LSST Informatics and Statistics Science Collaboration

Chad Schafer
Carnegie Mellon University
cschafer@cmu.edu

Overview

- One of the 10 LSST Science Collaborations
- http://issc.science.lsst.org
- "Core" Team:
  - Jogesh Babu, Penn State Univ.
  - Tamás Budavári, Johns Hopkins Univ.
  - Eric Feigelson, Penn State Univ.
  - Tom Loredo, Cornell Univ. (co-chair)
  - Chad Schafer, Carnegie Mellon Univ. (co-chair)
  - Sam Schmidt, UC Davis
  - Robert Wolpert, Duke Univ.

Objective

- To provide support to the other science collaborations, and to LSST in general, on challenging data analysis and handling questions
- Members are motivated by interesting, novel questions, the potential to develop new methodology, in addition to advancing LSST science goals

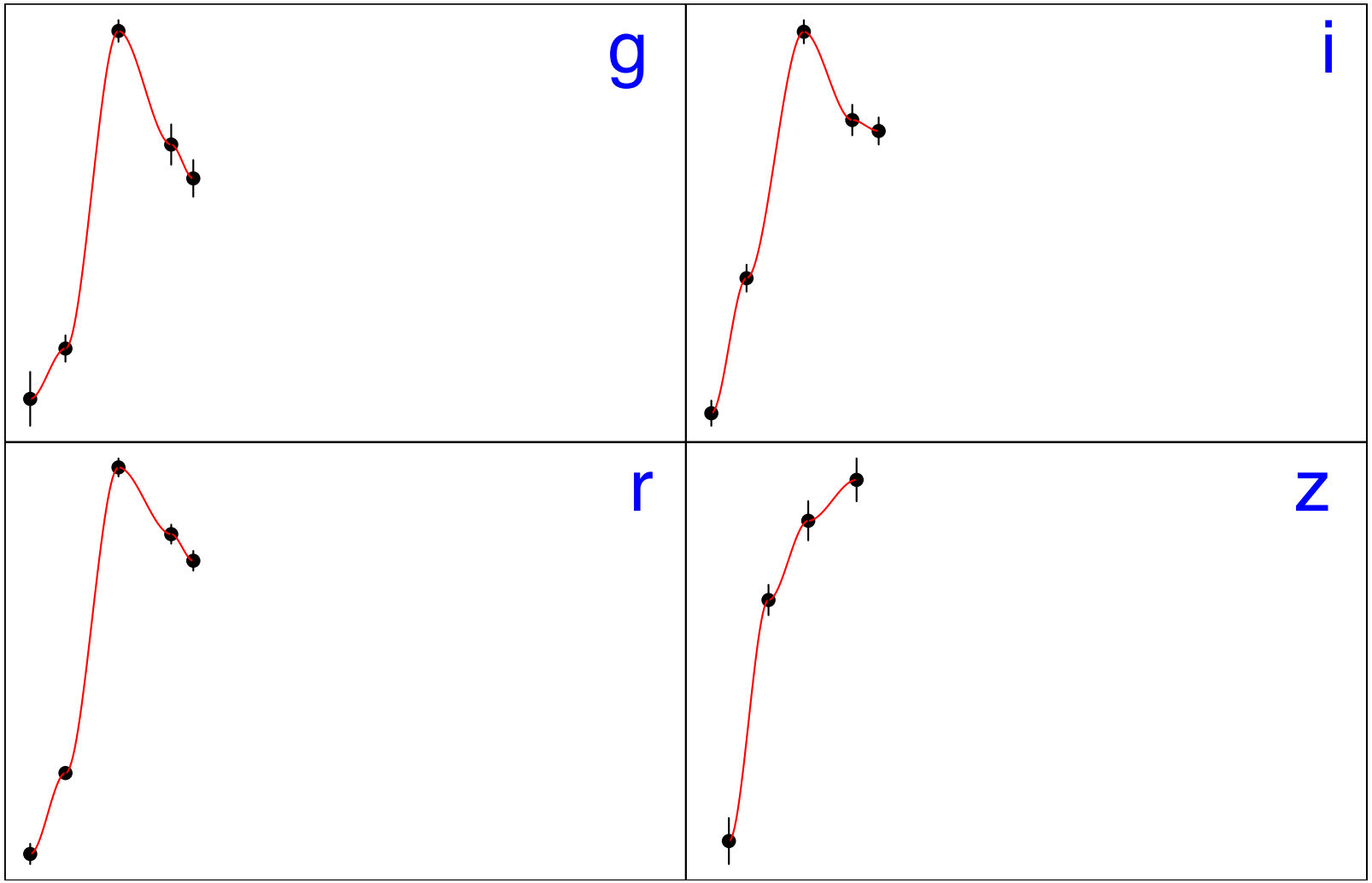
Our Membership

- Approximately 50 members – Consider Joining!
- Mix of astronomers and statisticians/machine learning experts
- Areas of Expertise
  - Dimension Reduction
  - Classification
  - Bayesian Methods
  - "Classical" Statistical Inference
  - Data Representation, Storage
  - Spatial Statistics
  - Image Analysis
  - …

How To Engage

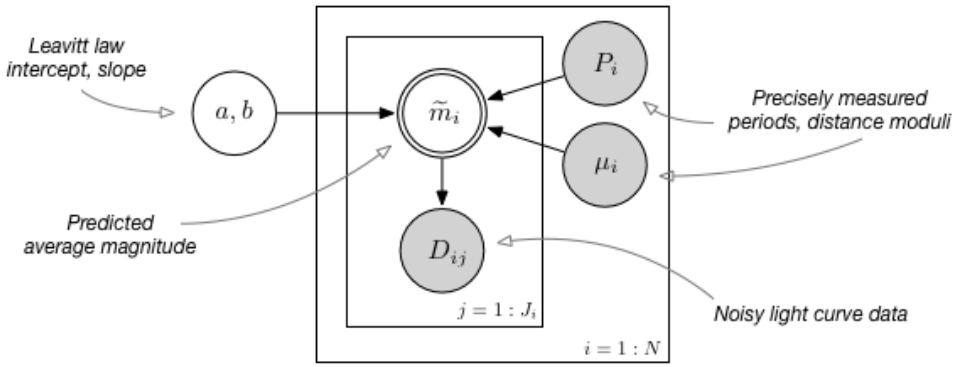
- Focused problems – isolate the key data challenge
- Sample Data Sets – “Challenges” – of varying complexity
- Contact me!

Example: SNe Classification Challenge

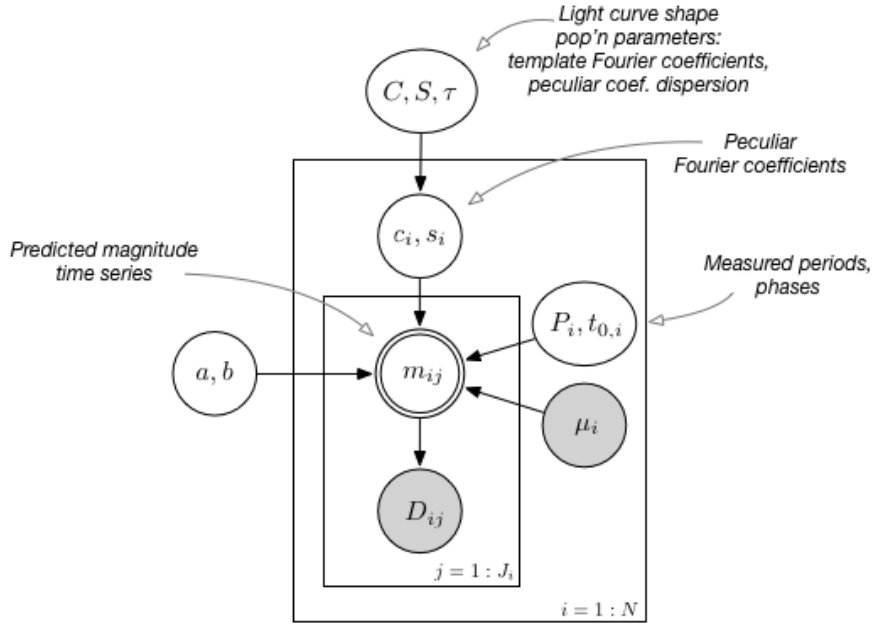


Example: FDA for Light Curve Analysis

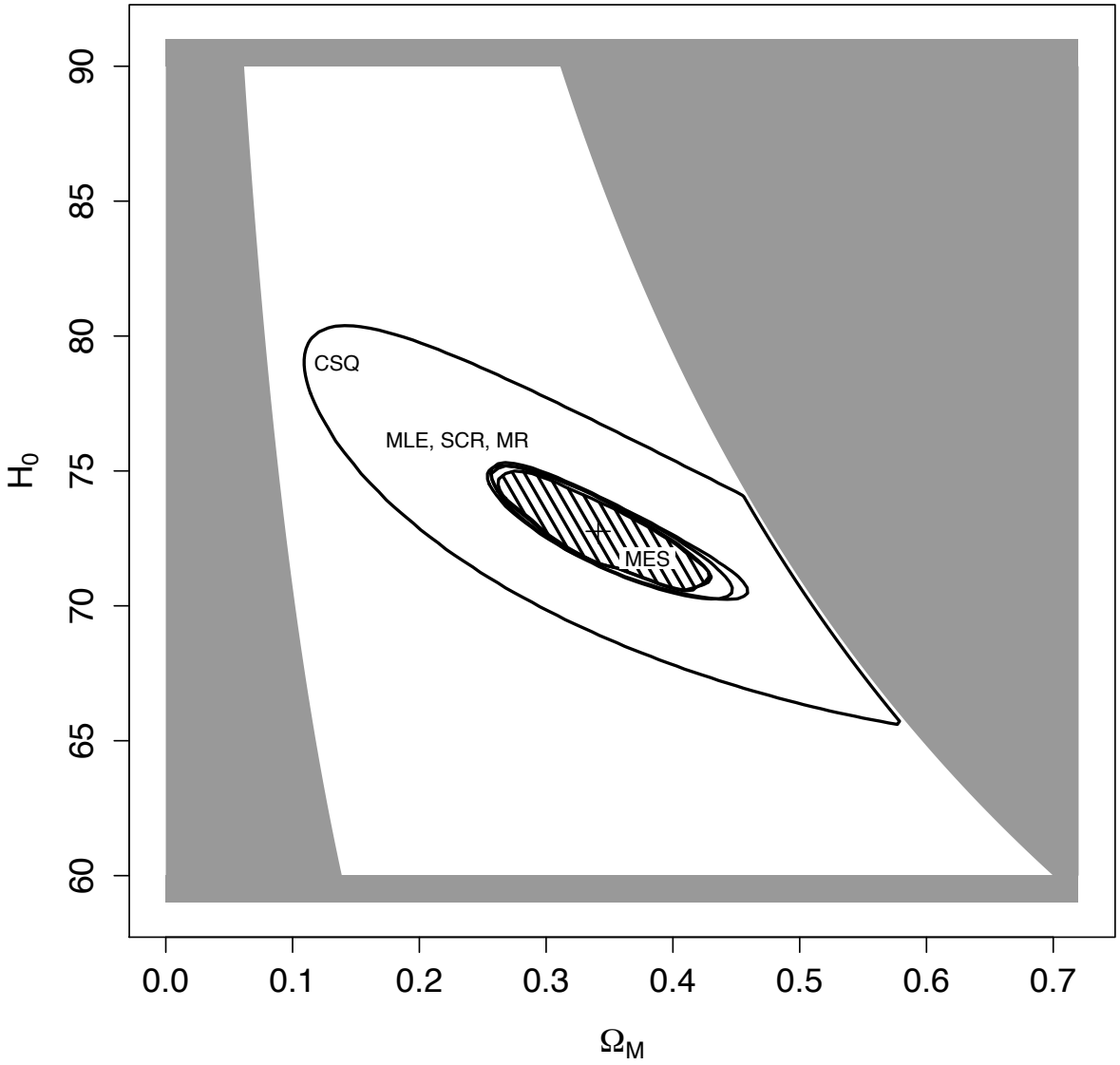
PLR calibration via regression



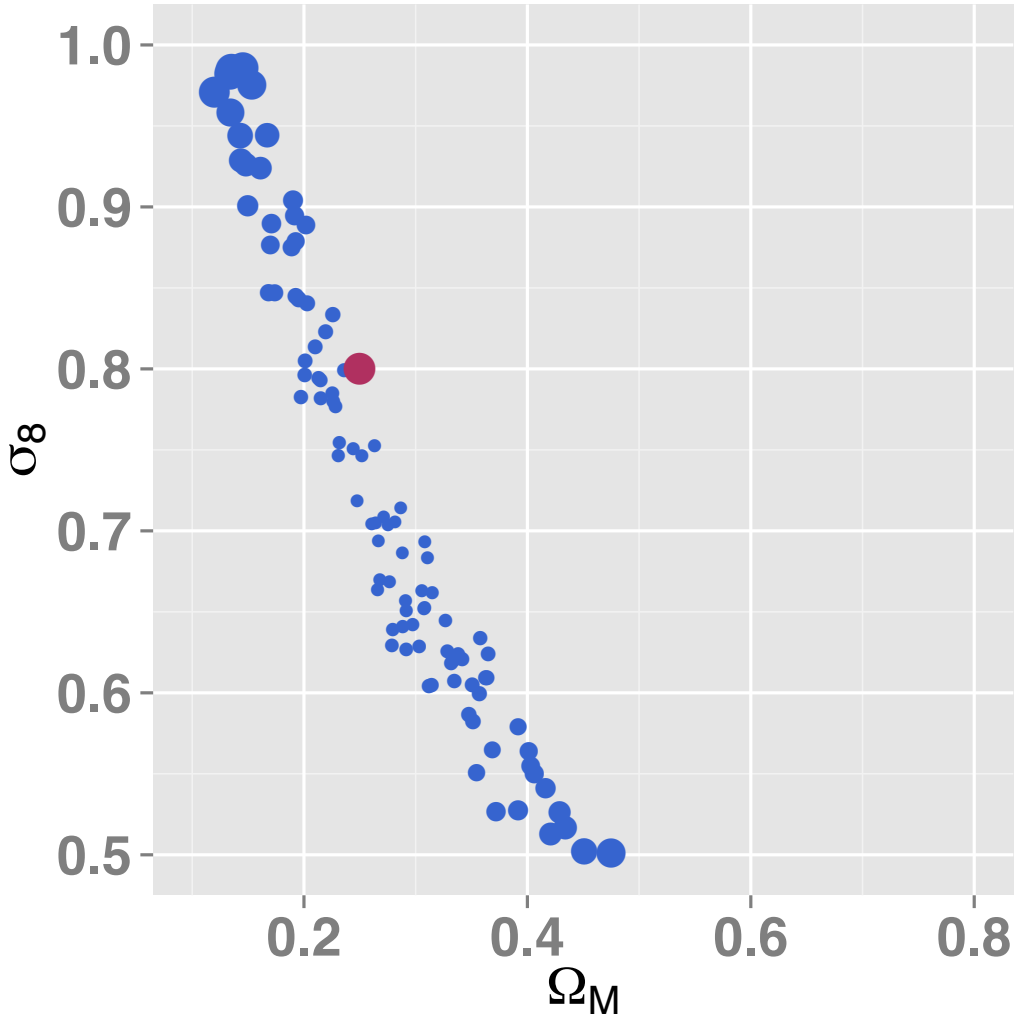
PLR calibration via FDA



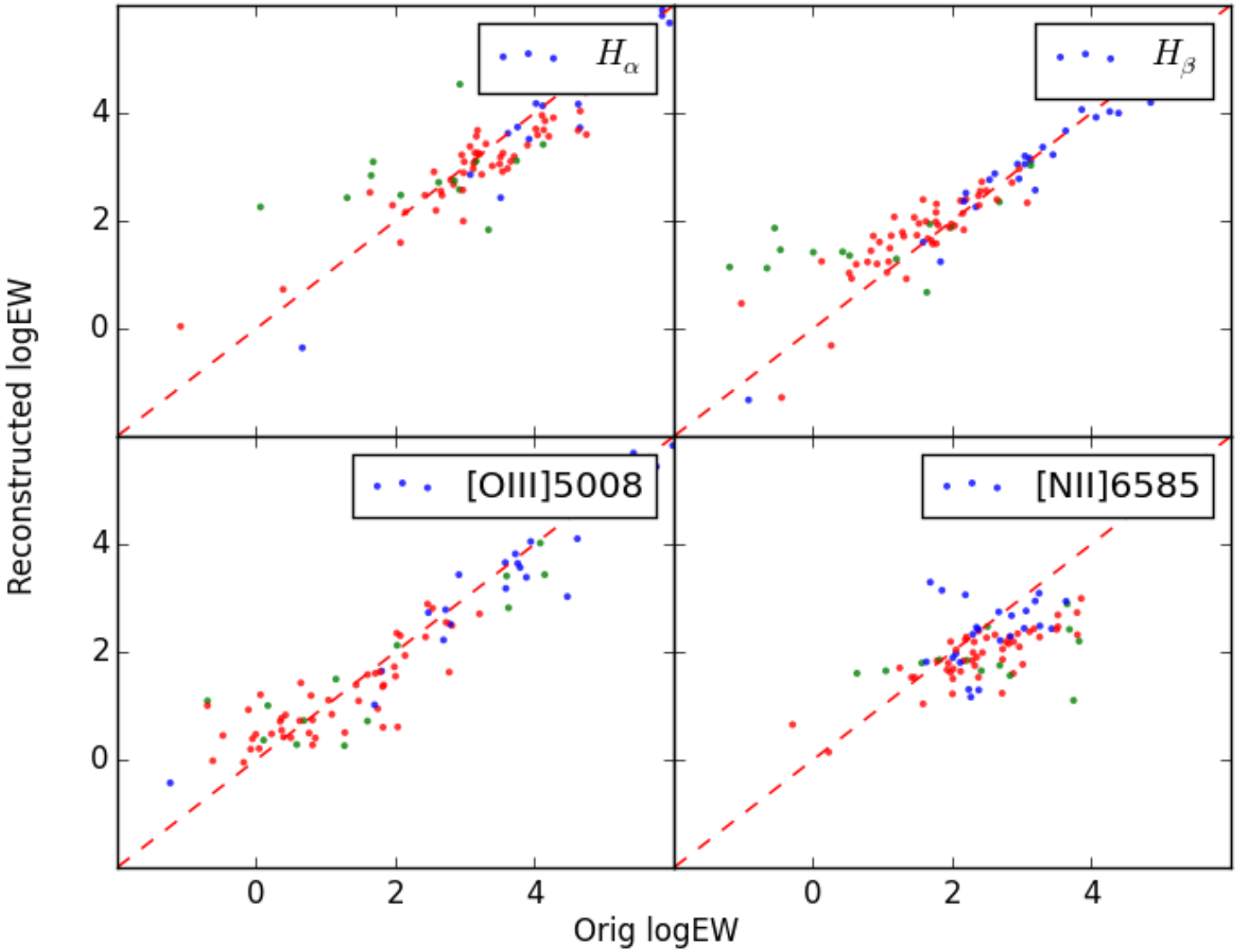
Example: Classical Parameter Inference



Example: Bayesian Parameter Inference



Example: Building Realistic Simulations



Statistical Challenges in Modern Astronomy VI

Image Credit: NASA, ESA

June 6 to 10, 2016 ◆ Carnegie Mellon University

scma6.org for more information

samsi
NSF Duke NCSU UNC NISS

2016-17 Program on Statistical, Mathematical and Computational Methods for Astronomy

The Statistical, Mathematical and Computational Methods for Astronomy Program focuses on the vast range of statistical and mathematical problems arising in modern astronomical and space sciences research, particularly due to the flood of data produced by both ground-based and space-based astronomical surveys at many wave-bands. To cope with the current and future needs of astronomy missions requires concerted efforts by cross-disciplinary collaborations involving astronomers, computer scientists, mathematicians and statisticians.

The research areas that form the main ingredients of the program include:

- Astronomical Simulations and Big data issues
- Exoplanets
- Functional Data Analysis
- Gravitational Wave Astrophysics
- High-performance computing for Bayesian inference and machine learning
- Lightcurve analysis/Time Domain Astronomy

For more details, visit www.samsi.info/ASTRO

Program Chair:
G. Jogesh Babu
(Penn State U.)

SAMSI Directorate Liaison:
Sujit Ghosh,ghosh@samsi.info
(SAMSI/NCSU)

National Advisory Committee Liaison:
Michael Stein
(U. Chicago)

www.samsi.info/astro

NSF-funded Mathematical Sciences Institute in Research Triangle, NC

Working Group on LSST-related Statistical Challenges

Opening Workshop: August 22-26, 2016

Possible extended, funded visits

Postdocs, graduate students, faculty

www.samsi.info/astro

Astro Working Group II: Synoptic Time Domain Surveys

Group Leaders: Ashish Mahabal (Astro, Caltech), Matthew Graham (Astro, Caltech), Chad Schafer (Stat, CMU), Soumen Lahiri (Stat, NCSU; co-chair), G. Jogesh Babu (Stat, PSU, co-chair)

Description: Time Domain Astronomy (TDA) has been getting richer in terms of datasets that span several years, many bands, and include dense and sparse light-curves for hundreds of millions of sources. The variety, volume etc. squarely fall in the Big Data regime, but the science questions that can be posed imply that standard, canned routines cannot be used except in trivial cases. The light-curves often have large gaps, are heteroskedastic, and the intrinsic variability - often poorly understood - adds an element of uncertainty when multi-band data are not obtained simultaneously. TDA can thus be viewed as the umbrella within which several large problems can be tackled. These span from Kepler-type planet search/characterization (also covered in other groups) to characterization of specific classes e.g. binary black-hole searches from Catalina Real-Time Transient Survey (CRTS) to searching flaring stars away from the plane of the Galaxy using light-curves as well as ancillary data from other sources like SDSS and WISE to name just a few. Combining different datasets is a fertile field in itself with the sum of the parts potentially being more than the whole, but not fully realized yet owing to lack of good methodology. Besides obtaining more sources of well understood types, the clustering in search for characterization naturally leads to outliers - not just individual interesting sources, but also entire subclasses.

Big Questions:

1. What mathematical and statistical approaches can be used to best characterize and quantify salient features of irregular, heteroscedastic, gappy time series, and to identify specific feature sets, templates and models? How can we identify many weak features or a few strong ones in such high dimensional time series Big Data?
2. What are the best methods for classifying time series incorporating auxiliary/covariate information? Are there specific domain-knowledge based features that can be identified to improve class discrimination?
3. How can significant outliers/anomalies and subclasses thereof be detected? Investigate correlated Functional Data techniques to detect outliers, anomalies and subclasses.
4. How can data sets from multiple surveys (with or without overlap in certain key parameters) be combined? Can we take a predictive model obtained in one survey and transform it into an accurate model for another survey?

What techniques apply to specific categories of time series: non-stationary, stochastic/deterministic, etc., e.g., can we assume ergodicity? Develop formal statistical tools/tests for assessing viability of simpler data structures assumptions.

Specific Objectives within TVS

- Create challenge classification sets
- Explore the “real time” classification problem
- …

For more information:

Chad Schafer cschafer@cmu.edu