RAFAEL MARTÍNEZ-GALARZA

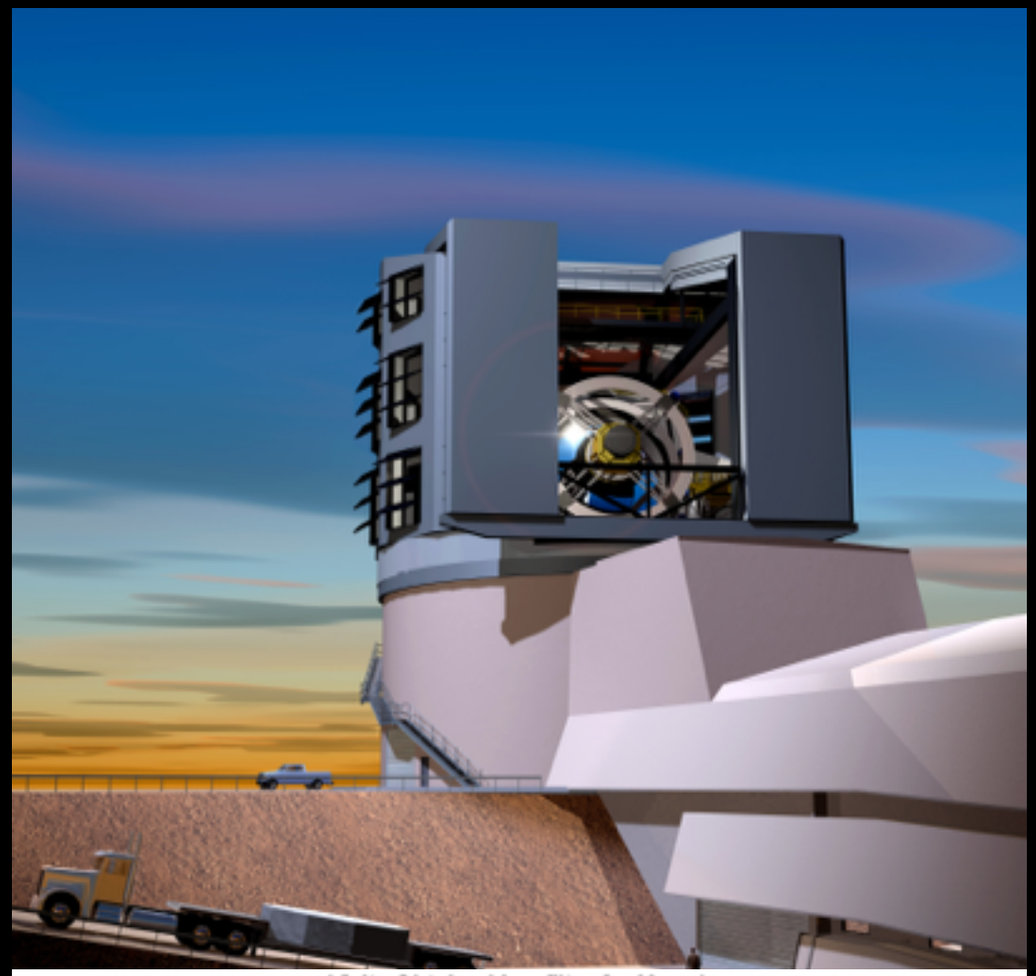
BUILDING A TRAINING SET FOR AN AUTOMATIC LSST LIGHT CURVE CLASSIFIER

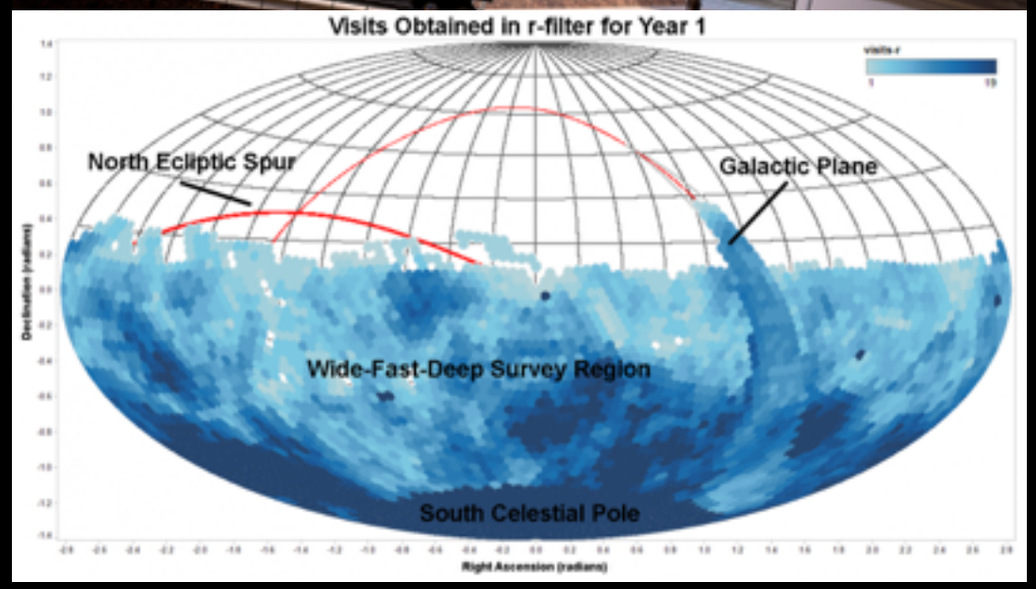
WITH: JAMES LONG, VIRISHA TIMMARAJU, ASHISH MAHABAL, VIVEK KOVAR AND THE SAMSI WG2

HARVARD-SMITHSONIAN
CENTER FOR ASTROPHYSICS

THE MOTIVATION: LSST IS COMING

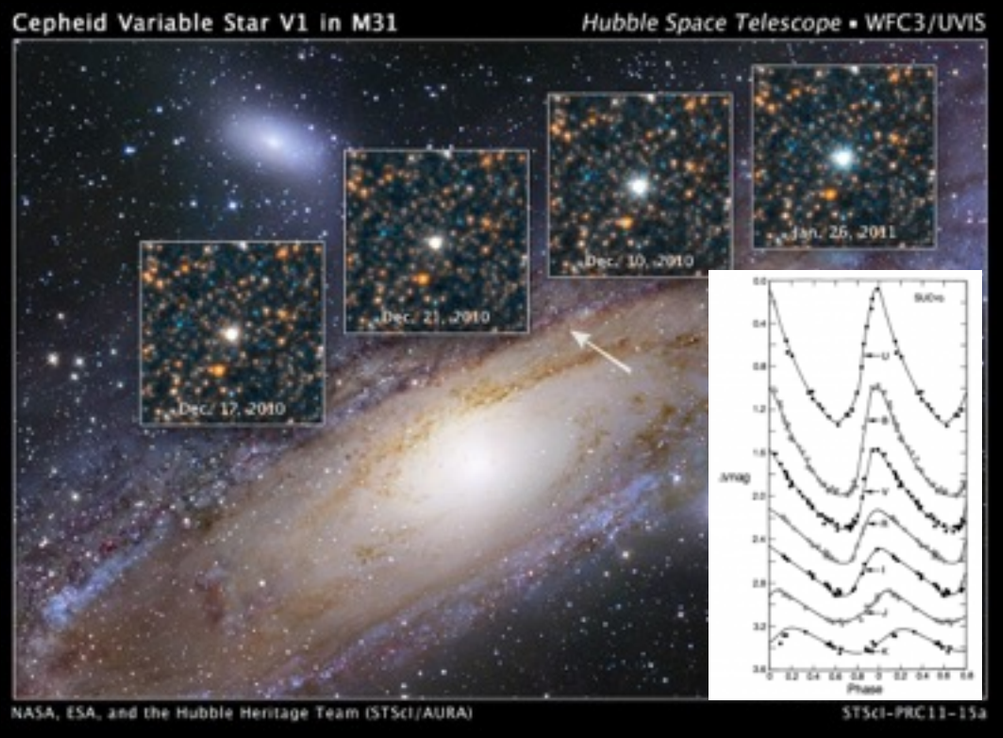
- The Large Synoptic Survey Telescope is a 8.4m reflector currently under construction in Chile (first light expected in 2021).
- Design concept: a survey that will take an image of every part of the entire visible sky every few nights, in six bands, for 10 years.
- Transients and variable stars: periodic and non-periodic variable sources will be studied in detail, and new types are expected at very short and very long timescales.

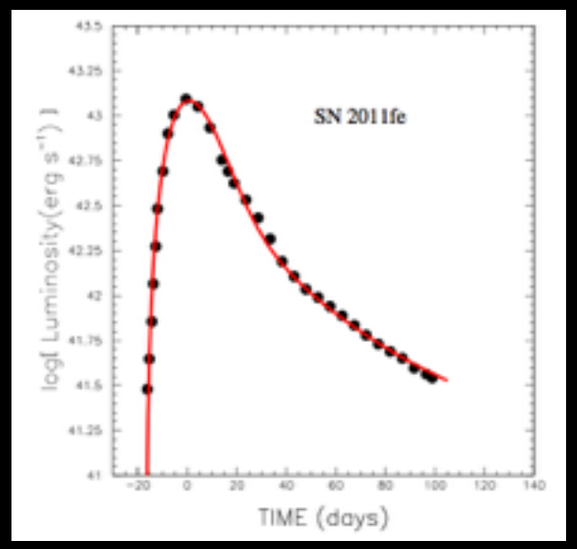


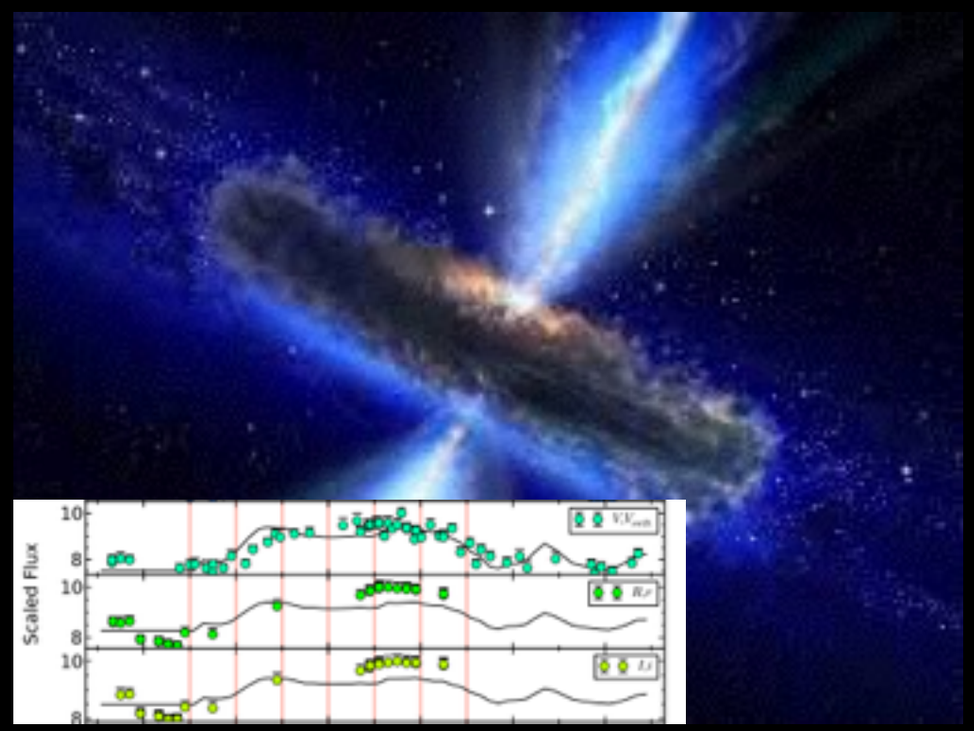


CHALLENGE: VARIABILITY IS DIVERSE

- Periodic (RR Lyrae stars, Cepheids)
  - Consistent in their periods and amplitudes.
- Quasi-periodic (Mira stars)
  - Dominating frequencies, but no consistency in phase or amplitude
- Stochastic (AGNs, QSOs)
  - Variability without any obvious patterns
- Transient (Supernovae, stellar flares, GRBs)
  - Short-time changes in flux, non periodic



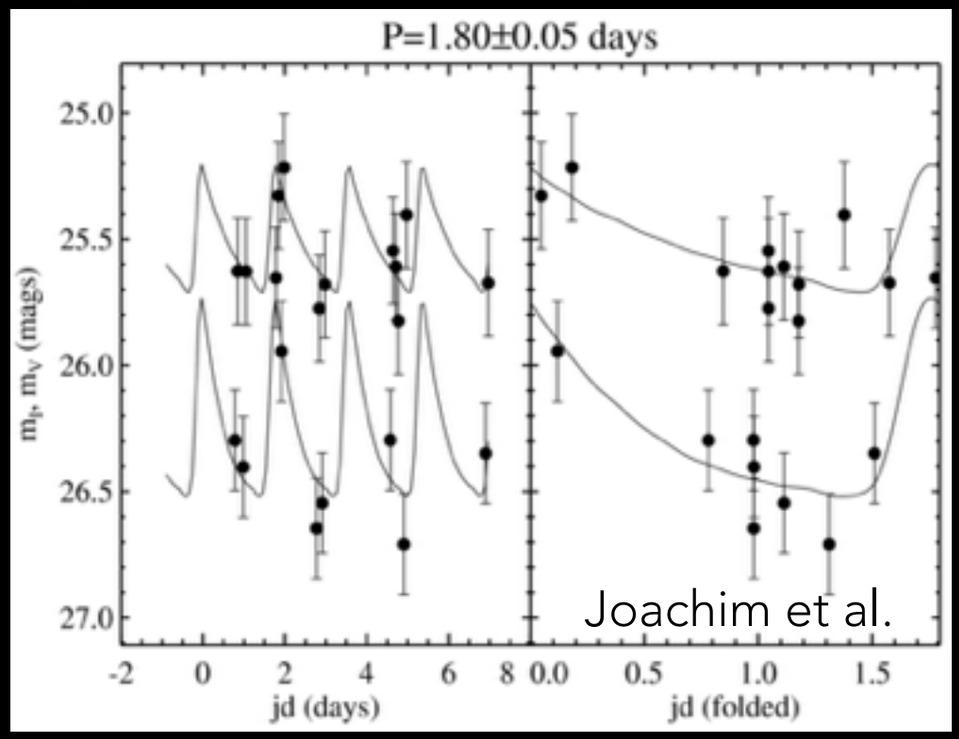




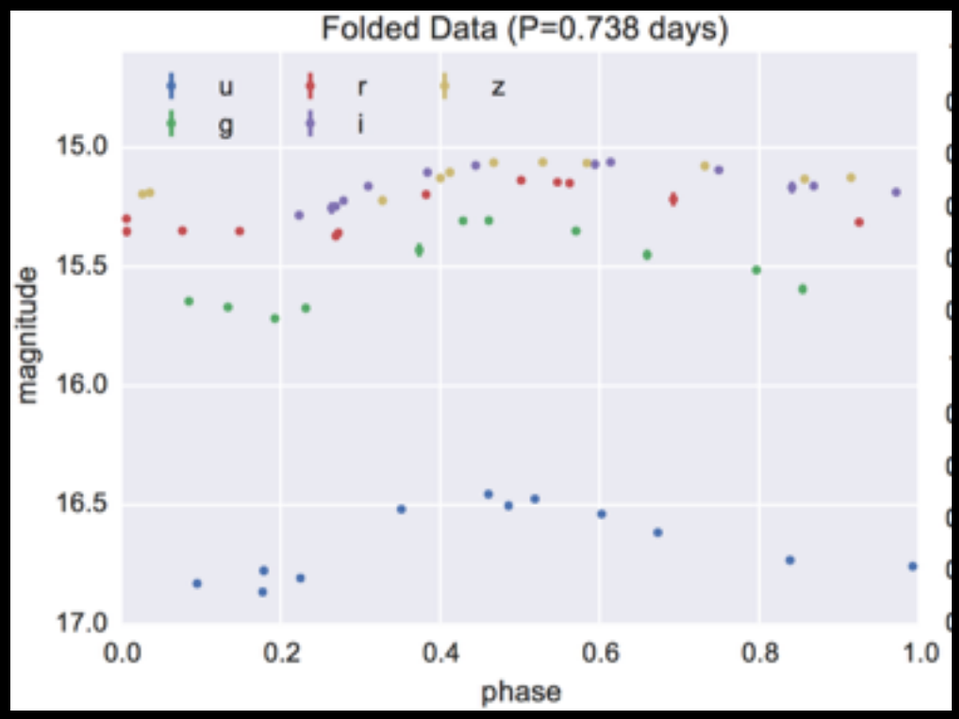
THE CLASSIFICATION CHALLENGE

- LSST will deliver the time stories of ~10^9 sources.
- Classification of sources according to their light curves becomes impossible for humans in reasonable times.
- Machine-learning algorithms can greatly help in this classification task (in principle).
- Algorithms can:
  - Learn functions that map the LCs features into class probabilities.
  - Detect outliers whose features stand out with respect to the full population.

Light curves will be both sparse…



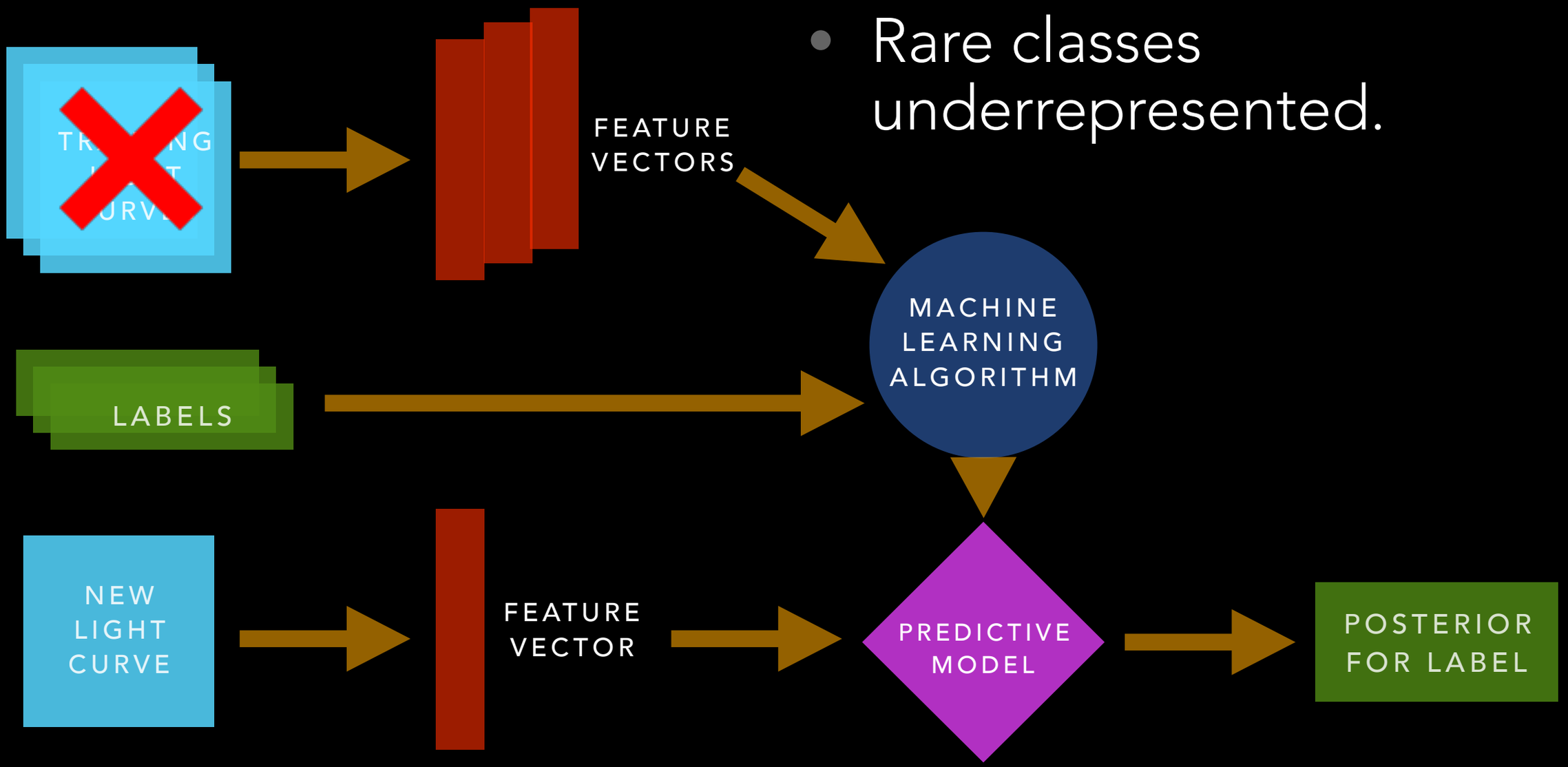
And non-simultaneous across filters



WHERE CAN THINGS GO WRONG?

1. Training set

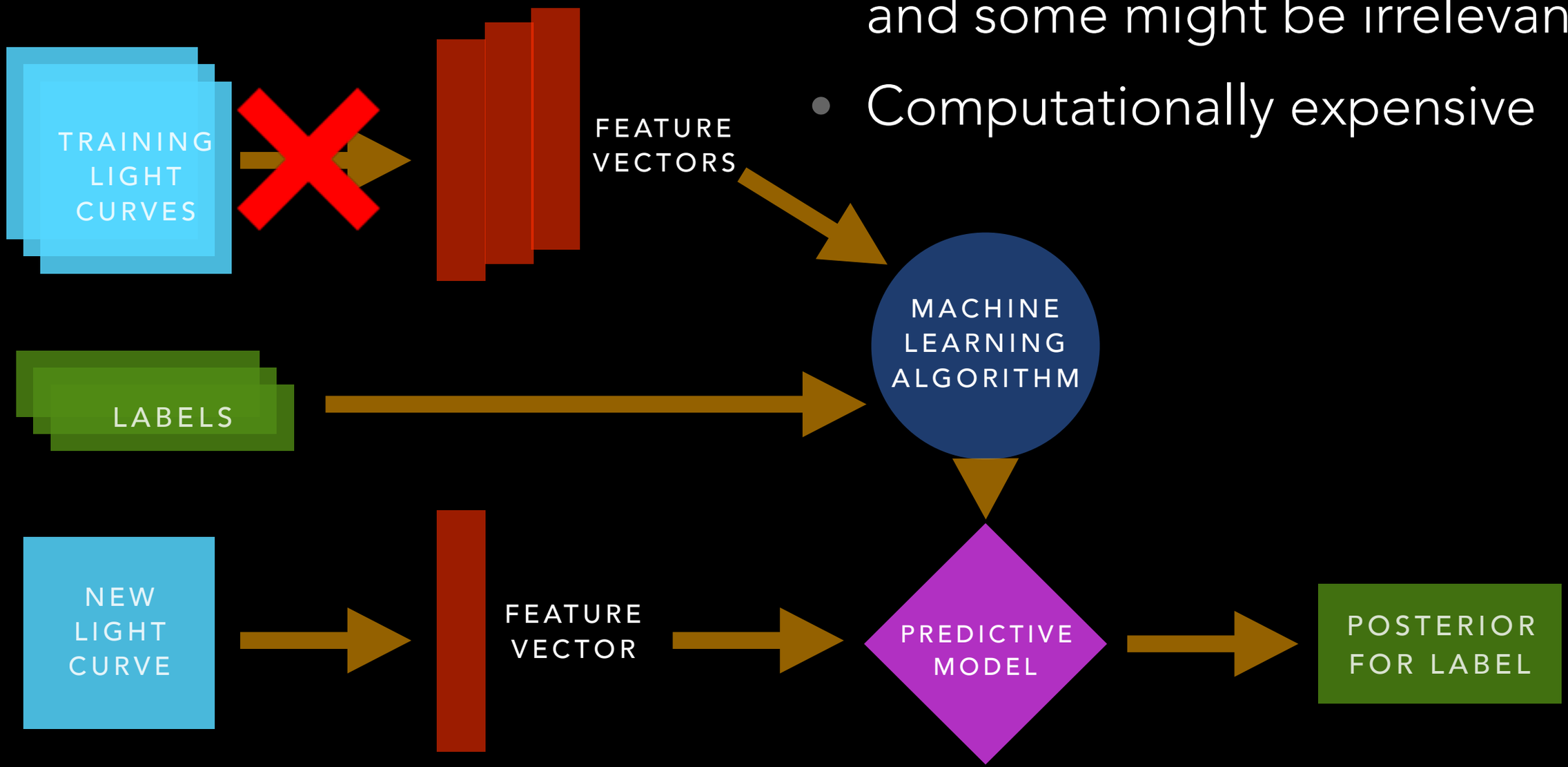
- Training set bias
- Only brightest or nearest sources have robust labels
- Rare classes underrepresented.



WHERE CAN THINGS GO WRONG?

2. Feature extraction

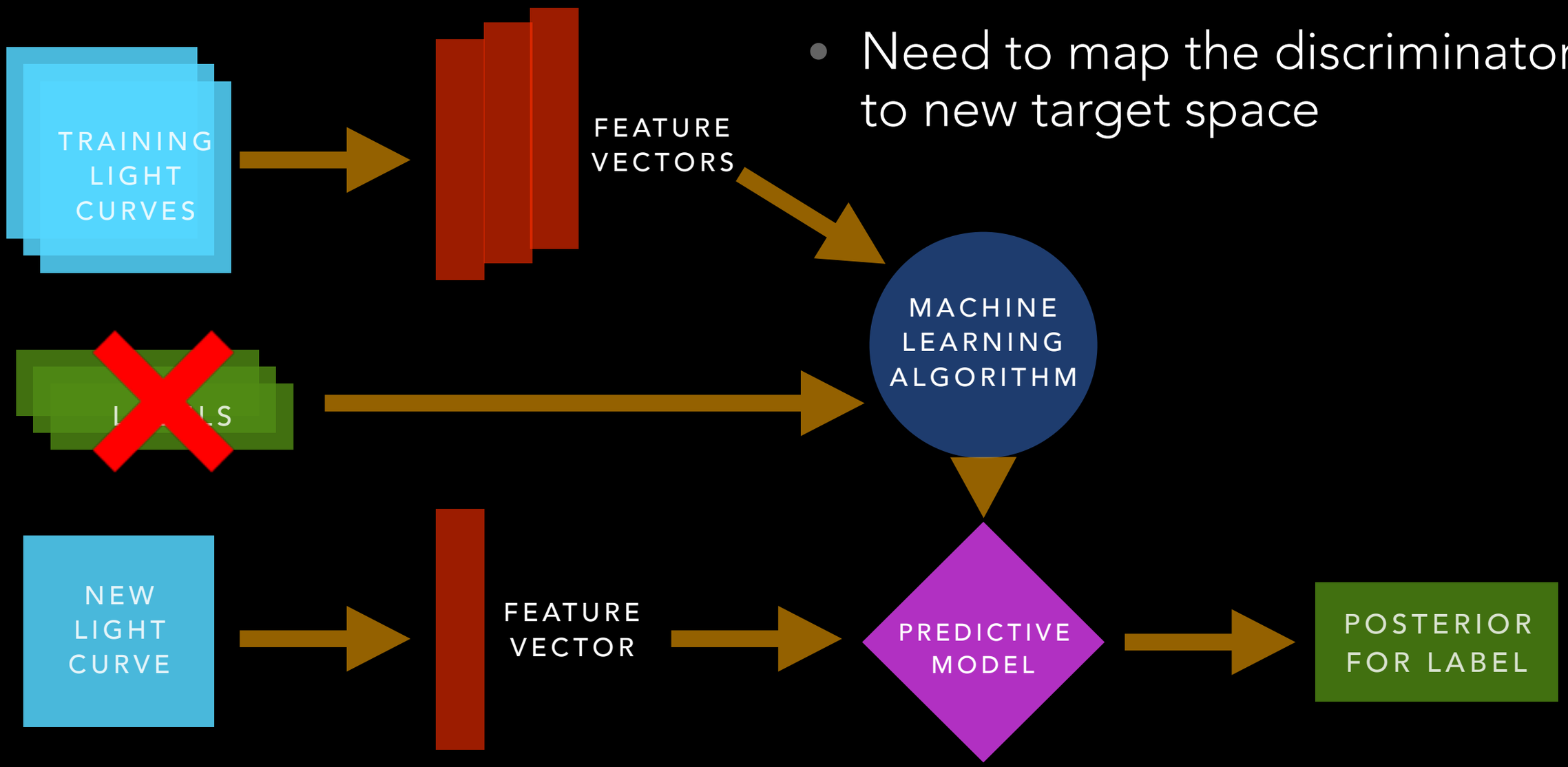
- Uneven time sampling.
- Noise.
- Features are domain specific and some might be irrelevant.
- Computationally expensive



WHERE CAN THINGS GO WRONG?

3. Features

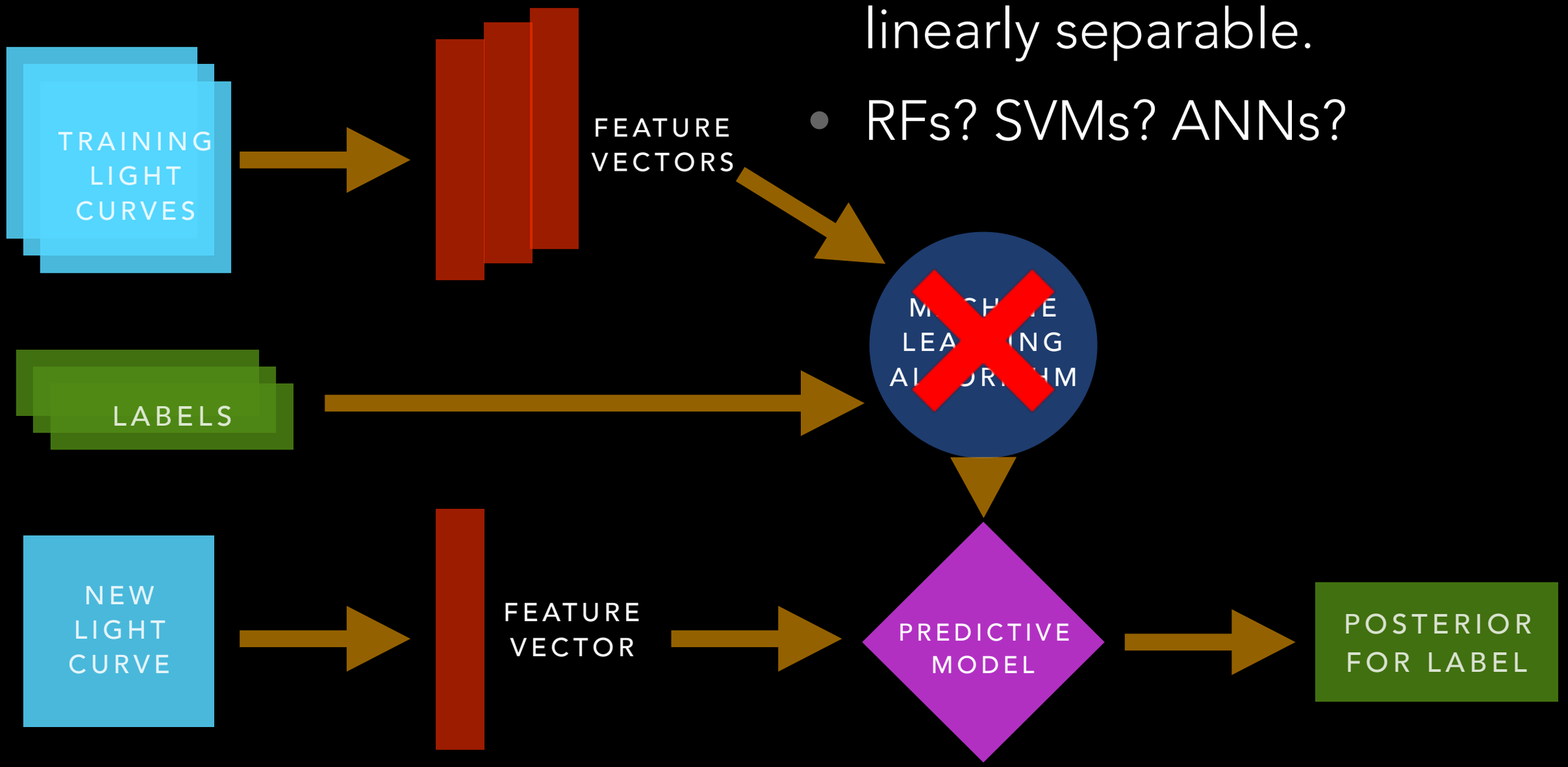
- Which is your "ground truth"?
- Labels might come from a different domain (i.e. a different survey)
- Need to map the discriminators to new target space



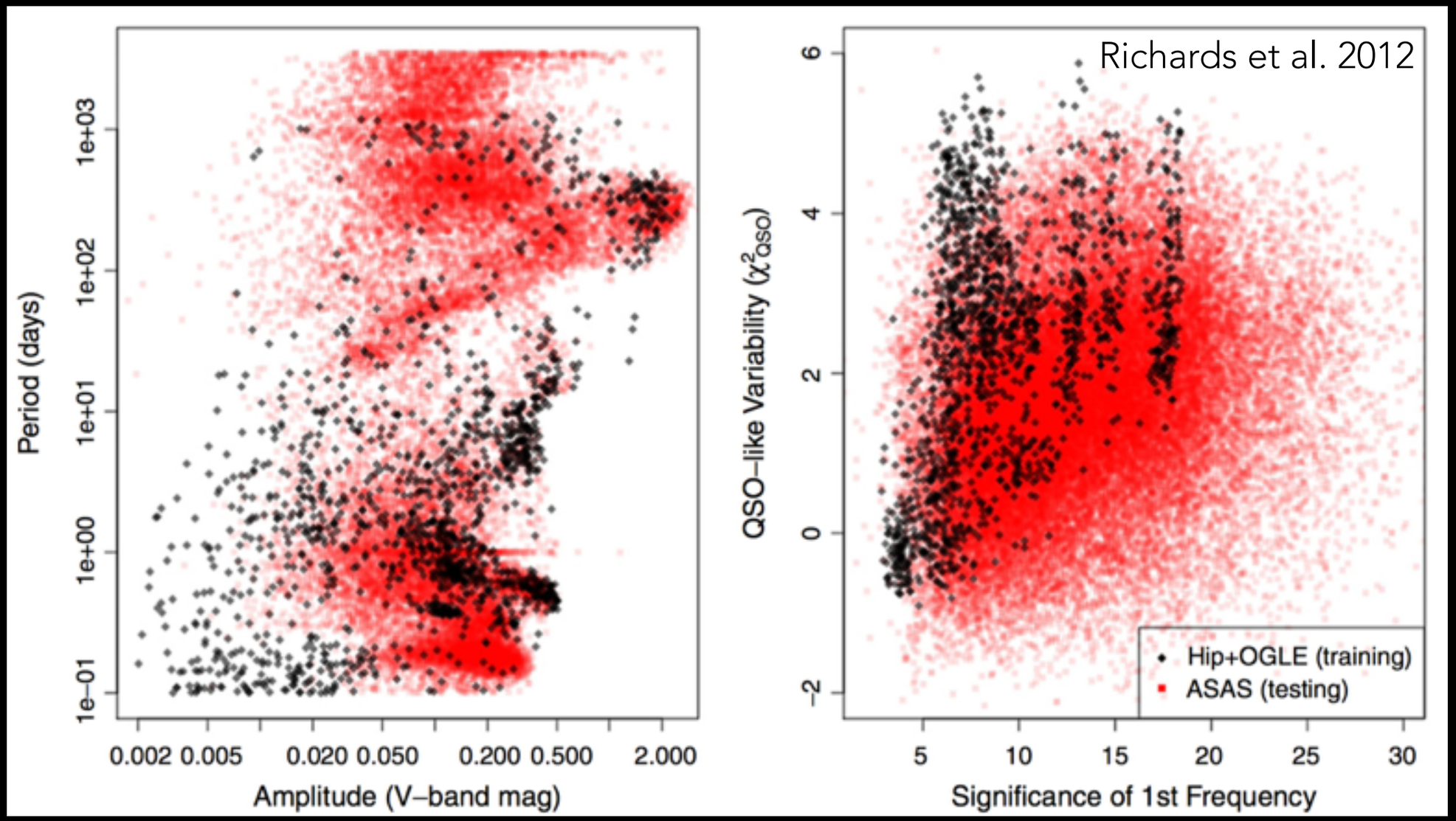
WHERE CAN THINGS GO WRONG?

4. Training

- Accuracy depends on method used.
- Multi-dimensional feature spaces where classes are not linearly separable.
- RFs? SVMs? ANNs?



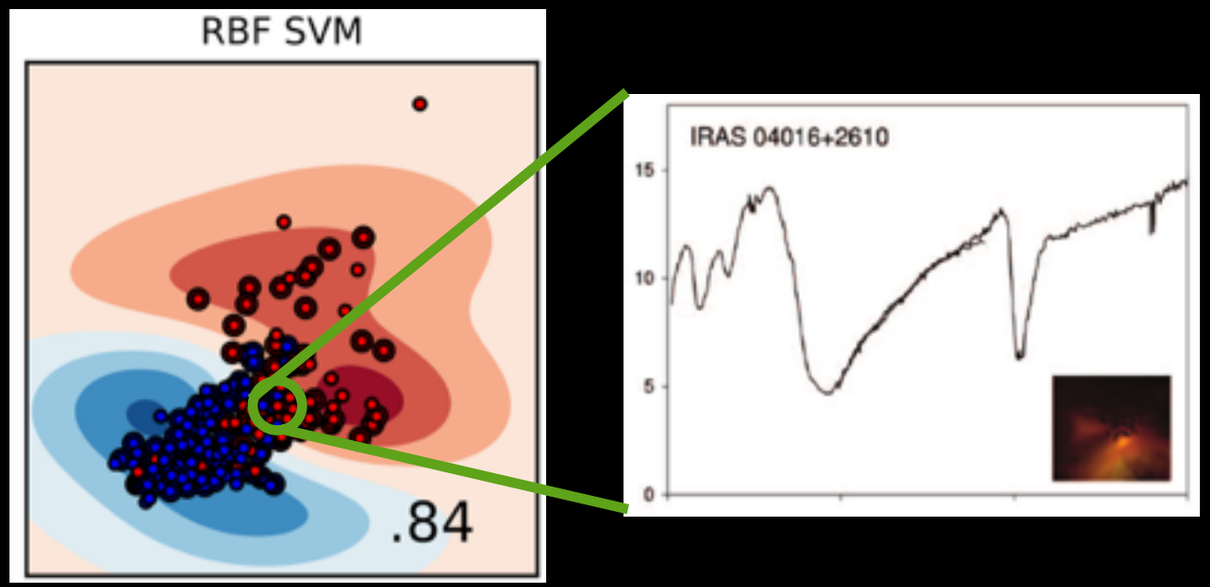
TRAINING SET BIAS



- Discrepancies in the period-amplitude plane: ASAS data has high density in the short period, high amplitude region. Testing data also has smaller values of the QSO-like variability metric.

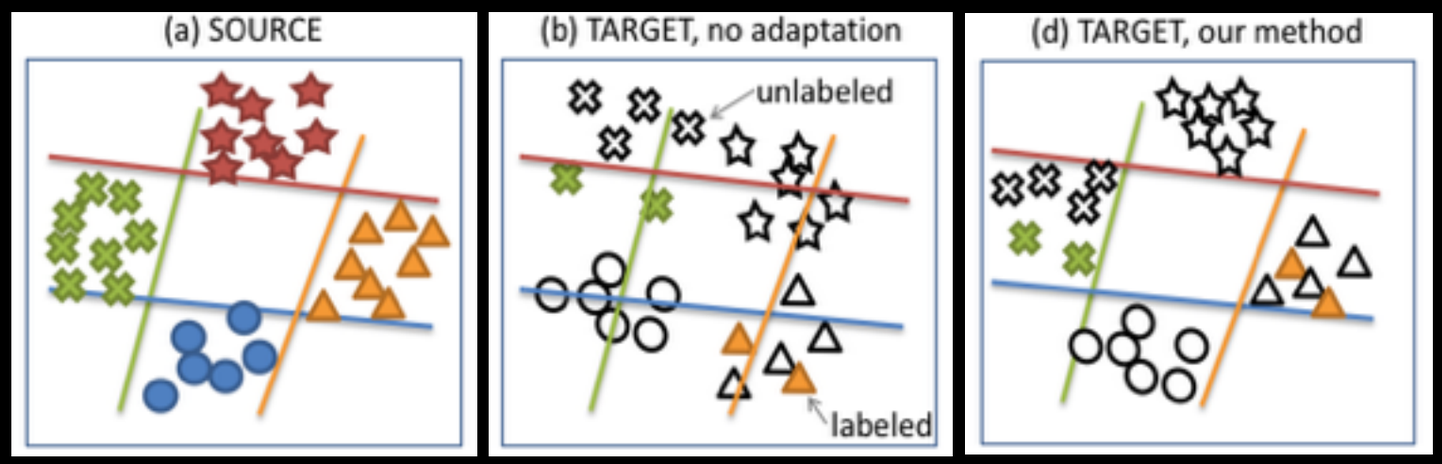
POSSIBLE APPROACHES TO TACKLE BIAS:

- Active Learning



1) Choose unlabelled sources in test set that would maximize the performance of the training if label was known. 2) Follow them up. 3) Add to the training set.

- Domain adaptation



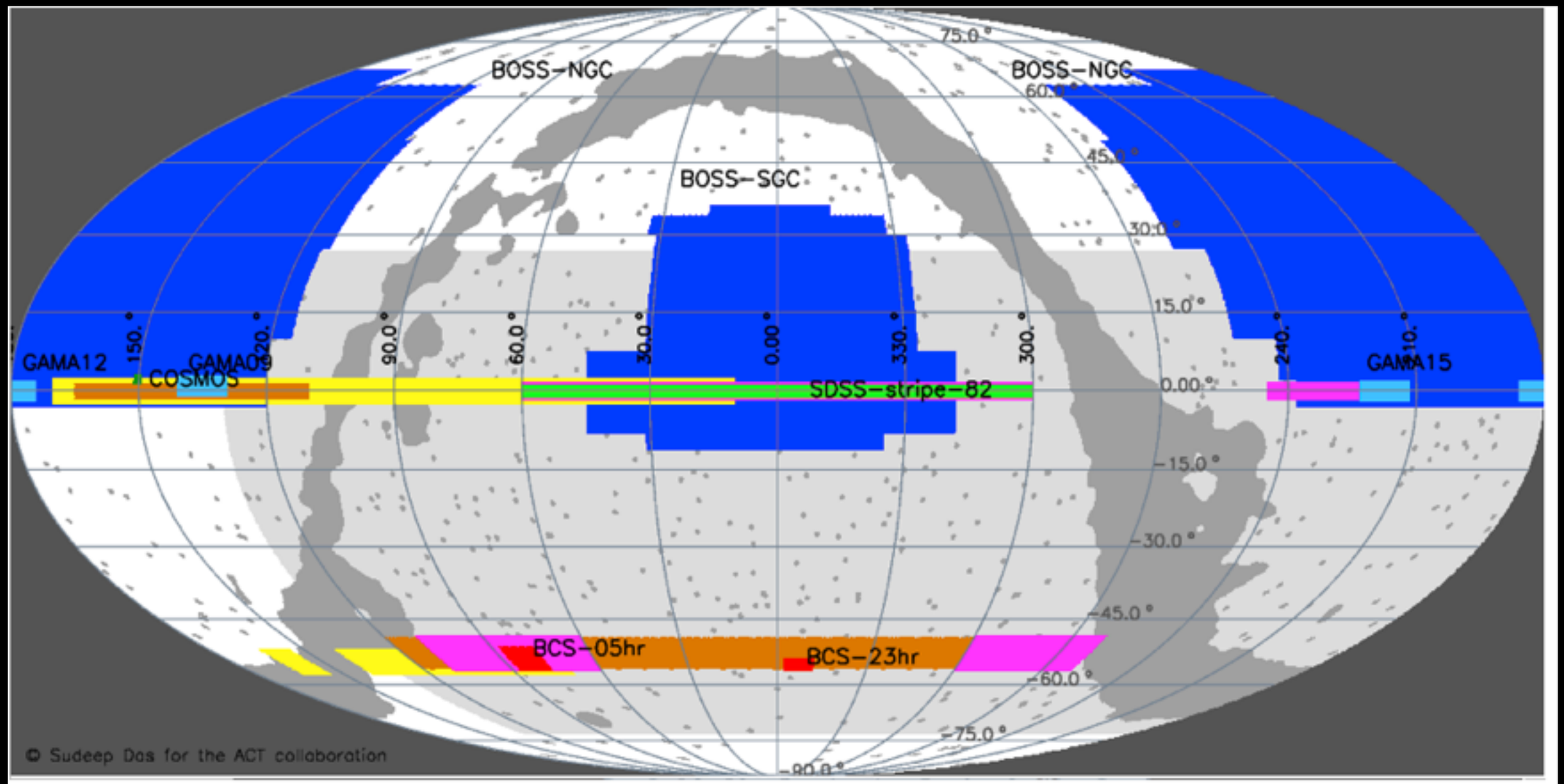
Transformation:

$T: D_{source}(\theta) \rightarrow D_{target}(\theta)$

Need only a fraction of labeled sources in the target datase

But we need to get started somewhere

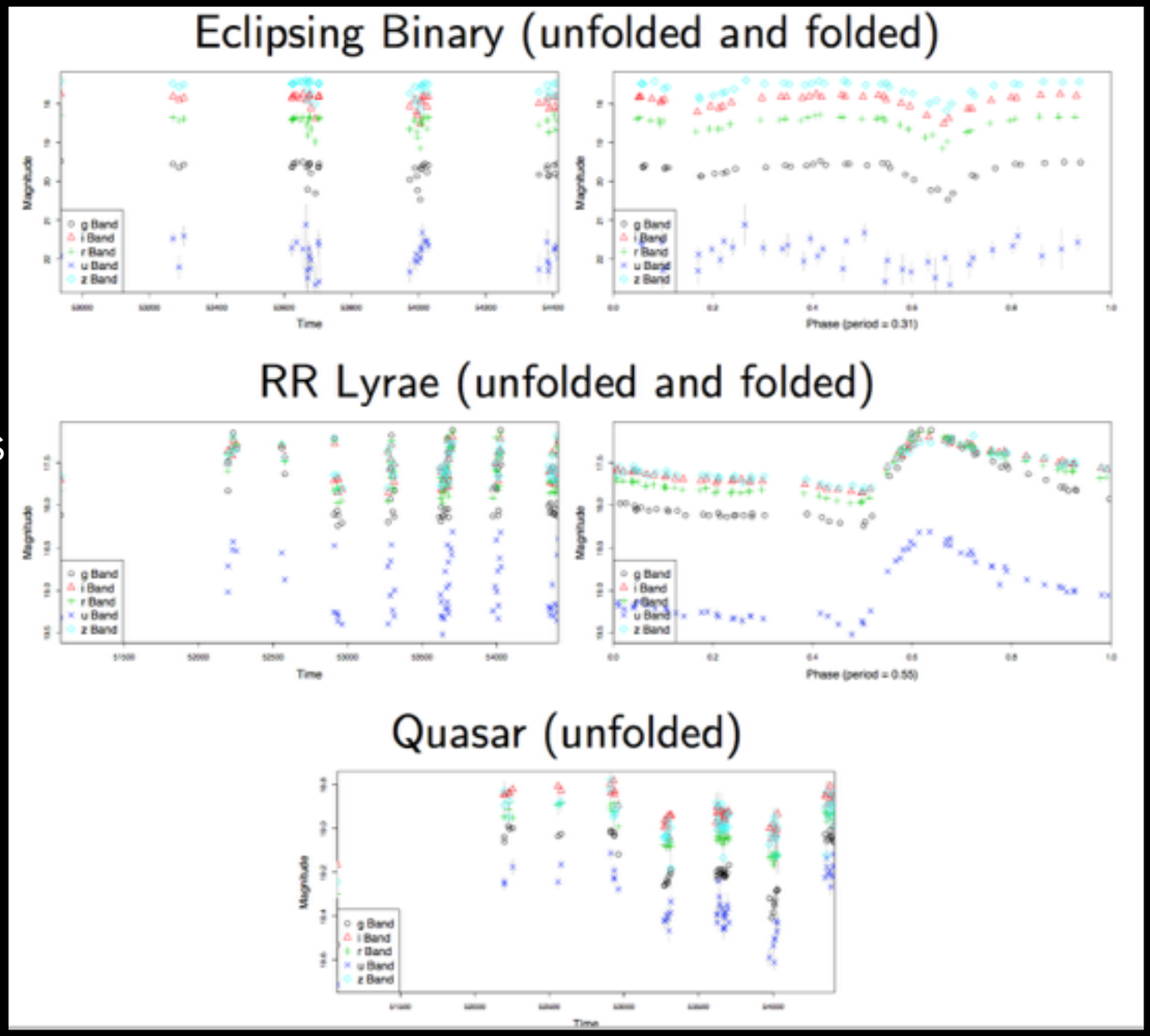
THE SDSS STRIPE 82



BASIC FACTS

- 5 bands: ugriz
- ~60,000 variable sources
- ~50 observations per band (but significant variance.)
- Photometry is roughly simultaneous across bands.
- Survey is deep: 2 mag deeper than regular SDSS obs.
- Cadence: on average, sources are re-observed every two days, followed by 5-day, 10-day, and yearly observations.

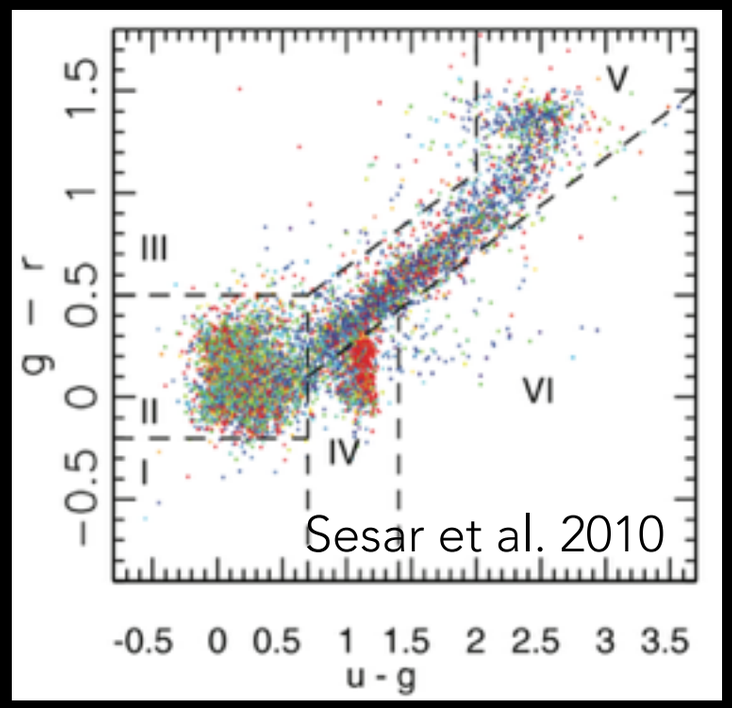
Example light curves:

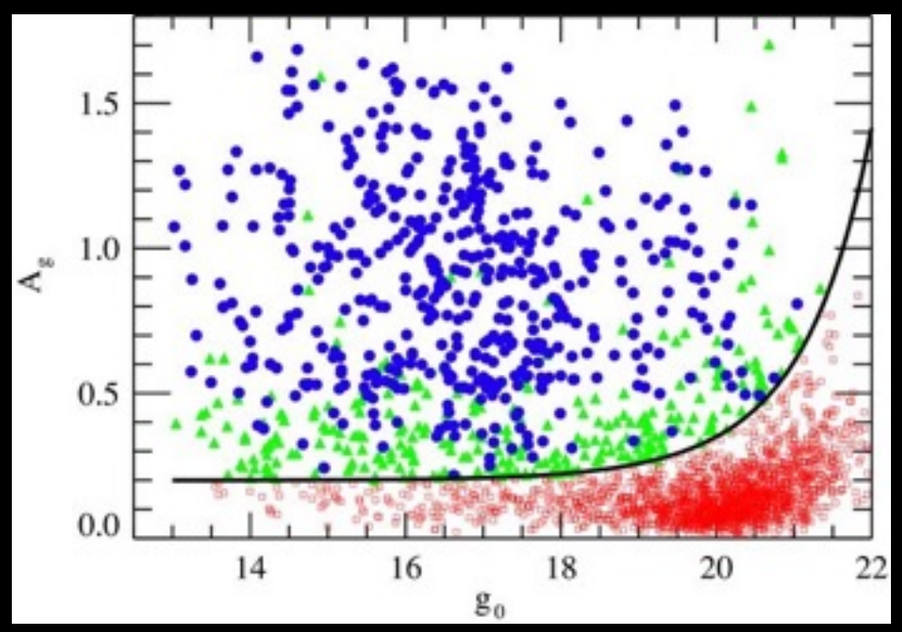


MERGING CLASSIFICATIONS

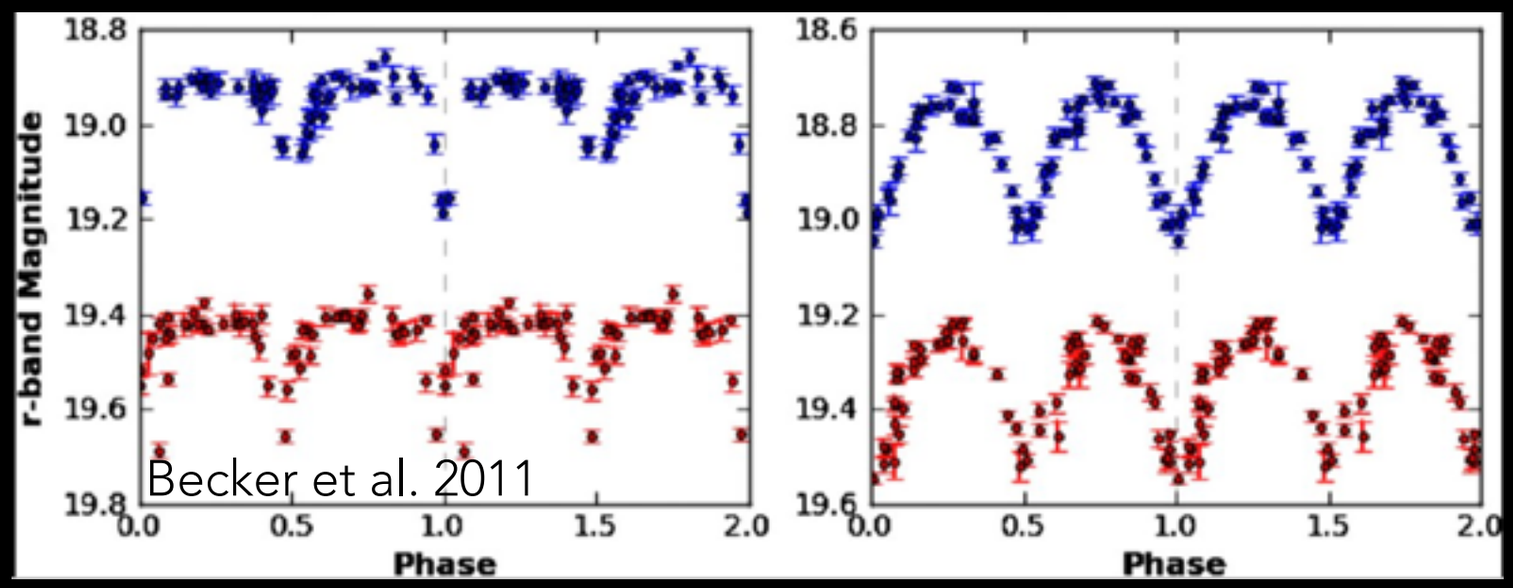
Stripe 82 sources have been independently labelled in previous studies
Many sources still unlabelled. Domain adaptation?

RR Lyrae:
Color cuts, template fitting

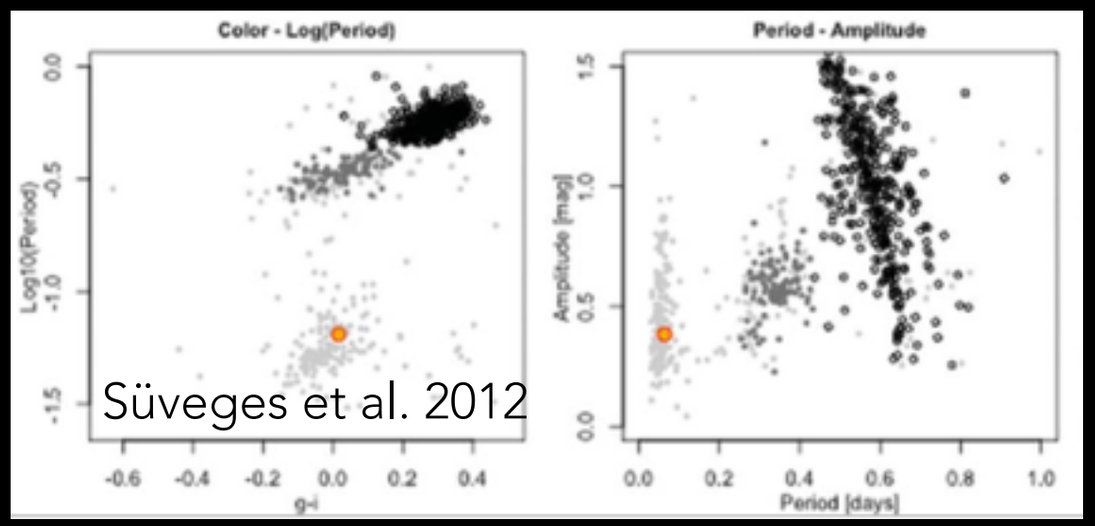




Eclipsing binaries
Spectra and LC shape



HADS
Visual inspection + PCA for colors
+ Random forest classifier



QSOs:
Spectra

MULTI-BAND PERIOD DETERMINATION

- We use the multi-band periodogram (van der Plas & Ivezic, 2015) to estimate periods.

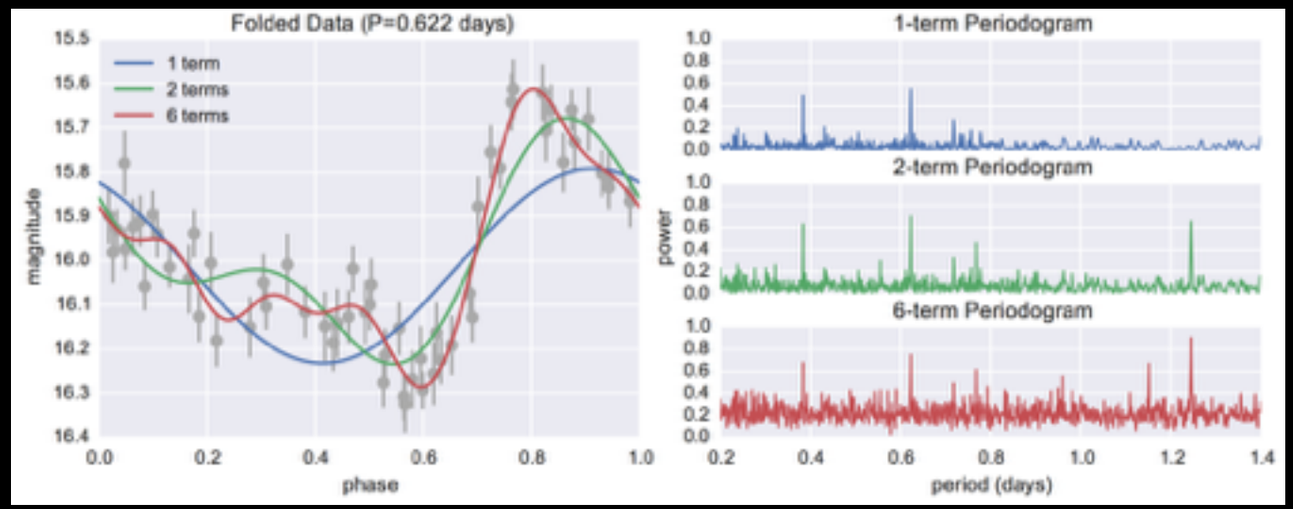
- Outperforms existing methods, specially for non-simultaneous, sparsely sampled multi-band LCs.

- Method is linear on the θ parameters, and thus it is fast.

- Regularization is the key to allow multi-band analysis, and to avoid overfitting.

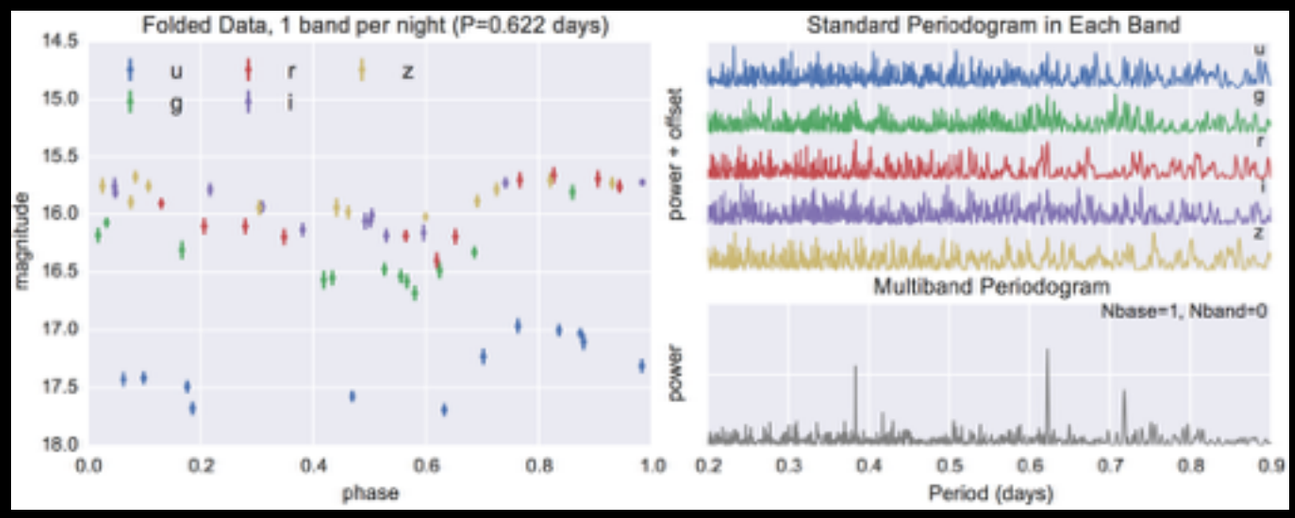
Single band (Lomb-Scargle):

$$y(t|\omega,\theta) = \theta_0 + \sum_{n=1}^{N} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right].$$



Multi-band:

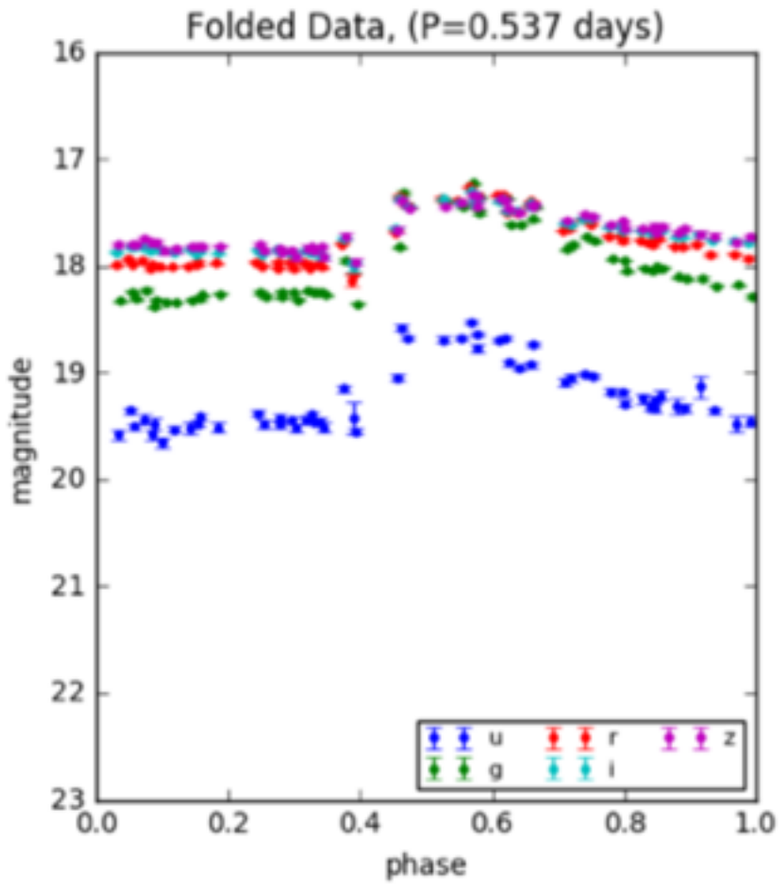
$$y_k(t|\omega,\theta) = \theta_0 + \sum_{n=1}^{M_{base}} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right] + \theta_0^{(k)} + \sum_{n=1}^{M_{band}} \left[\theta_{2n-1}^{(k)}\sin(n\omega t) + \theta_{2n}^{(k)}\cos(n\omega t)\right].$$

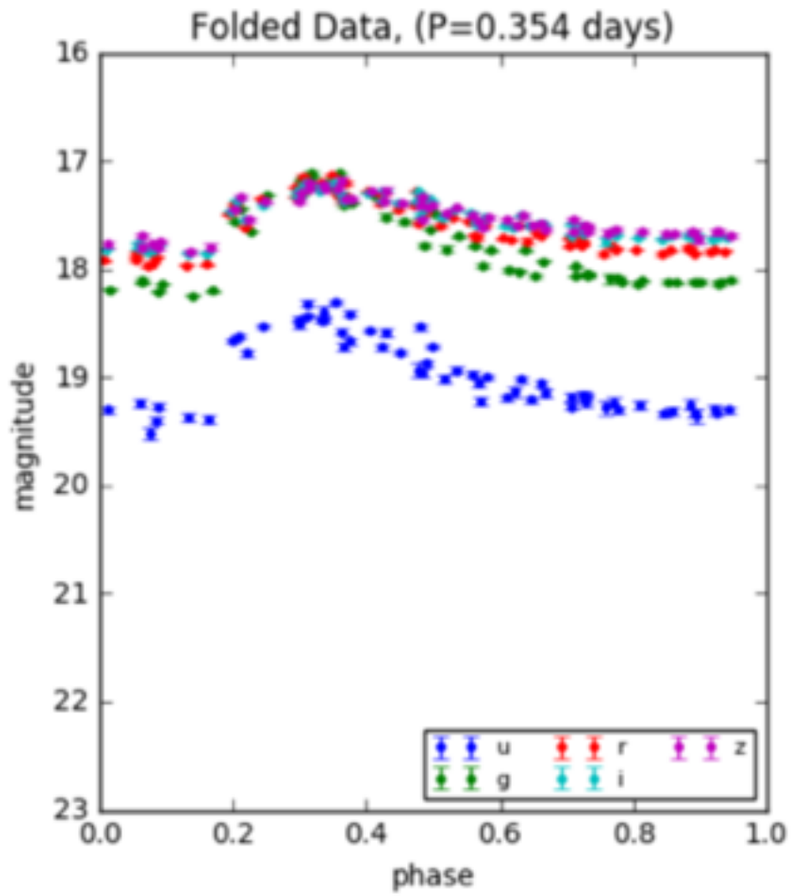


RESULTS ON STRIPE 82 SOURCES

Period Extraction

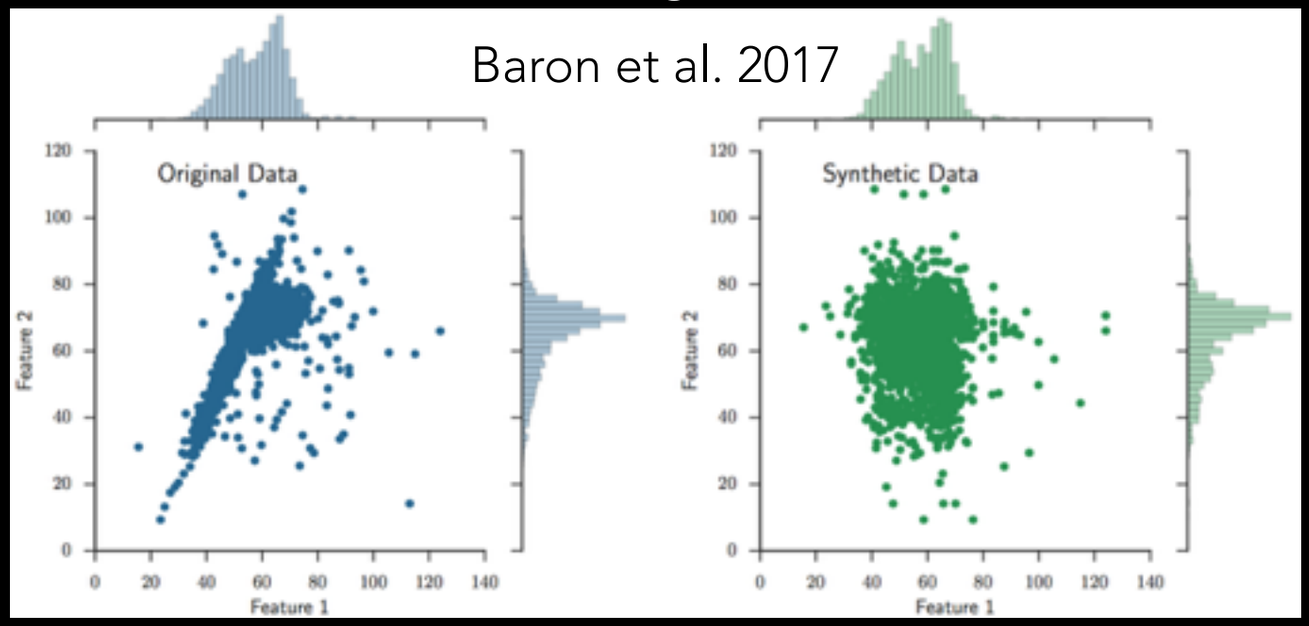
Lomb Scargle Multiband: Finding periods for randomly sampled multiband light curves like LSST.



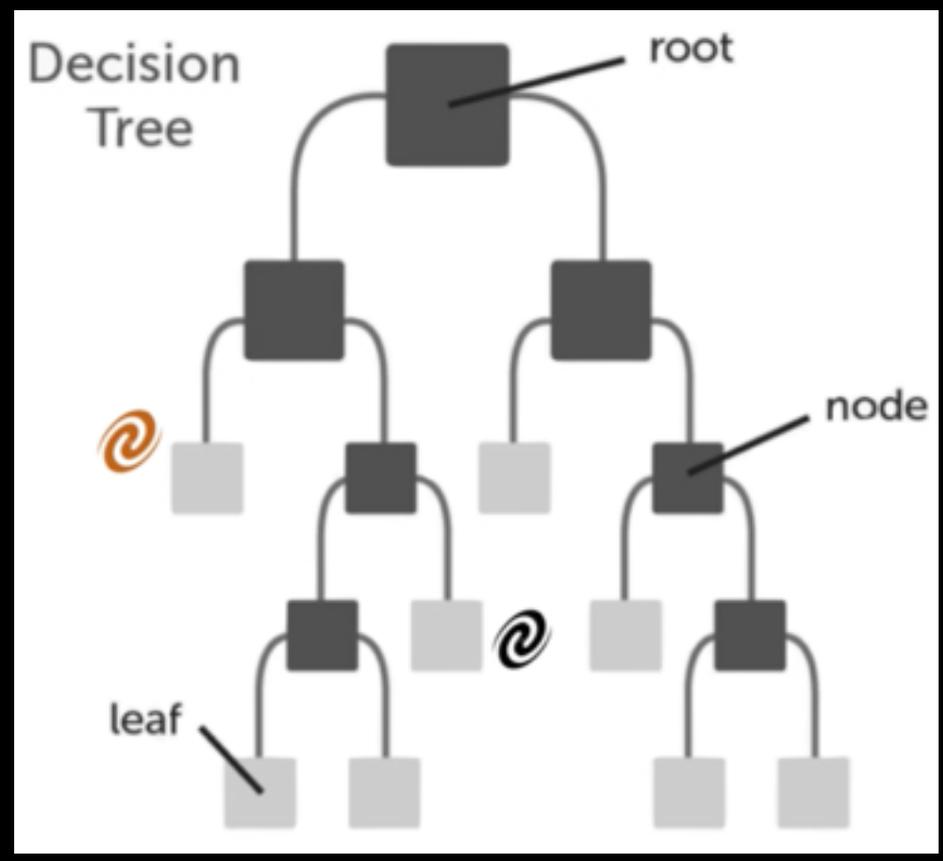


WHAT IF WE CAN FORGET ABOUT THE LABELS? Outlier detection with unsupervised random forest

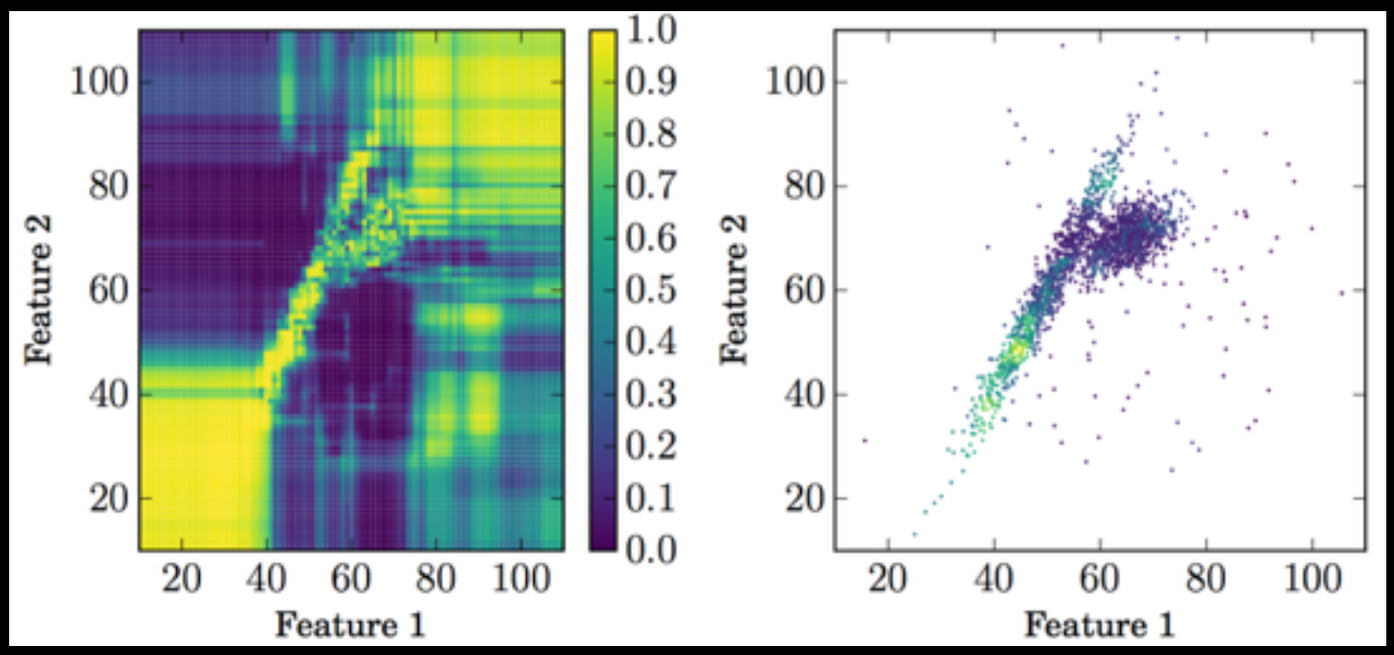
1. Sample from marginal distributions



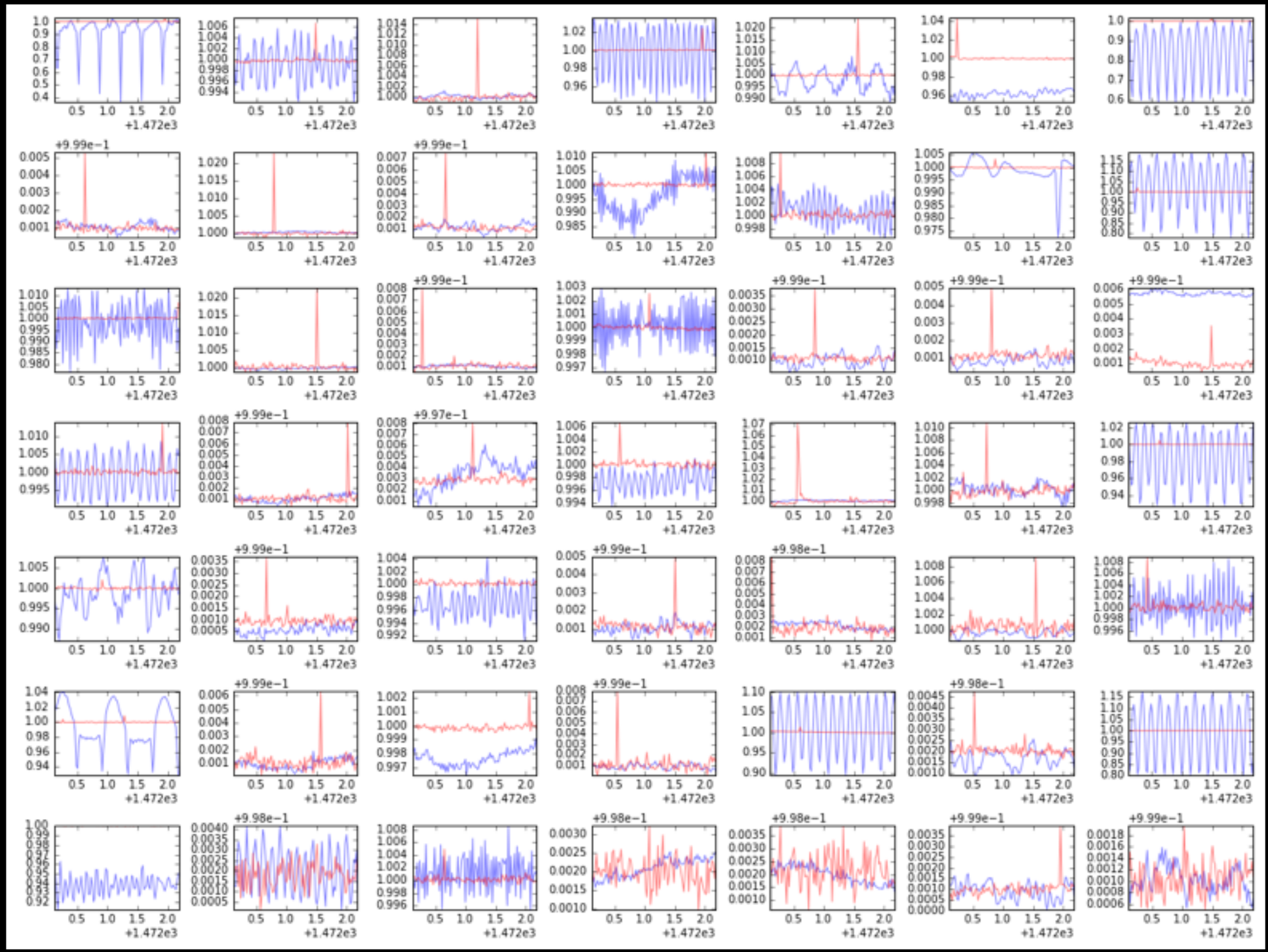
2. Train RF classifier with real and synthetic as classes



3. Find weird objects with respect to metric



FINDING THE WEIRDEST TRANSITS IN KEPLER LIGHT CURVES



CHALLENGING QUESTIONS

- Can we leverage results from other surveys (e.g., Catalina) to perform domain adaptation? Does it make sense to design a domain adaptation challenge?

- Does it make sense to pursue a follow up campaign of Stripe 82 sources to perform active learning?

- Is multi-band periodogram a reliable way to estimate periods in LSST-like data? What are the most important features besides from the period?

- How long into the LSST survey do we need to wait before outlier detection methods can identify unexpected variability types?