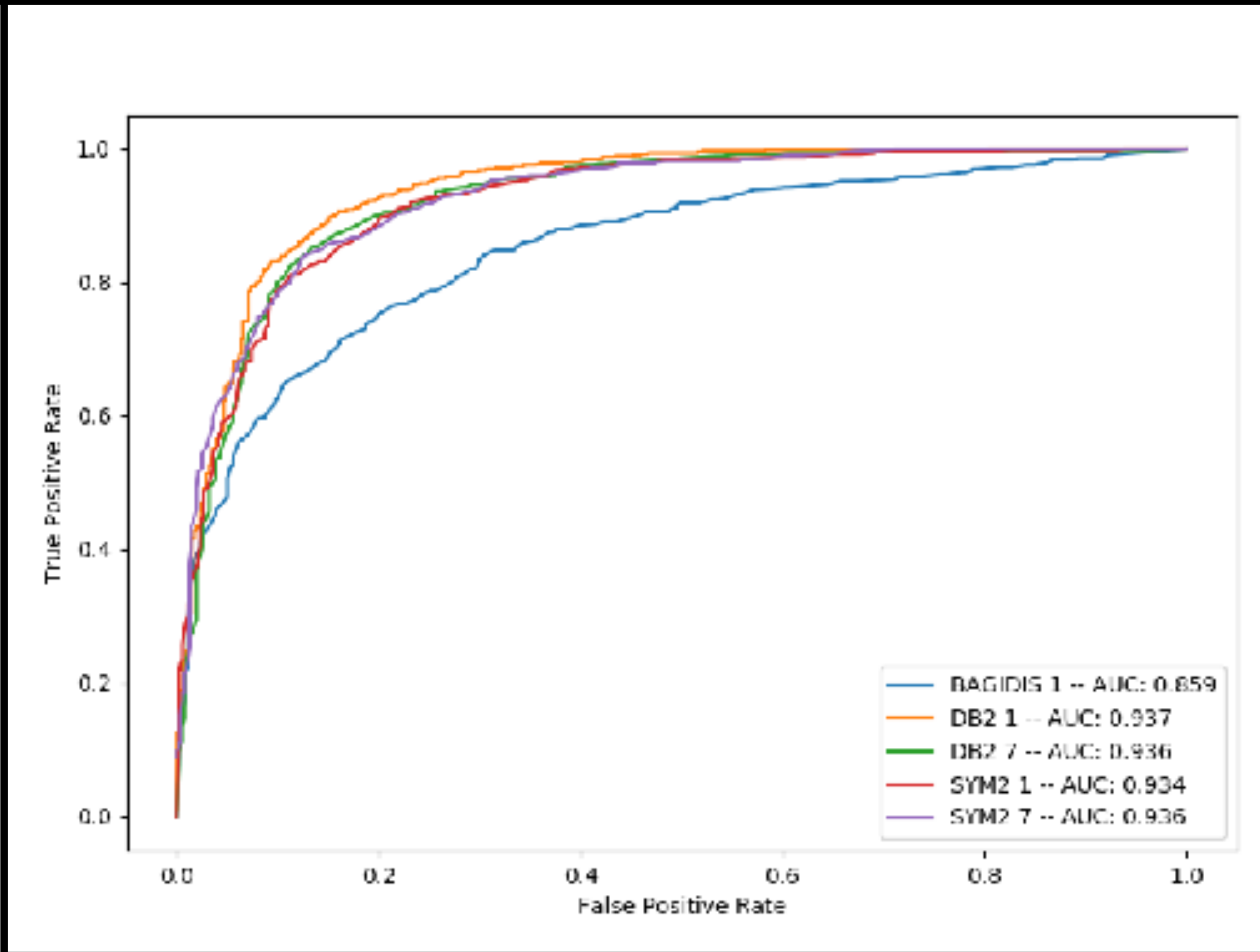


		Predicted Classification	
		Positive prediction	Negative prediction
True Classification	Positive class	True Positive (TP)	False Negative (FN)
	Negative class	False Positive (FP)	True Negative (TN)

CLASSIFICATION RESULTS



Simulated data: SNPhotCC



Real data: OSC

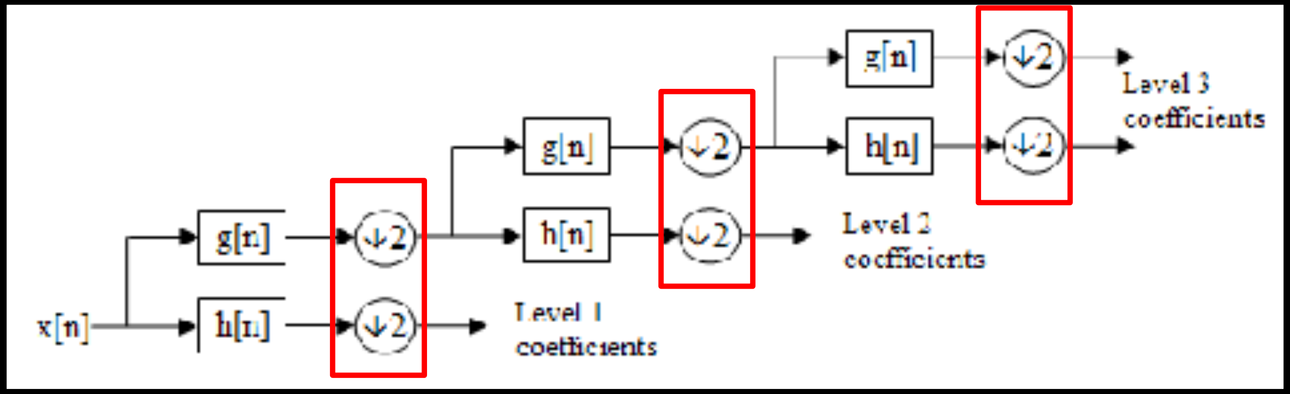
OUTPUT

YOU NOW GET TO BUILD THIS IN 1.5 HOURS!

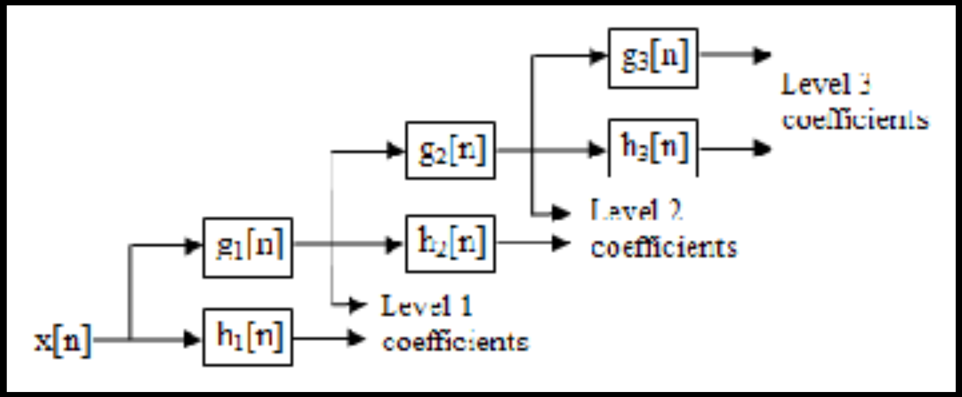
STATIONARY WAVELET TRANSFORM

- Eliminate sub-sampling (illustrated in red box) of the Discrete Wavelet Transform
  - Leads to a highly redundant set of coefficients - not ideal. Use dimensionality reduction after.
  - Advantage is that SWT is "approximately" translationally invariant

Discrete Wavelet Transform



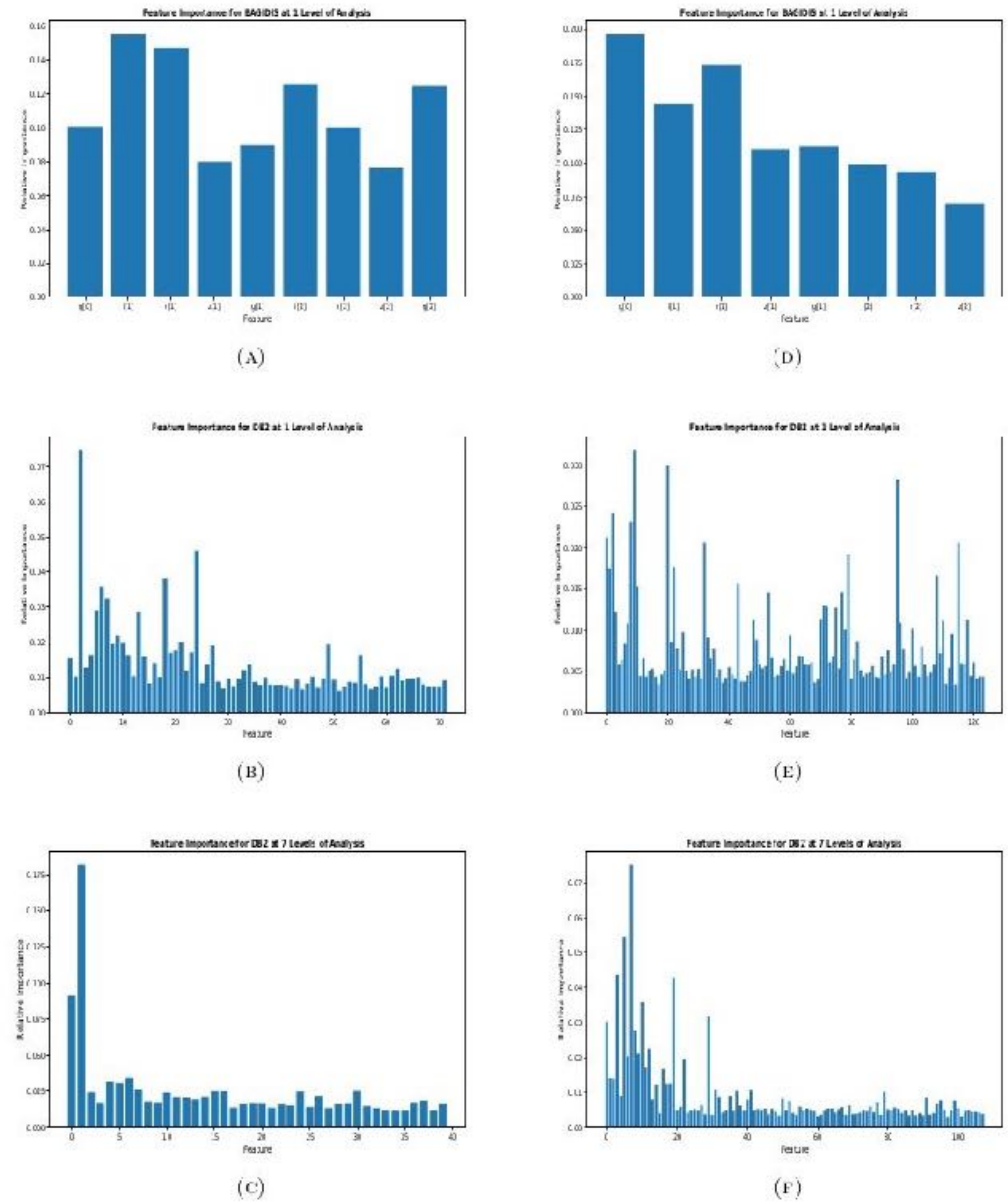
Stationary Wavelet Transform



Faster to compute than BAGIDIS Nlog(N) vs N^2

FEATURE IMPORTANCES

- Use physical intuition (i.e. worry about your data) to decide what base features to compute
    - ML algorithms will faithfully reproduce biases in your training set
- Pick derived features based on what has the highest weight wherever possible
- You can introduce new biases if you do this without thought - example using apparent magnitude as a feature



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