RAFAEL MARTÍNEZ-GALARZA

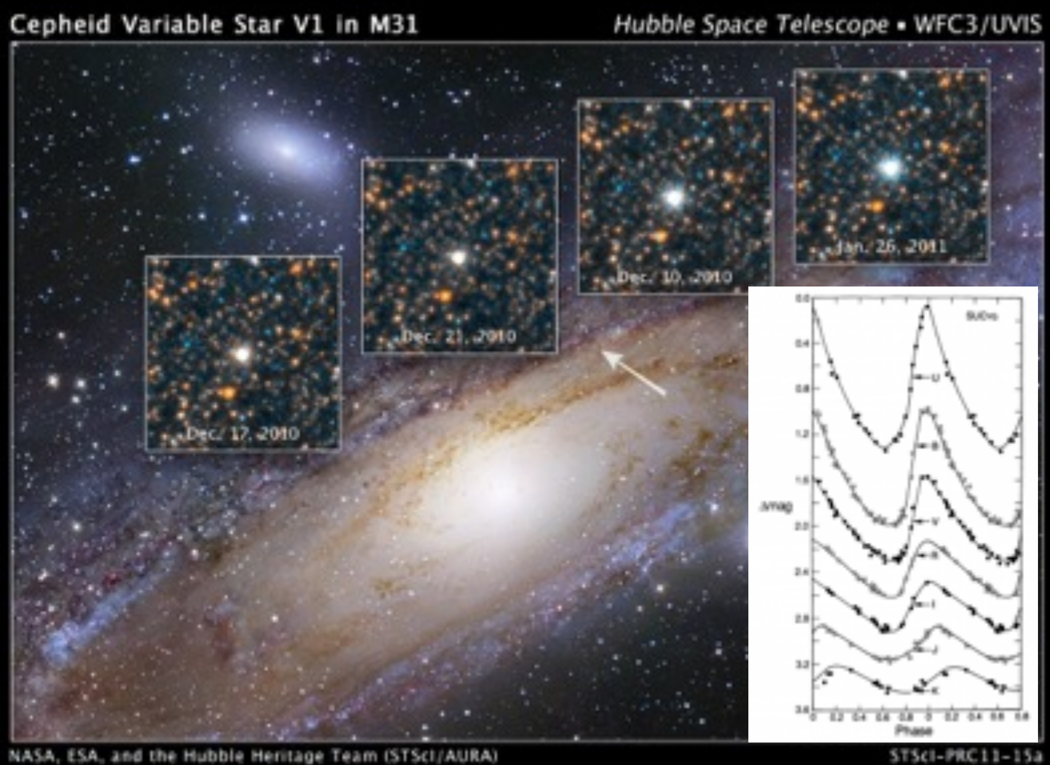
BUILDING A TRAINING SET FOR AN AUTOMATIC (LSST) LIGHT CURVE CLASSIFIER

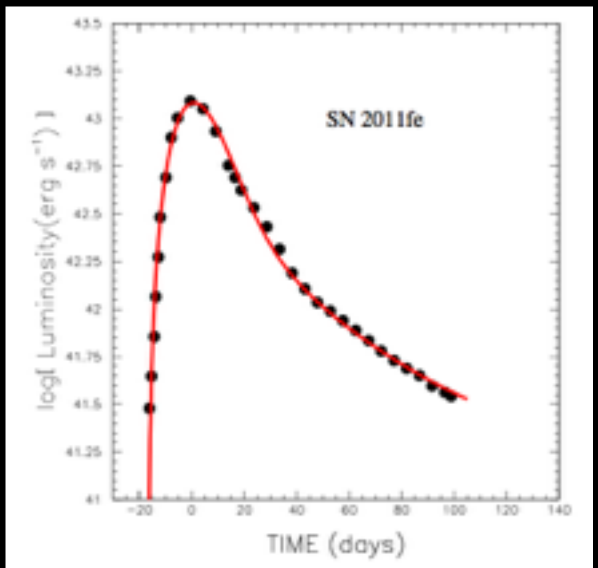
WITH: JAMES LONG, CHRISTINA LINDBERG, VIRISHA TIMMARAJU, JACKELINE MORENO, ASHISH MAHABAL, VIVEK KOVAR AND THE SAMSI WG2

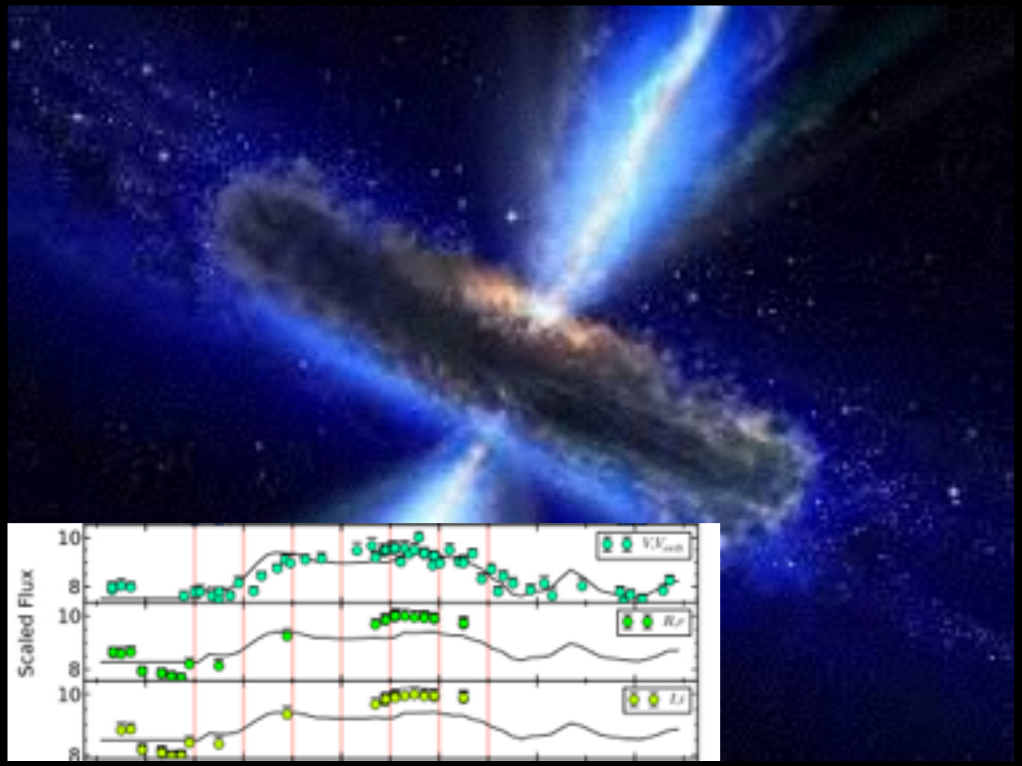
HARVARD-SMITHSONIAN CENTER FOR ASTROPHYSICS

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



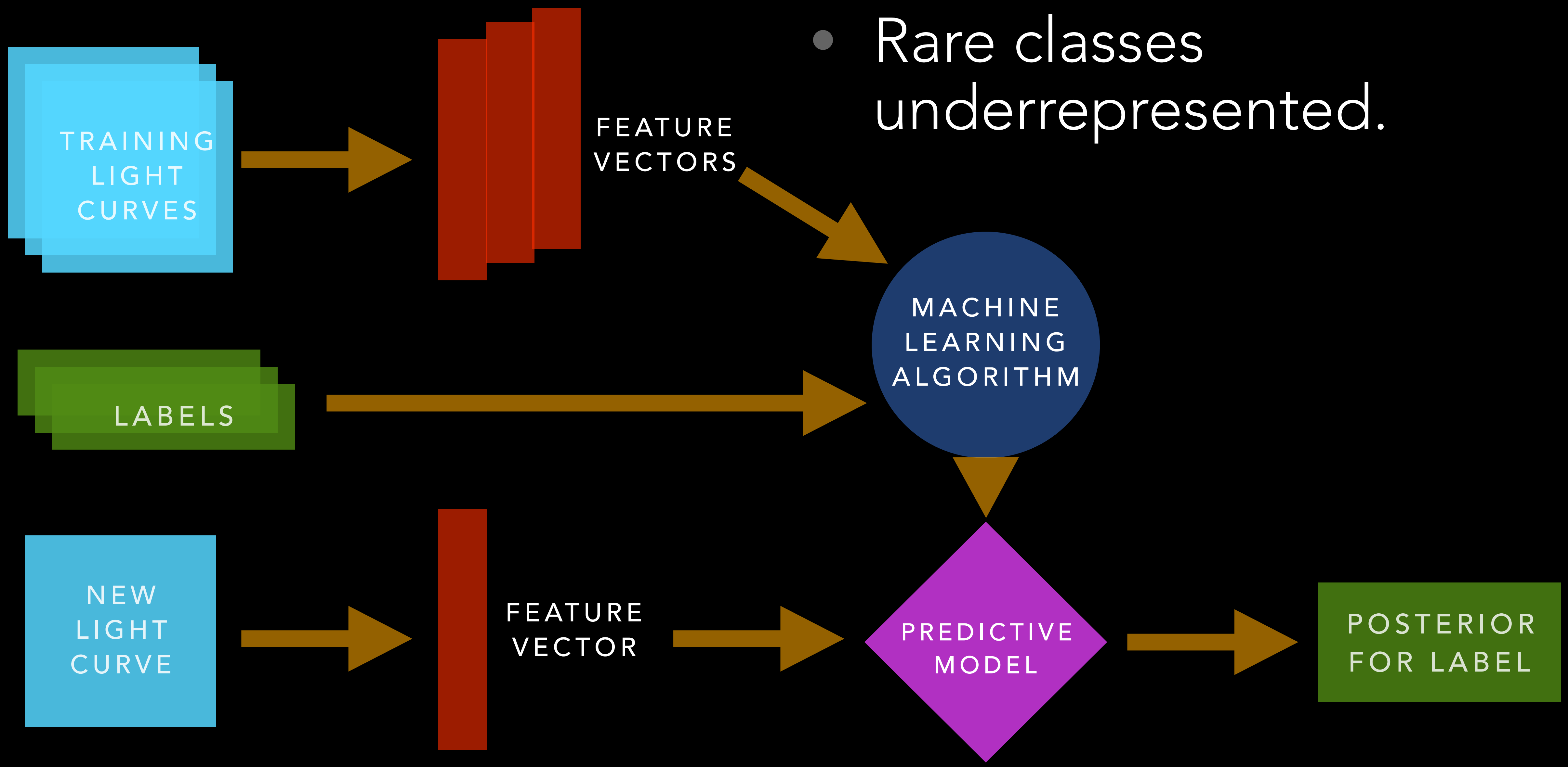




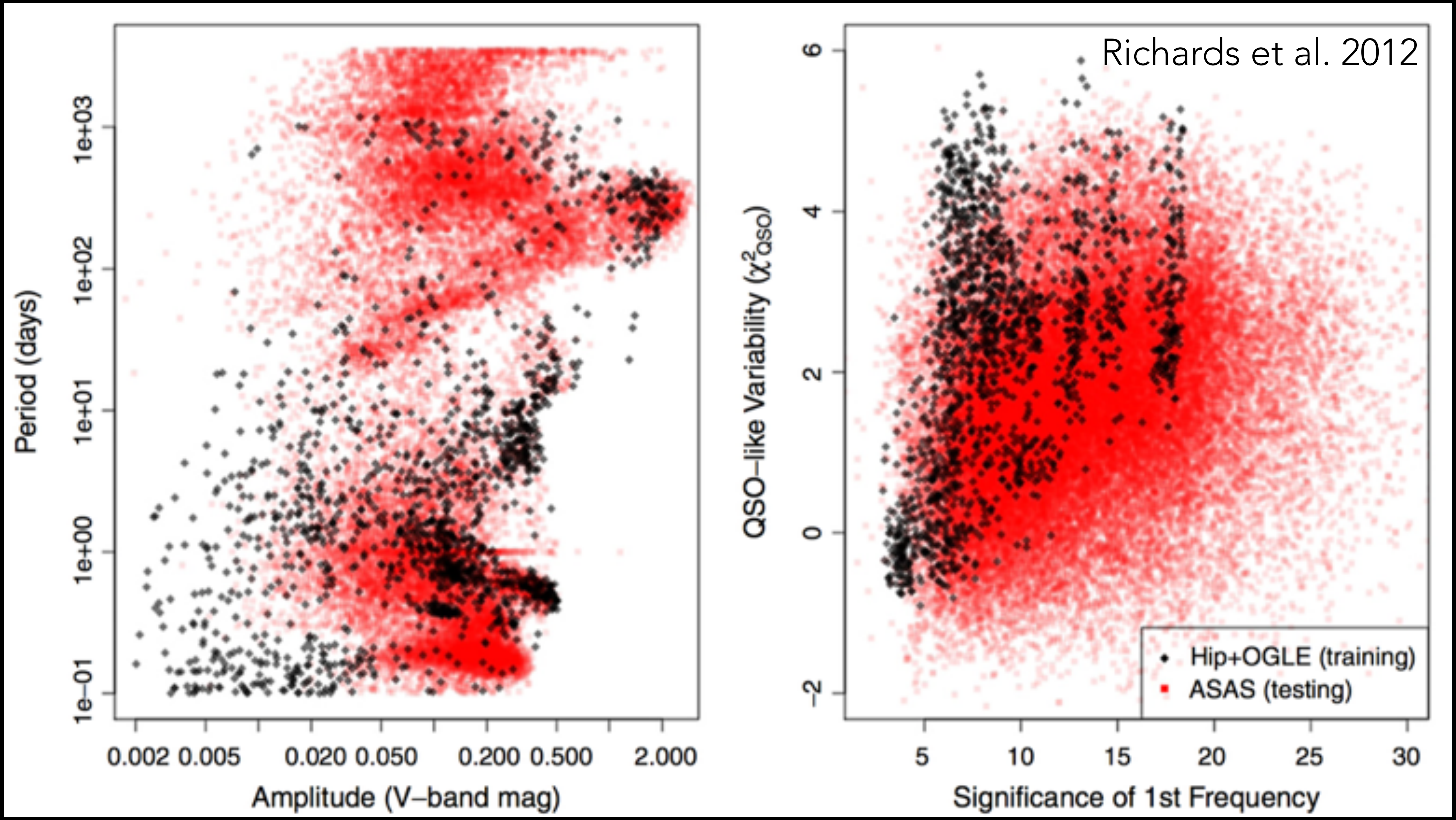
WHERE CAN THINGS GO WRONG?

1. Training set

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



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.
- WE SHOULD BE ABLE TO MAKE PLOTS LIKE THIS FOR THE MODELS¡¡

THE samsi ASTRO PROGRAM

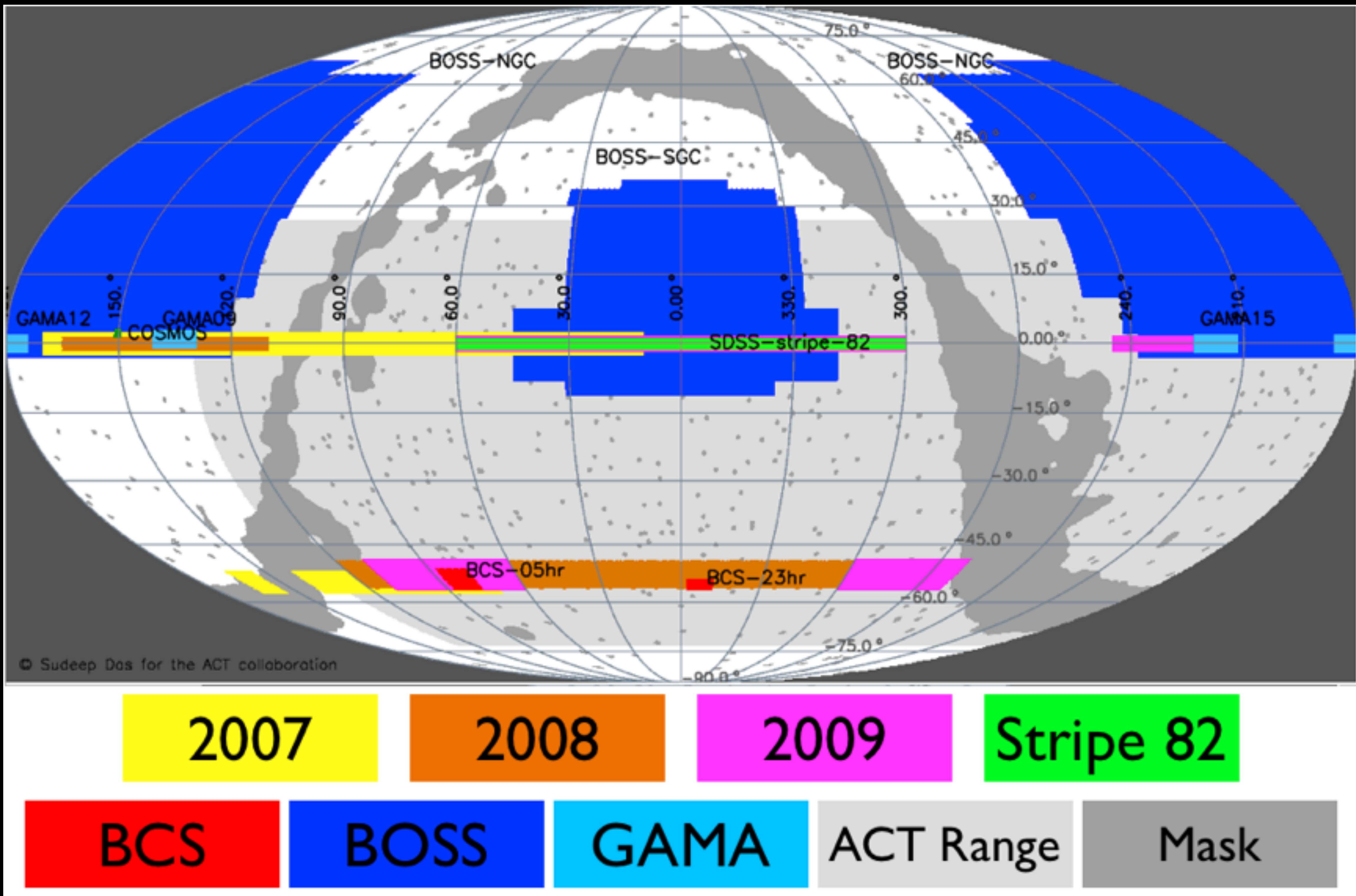
NSF Duke NCSU UNC

- Program on Statistical, Mathematical and Computational Methods for Astronomy.
- Several working groups, one of relevance to us:
  - Working Group II: Synoptic Time Domain Surveys
    - Subgroup 1: Data Challenge
  - Big questions:
    - Statistical approaches to characterize and quantify features. This should be applicable to data AND models.
    - Are there specific domain-knowledge based features that can be identified to improve class discrimination?
    - Advantages of a data-approach to the challenge

DATA VS. MODELS. PROS AND CONS

- How realistic are models? Do we have models for all kinds of transients, periodic, and stochastic sources? Do they properly account for outliers?
- Survey datasets can be complementary to models.
- But with models we know (in principle) the ground truth and can simulate any cadence.
- Can we somehow combine data and models to produce a more robust challenge? By attempting classification of datasets with a model-trained classifier? Or by checking models against outliers?

THE SDSS STRIPE 82

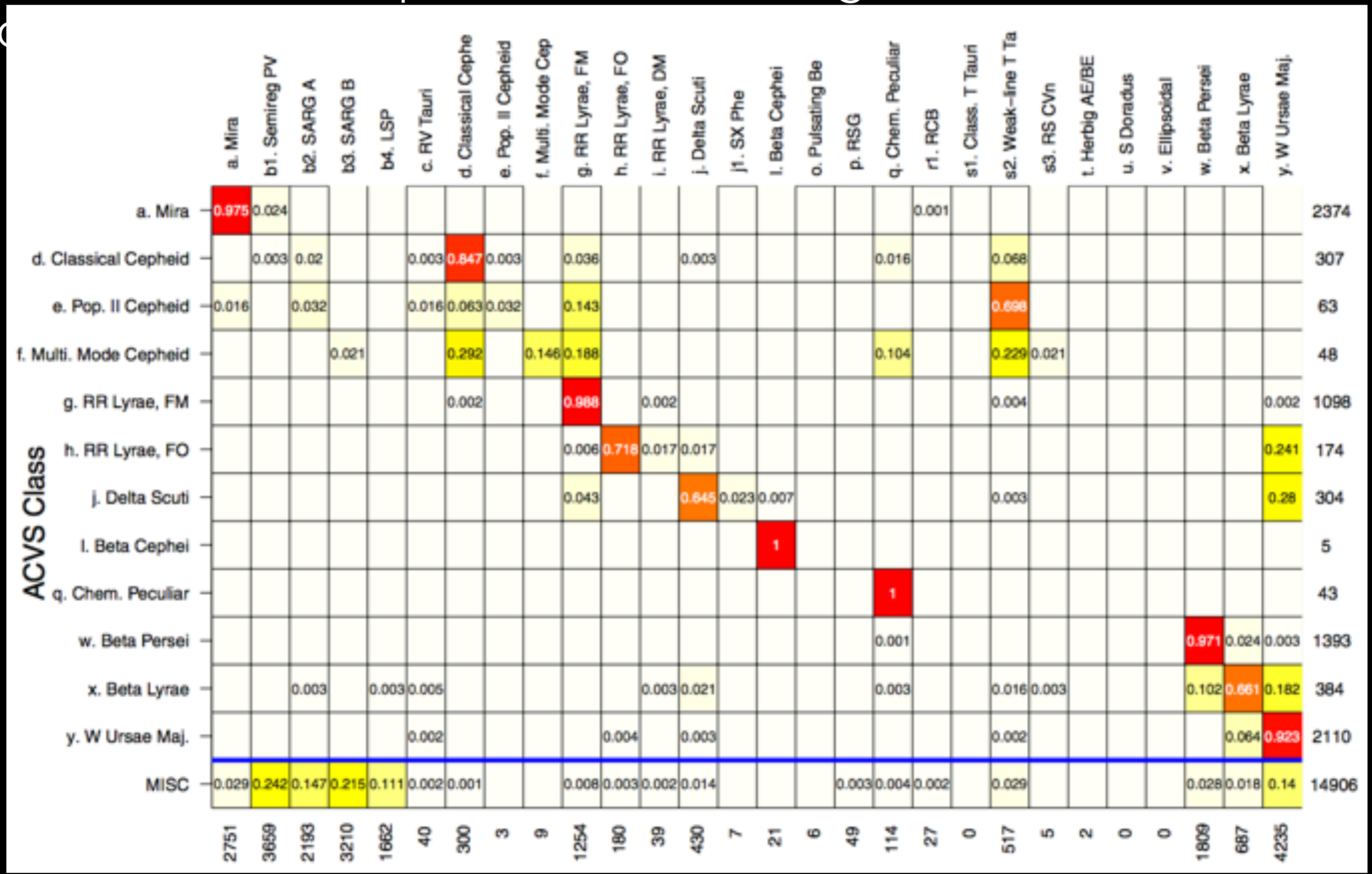


WE ARE BUILDING A TRAINING/TEST SET USING STRIPE 82 SOURCES

- The catalog has ~60K light curves in bands u,g,r,i,z, with about ~50 observations per LC.
- We have a github repository with code to download the dataset, gather existing literature labels, merge the classifications, and split the dataset into training and testing sets: https://github.com/jpl2116/stripe82-class
- We have also tested code to:
  - Inspect variability of sources, and make a census of the different source classes (QSOs, RR Lyrae, Delta Scuti, eclipsing binaries, etc.)
  - Perform feature extraction
  - Test supervised and unsupervised classification methods (random forests, K-means, clustering) - Next talk by Virisha.
  - Identify outliers, and discover the weirdest objects.

SOME NUMBERS

- Our catalog has ~60K sources.
- We have identified labels for ~10% of those sources. Here is the break up of the sources with labels:
  - QSOs: 86%
  - RR Lyrae: 8%
  - ew+ea+eb: 4.1%
  - Delta Scuti: 1.5%
- We are currently merging our Stripe 82 catalog with the CRTS, and the Richards et al. probabilistic catalog.



A TOOL FOR FEATURE EXTRACTION

Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Feature Analysis for Time Series

Authors: Isadora Nun isadoranun@g.harvard.edu , Pavlos Protopapas pavlos@seas.harvard.edu

Contributors: Daniel Acuña, Nicolás Castro, Rahul Dave, Cristobal Mackenzie, Jorge Martinez, Adam Miller, Karim Pichara, Andrés Riveros, Brandon Sim and Ming Zhu

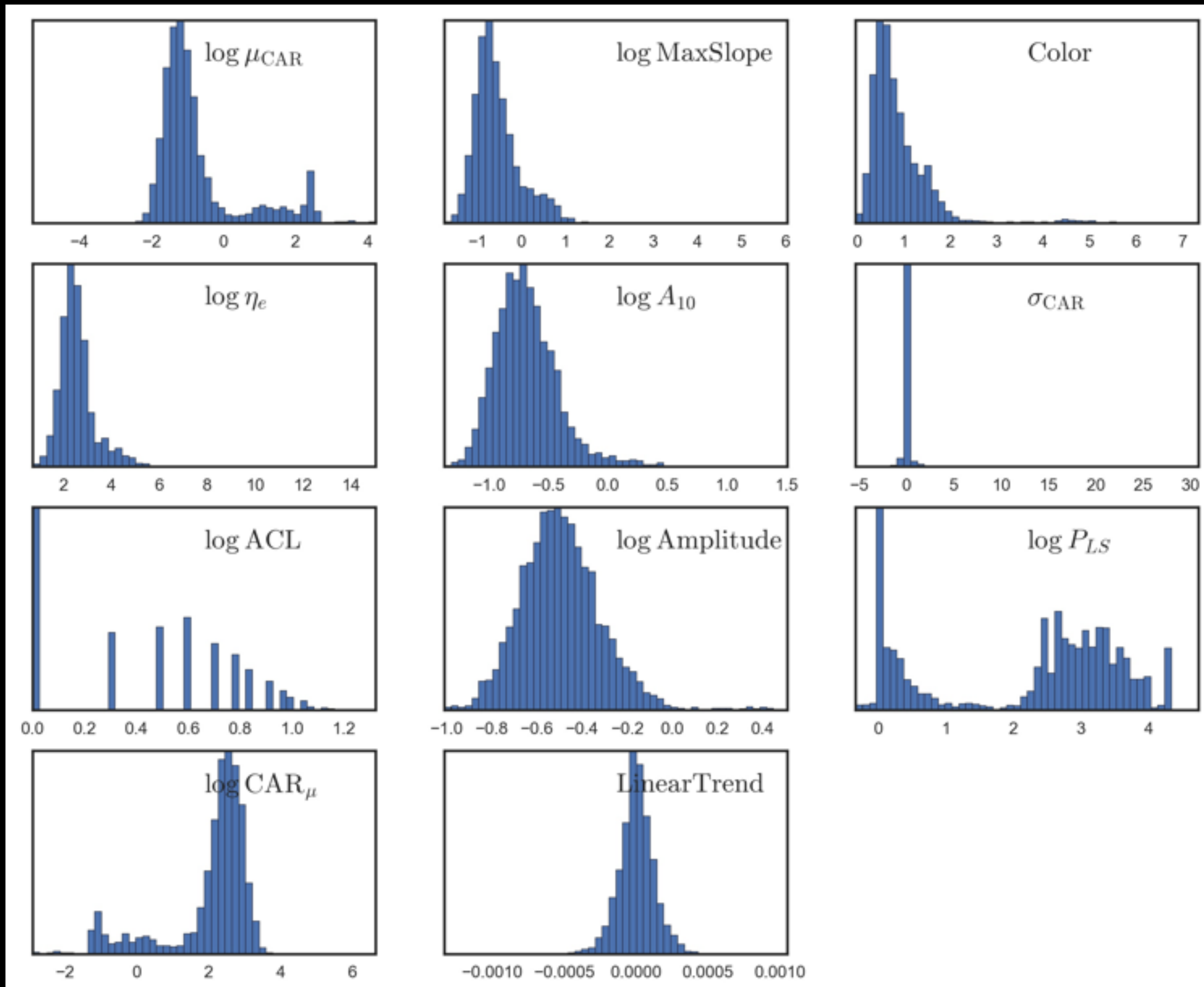
We want to improve this:

See: http://isadoranun.github.io/tsfeat/FeaturesDocumentation.html

EXTRACTING FEATURES FROM IRREGULAR TIME SERIES

TYPE	EXAMPLES	
VARIABILITY	$\eta = \frac{1}{(N-1)\,\sigma^2} \sum_{i=1}^{N-1} (m_{i+1} - m_i)^2$	$\kappa = \frac{N\,(N+1)}{(N-1)\,(N-2)\,(N-3)} \sum_{i=1}^{N} \left(\frac{m_i - \hat{m}}{\sigma}\right)^4 - \frac{3\,(N-1)^2}{(N-2)\,(N-3)}$
PERIODICITY	$y(t\|\omega,\theta) = \theta_0 + \sum_{n=1}^{N} [\theta_{2n-1} \sin(n\omega t) + \theta_{2n} \cos(n\omega t)]\,.$	$A_{i,j} = \sqrt{a_{i,j}^2 + b_{i,j}^2}$ $\mathrm{PH}_{i,j} = \arctan\left(\frac{b_{i,j}}{a_{i,j}}\right)$
REGRESSION	$dX(t) = -\frac{1}{\tau} X(t) dt + \sigma_C \sqrt{dt}\epsilon(t) + b dt$ for $\tau, \sigma_C, t \geq 0$	CAR(1) MODELS
MULTIBAND PROPERTIES	COLOR	$I = \sqrt{\frac{1}{n(n-1)}} \sum_{i=1}^{n} \left(\frac{b_i - \hat{b}}{\sigma_{b,i}}\right) \left(\frac{v_i - \hat{v}}{\sigma_{v,i}}\right)$

FEATURE EXTRACTION



RESULTS ON STRIPE 82 SOURCES

Period Extraction

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