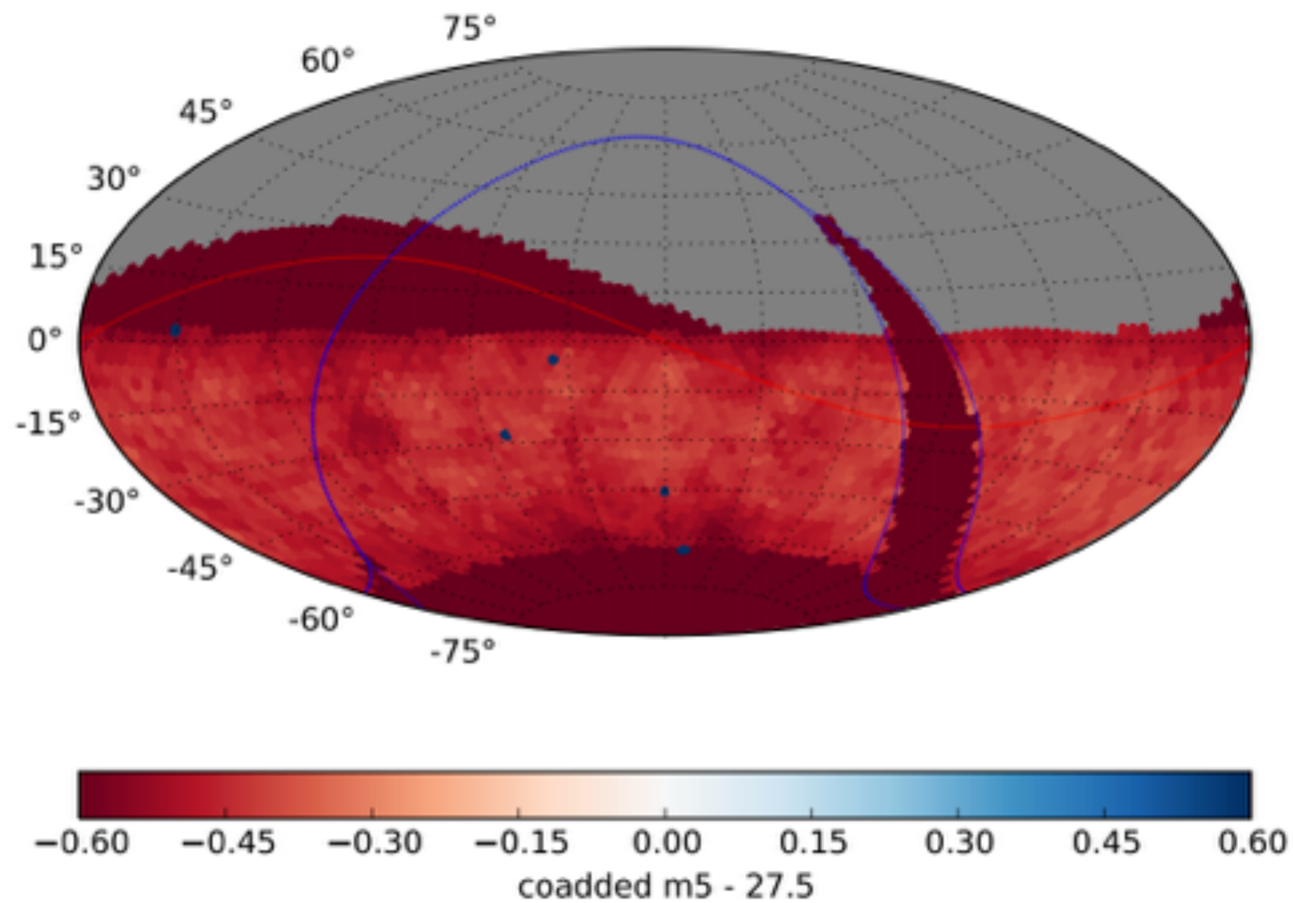
Challenges for Real-time and Archival Time domain astronomy



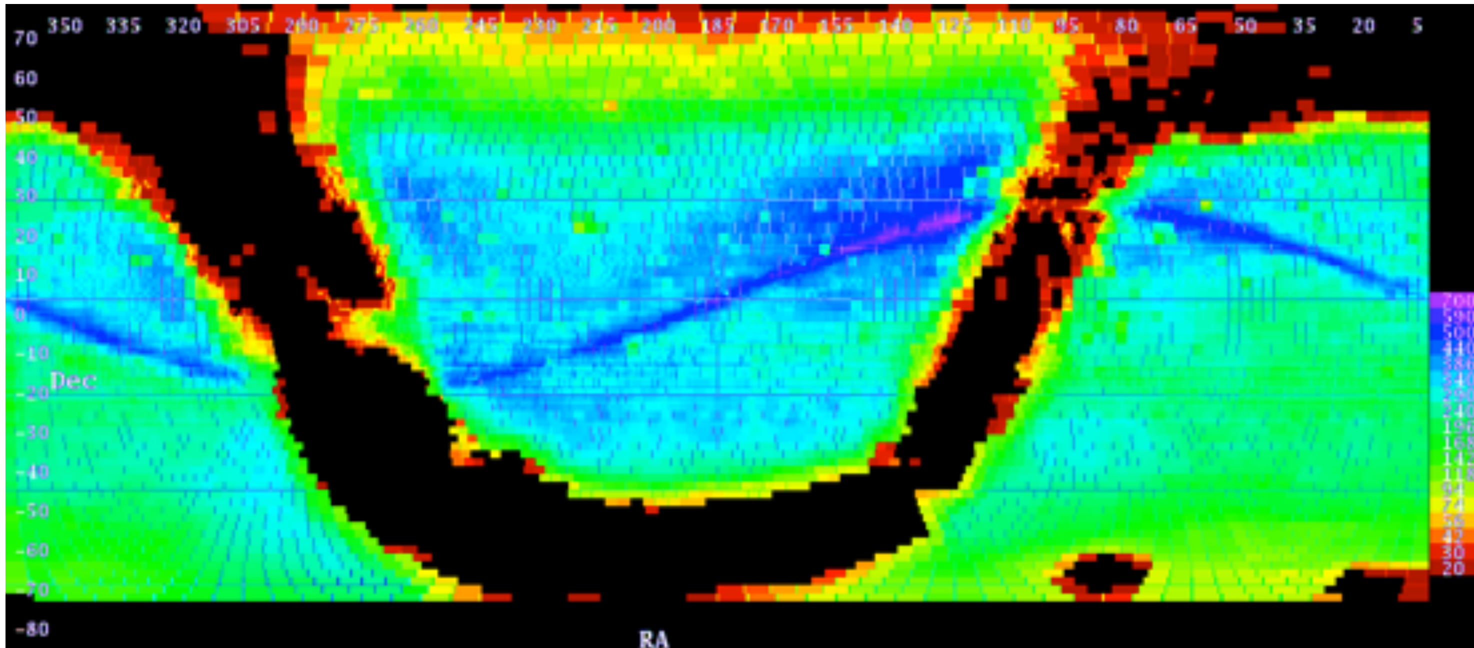
Ashish Mahabal
Center for Data-Driven Discovery, Caltech
ICTS-SAMSI meeting, 20 Mar 2017

Outline

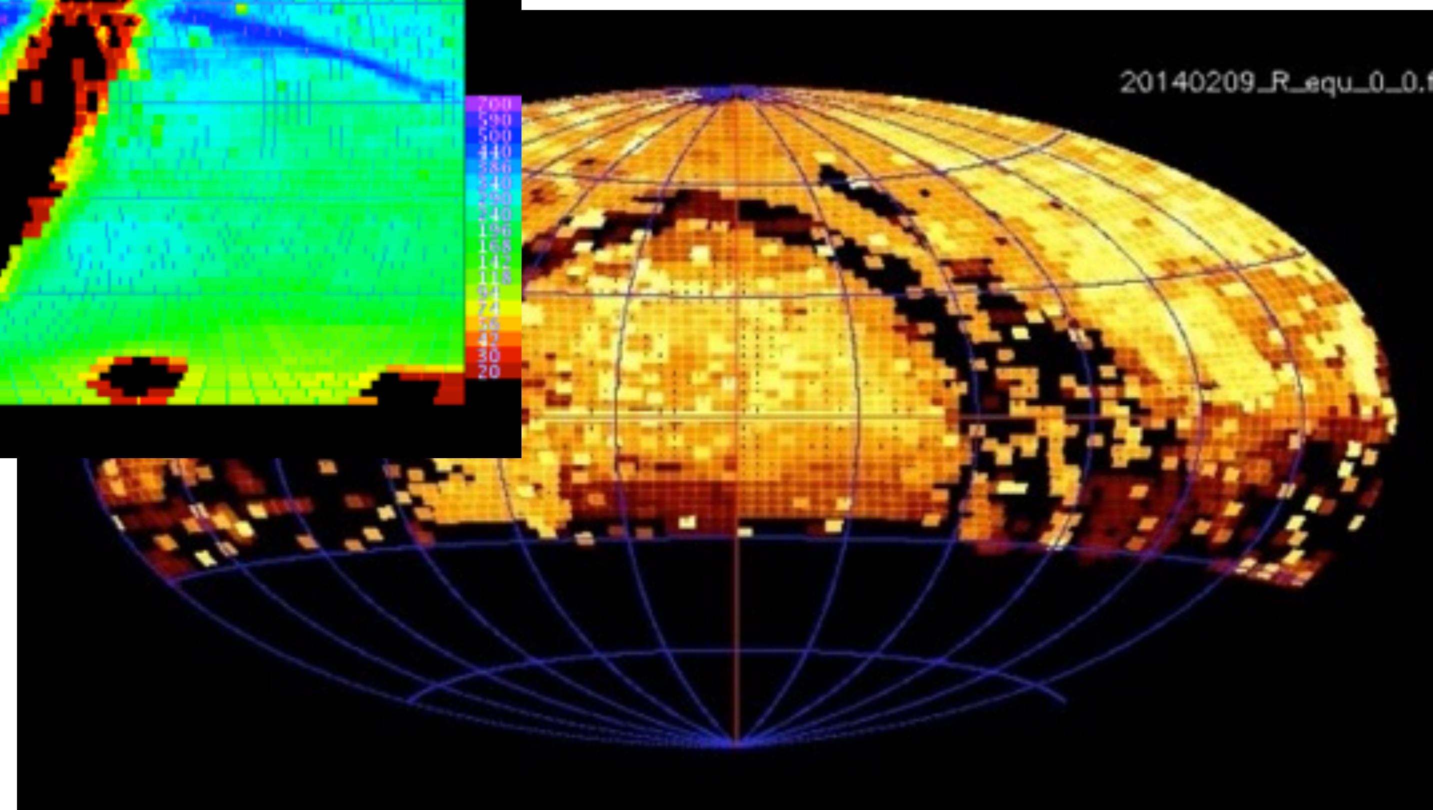
- Overview of surveys and data-holdings
- Data types: light-curves, images (e.g asteroid streaks), ancillary data, spectra
- Variability searches (Stetson J nice, but per area tuning may be needed)
- Period finding
- Combining surveys
- Combining classifiers

From snapshots to (slow) movies of the sky

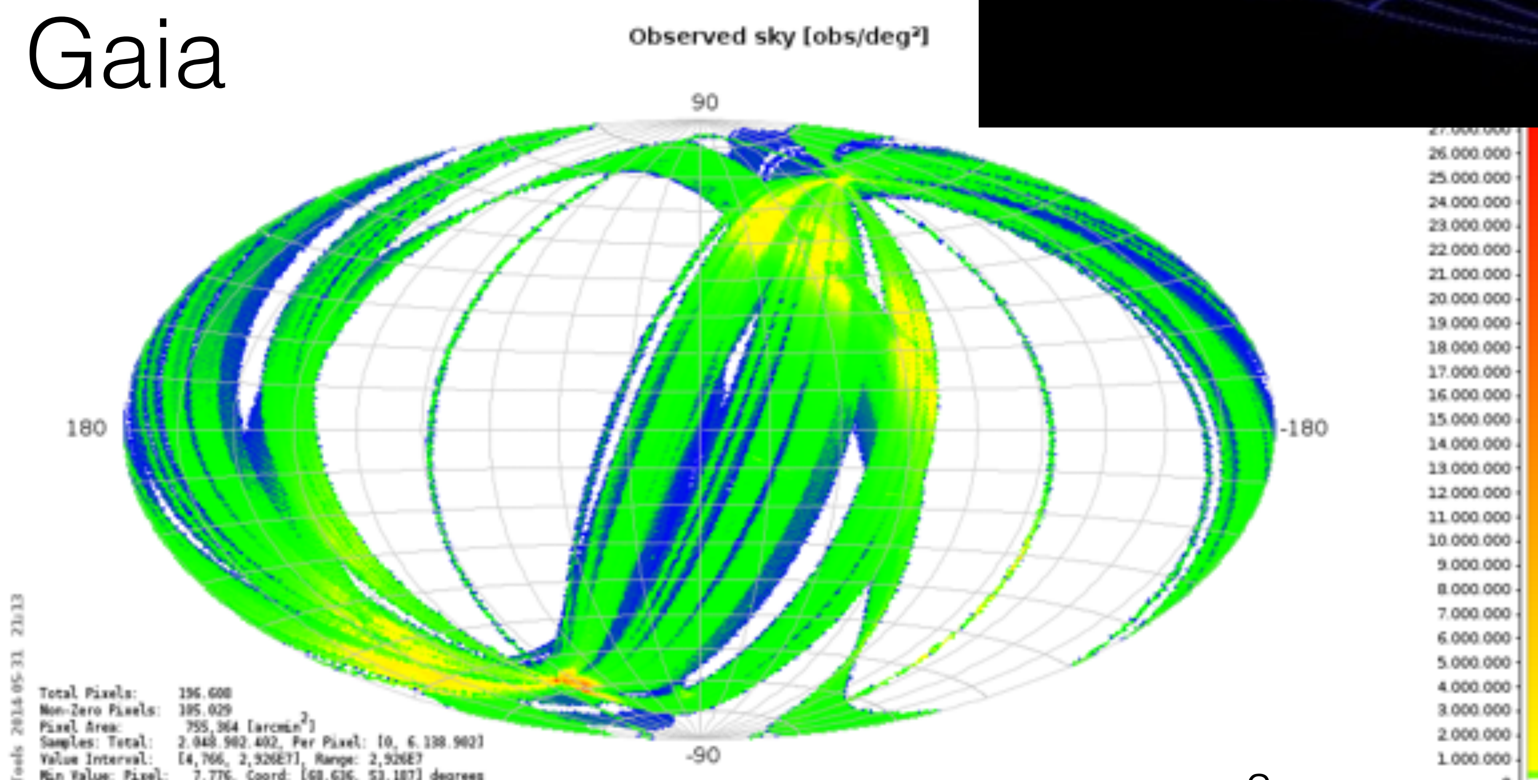
CRTS

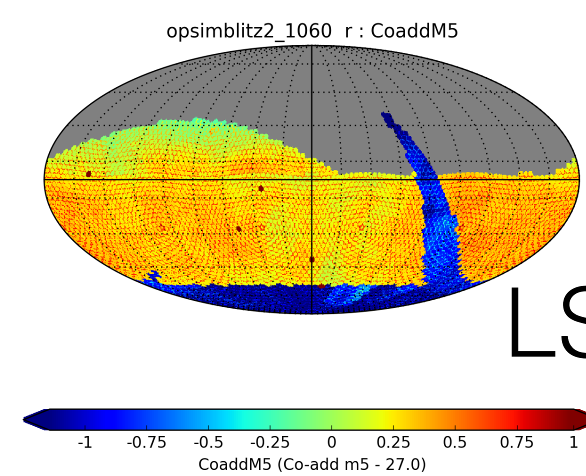


PTF



Gaia





LSST

What do survey's do?


- Pick low-hanging fruit
  - select best objects, easy science
  - get spectroscopy
- That does push the envelope
  - but also leaves gaps

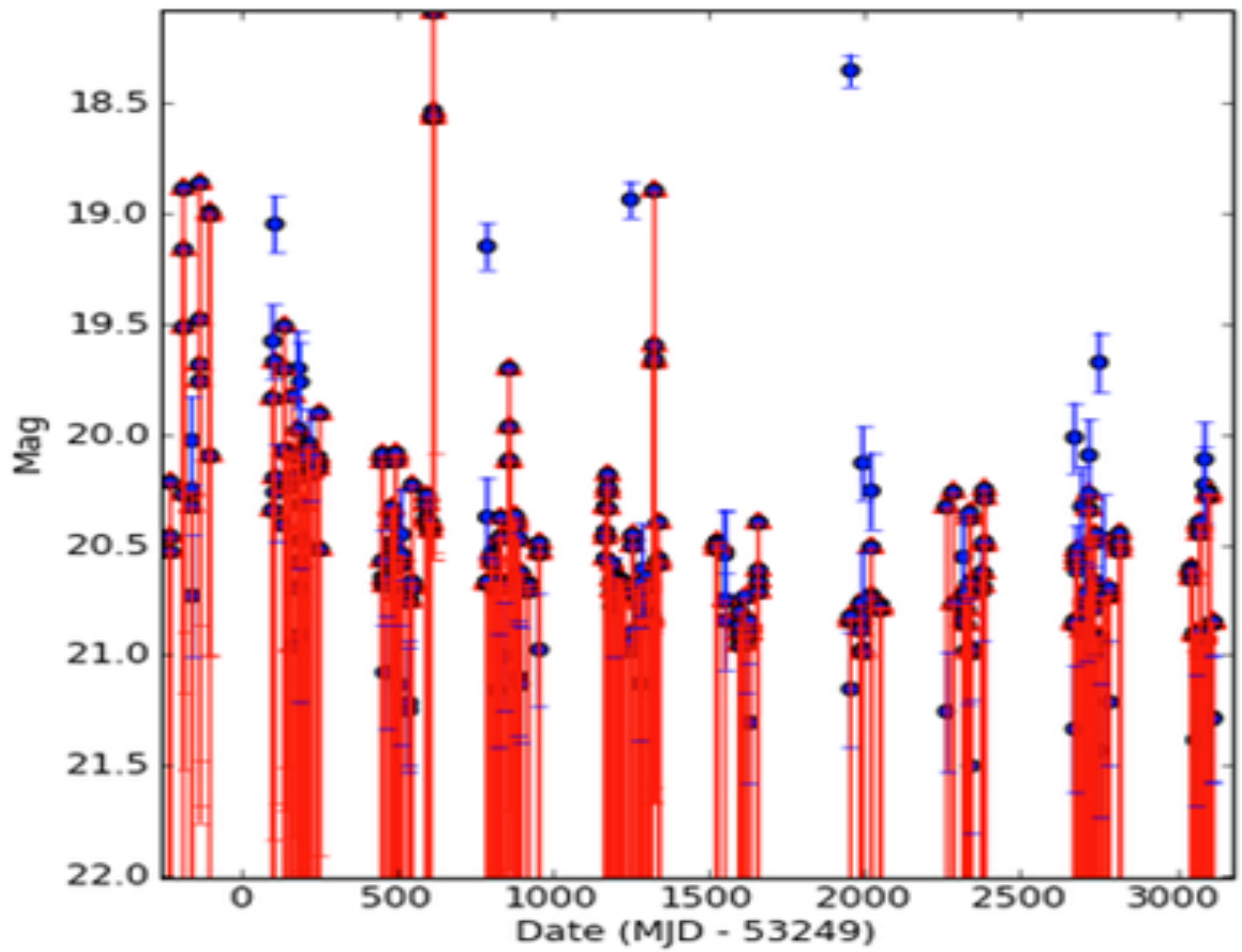
LSST:
1000 30-sec epochs
10 years
3*10^4/3*10^8
1mm in 10m

Today's holdings

- CRTS (public, open filter, lumpy cadence for asteroids):
  - ~500M light-curves over > 12 years
  - ~1B cutouts; coadds
- PTF (partly public; multi-filter, mixed cadence):
  - few hundred M lc over few years
  - Dataset with streaks for near-by asteroids
- Pan-STARRS/Gaia (multi-filter): small data releases
- Kepler (small area, non-sparse)
- Pulsar Timing arrays …

Properties of light-curves

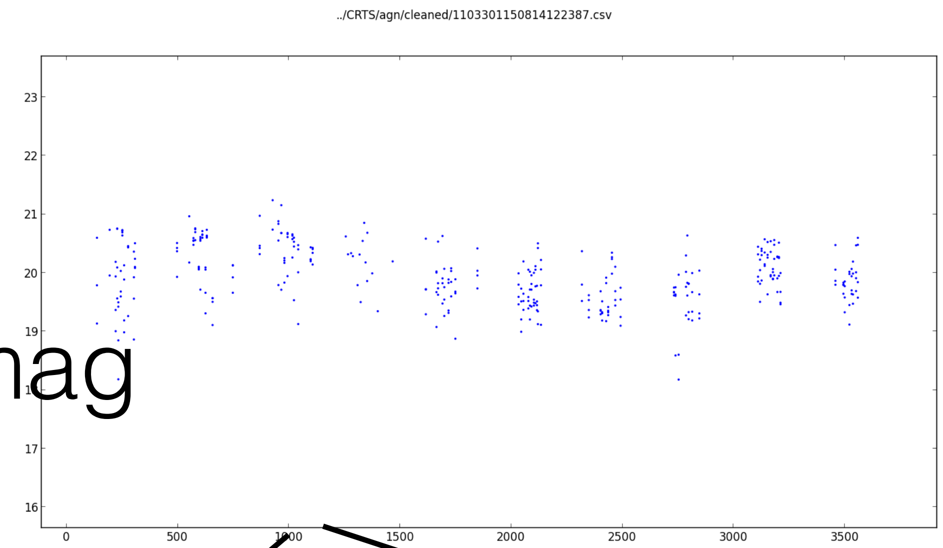
- Gappy
- Irregular
- Heteroskedastic



Reasons:

- **expense, rotation/revolution of Earth, moon**
- **science objectives, weather, moon**
- **weather, moon, airmass**

errors ignored by many methods

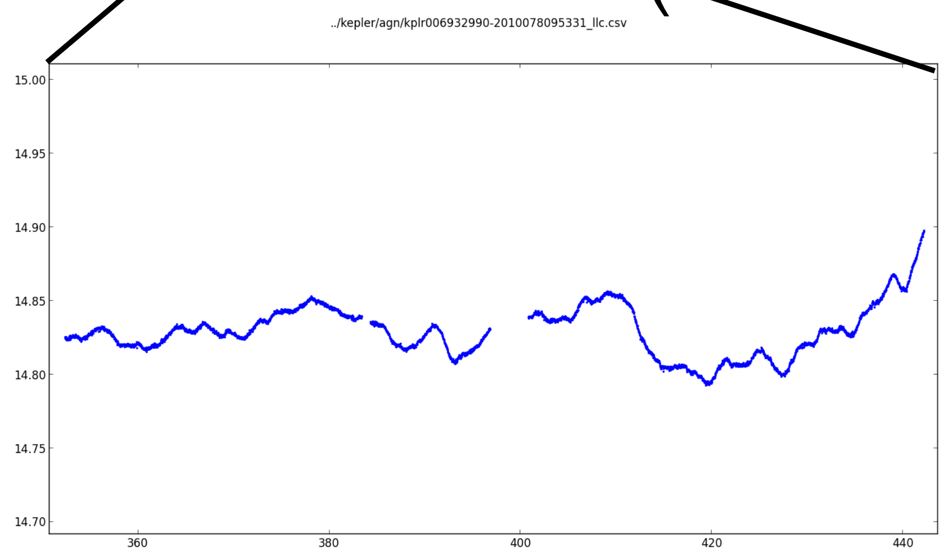


mag

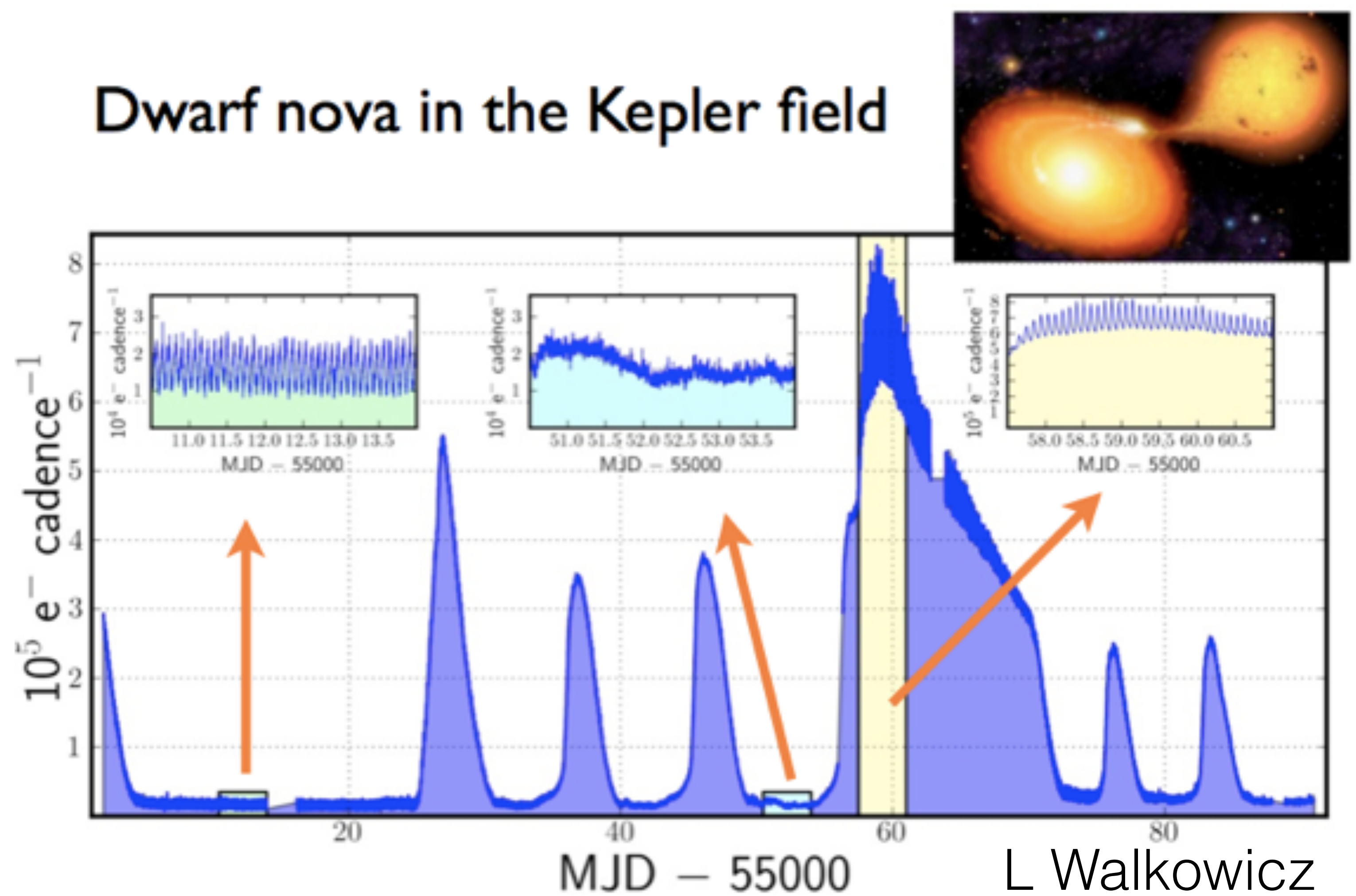
CRTS

Kepler - small area non-sparse

Time (3000+ days)



~100 days



L Walkowicz

Statistical characteristics

Richards et al. (non-sparse OGLE-Hipparcos time-series) 2011

skew

small_kurtosis

std

beyond1std

stetson_j

stetson_k

max_slope

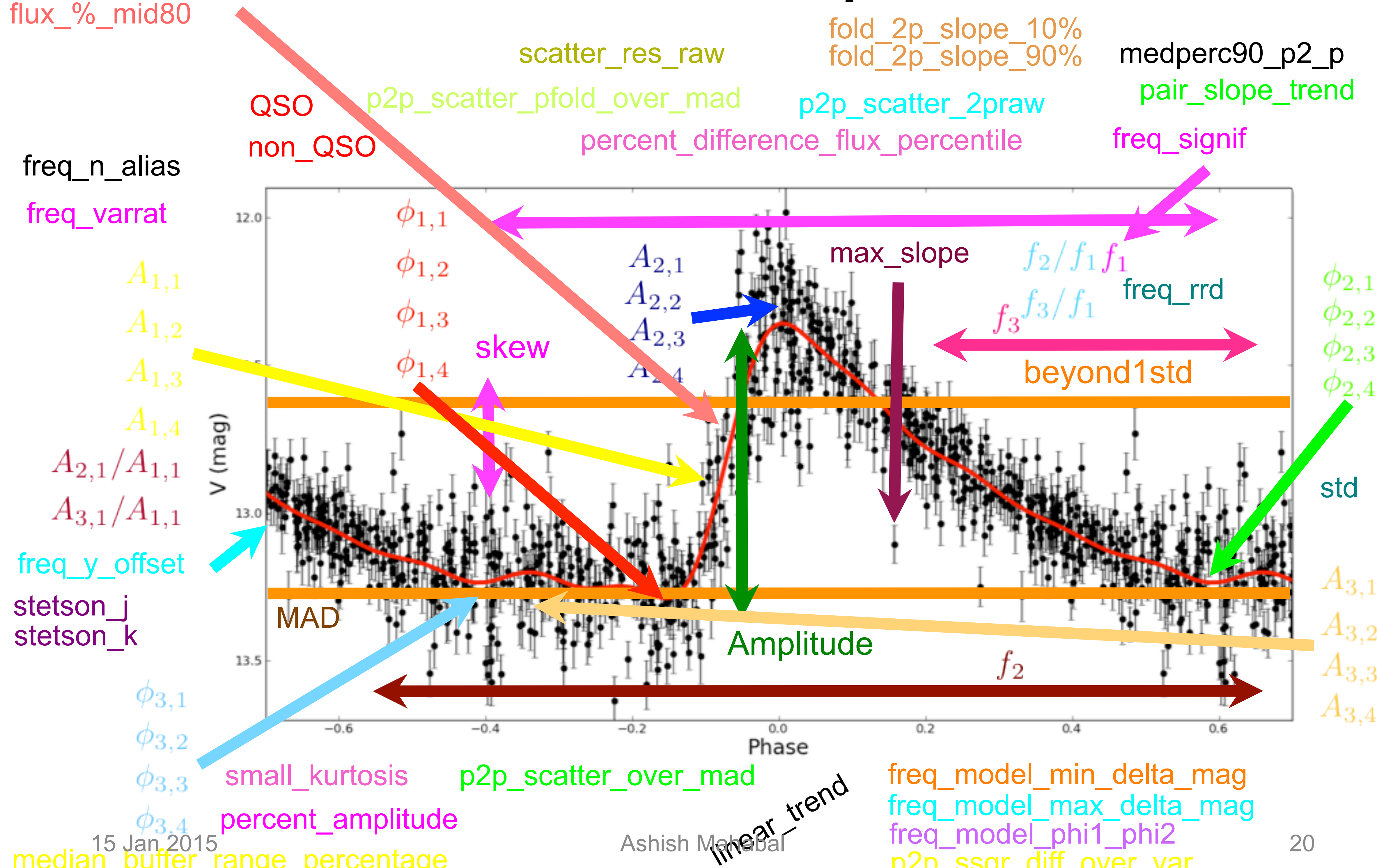
amplitude

Short name	Data type	Summary
amplitude	float	0.5 * (mag_{max} - mag_{min})
beyond1std	float	p(|(mag - <mag>)| > σ)
flux_percentile_ratio_mid20	float	($flux_{60}$ - $flux_{40}$) / ($flux_{95}$ - $flux_5$)
flux_percentile_ratio_mid35	float	($flux_{67.5}$ - $flux_{32.5}$) / ($flux_{95}$ - $flux_5$)
flux_percentile_ratio_mid50	float	($flux_{75}$ - $flux_{25}$) / ($flux_{95}$ - $flux_5$)
flux_percentile_ratio_mid65	float	($flux_{82.5}$ - $flux_{17.5}$) / ($flux_{95}$ - $flux_5$)
flux_percentile_ratio_mid80	float	($flux_{90}$ - $flux_{10}$) / ($flux_{95}$ - $flux_5$)
linear_trend	float	b where mag = a * t + b
max_slope	float	max(|(mag_{i+1}-mag_i)/(t_{i+1}-t_i)|)
mad	float	med(flux - $flux_{med}$)
median_buffer_range_percentage	float	p(|flux - $flux_{med}$| < 0.1 * $flux_{med}$)
pair_slope_trend	float	p($flux_{i+1}$ - $flux_i$ > 0; i = n-30, n)
percent_amplitude	float	max(|f_{max} - f_{med}|, |f_{min} - f_{med}|)
pdfp	float	($flux_{95}$ - $flux_5$) / $flux_{med}$
qso	4x1	var_{qso}
skew	float	μ_3/σ^3
small_kurtosis	float	μ_4/σ^4
std	float	σ
stetson_j	float	var_j(mag)
stetjon_k	float	var_k(mag)

lightcurve characterization service: http://nirgun.caltech.edu:8000/

Many features
- not all are independent

Adam Miller



Stetson Stats

Welch-Statson
1996PASP..108..851S

$$I=\sqrt{\frac{1}{n(n-1)}}\sum_{i=1}^{n}\left(\frac{b_i-\bar{b}}{\sigma_{b,i}}\right)\left(\frac{v_i-\bar{v}}{\sigma_{v,i}}\right),$$

Pairwise observations in 2 filters

$$J=\frac{\Sigma_{k=1}^{n} w_k \operatorname{sgn}(P_k)\sqrt{|P_k|}}{\Sigma_{k=1}^{n} w_k},$$

Pairwise observations (single filter)

$$K=\frac{1/N \ \ \Sigma_{i=1}^{N}|\delta_i|}{\sqrt{1/N \ \ \Sigma_{i=1}^{N}\delta_i^2}},$$

No pairing required

$$L=\left(\frac{JK}{0.798}\right)\left(\frac{\Sigma w}{w_{\mathrm{all}}}\right)$$

Combined for thresholding

CRTS variables

Drake et al. 2014

- 150M sources from a few thousand “fields”
- ~5.5M variables after filtering using per field J
- ~50K periodic (LS False Alarm Probability < 10^-5; M_t thresholds)
    - 15 classes

M_t: Fraction of time below median (Kinemuchi et al. 2006)

with Djorgovski (PI), Drake (PI), Graham, Donalek

Q: Amplitude variations

$$Q = \frac{(\mathrm{RMS}^2_{\mathrm{resid}} - \sigma^2)}{(\mathrm{RMS}^2_{\mathrm{raw}} - \sigma^2)}, \tag{6}$$

where $\mathrm{RMS}_{\mathrm{raw}}$ and $\mathrm{RMS}_{\mathrm{resid}}$ are the RMS values of the raw light curve and the phase subtracted light curve, respectively, whereas σ is the estimated uncertainty including the systematics (e.g., Section 3.3). Testing on si-

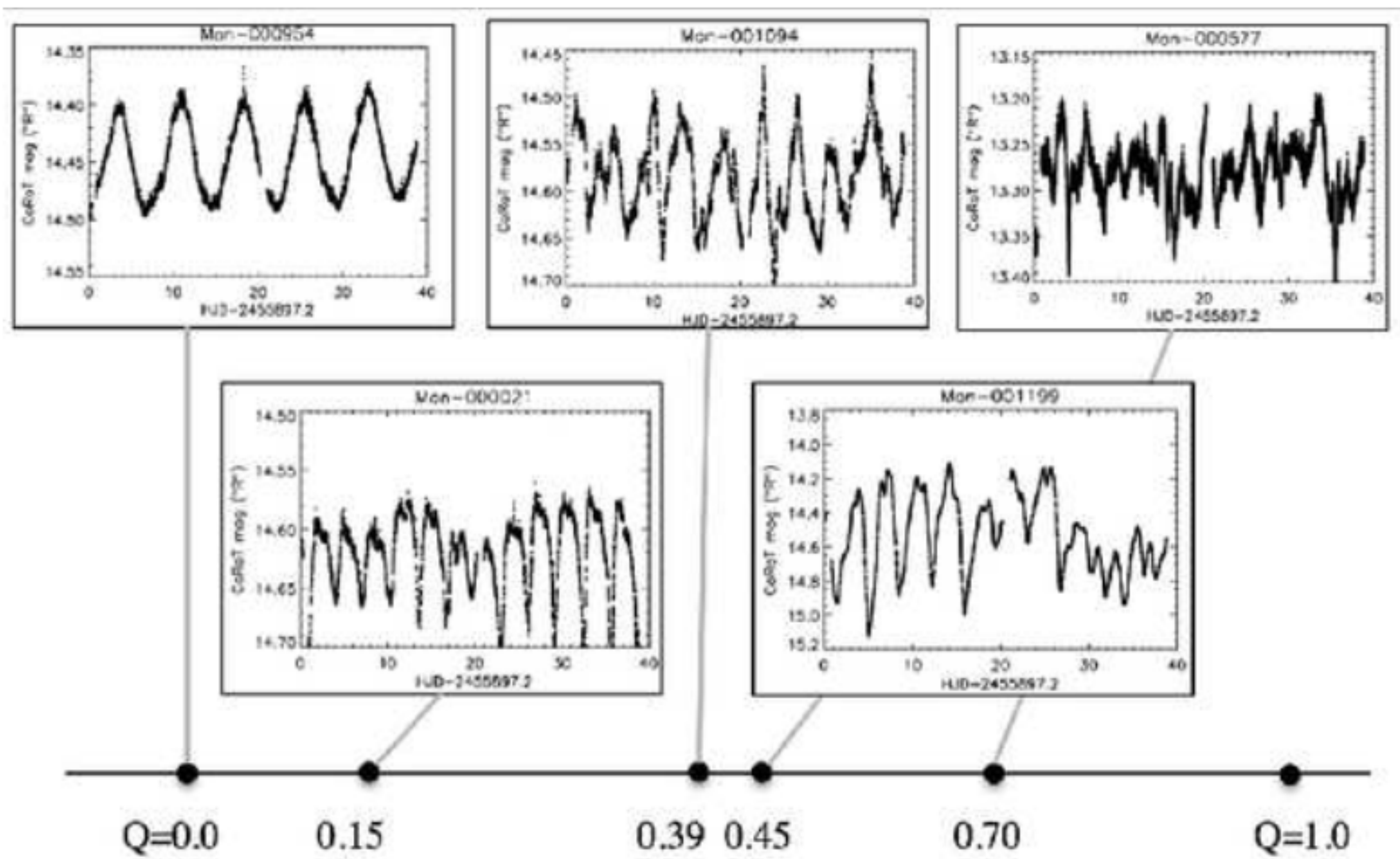


FIG. 29.— *CoRoT* light curves with representative values of the Q parameter, ranging from periodic (Q=0–0.15) to quasi-periodic (Q=0.15–0.5), to aperiodic $Q > 0.5$.

Cody et al. 2014

$$M = (< d_{10\%} > - d_{\mathrm{med}})/\sigma_d, \tag{7}$$

M: Bursters and dippers

where $< d_{10\%} >$ is the mean of all data at the top and bottom decile of light curve, d_{med} is the median of the entire light curve, and σ_d is its overall RMS.

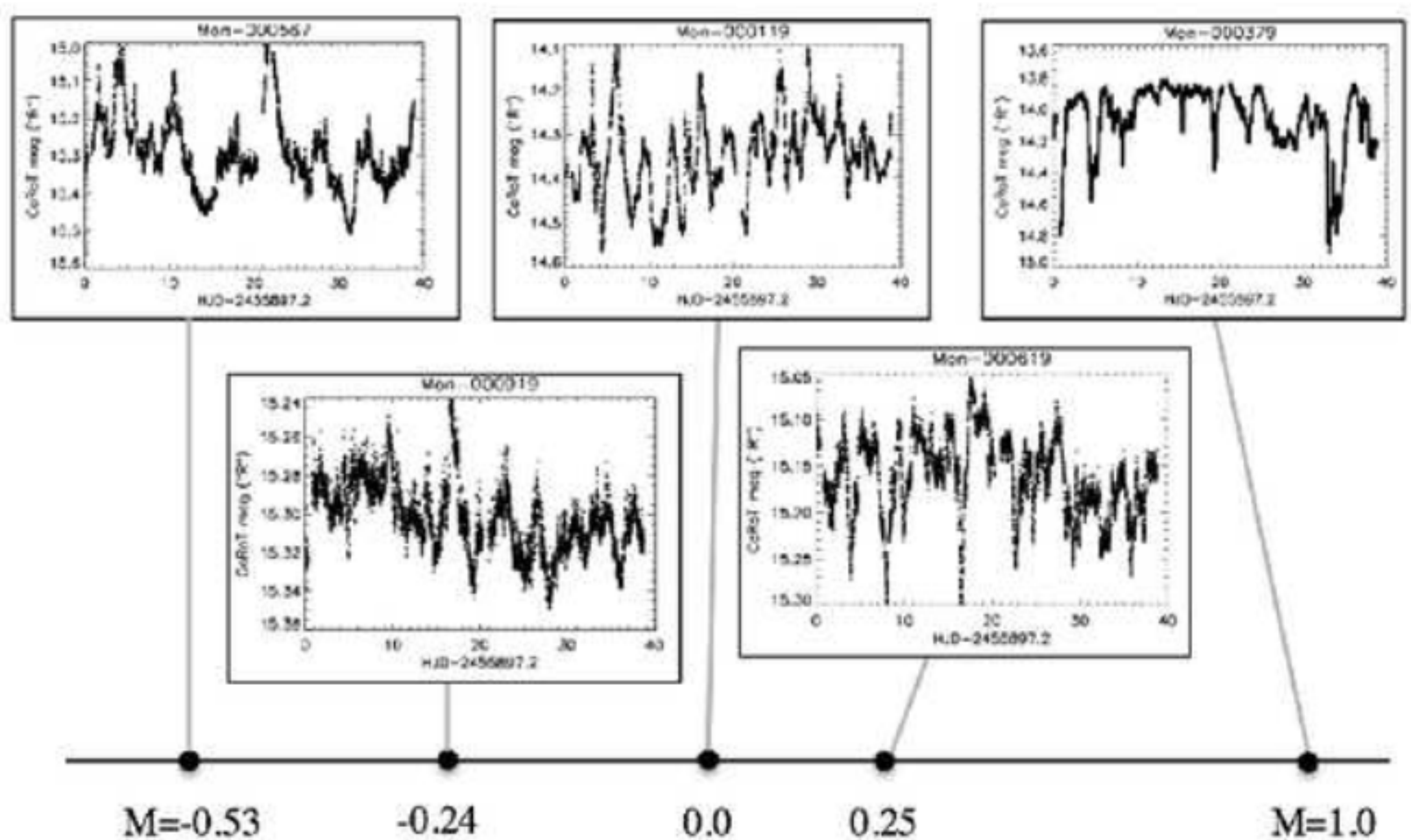


FIG. 30.— *CoRoT* light curves with representative values of the M parameter, ranging from bursting ($M < -0.25$) to symmetric (M=-0.25–0.25), to dipping $M > 0.25$.

Cody et al. 2014

Q-M plane

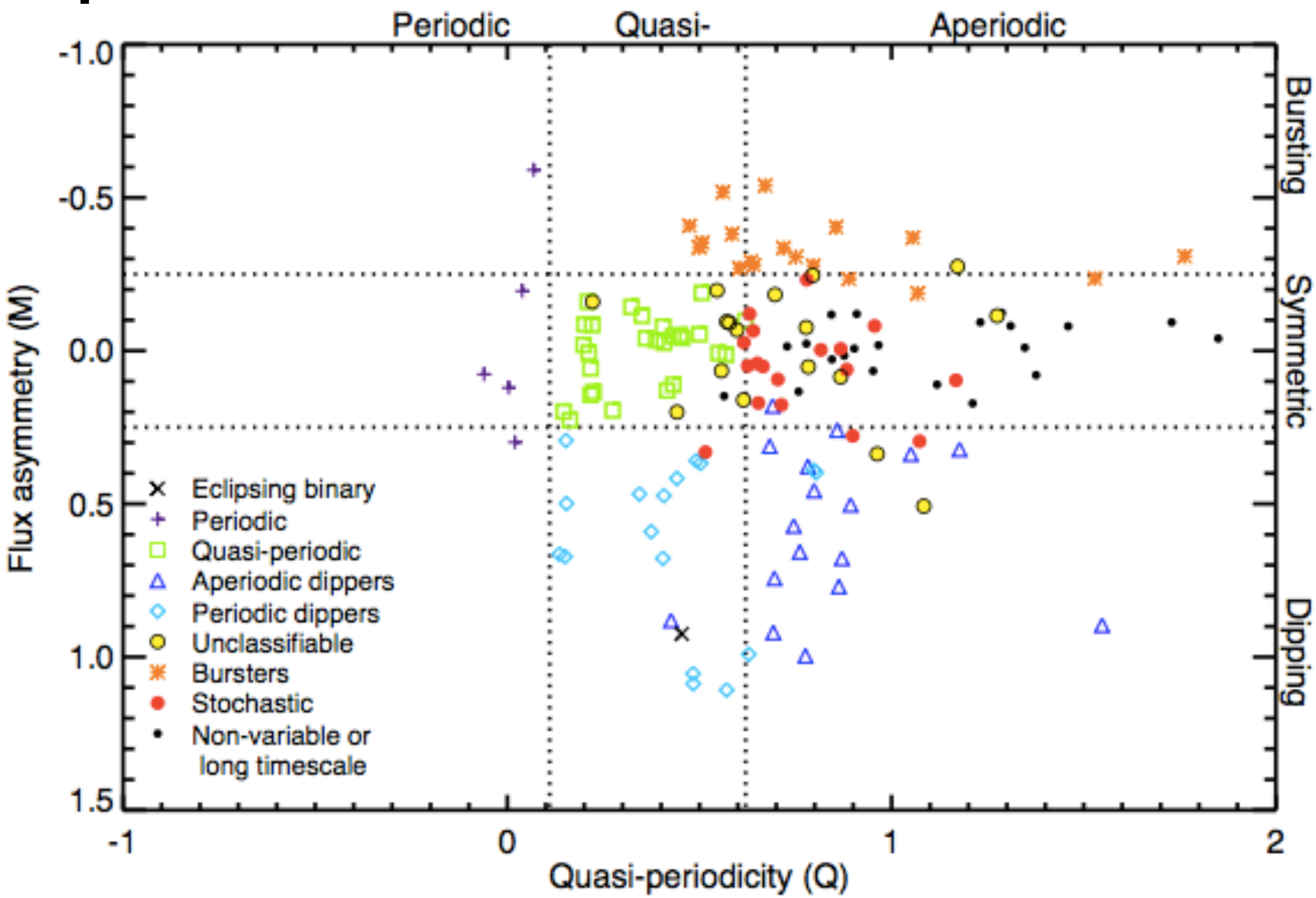
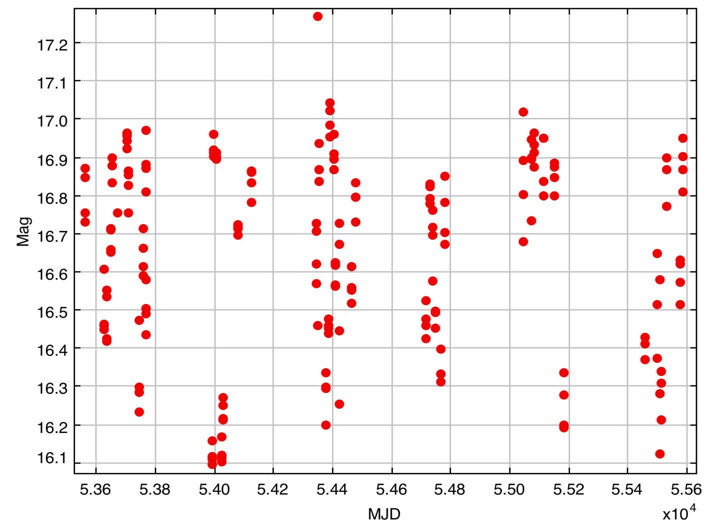


FIG. 31.— *Top*: Light curve morphology classes, as divided by the quasi-periodicity (Q) and flux asymmetry (M) parameters for optical light curves from *CoRoT* in our disk-bearing sample. Color coding indicates the variability classification chosen by eye, before statistical assessment. The eclipsing binary is not strictly periodic because its light curve contains aperiodic fluctuations out of eclipse. *Bottom*: Same

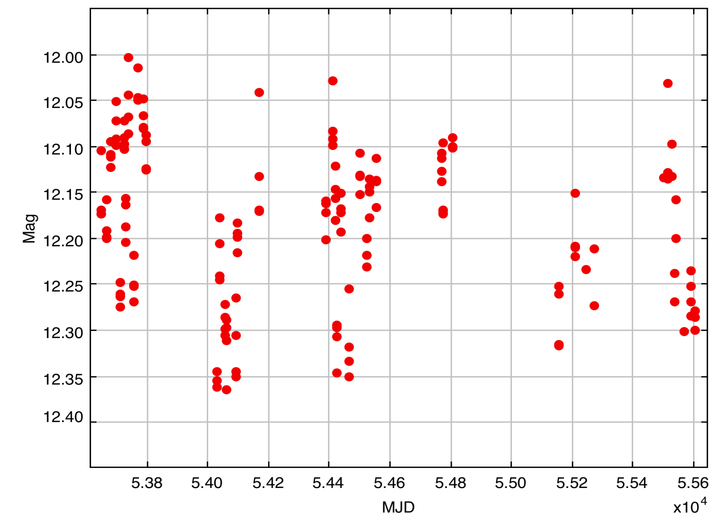
Cody et al. 2014

Features for RR Lyrae and W UMa

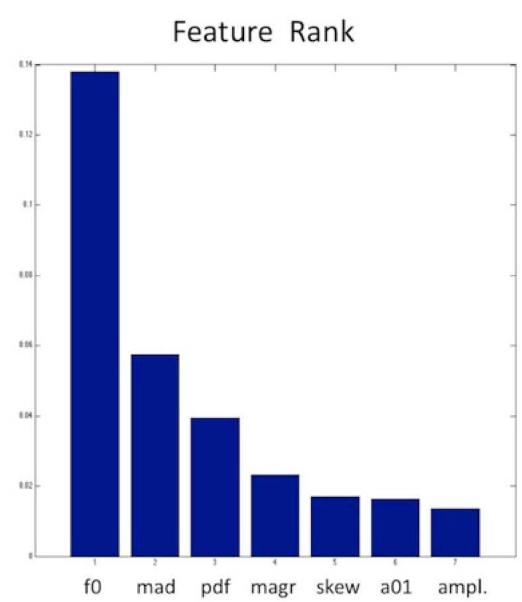
Rank features in the order of classification quality for a given classification problem, e.g., RR Lyrae vs. WUMa

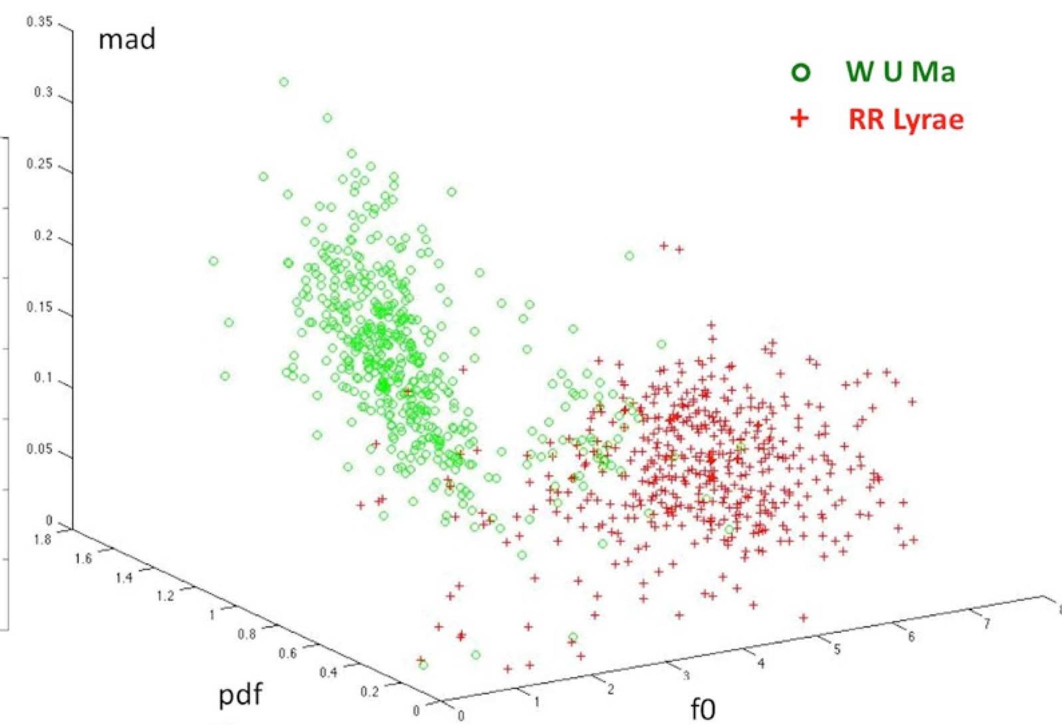


RR Lyrae



Eclipsing binary (W U Ma)

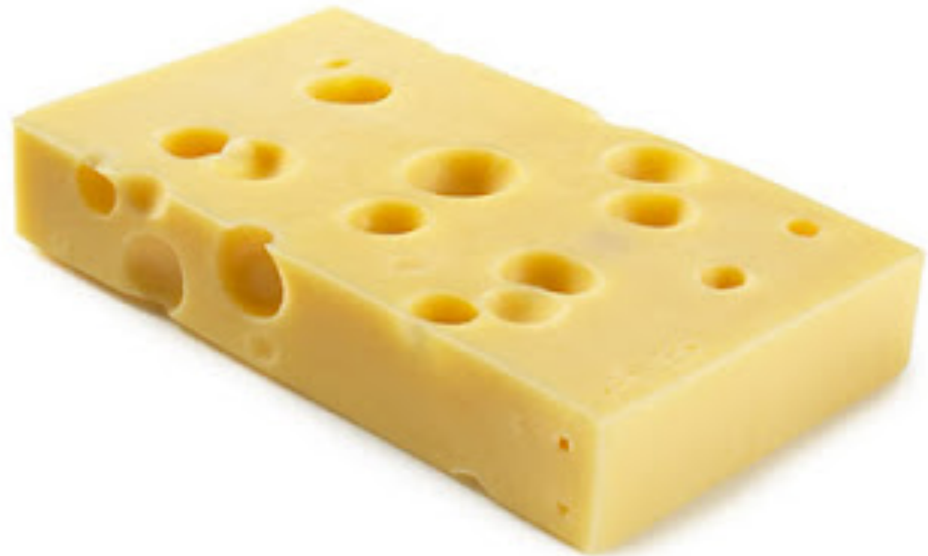




Challenge: A Variety of Parameters

- Discovery: magnitudes, delta-magnitudes
- Contextual:
  - Distance to nearest star
  - Magnitude of the star
  - Color of that star
  - Normalized distance to nearest galaxy
  - Distance to nearest radio source
  - Flux of nearest radio source
  - Galactic latitude
- Follow-up
  - Colors (g-r, r-I, i-z etc.)
- Prior classifications (event type)
- **Characteristics from light-curve**
  - **Amplitude**
  - **Median buffer range percentage**
  - **Standard deviation**
  - **Stetson k**
  - **Flux percentile ratio mid80**
  - **Prior outburst statistic**

Not all parameters are always present leading to swiss-cheese like data



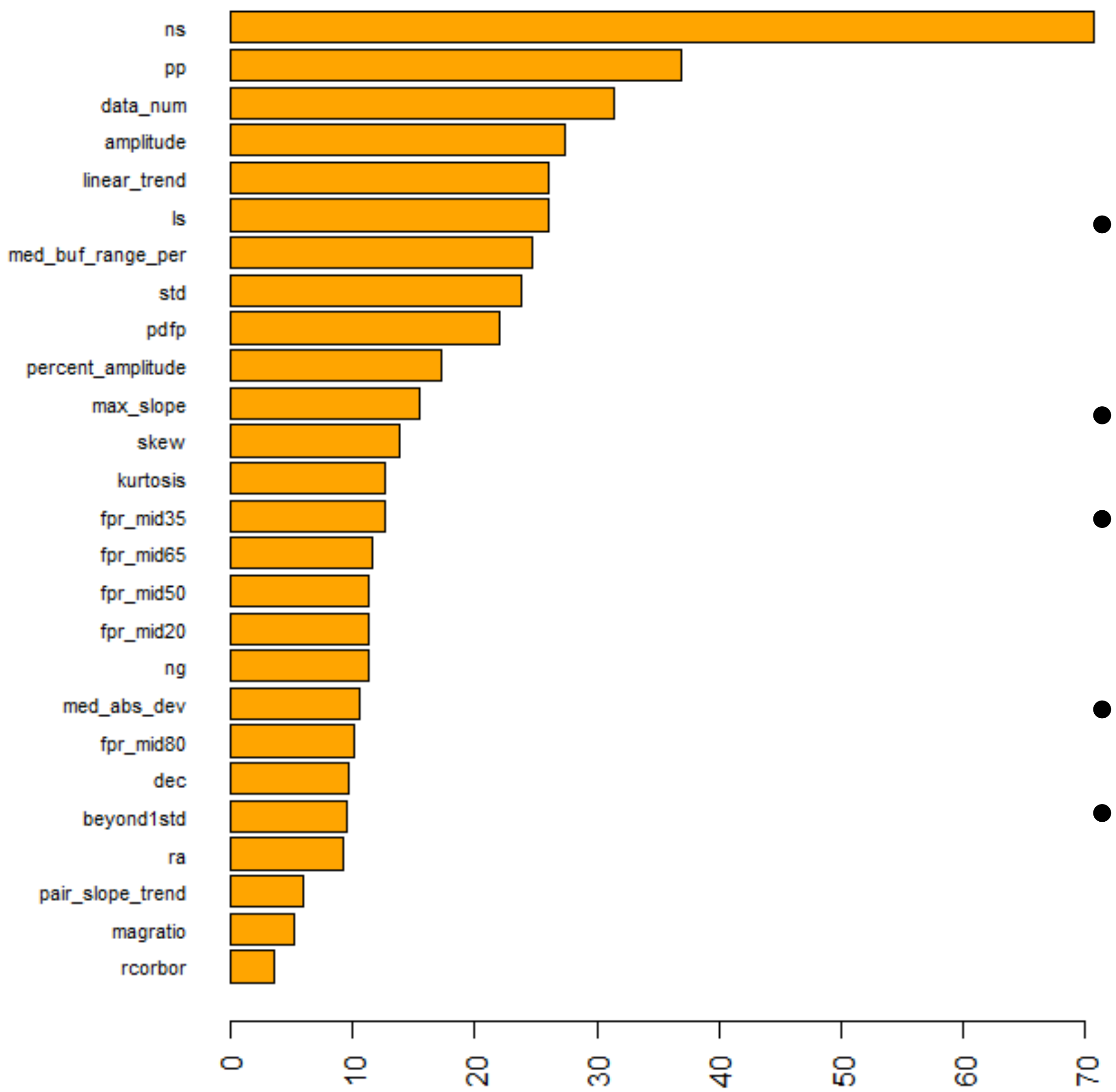
http://ki-media.blogspot.com/

Measures from Feigelson and Babu (Graham)

New lightcurve-based parameters: (Faraway)

•Whole curve measures
•Fitted curve measures
•Residual from fit measures
•Cluster measures
•Other

Feature selection strategy



- Fast Relief Algorithm (wt and threshold)
- Fisher Discriminant Ratio
- Correlation based Feature Selection
- Fast Correlation Based Filter
- Multi Class Feature Selection

Donalek, .., Mahabal, … arxiv:1310.1976

A variety of parameters - choose judiciously

Discovery; Contextual; Follow-up; Prior Classification …

Whole curve measures

Median magnitude (mag); mean of absolute differences of successive observed
magnitude; the maximum difference magnitudes

Fitted curve measures

Scaled total variation scaled by number of days of observation; range of fitted curve;
maximum derivative in the fitted curve

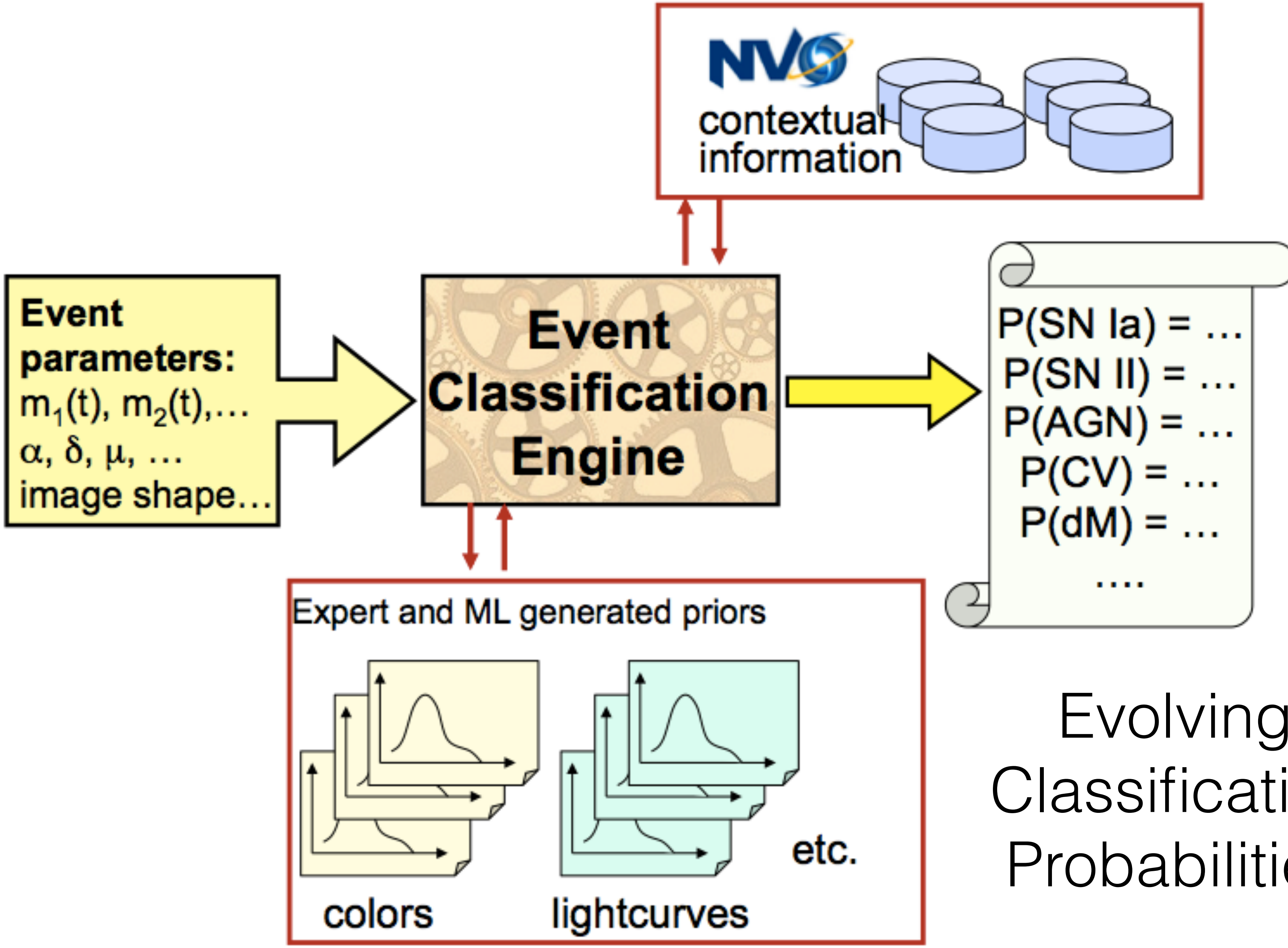
Residual from fit measures

The maximum studentized residual; SD of residuals; skewness of residuals;
Shapiro-Wilk statistic of residuals

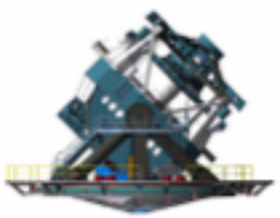
Cluster measures

Fit the means within the groups (up to 4 measurements); and then take the
logged SD of the residuals from this fit; the max absolute residuals from this fit;
total variation of curve based on group means scaled by range of observation

A few years ago …



Evolving Classification Probabilities



LSST data volume and scientific yields

Raw Data

- Two 6.4-gigabyte images (one visit) every 39 seconds (15TB per night)
- ~1000 visits each night, ~300 nights a year
- Up to 450 calibration exposures per day

Level 1

- Can detect >10 million real time events per night, for 10 years
- Changes detected, transmitted, within 60 seconds of the observation

Level 2

- Observe ~38 billion objects (24B galaxies, 14B stars)
- Collect ~5 trillion observations ("sources") and ~32 trillion measurements ("forced sources") in a 20 PB catalog

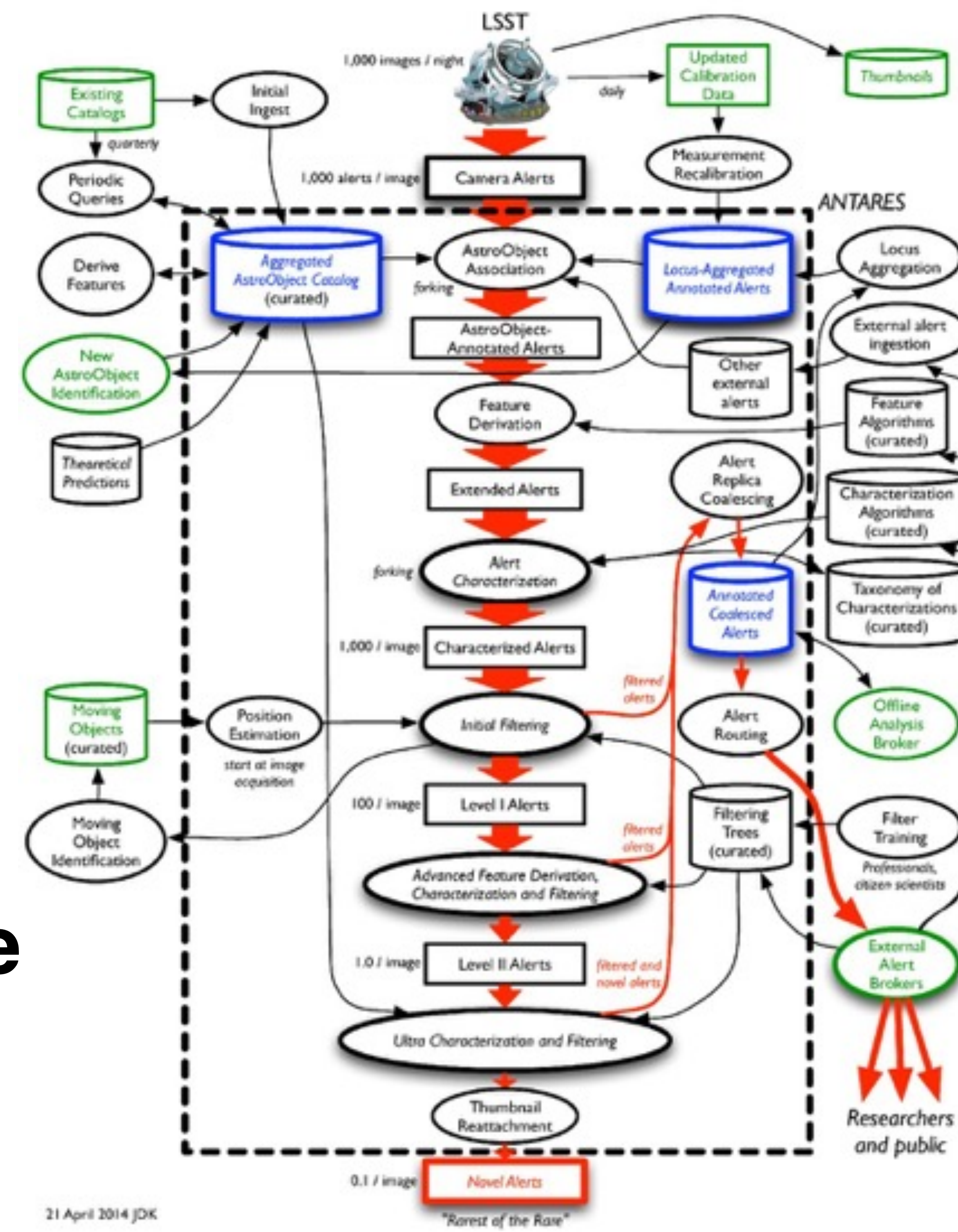
Level 3

- User databases and workspaces ("mydb")
- Making the LSST software available to end-users
- Feeding the data back to the community

Antares:
Prototype broker for LSST

LSST images
far deeper
more filters
equally sparse!

10^7 transients



10^3 rare transients

Saha et al
1409.0056

Challenge: Characterize rare transients quickly with minimal data

LS by Many names

$$\phi(t) = A \sin \omega t + B \cos \omega t + C.$$

sines + cosines < n
generalized version fits for mean
(rather than using mean = 0 through subtraction)

- Lomb-Scargle periodogram is a least squares sinusoid fit (Least Squares Spectral Analysis)
- Matching Persuit

Entropy based period finding

Graham et al. 2013

$$H_0 = -\sum_{i=1}^{k} \mu_i \ln(\mu_i) \quad \forall \mu_i \neq 0,$$

Counts in k-partitions after phasing

1-day aliasing!

$$H_c = \sum_{i,j} p(m_i, \phi_j) \ln\left(\frac{p(\phi_j)}{p(m_i, \phi_j)}\right),$$

Counts in partitions after
phasing in time
and binning in mags

Challenge: better period finding for sparse time series

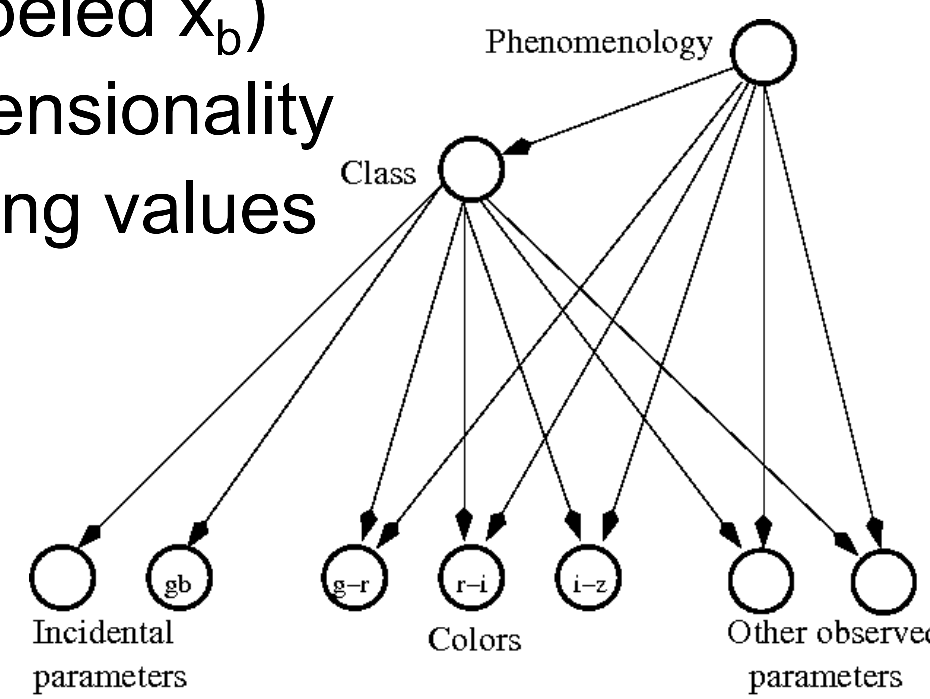
Various classification/clustering methods

- Support Vector Machines
- Self Organizing Maps
- Minimal Spanning Trees
- t-SNE
- Decision trees
- Random Forests
- …

Naïve Bayes

$$P(y=k \mid x) = P(x \mid y=k)P(k)/P(x) \propto P(k)P(x \mid y=k) \approx P(k)\prod_{b=1}^{B} P(x_b \mid y=k)$$

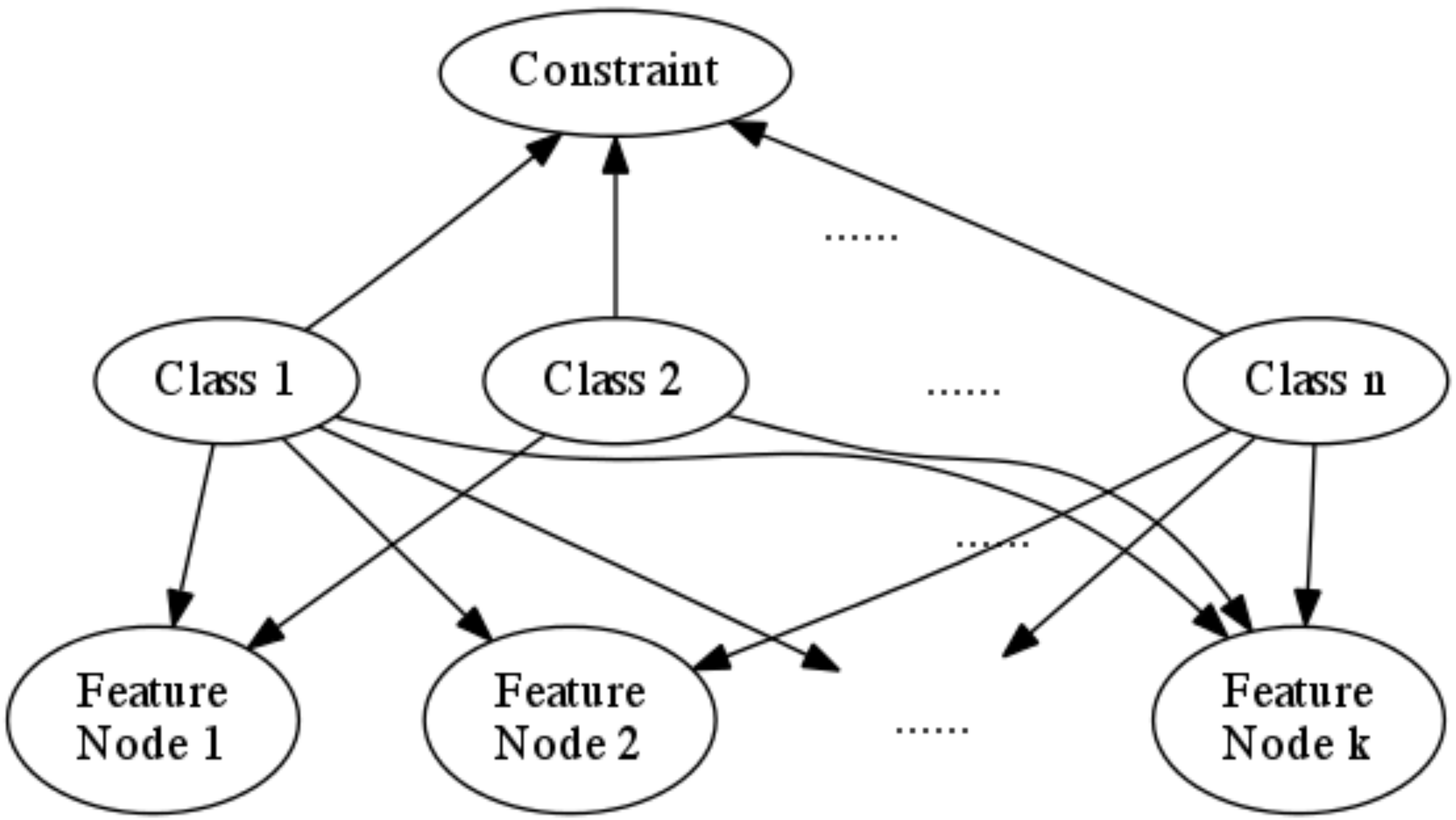
- x: feature vector of event parameters
- y: object class that gives rise to x (1<y<k)
- Certain features of x known: (position, flux)
- Others will be unknown: (color, delta-mag)
- Assumption: based on y, x is decomposable into B distinct independent classes (labeled x_b)
- This helps with the curse of dimensionality
- Also allows us to deal with missing values



BN with P60 follow-up colors:
CV/SN classification ~80% with single epoch

Radio, for instance

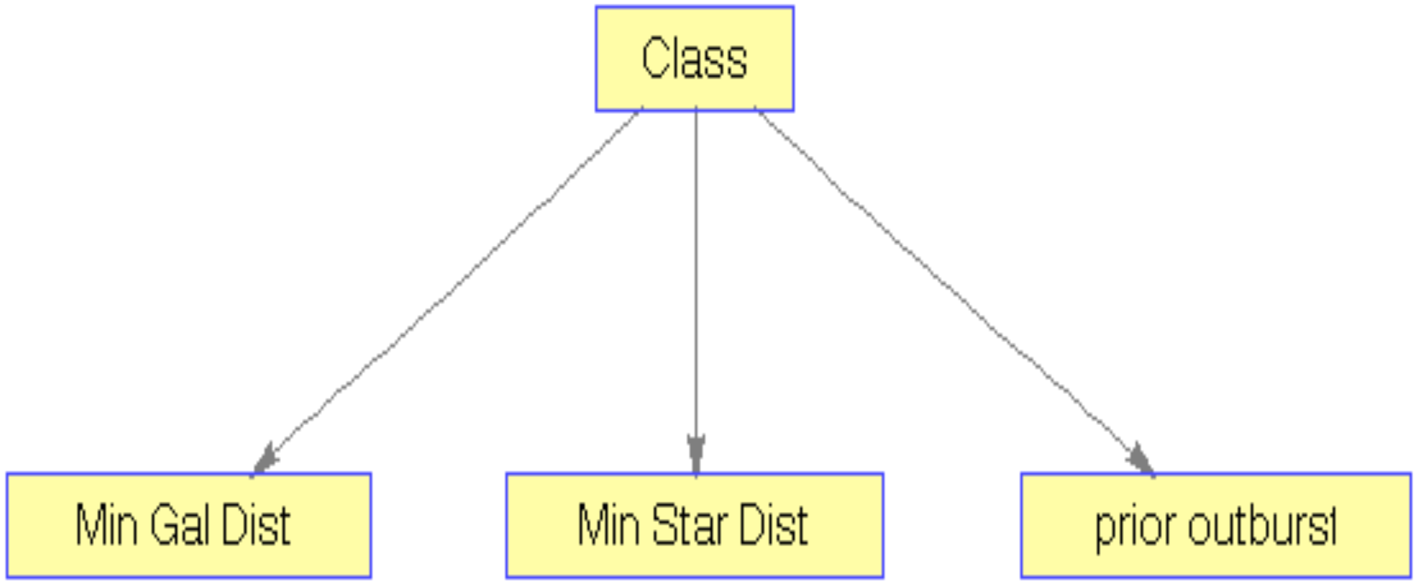
The aim …



n	G(n)
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1.1 x 10^9
8	7.8 x 10^11
9	1.2 x 10^15
10	4.2 x 10^18

As number of parameters grow, search space increases super-exponentially

SN v. non-SN

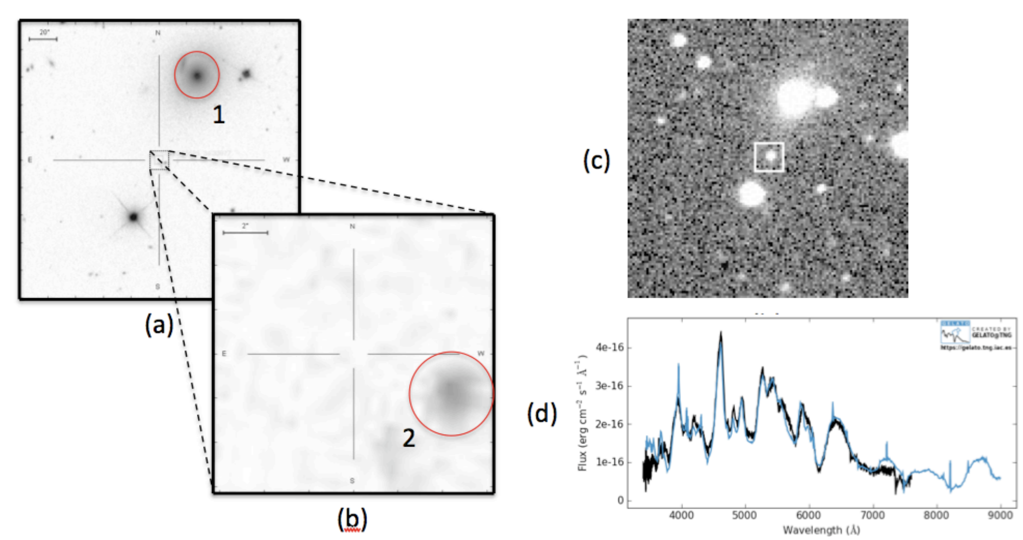


normalized

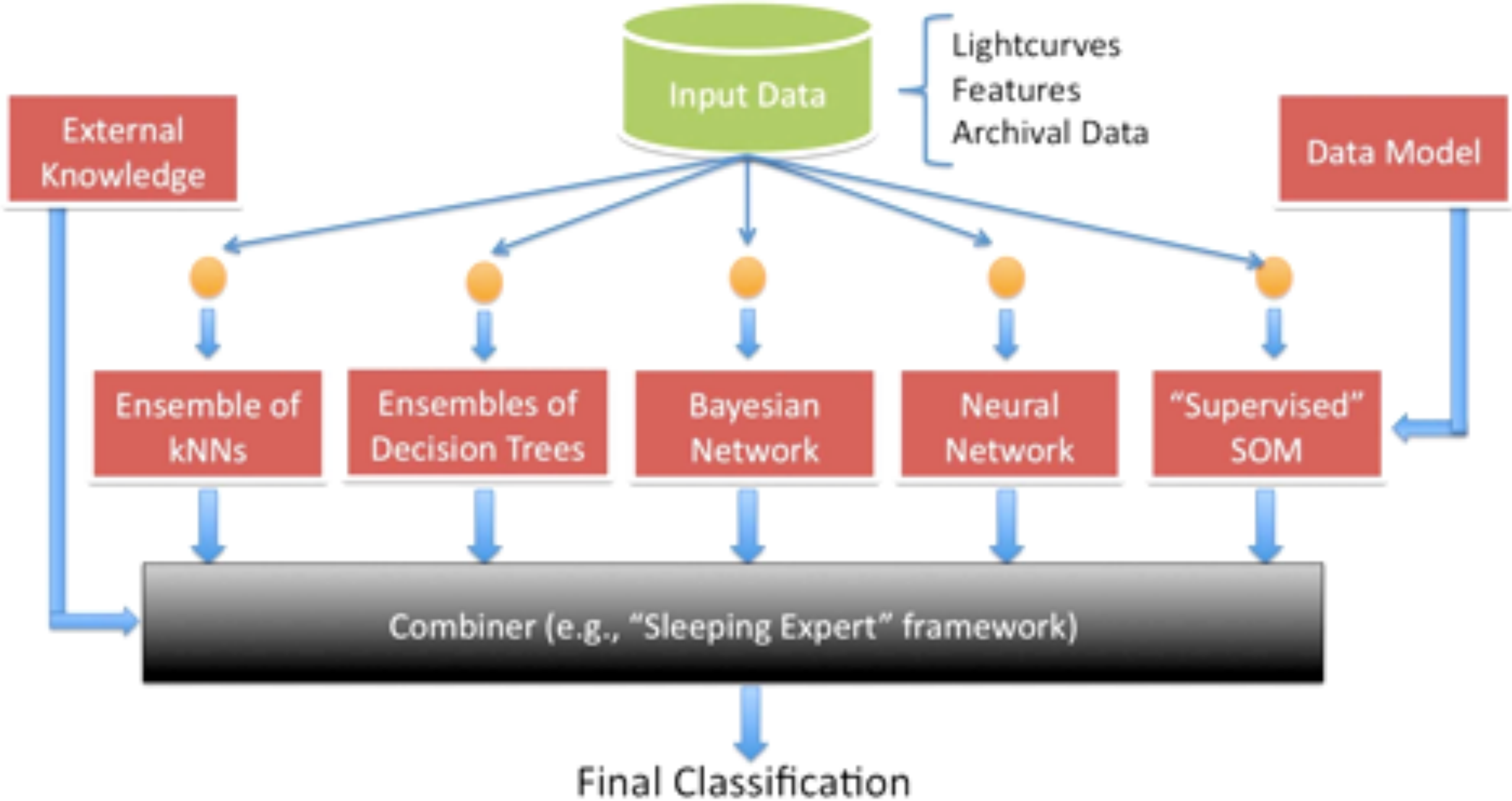
based on peaks

$$\left(\frac{1}{t_{span}}\left(\frac{1}{N}\Sigma_i w_i (p_i - p_m)^2\right)^{1/2}\right.$$

Challenge: Designing domain knowledge assisted features



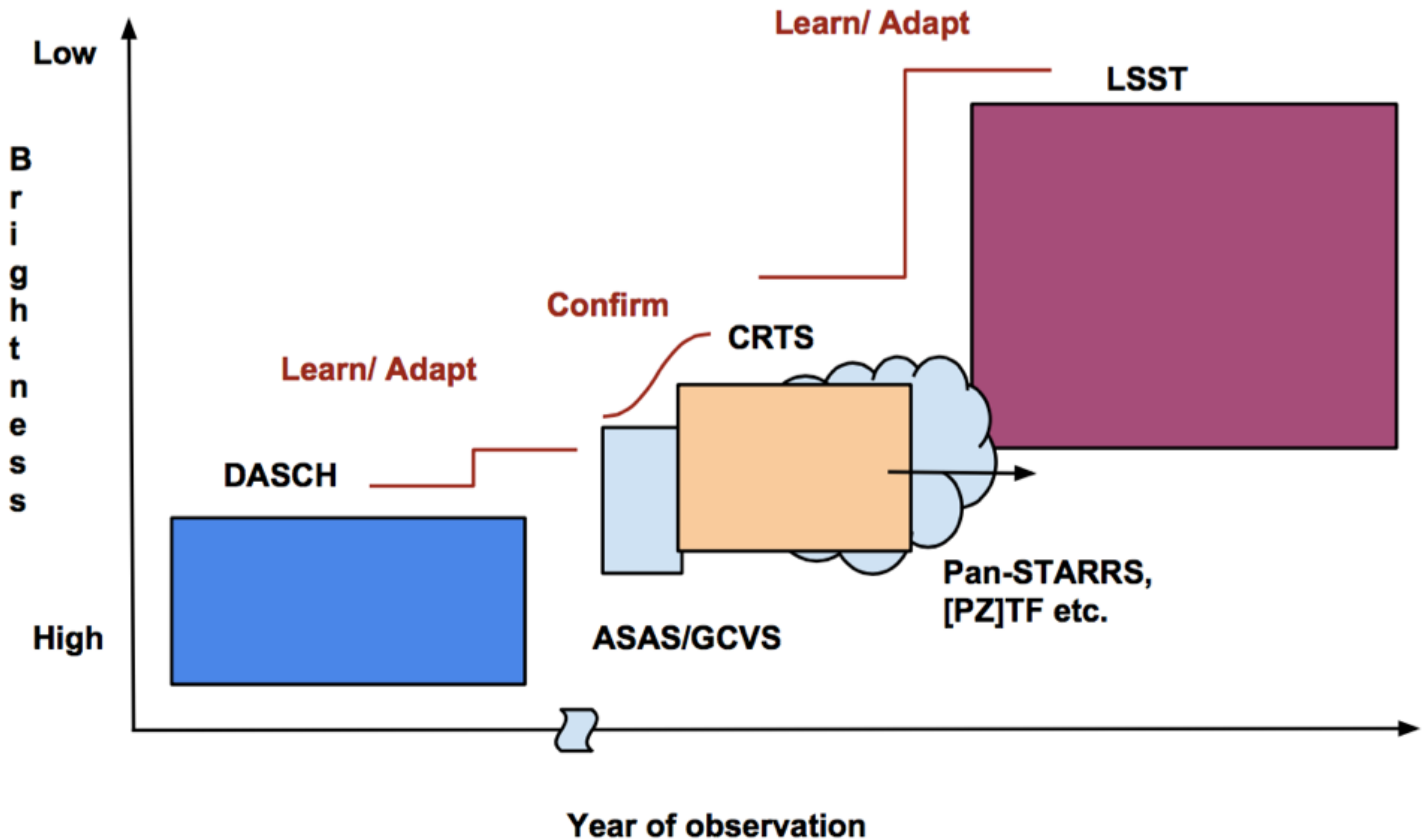
Metaclassification: An optimal combining of classifiers



Exploring a variety of techniques for an optimal classification fusion:
Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit, Sleeping Expert…

Mahabal, Donalek

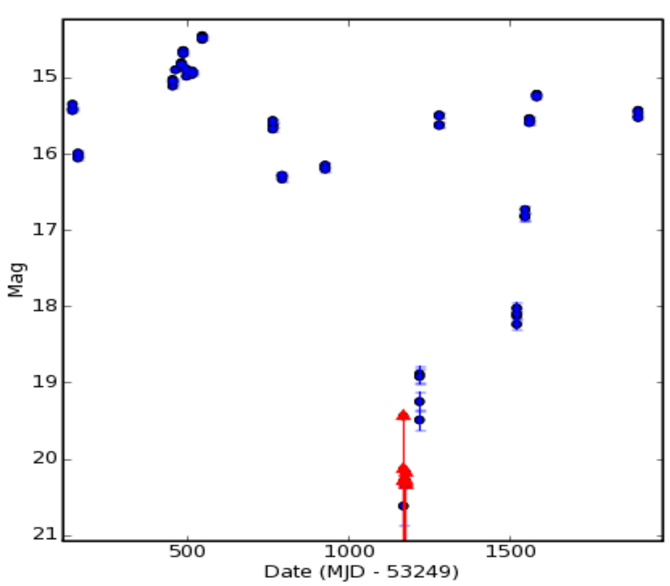
We need to be able to look at all surveys holistically

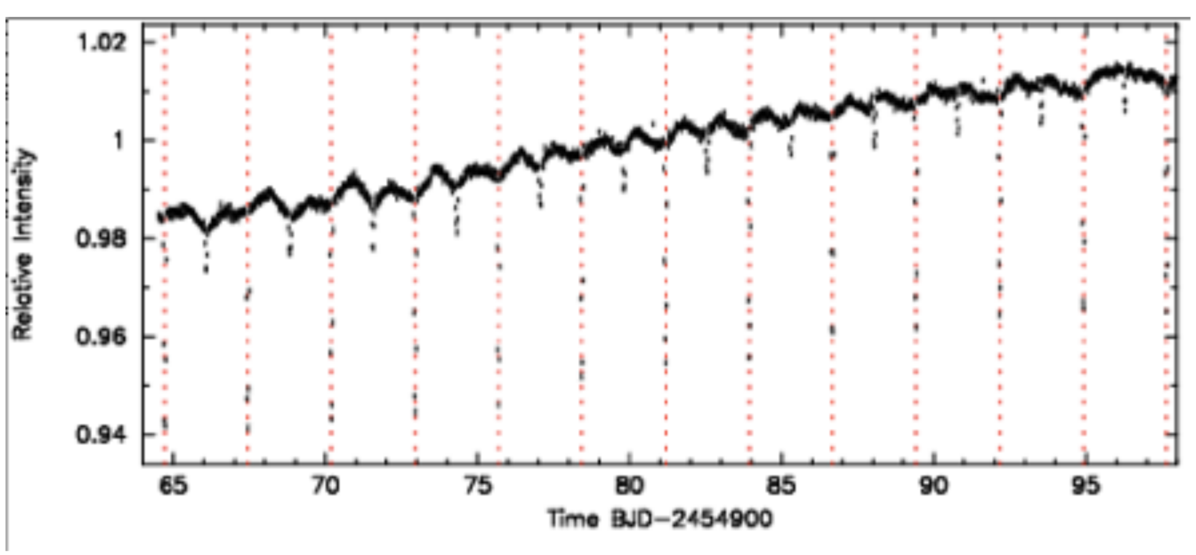


This shows just two dimensions

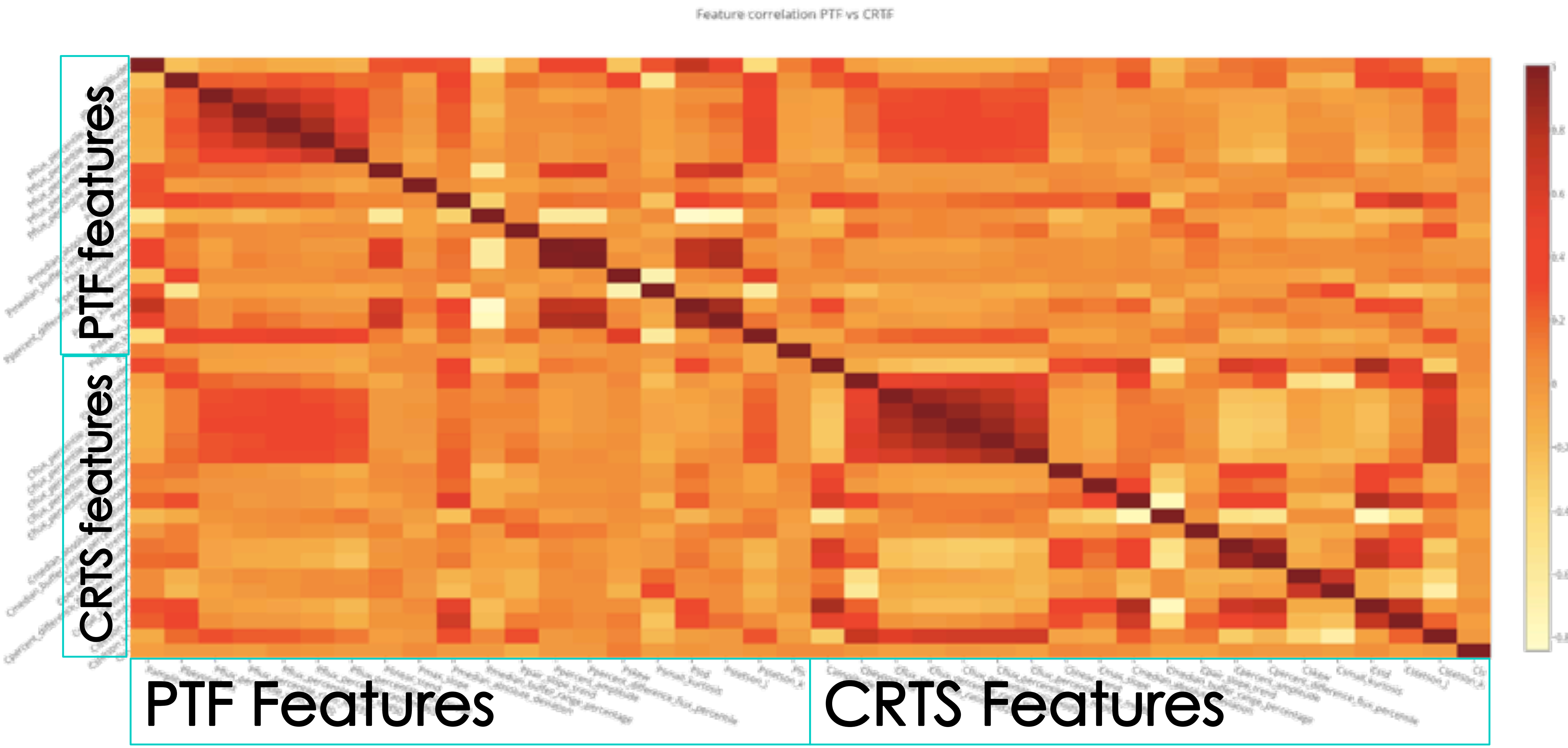
Why Domain Adaptation?

- Surveys differ in depth (aperture), filters, cadence
- Same (type of) objects produce different statistical features (skew, median absolute deviation etc.)
- Learning tends to be done on each survey separately - leading to unnecessary delays
- DA helps build on the otherwise untapped intersurvey synergy (think DASCH -> CRTS/ZTF/Kepler -> LSST)

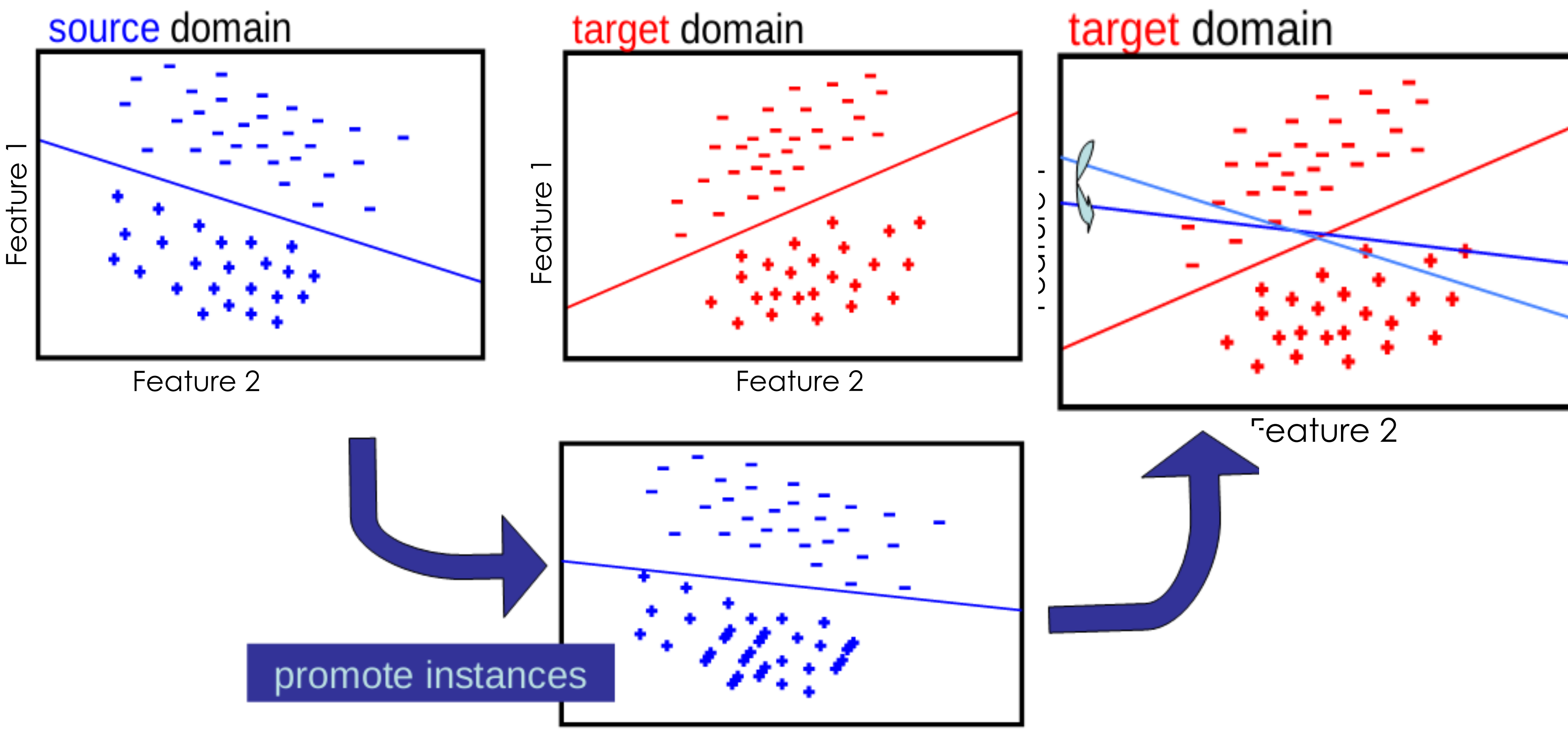




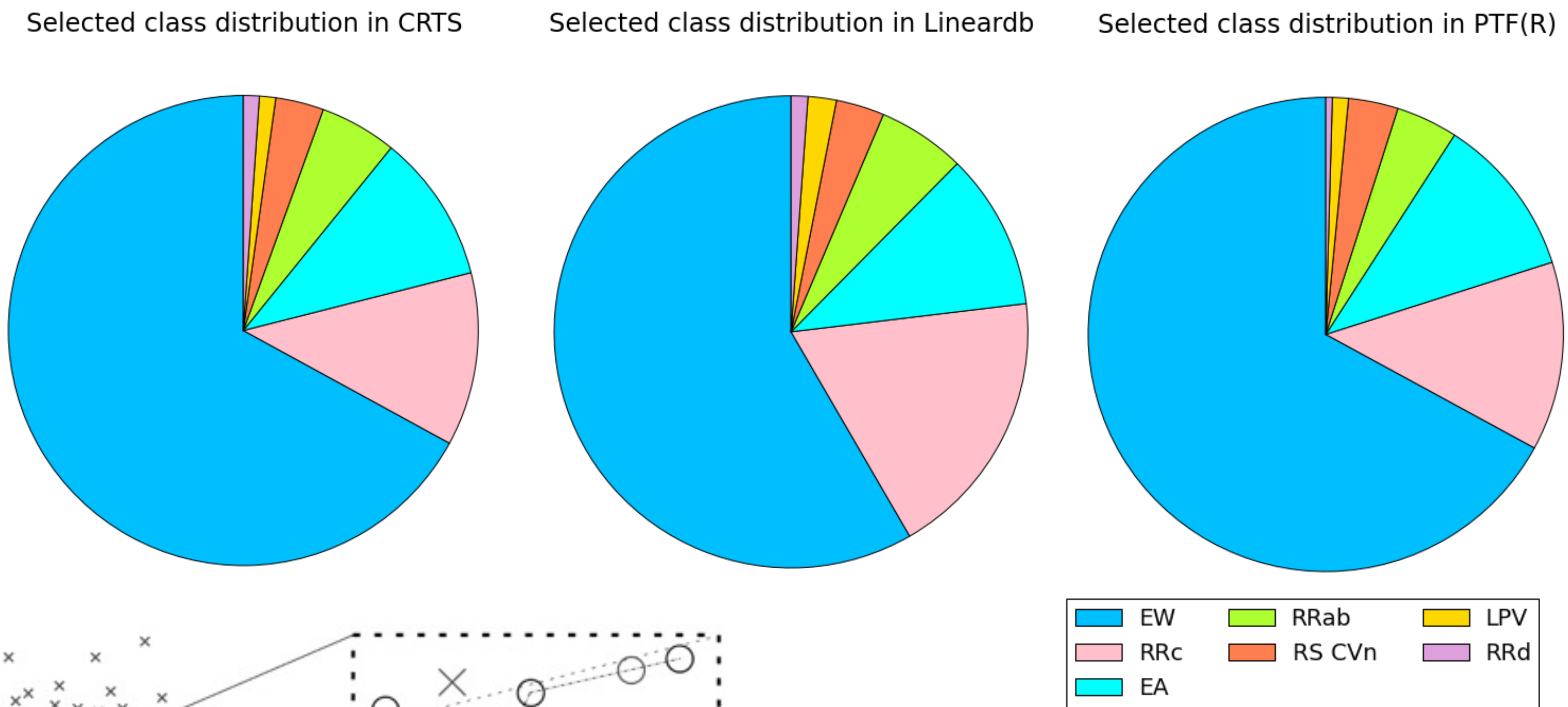
Feature Correlations

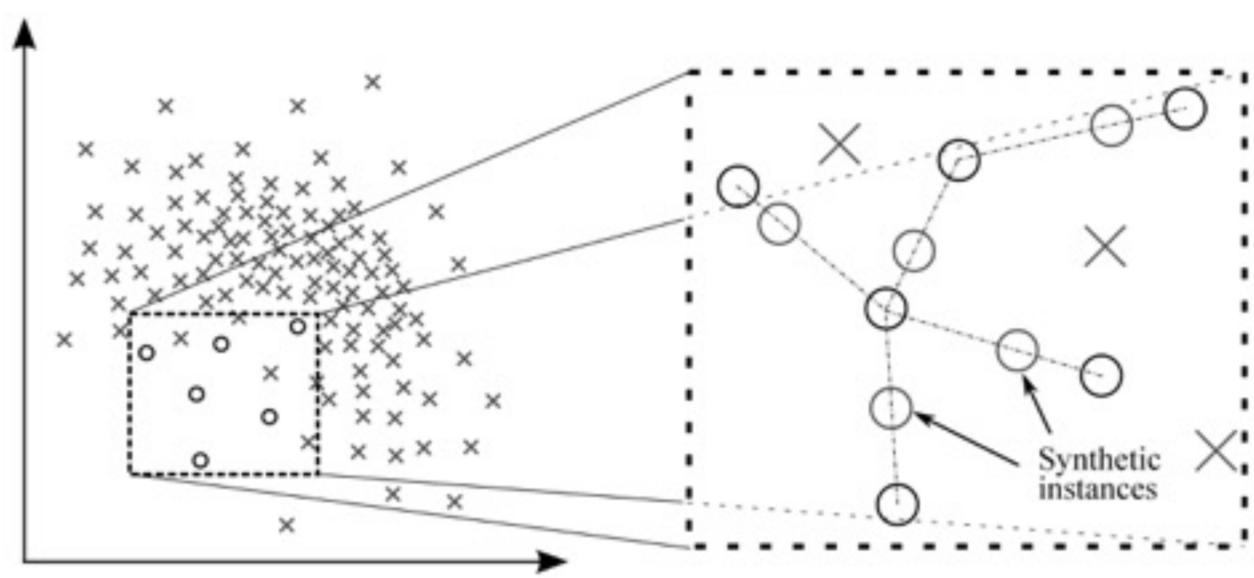


If you had just two features



50K Variables from CRTS





Drake et al. 2014

SMOTE and
Sampling with replacement
used to take care of unbalancedness

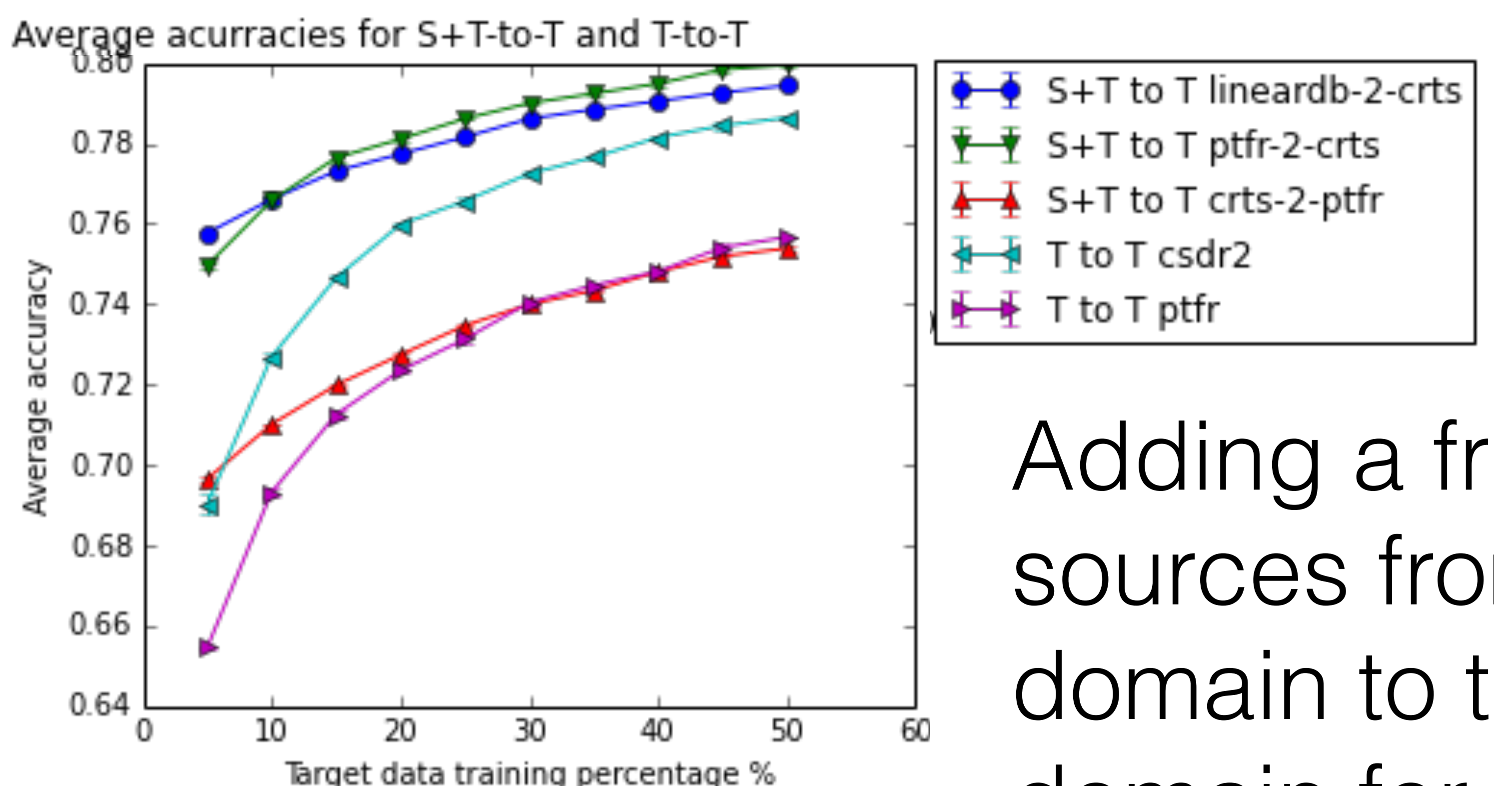
Co-Domain Adaptation

- Slow adaptation from S to T
- Add best target objects in each round
- Elect shared S and T subsets from training and unlabelled data (Chen et al. 2011)

$L = D_S \cup D_T^l$

$U = D_T^u$

Average Accuracy



Target Data Training %

Adding a fraction of sources from the target domain to the source domain for training improves performance

Not covered

- Next observation
- Ancillary information inclusion
- lightcurve decomposition
- Data challenge
- Outlier detection
- Designing new features

Summary of challenges

1. **Characterize/Classify as much with as little data as possible**
2. **Only a small fraction are rare - find/characterize them early**
3. **A variety of parameters - choose judiciously**
4. **Real-time computation is required - find ways to make that happen**
5. **Metaclassification - combining diverse classifiers optimally**