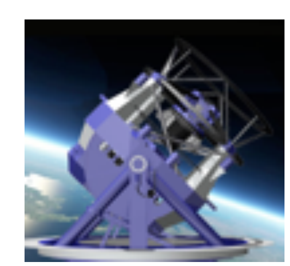
SAMSI ASTRO WG2 and LSST Informatics

Ashish Mahabal
aam at astro.caltech.edu
Center for Data Driven Discovery, Caltech
Co-Chair, LSST Transients and Variable Stars SC

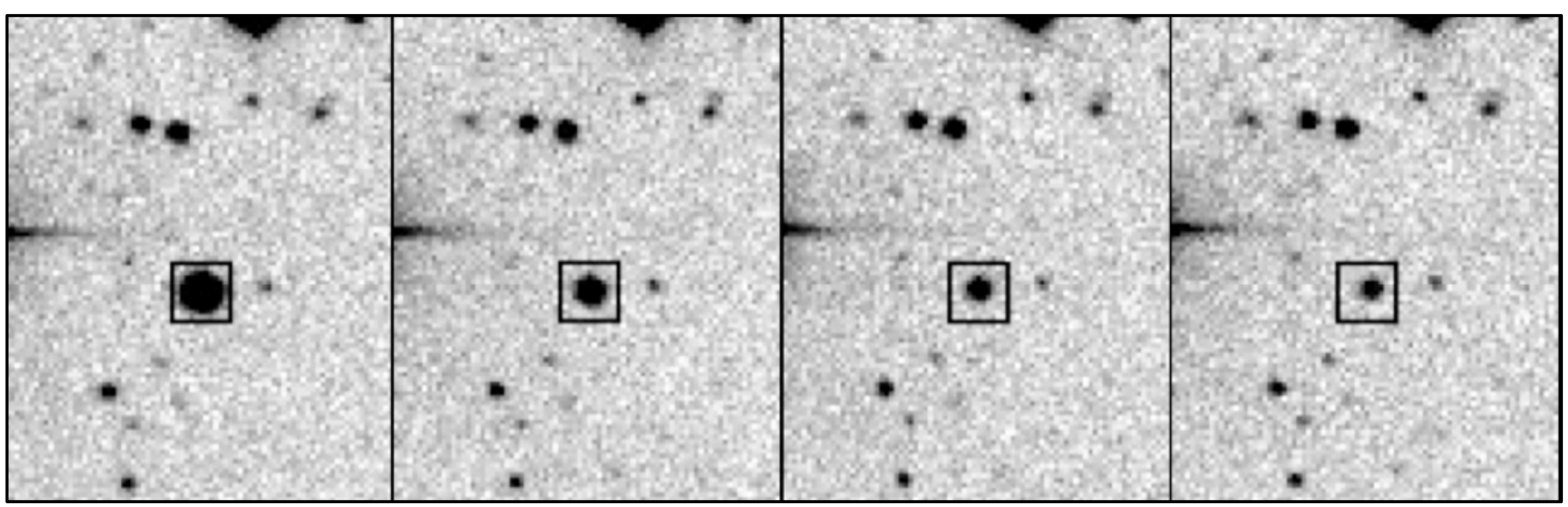


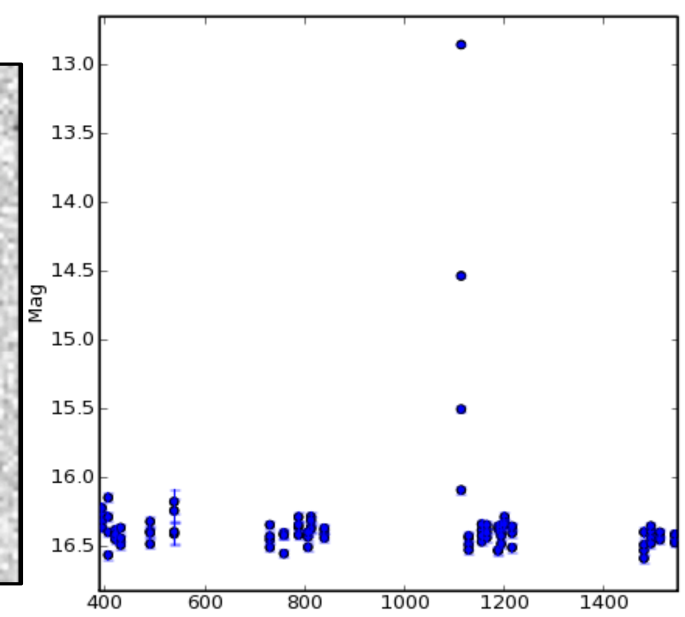
Outline

- Tom Loredo already spoke about the overall ASTRO program

- A few weeks back Federica Bianco spoke about LSST TVS

- **WGII: Synoptic Time Domain Surveys**





WG2 subgroups

Overall leaders:
Ashish Mahabal
Jogesh Babu

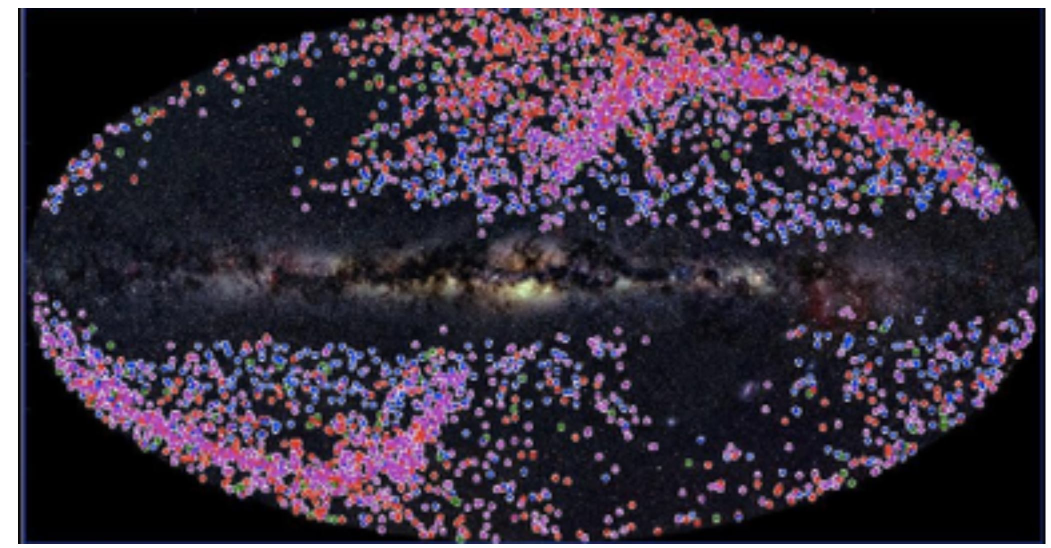
1. Data Challenge
2. Designer Features
3. Scheduling Obs
4. Interpolating Lightcurves
5. Incorporating Non-Structured Ancillary Info
6. Outlier Detection
7. Domain Adaptation
8. Lightcurve Decomposition

Interconnectivity of the subgroups

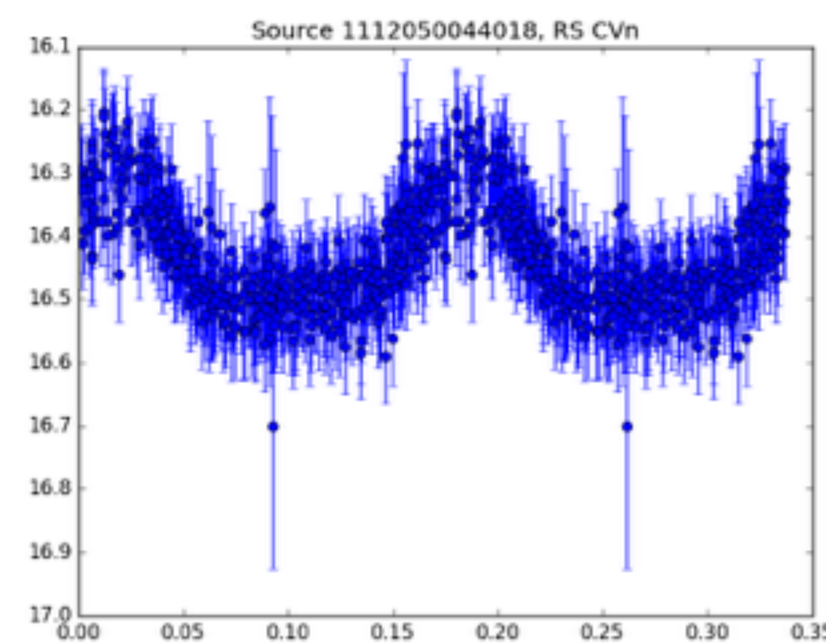
~25 members
Opening Workshop
biweekly telecons
Follow-up meetings
Connection to LSST "community"

Intricacies of a data challenge

- SNe data challenge (Kessler et al.)
  - full light-curves
  - first six data points



- Great3 challenge (Cosmology)
- Kaggle (Widely popular platform)
- Our plans: new challenge



Transients and brokers

- Expected rate: 1-10 million transients per night
- Majority will be well understood classes
- Early characterization crucial to follow-up rare classes
- Two-tiered challenge to ensure astronomers and non-astronomers participate
- Challenge: Gappy, sparse, heteroscedastic lightcurves

10^7 transients

10^3 rare transients

Data challenge details

- Possible Datasets:
  - Catalina Real-Time Transient Survey
  - MACHOs survey
  - OGLE
  - Pan-STARRS
  - PTF
  - SDSS STRIPE82

Simulations
Theory

Lead: Rafael Martinez-Galarza

Peter Freeman
Matthew Graham
Shashi Kanbur
Vivek Kohar
James Long
Ashish Mahabal
Wenlong Yuan

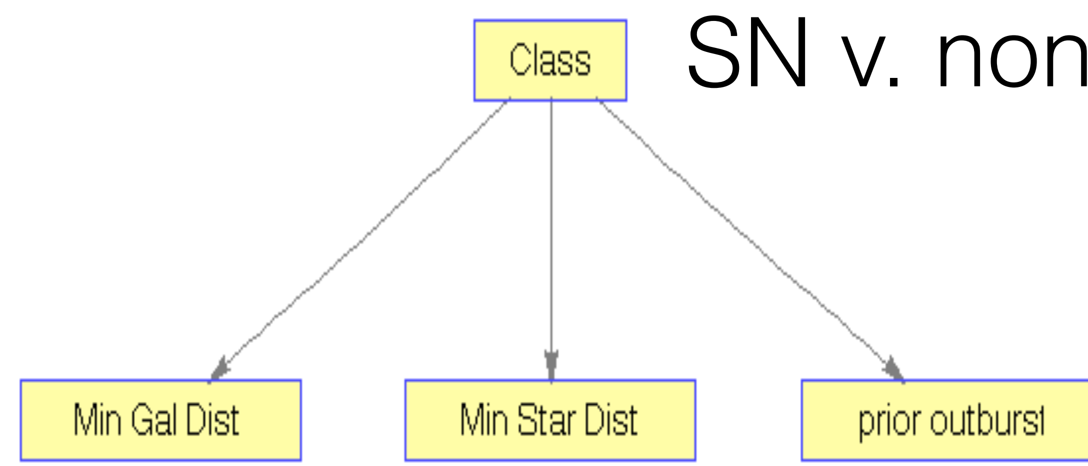
https://community.lsst.org/t/data-challenge-to-characterize-transient-and-variable-objects/1061/14

Designer features

Matthew Graham
Ashish Mahabal

- Supernova from just archival information
- R Cor Bor plateaus
- Role of ancillary data (e.g. archival radio source)

Also based on lightcurve decomposition



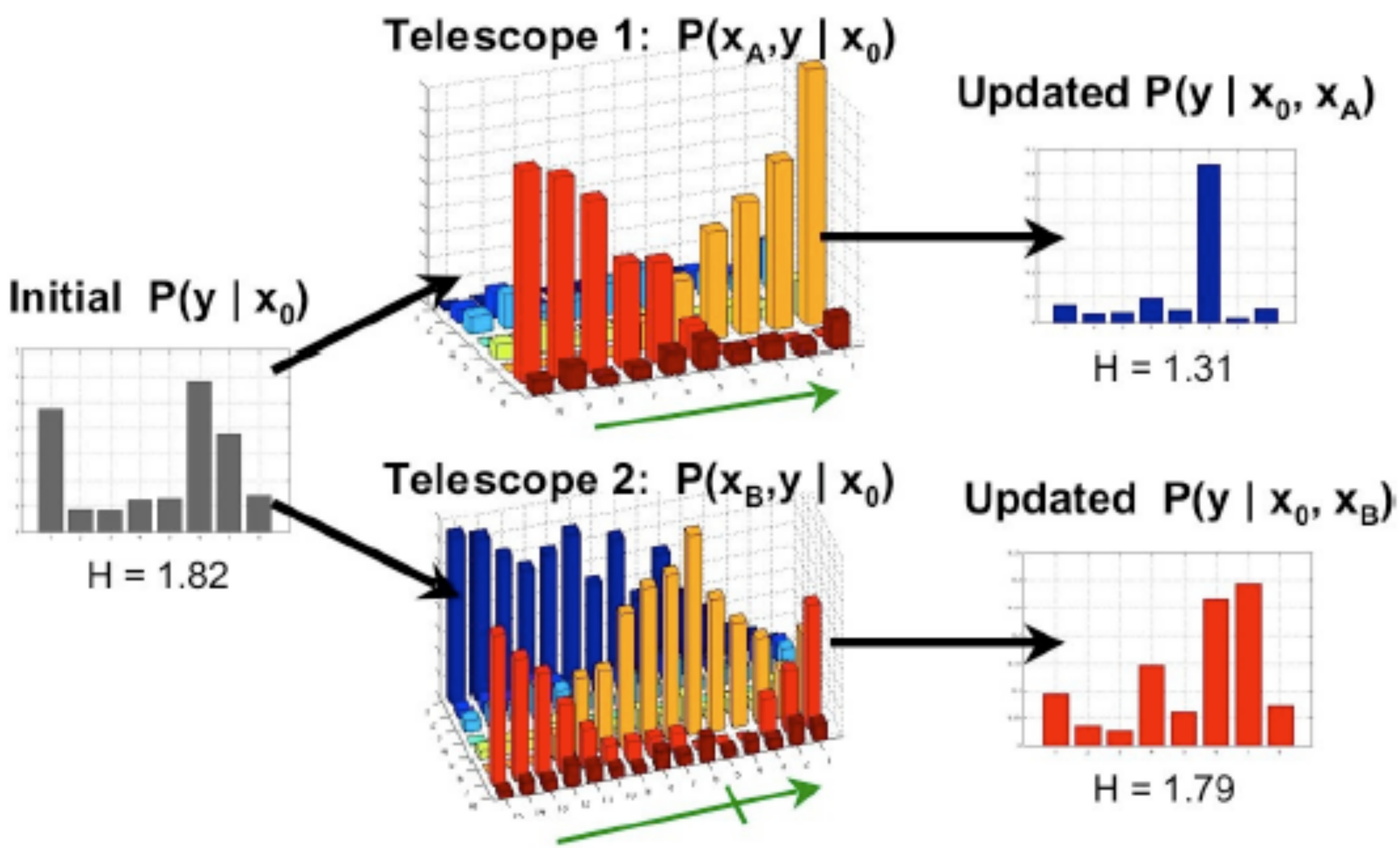
SN v. non-SN

normalized

based on peaks

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

Scheduling observations



A possible bayesian approach

Scheduling observations

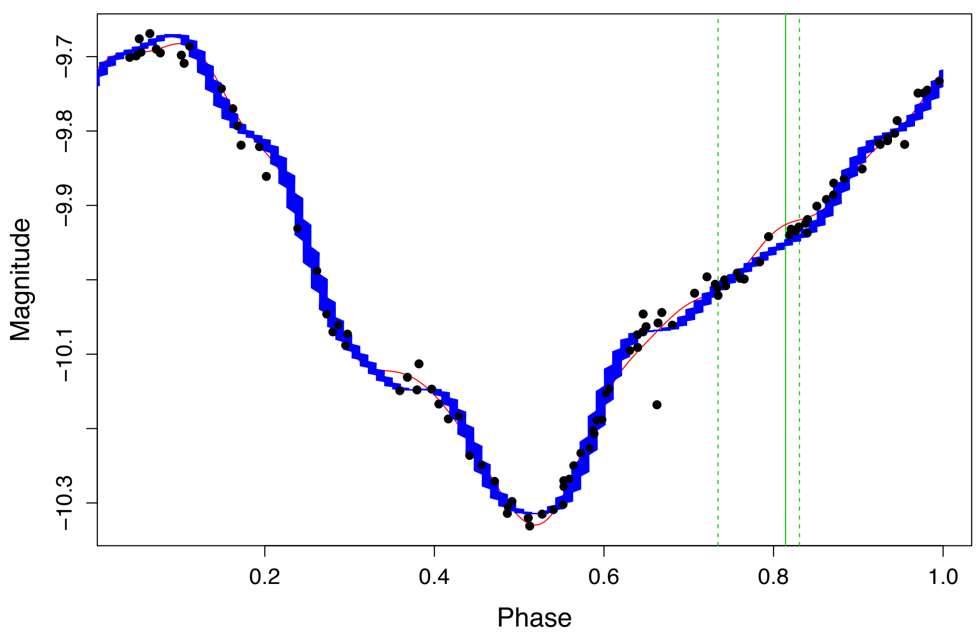
Lead: David Jones

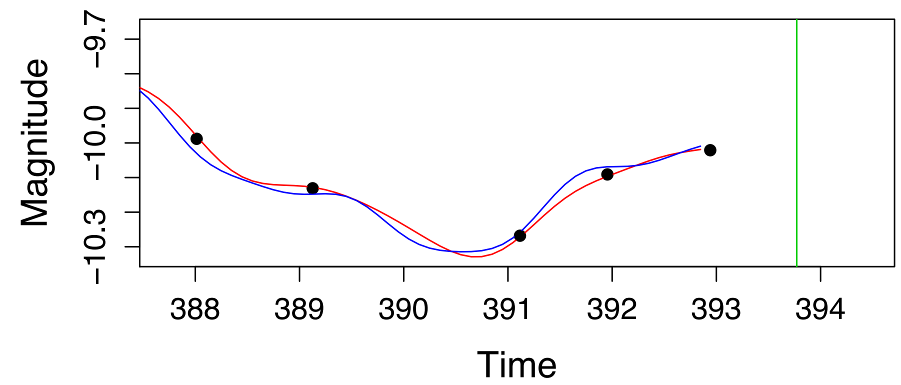
Sujit Ghosh, James Long,
Zhenfeng Lin, Ashish Mahabal

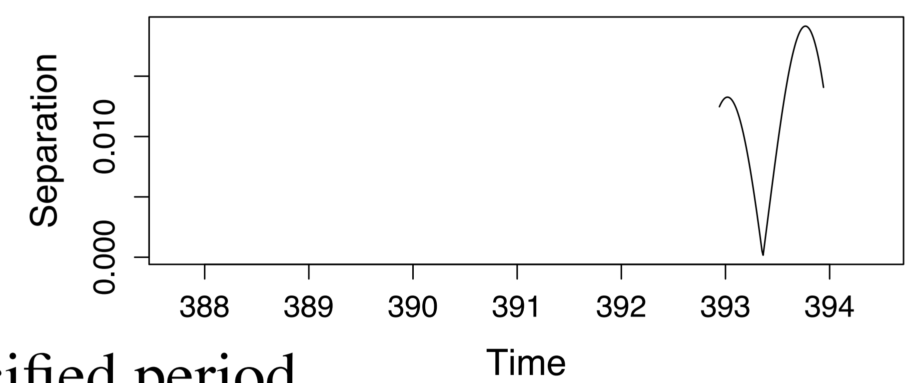
- Basis models for lightcurves (computationally efficient approx. to GPs)
- Basis coefficients have different prior for each class
- Training / prior construction step: use Stan to fit Bayesian hierarchical model that shares information between lightcurves of the same class
- For a new lightcurve: get posterior draws of "separation" (can be chosen) between models at different future observation times

Scheduling observations

Toy Cepheid example







Class / Model 1: basis model with correct period
Class / Model 2: basis model with slightly misspecified period

Left: solid green line shows the optimal (posterior mean) time for a new observation in a one day interval indicated by vertical dashed lines. Red and blue curves show current posterior mean fits for models 1 and 2.

Right: top shows the optimal observation time with the two model means plotted for a single posterior draw of the parameters. Bottom shows the corresponding posterior draw of the separation between the model means

Interpolating light-curves

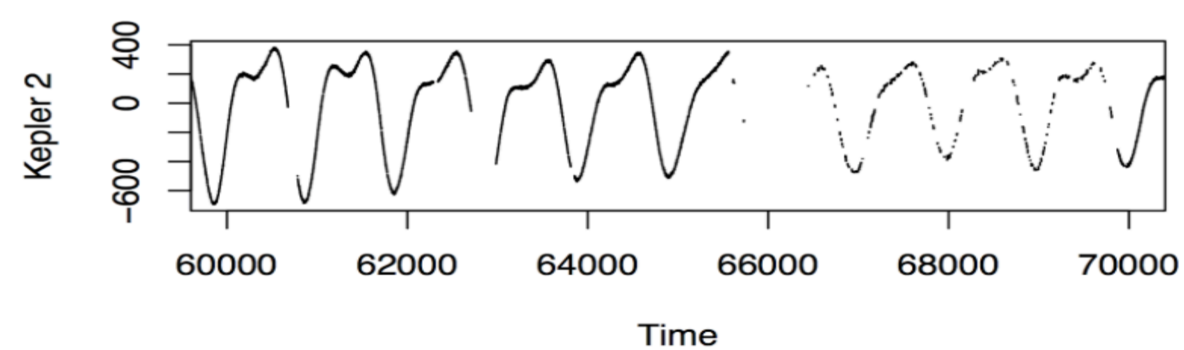
Fourier decomposition (Bharadwaj, 2015)
PCA (Deb and Singh, 2010)
Emperical mode (Wysocki et al 2016)
Non-linear mode (Latsenko et al 2012)
Dynamical systems theory

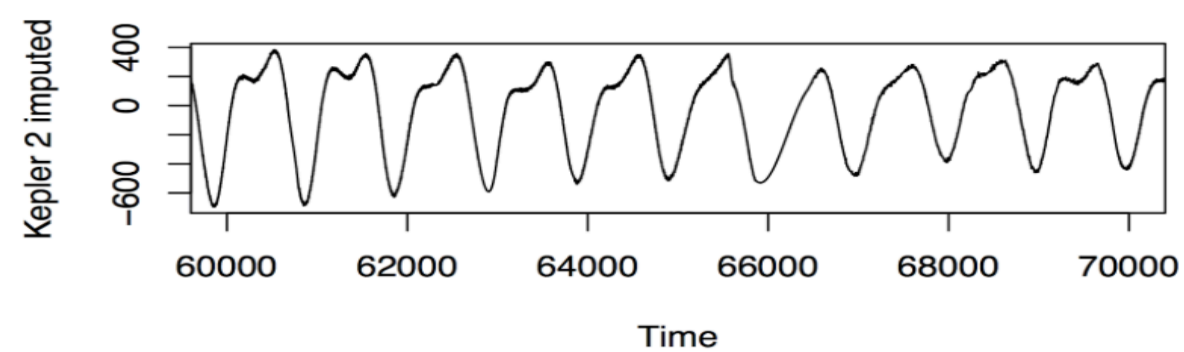
R methods:
Amelia, ImputePSF, mtsdi
ARIMA autoregressive models
Gaussian state space models

na.kalman of imputeTS (Arima 0,2,2
Kalman filter max likelihood
seasonal component

Lead: Shashi Kanbur
Erik Feigelson
Vivek Kohor
Rafael Garrido Haba

KIC 007609553





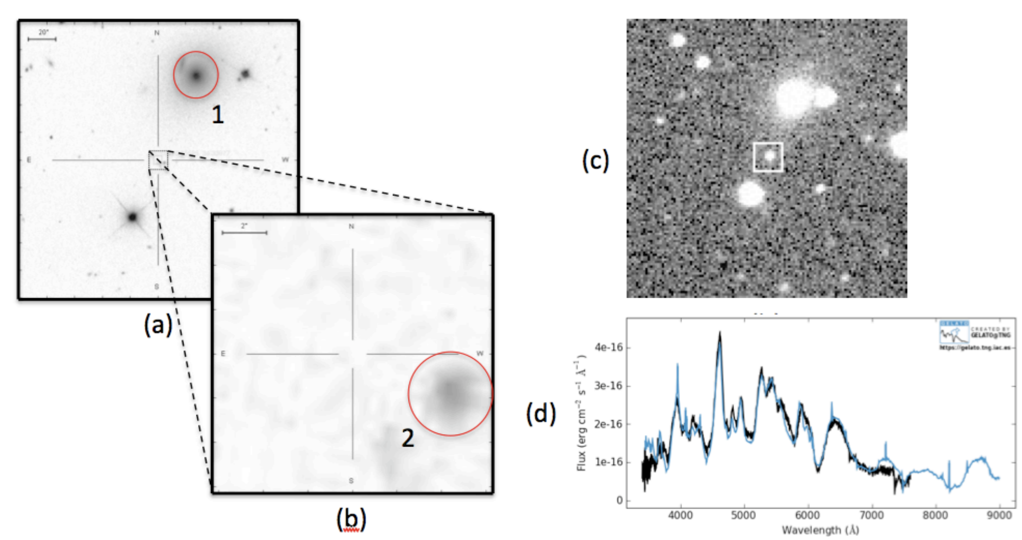
29.4 min cadence

Incorporating ancillary info

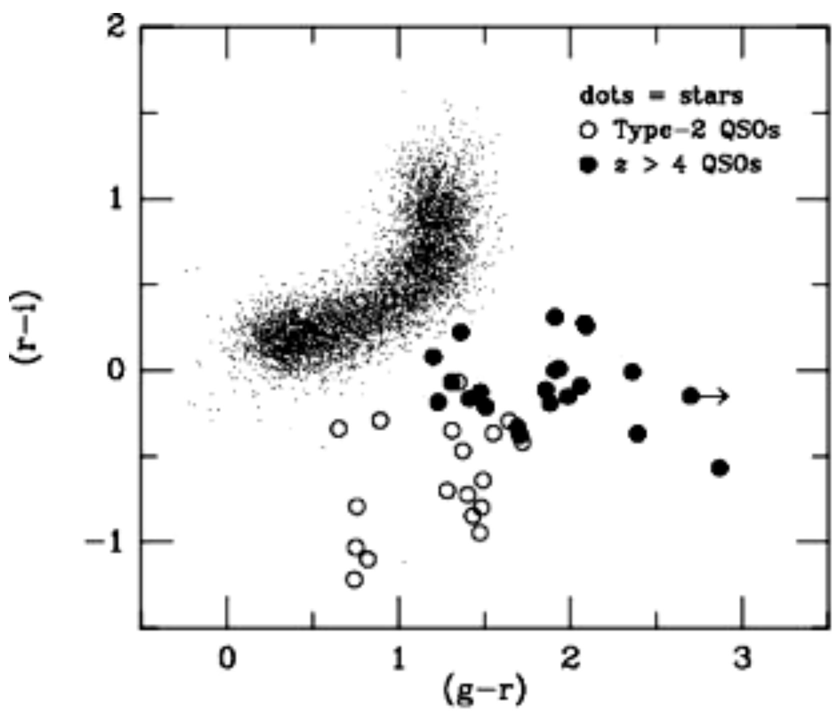
Lead: James Lang
David Jones
Ashish Mahabal

- Parameters like
  - Galactic latitude (Galactic versus extra-galactic)
  - Nearest galaxy (Supernova versus non-)
  - Nearest radio source (blazar or not)

Natural language
Best guesses



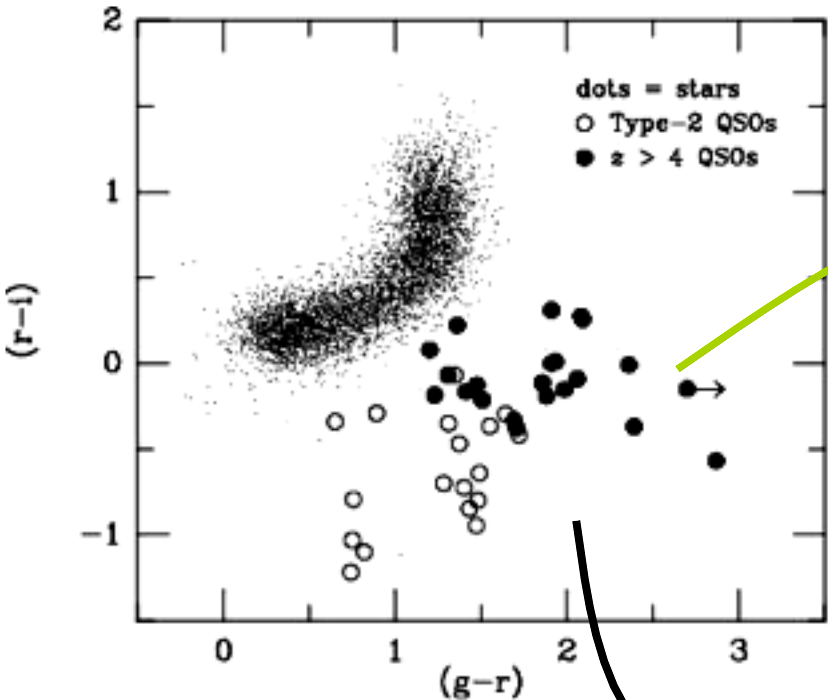
Outlier detection



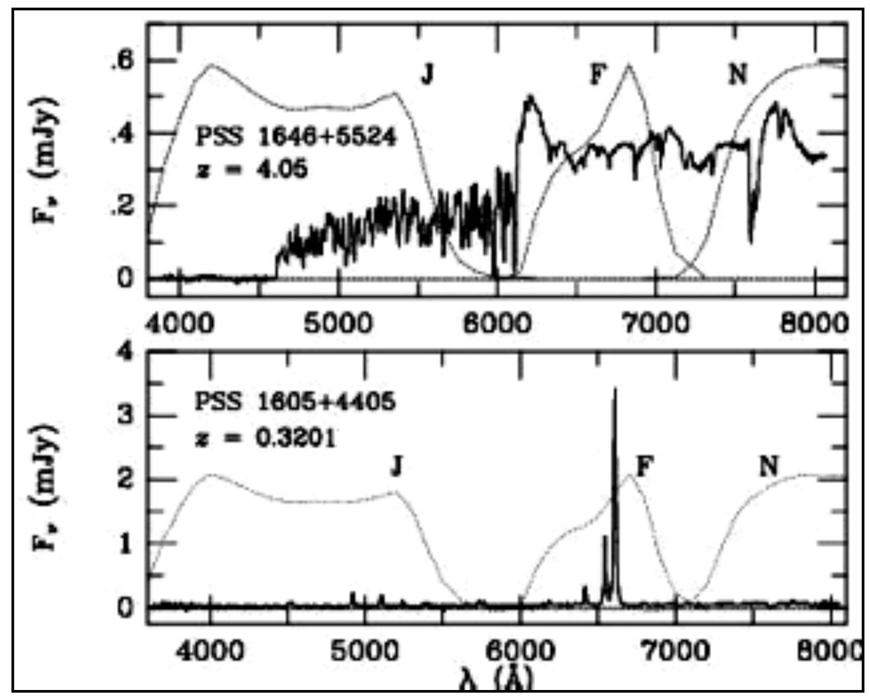
- The importance: new species, new subspecies
- New physics

Tests:
Gaussianity:
Dimensionality:
Local Outliers (Hierarchical):

Outlier detection



hi-z qso



type-II qso

Ashish Mahabal
Soumendra Lahiri
Jogesh Babu

Matthew Graham,
David Jones,
Zhenfeng,
Ji Meng

Methods:

Clustering: objects not belonging to any cluster are outliers.
(noise, natural distribution in the dimensions considered)
Model-based: Separate objects by goodness-of-fit
Mixture of Experts

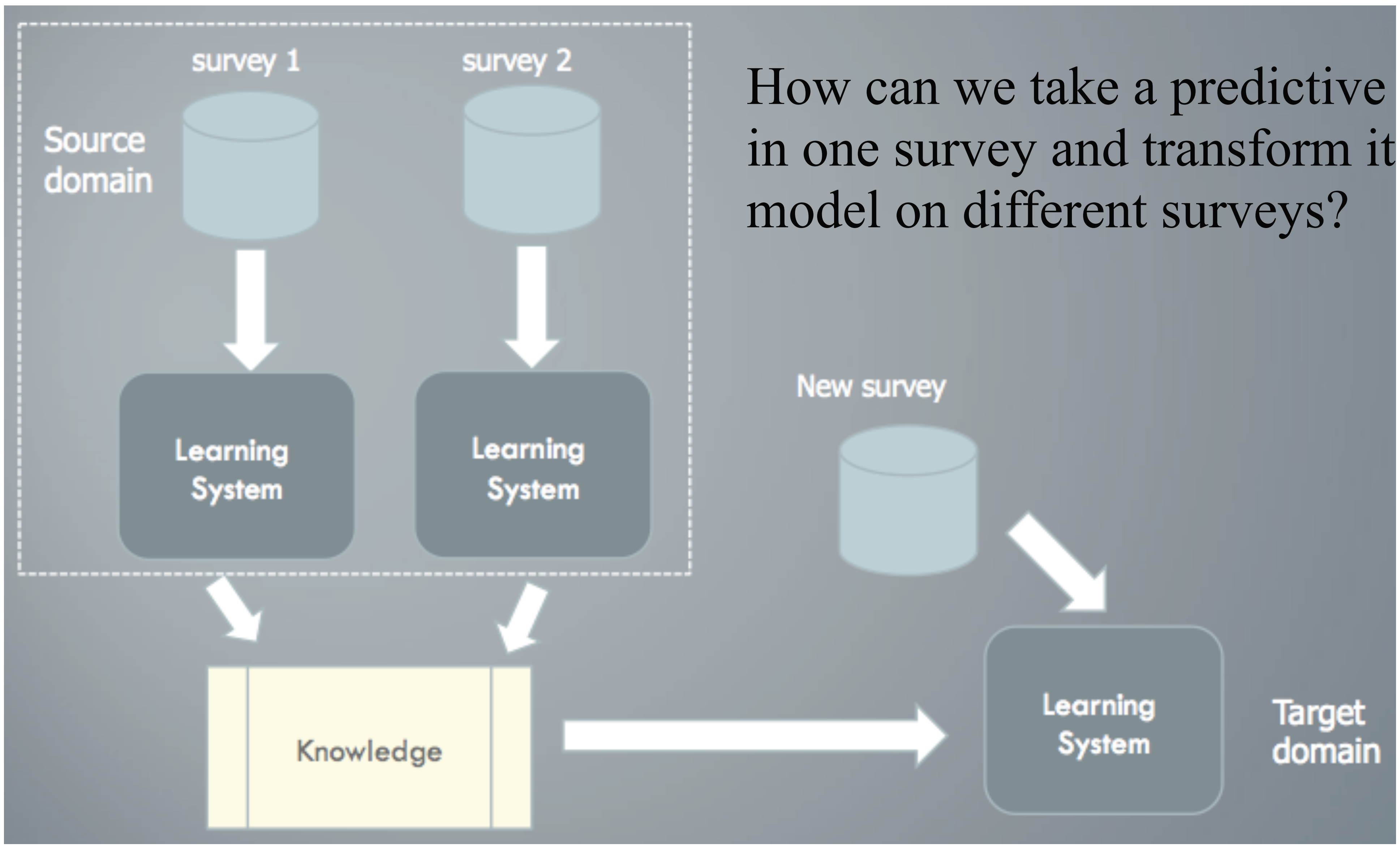
Domain Adaptation to Learn Predictive Models Across Astronomical Surveys

Lead: Ricardo Vilalta

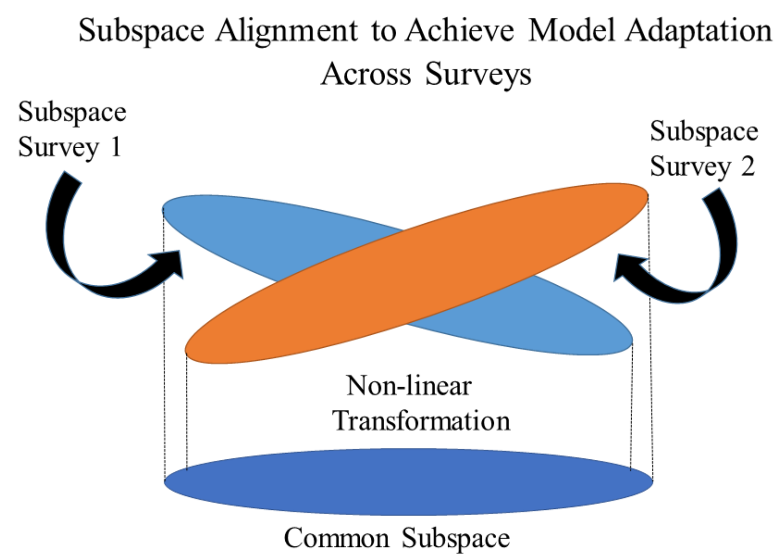
Jogesh Babu
Ashish Mahabal
Ji Meng

How can we exploit information from multiple surveys simultaneously to obtain more accurate predictive models?

How can we take a predictive model obtained in one survey and transform it into an accurate model on different surveys?



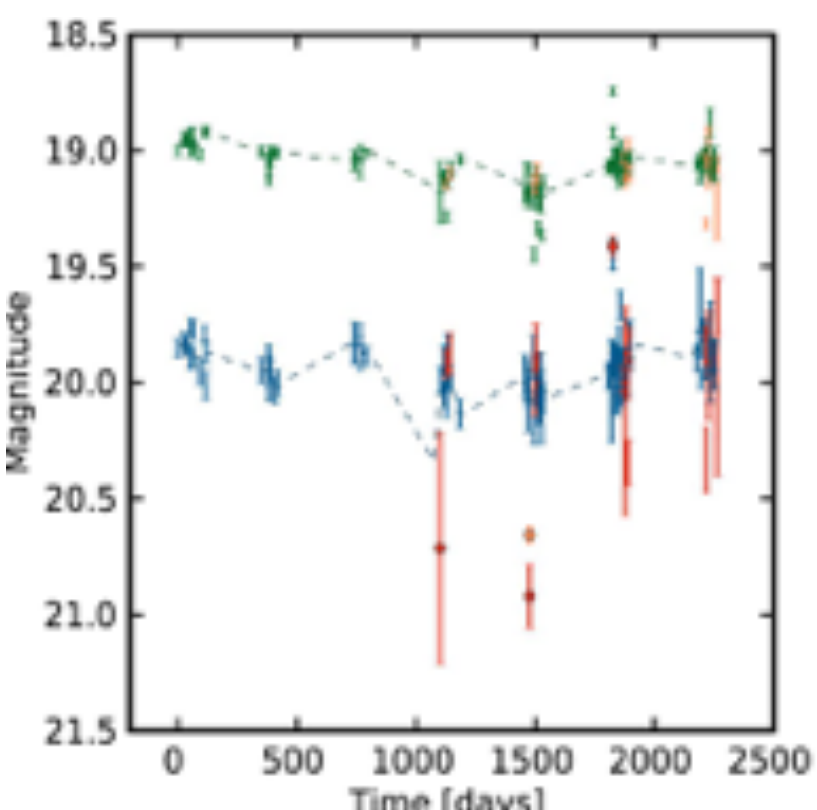
Model Adaptation …

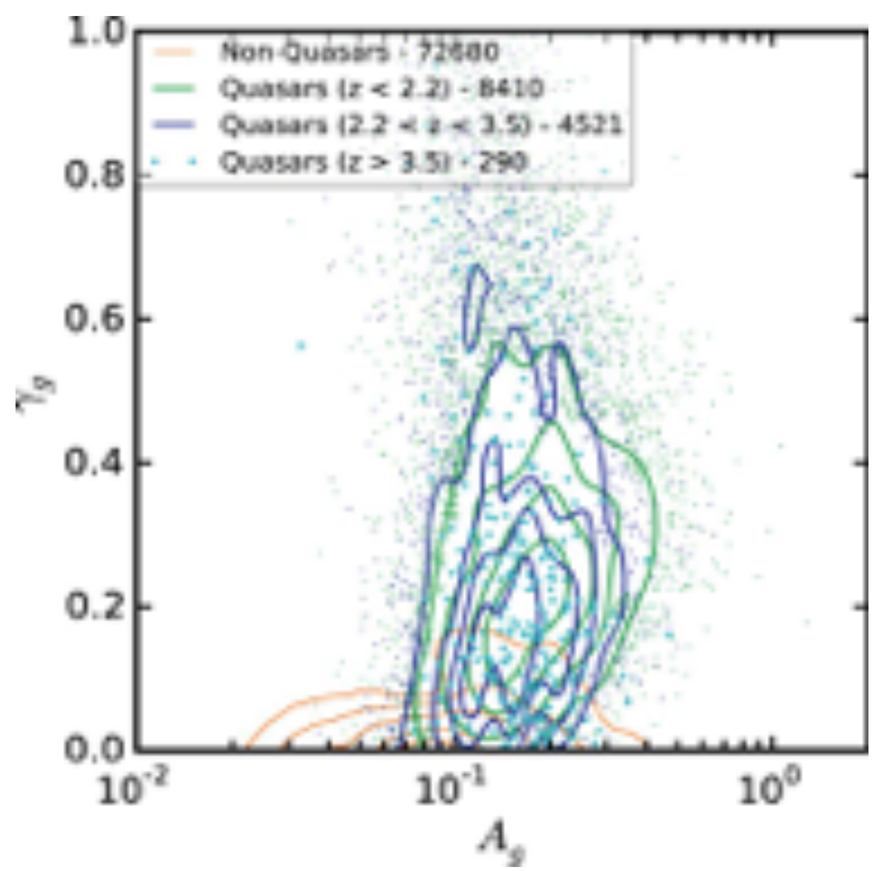


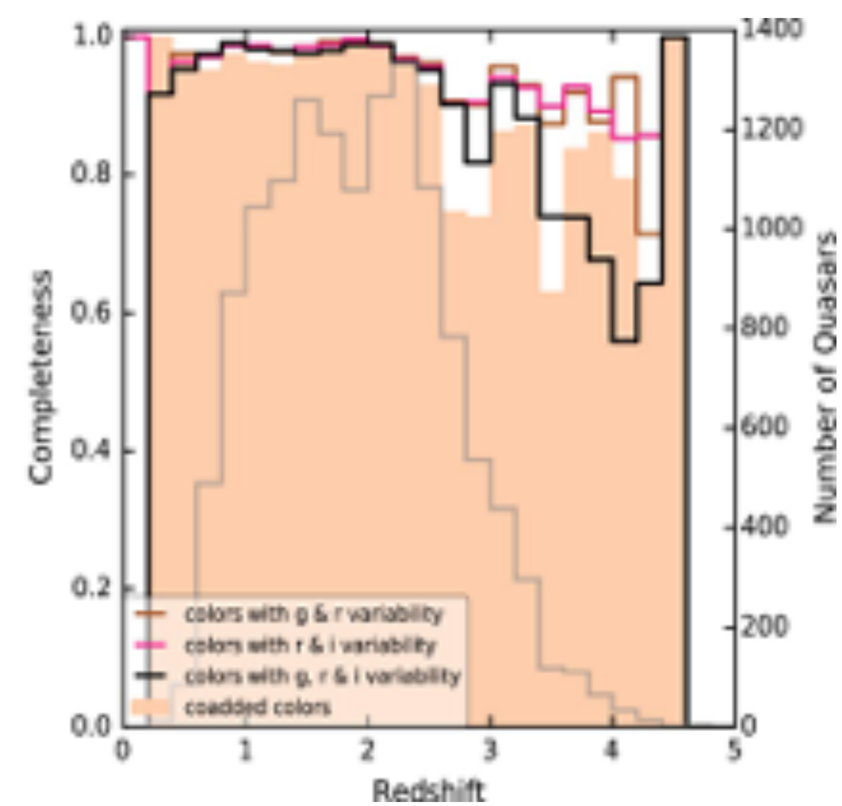
Find a common subspace where source and target domains overlap. Once source and target are mapped into a common subspace, a model trained on the source domain can be used on the target domain.

Lightcurve decomposition

To characterize data with a random component, a trend or cyclic variability of interest
To classify objects based on lightcurve signatures or parametrizations of changes in brightness (Peters et al. (2016), Schmidt et al. (2010), MacLeod et al. (2011))





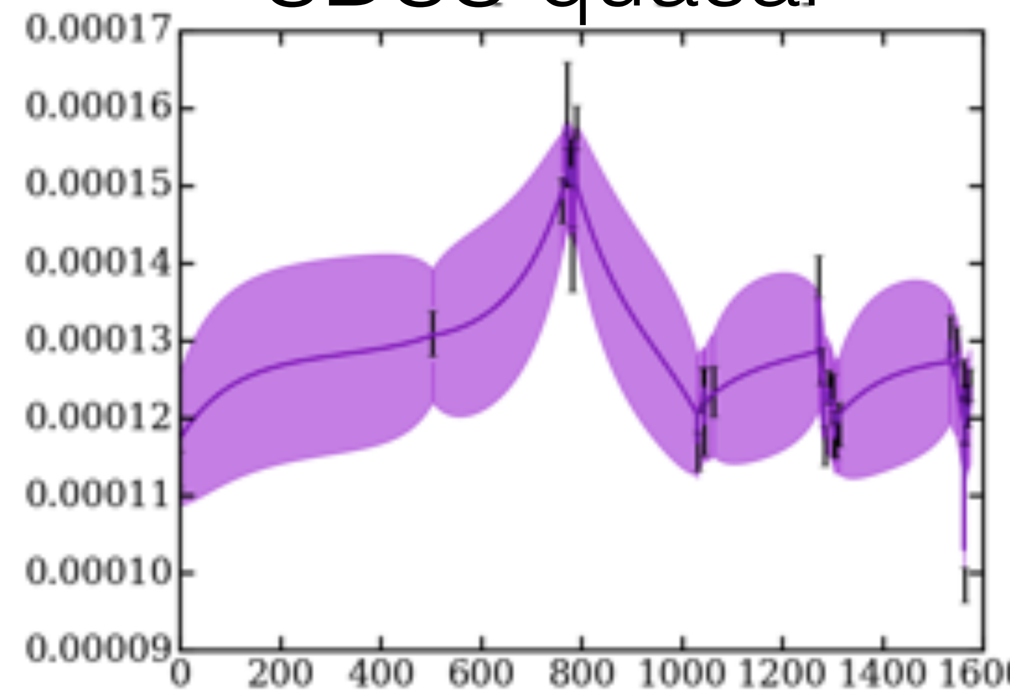


Lightcurve decomposition

Lead: Jackeline Moreno

Garrido
Sujit Ghosh
Matthew Graham
Shiyuan He
David Jones
Shashi Kanbur
Vivek Kohar
Soument Lahiri
Ashish Mahabal

SDSS quasar



time (days)

This group is taking a closer look at

CARMA (auto-correlated behavior at various timescales + random disturbances)

CARIMA (non-stationary process)

CARFIMA (long memory process)

Continuous time models are necessary for irregularly sampled data like that which will be taken by LSST

Summary

Please join the fun!

- Interconnectedness of the work
    - Classification is one of the over-arching themes
    - Nature of light-curves: filling gaps, decomposing them, features to separate classes, subspaces to match cadences, determining outliers, incorporating ancillary information and determine best times to classify the sources
    - That is the grand (data-)challenge

Informatics contacts: Tom Loredo, Chad Schafer
TVS contacts: Ashish Mahabal, Federica Bianco

aam at astro.caltech.edu