Astrostatistics: Opening the Black Box

Jake VanderPlas

11-10-2015

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Big Data in Astronomy:

Annie Jump Cannon processed 300,000 stellar spectra in her lifetime… *by hand!*

Big Data in Astronomy:

Annie Jump Cannon processed 300,000 stellar spectra in her lifetime… *by hand!*

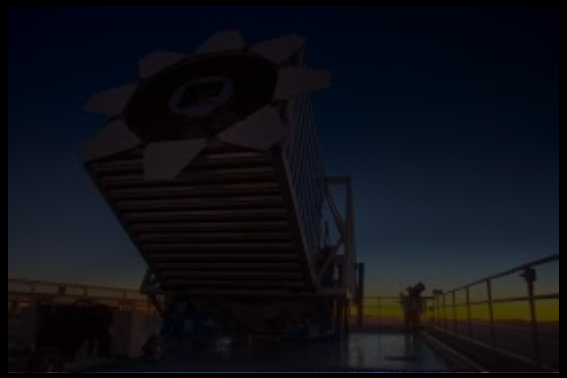
SDSS gathered ~3 million spectra in 10 years

~30,000 GB catalog over a decade

Big Data in Astronomy:

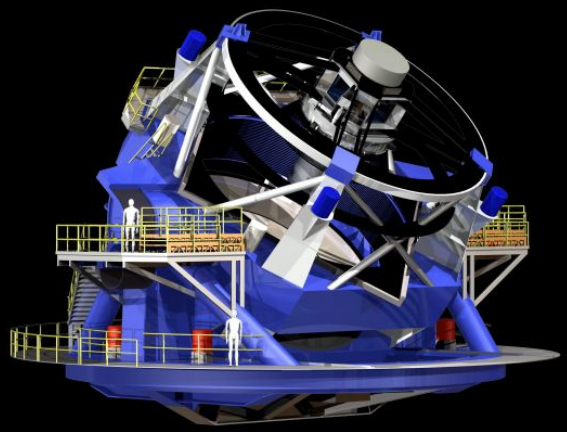
Annie Jump Cannon processed 300,000 stellar spectra in her lifetime… *by hand!*

SDSS gathered ~3 million spectra in 10 years



~30,000 GB catalog over a decade

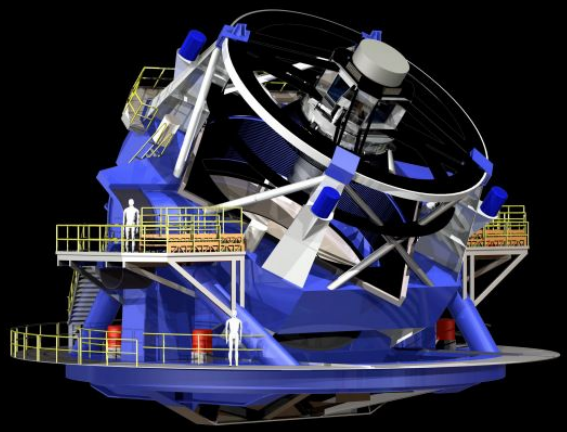
LSST will do an SDSS-scale photometric survey *every night* for 10 years!



Astronomy's Data Revolution:

Orders-of-magnitude growth in data requires many new statistical and algorithmic approaches.

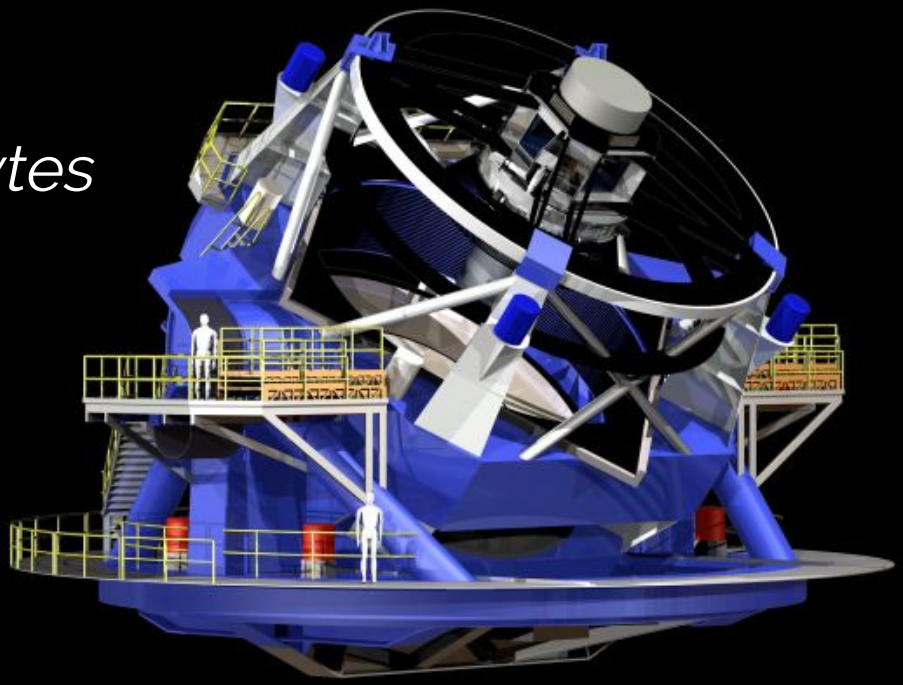
We should expect the jump from current data to LSST to be no different.



Large Synoptic Survey Telescope (LSST)

Exemplar of the new data-intensive astronomy

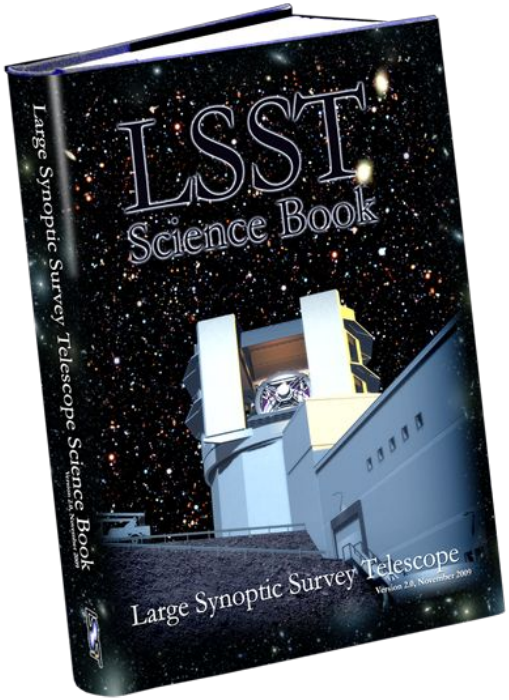
- photometry of the full southern sky every 3-4 nights *for 10 years*
- ugrizy multiband data
- 30,000GB *per night*
- Final catalog: 100s of *Petabytes*
- ~1000 observations per field



LSST Science Book

~600 Pages, 245 authors, nearly every astronomy sub-domain represented.

Scope of the dataset will be *transformative*.



But challenges abound: survey data designed to be **generally useful** is rarely **optimal for *your* science.**

Your favorite methods may not work anymore. . .

. . . enter **AstroStatistics**

http://www.lsst.org/scientists/scibook

Astrostatistics *(n.)*

> The application of Statistics to the study and analysis of Astronomical Data
>
> — Wiktionary

Astrostatistics *(n.)*

The adaptation of standard methods — and development of new ones — for use with modern **large, noisy,** and/or **heterogeneous** datasets.

— JTV

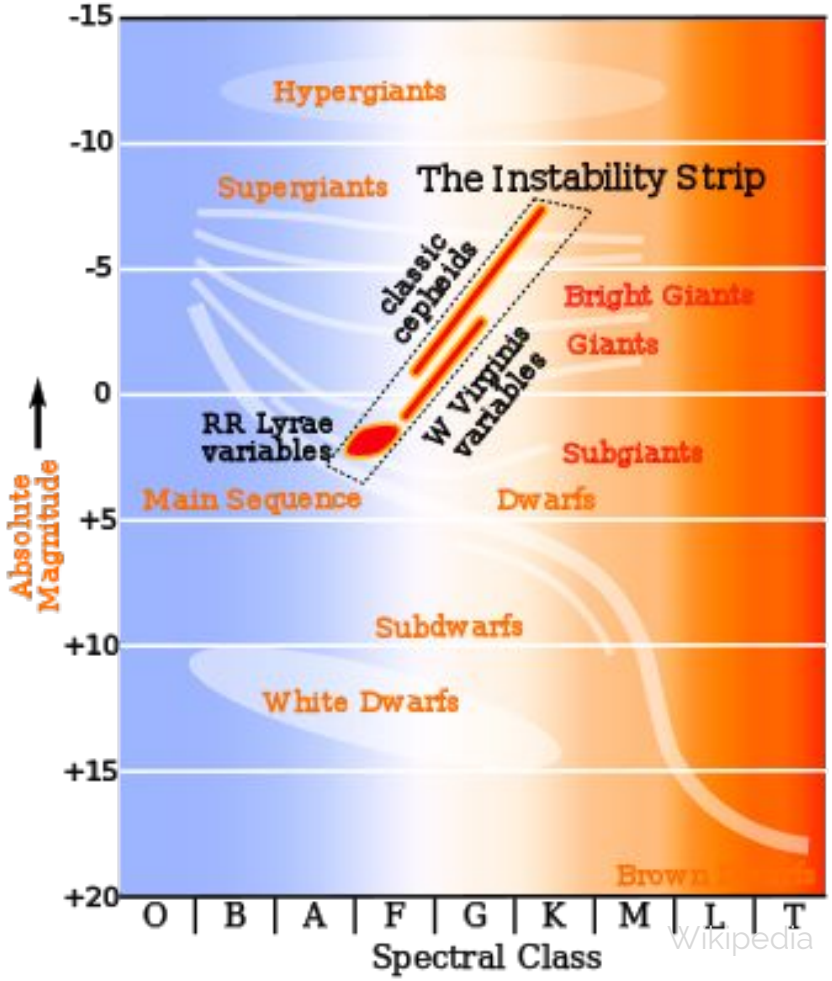
Astrostatistics Case Study: Mapping the Milky Way with RR Lyrae

Jake VanderPlas

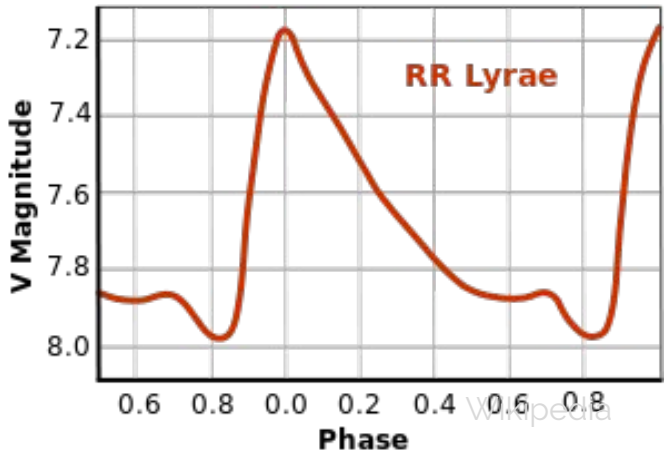
UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Background: RR Lyrae-type Stars

A particular class of variable star:



Easily detectable via distinct lightcurve shape:



Standard Candles:
Direct tracer of distance!

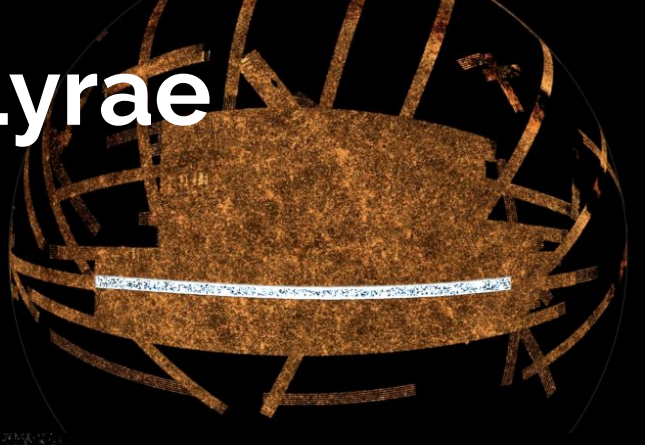
$$M_V = (0.23 \pm 0.04)\,[Fe/H] + (0.93 \pm 0.12)$$

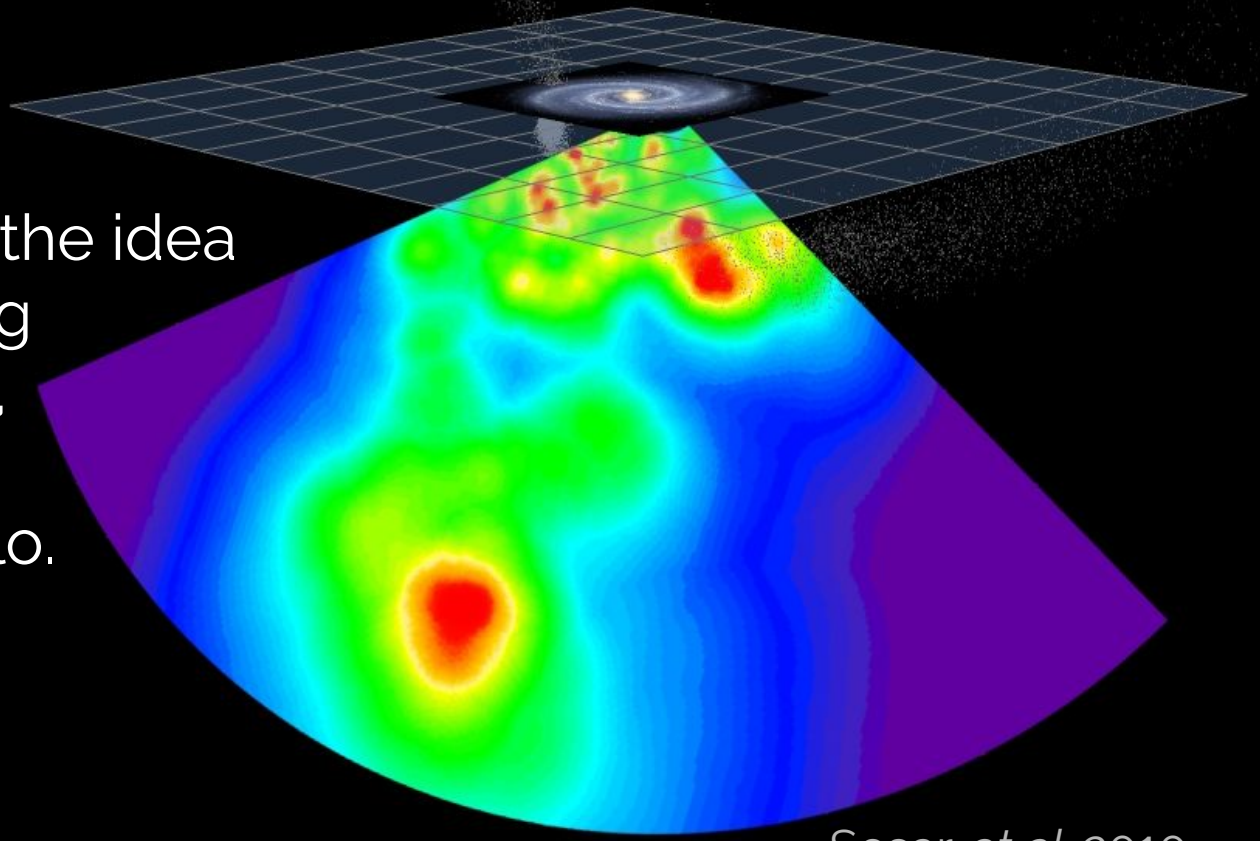
(Chaboyer *et al.* 1999)

Mapping the MW with RR Lyrae

SDSS II Stripe 82:

- 483 RR Lyrae to r~22
- 300 deg^2
- d ~ 100 kpc





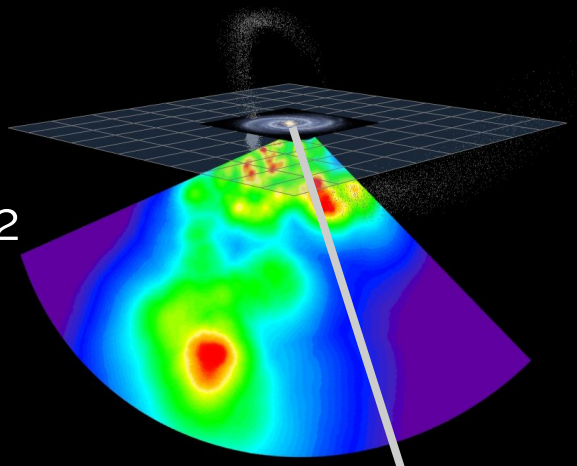
Analysis supports the idea of an early-forming smooth inner halo, and late-forming accreted outer halo.

Sesar *et al.* 2010

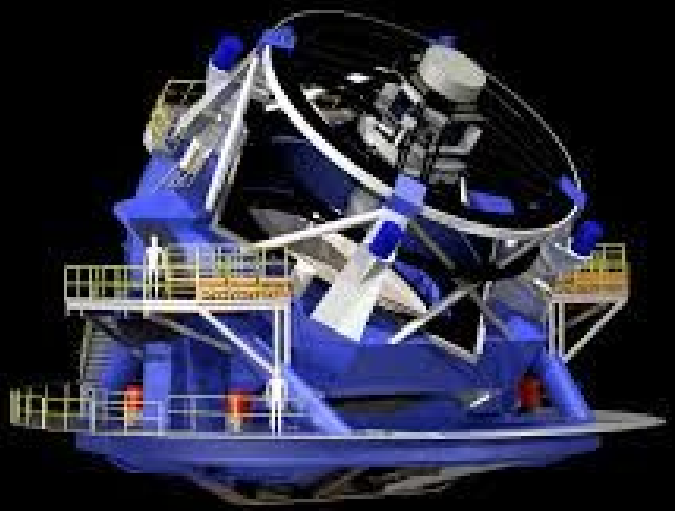
RR Lyrae in LSST

SDSS II
300 deg²
r ~ 22 mags
d ~ 100 kpc
483 RR Lyrae

LSST
~20,000 deg²
r ~ 24 mags
d ~ 300 kpc
> 30,000 RR Lyrae??
(nobody knows!)



? ? ?



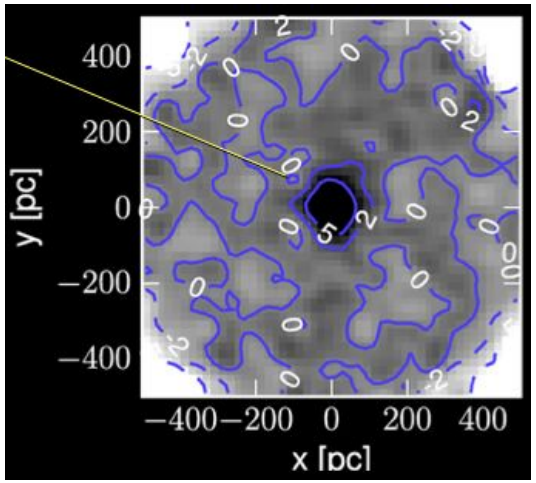
Science with RR-Lyrae

Every MW Satellite with time-series available has ≥ 1 observed RR Lyr

Object	N_{RRab}[a]	⟨[Fe/H]⟩
CVn I	18	−1.962
Herc	6	−2.518
For	396	−1.025
Dra	123	−1.946
Leo IV	3	−2.363
Sex	26	−1.966
Leo I	47	−1.450
Leo II	103	−1.670
UMi	47	−2.112
Scl	129	−1.726
Boo I	7	−2.531
ComBer	1	−2.640
CVn II	1	−2.444
UMa I	5	−2.334
UMa II	1	−2.357
Seg2	1	−2.257

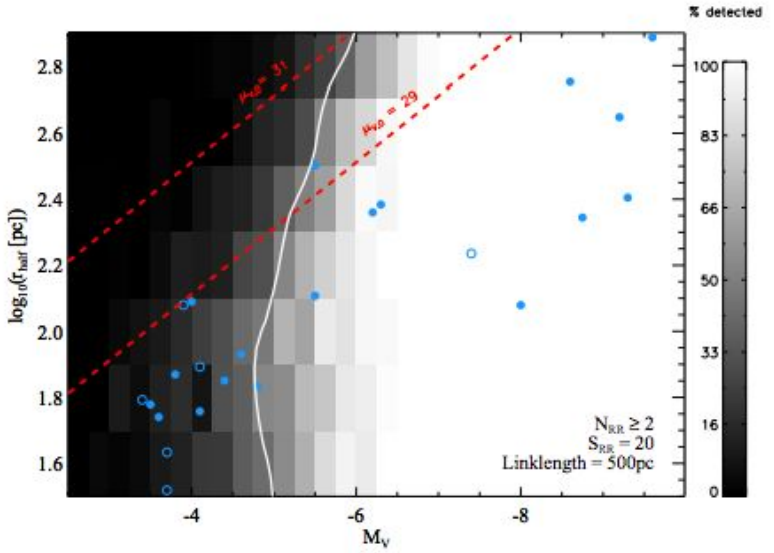
(Boettcher *et al.* 2013, Table 4; See also Baker & Willman 2015)

A single halo RR-Lyr can indicate structure:



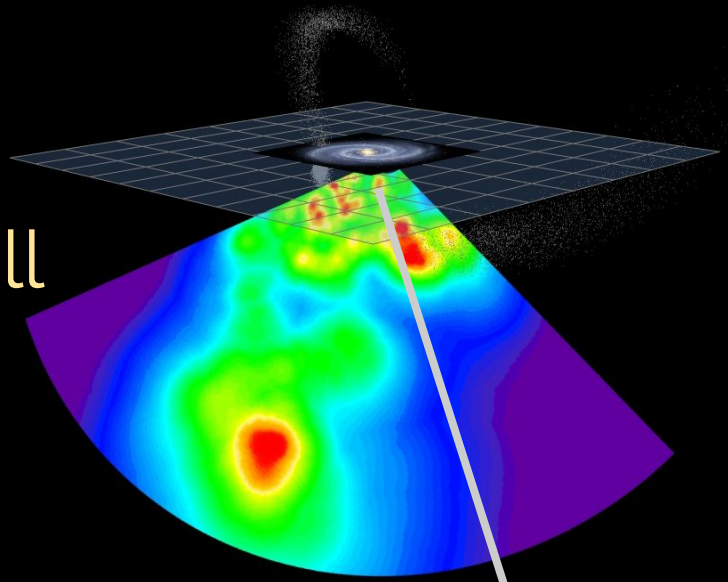
Sesar *et al.* 2013

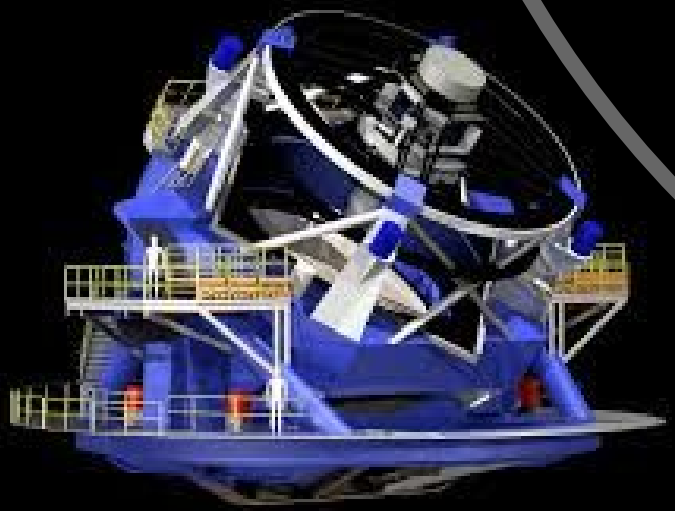
Two RR-Lyr past ~100kpc almost *certainly* indicate a MW Satellite!



Baker & Willman 2015

In other words: any single distant RR Lyrae detected will *almost certainly* yield new constraints on MW potential, formation history, etc.





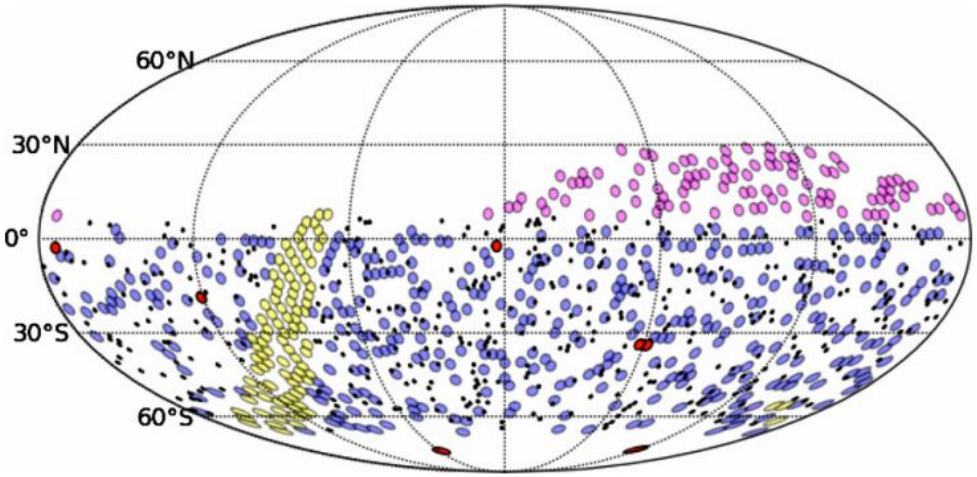
How to find RR Lyrae

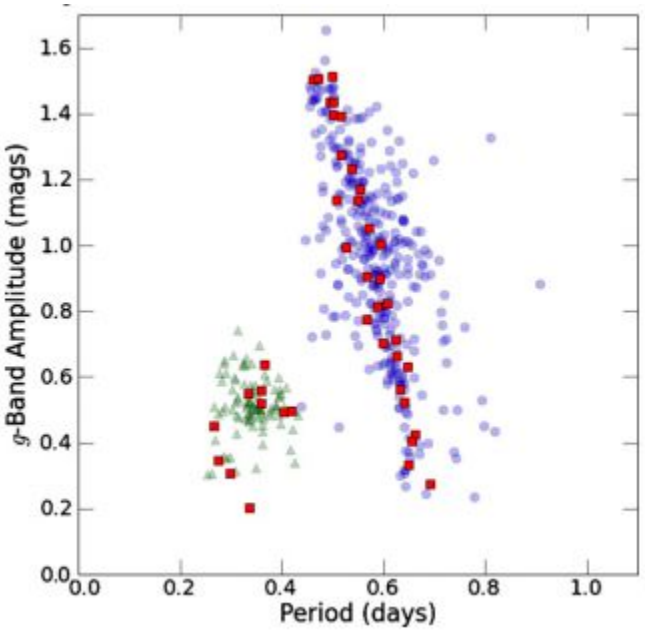
1. Gather time-series observations
2. Detect periodic objects
   - Lomb-Scargle Periodogram
   - Supersmoother
   - AoV Periodogram
   - CARMA models
   - etc.
3. Fit Templates at matching periods
4. Do Science!!!

If only it were that straightforward...

Oluseyi 2012 Simulations:

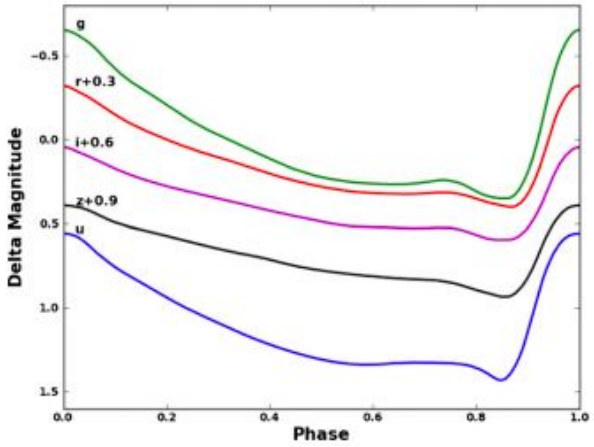
Detailed simulation of RR Lyrae observations in 10 years of LSST

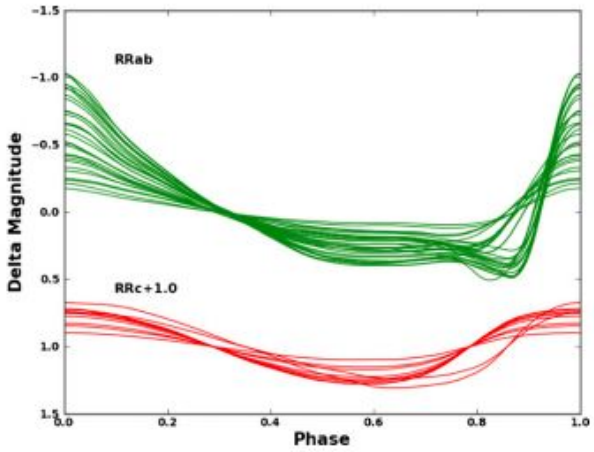




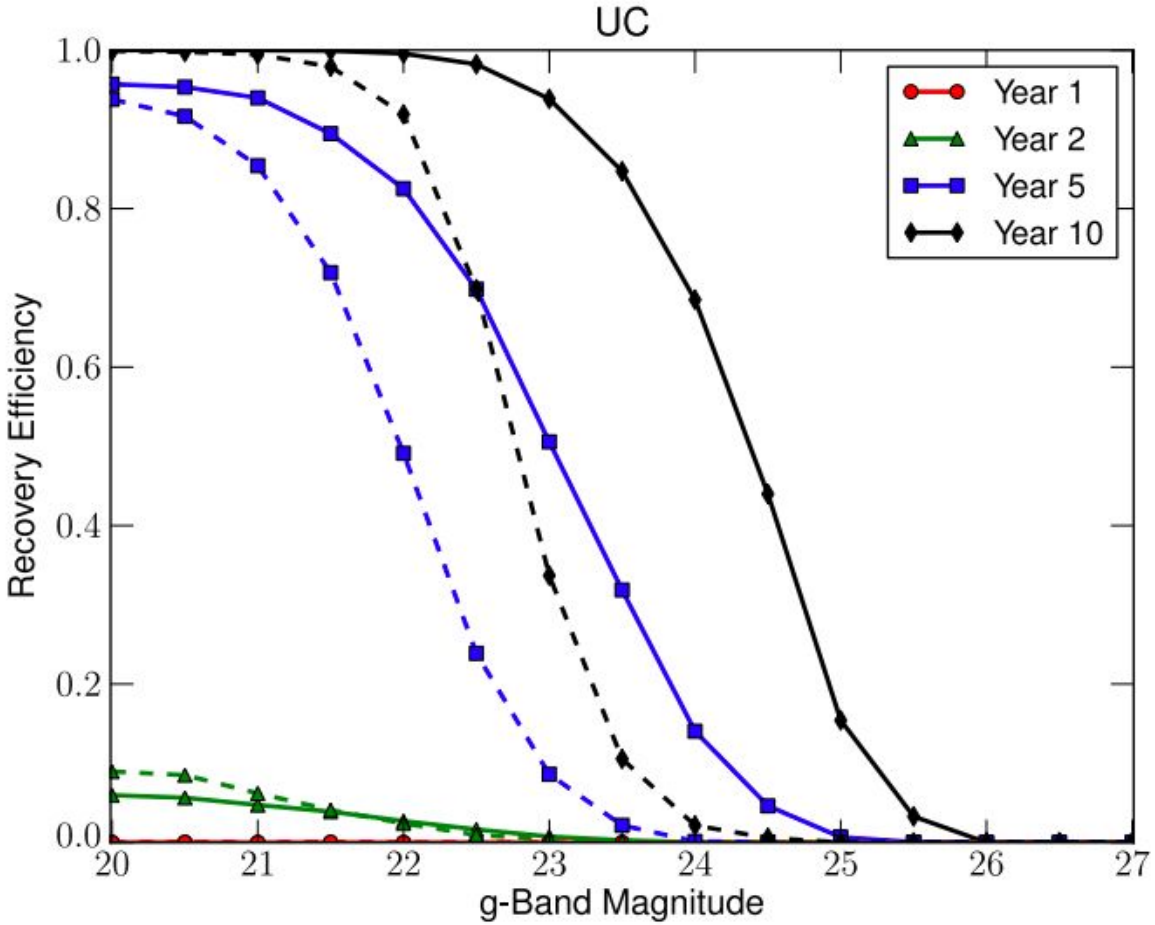
(Primarily based on Stripe82 sample, Sesar *et al.* 2010)

Best available data on RR-Lyr templates, populations, LSST cadence, etc.





Oluseyi 2012 Simulations:

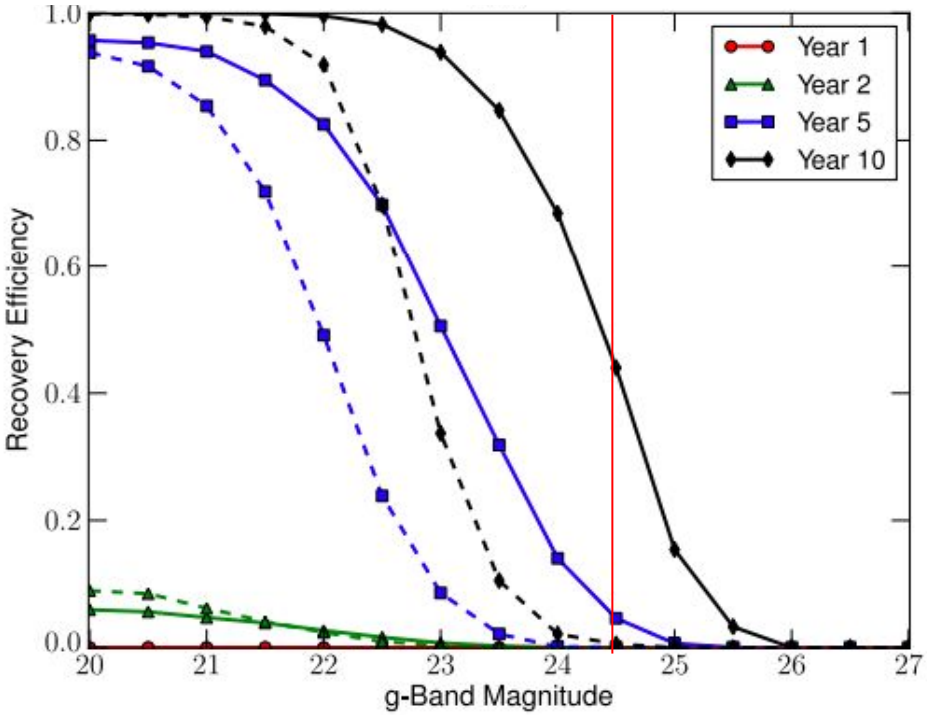


Faintest RR-Lyrae: Pessimistic period recovery even with 5-10 years of LSST data!

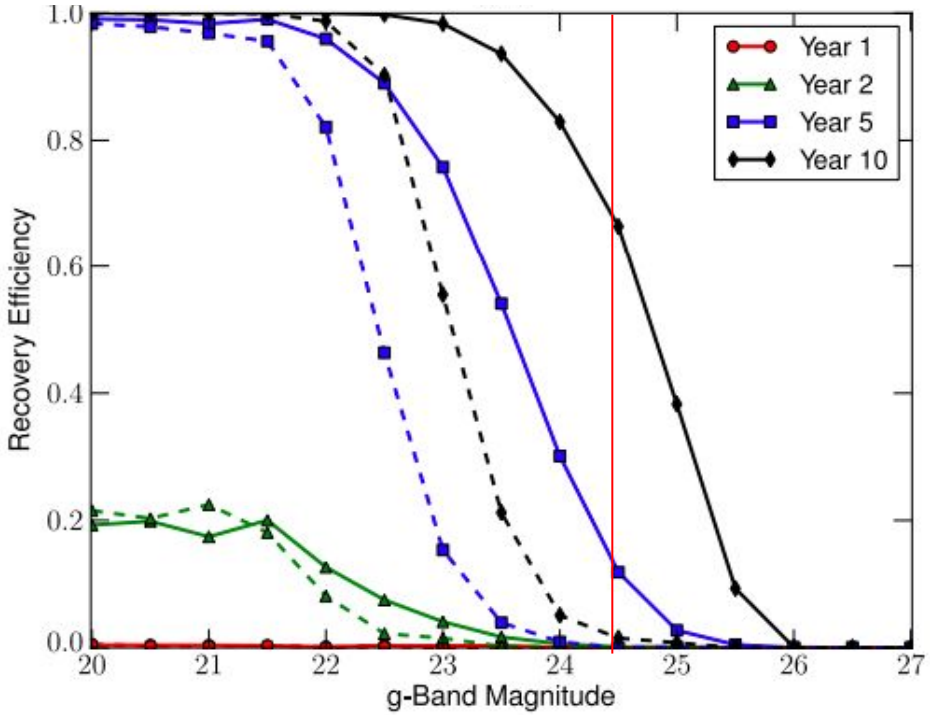
solid lines = RRab; dashed lines = RRc
Results for Universal Cadence fields

Oluseyi 2012 Simulations:

Universal Cadence



Overlap Regions

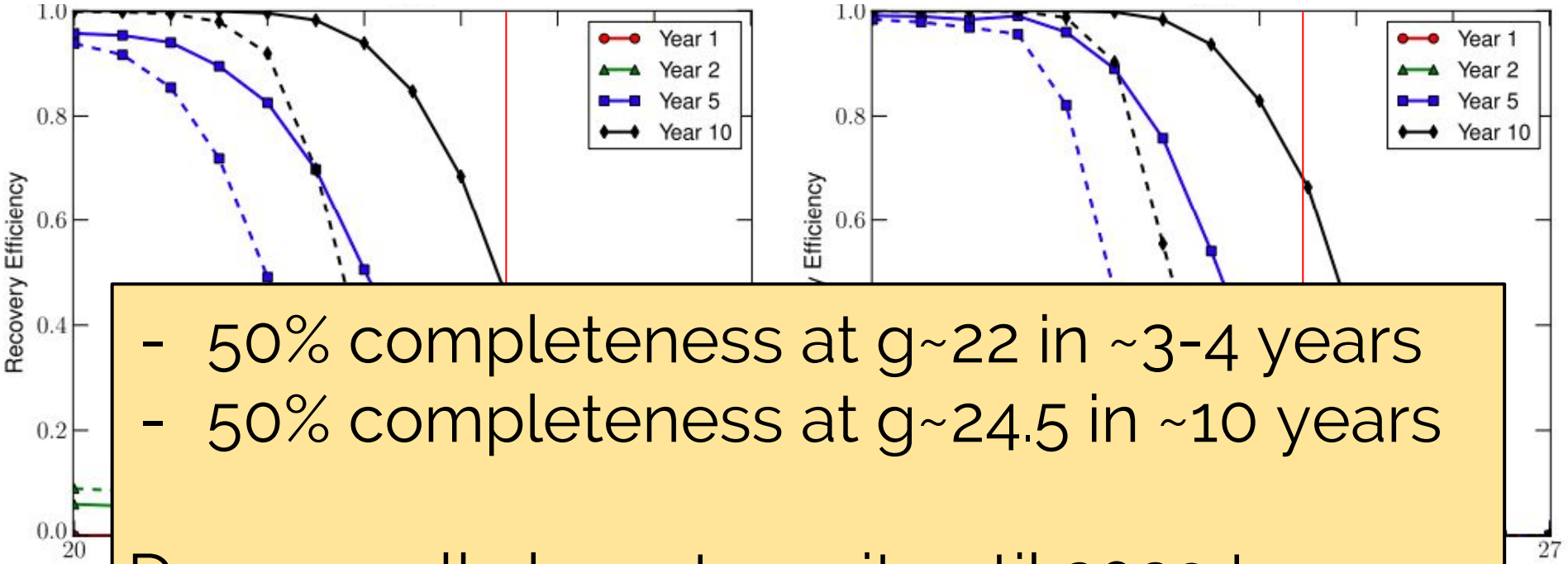


solid lines = RRab; dashed lines = RRc

Oluseyi 2012 Simulations:

Universal Cadence

Overlap Regions



- 50% completeness at g~22 in ~3-4 years
- 50% completeness at g~24.5 in ~10 years

Do we really have to wait until 2029 to detect faint MW dwarfs with LSST?

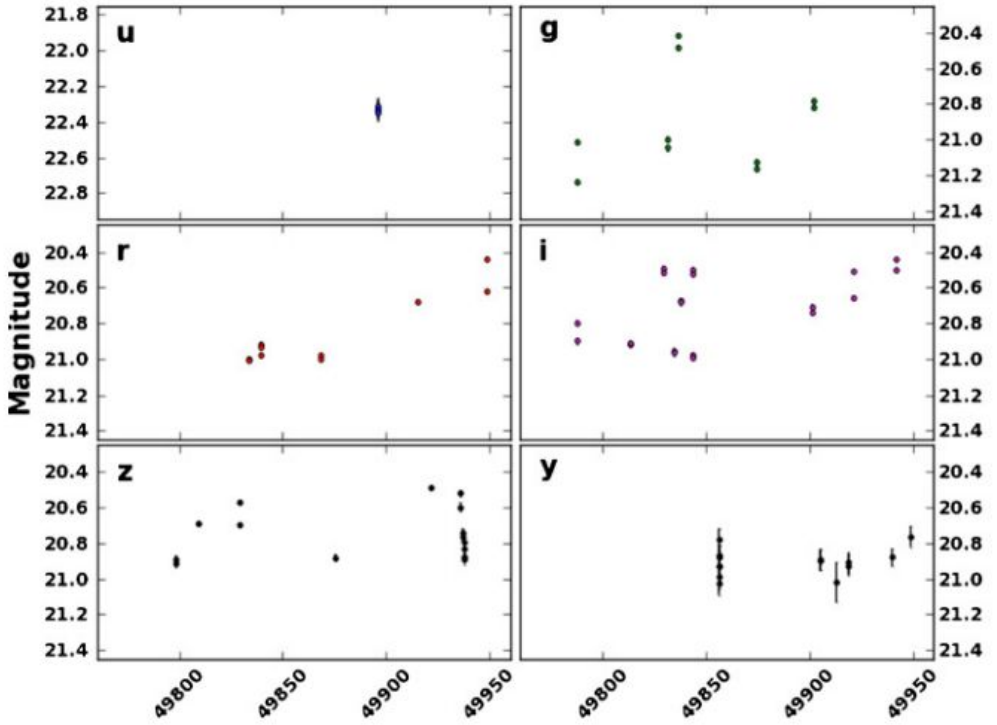
LSST is *not designed* for RR Lyr Detection!

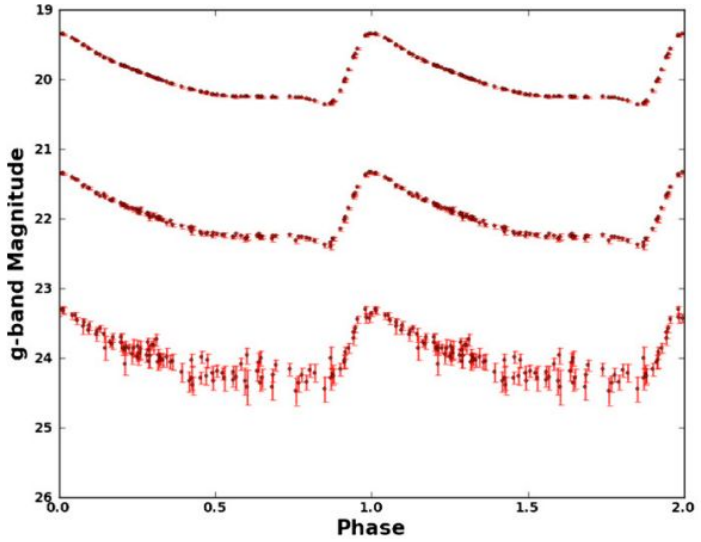
Sparsity: ~one visit every ~three nights (cf. 0.4-1.0 day period of RR Lyrae)

Heterogeneity: only *one band* (ugrizy) per visit

Noise: Interesting objects near the detection limit

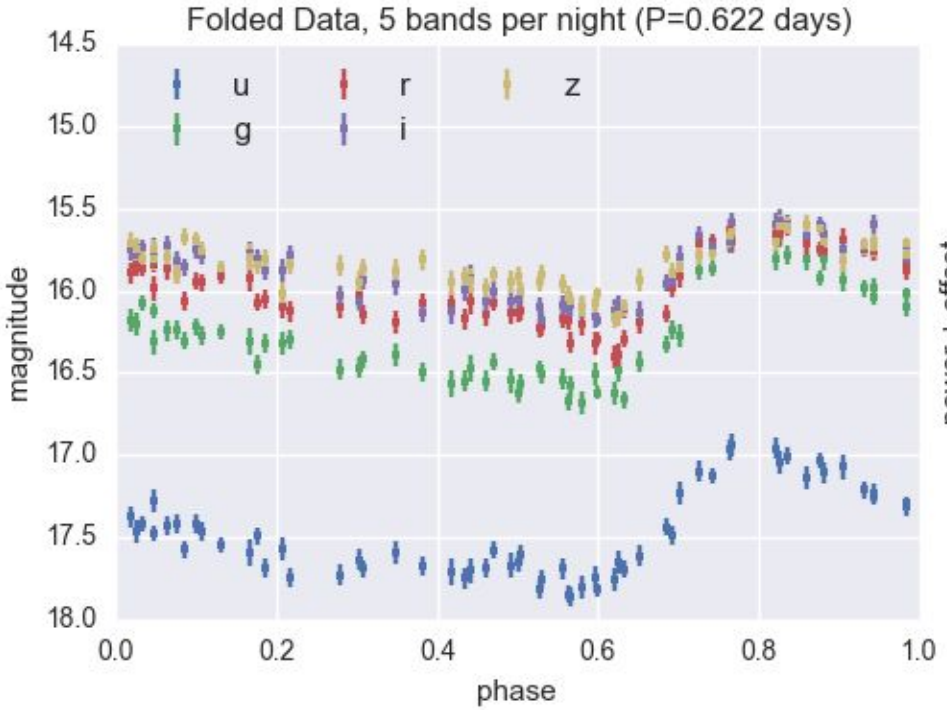
Data Size: Expensive period searches untenable (~30sec budget per object)

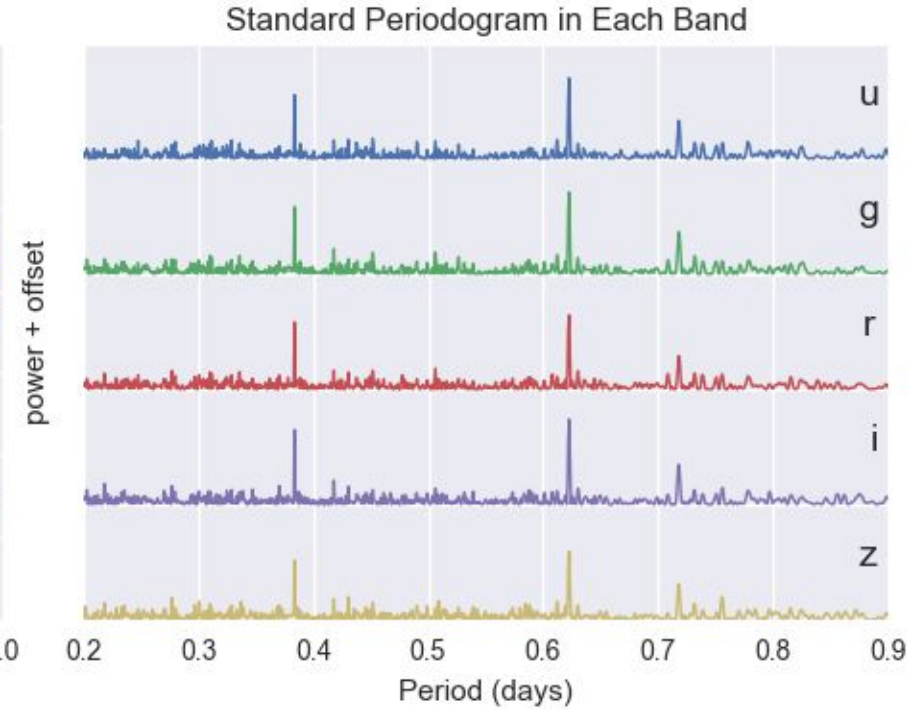




Quick Test: (SDSS-like data)

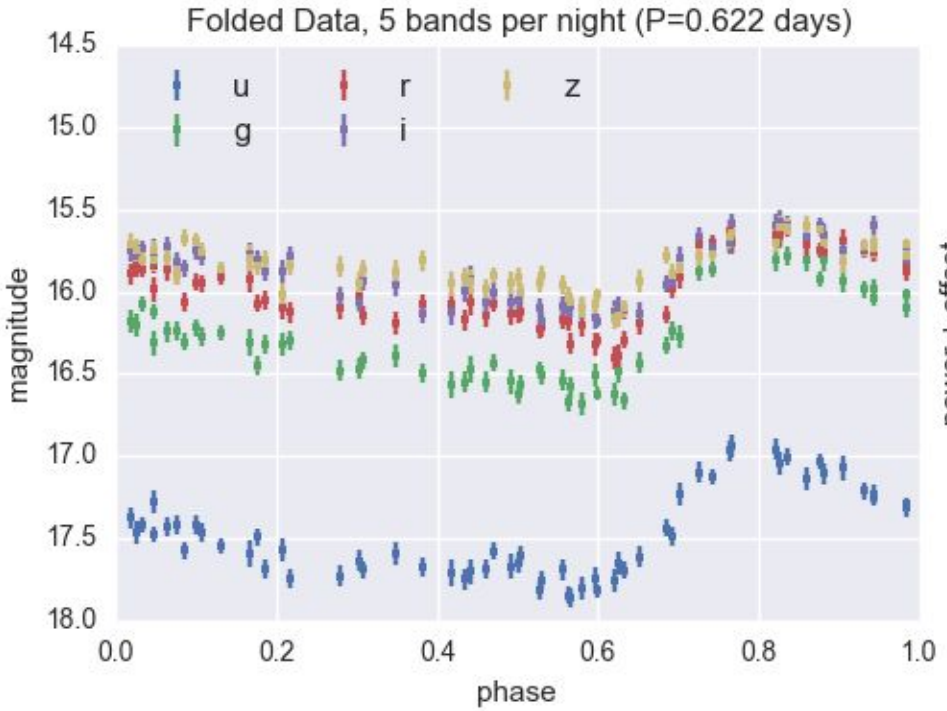
- Standard Lomb-Scargle Periodogram
- 60 visits over 6 months; 5 bands per visit

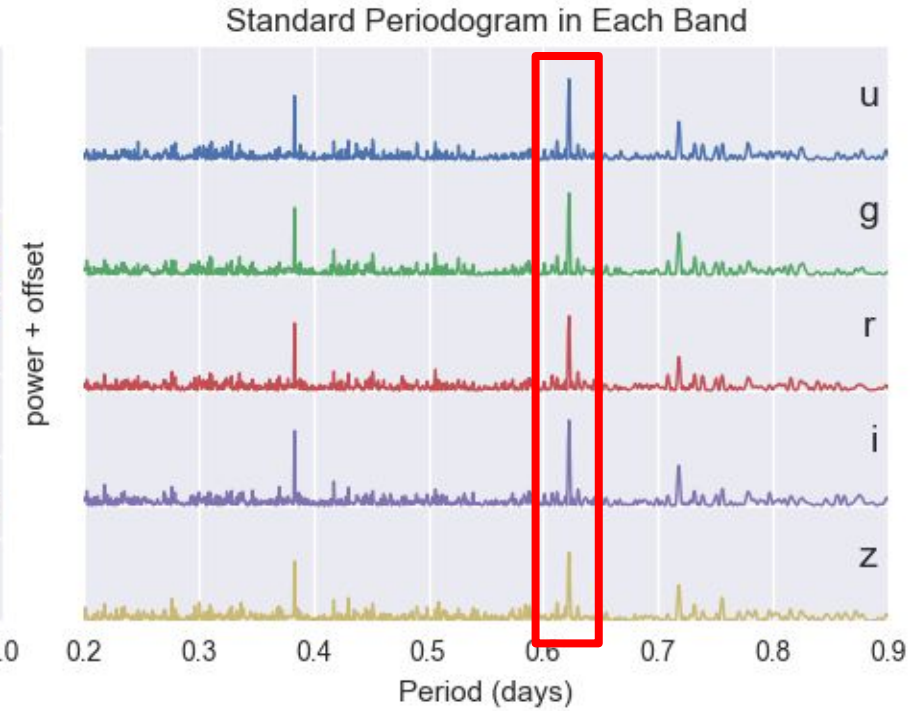




Quick Test: (SDSS-like data)

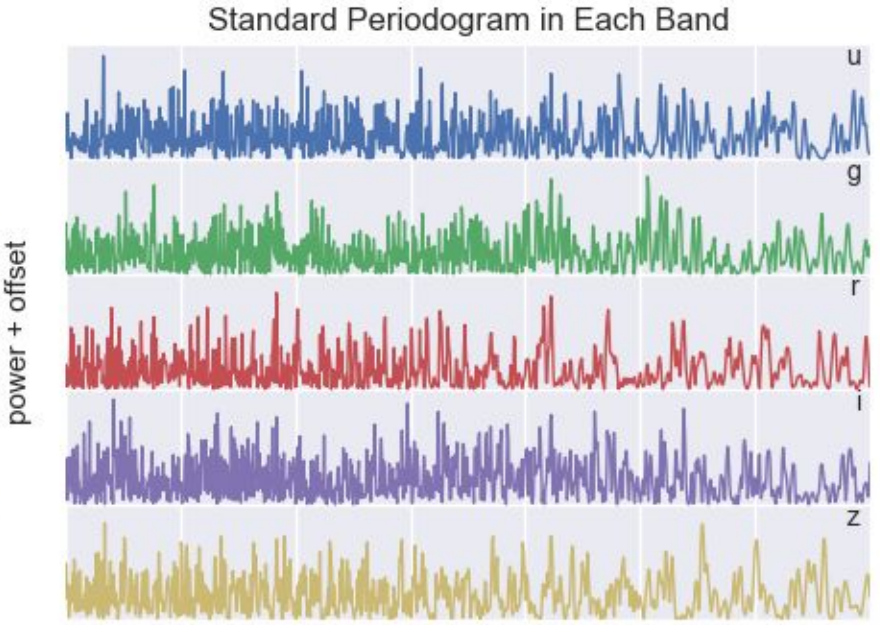
- Standard Lomb-Scargle Periodogram
- 60 visits over 6 months; 5 bands per visit





Quick Test: (LSST-like data)

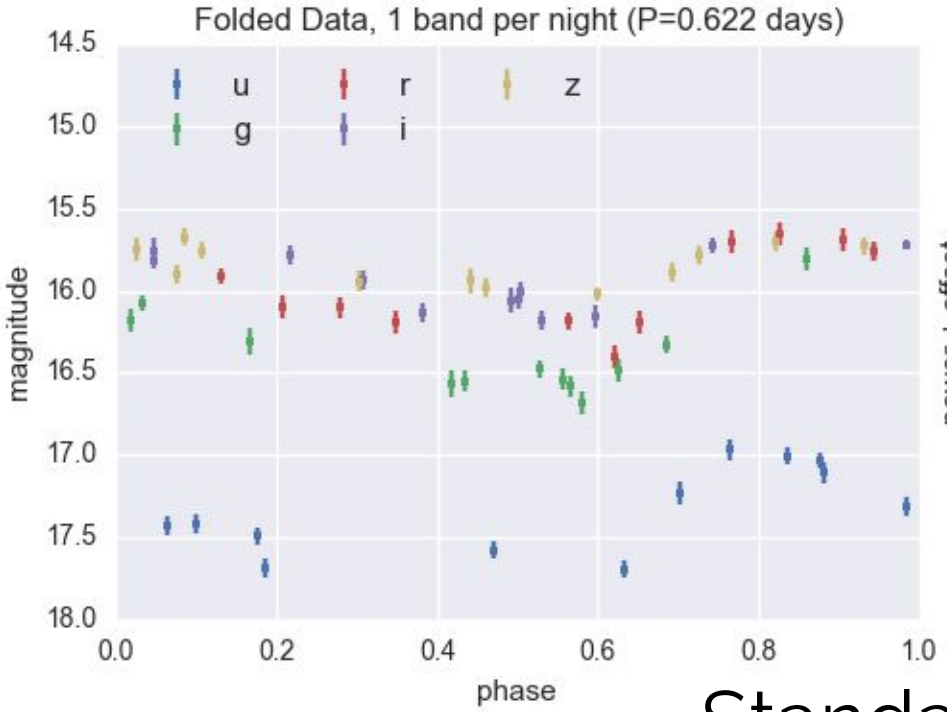
- Standard Lomb-Scargle Periodogram
- 60 visits over 6 months; single band each visit

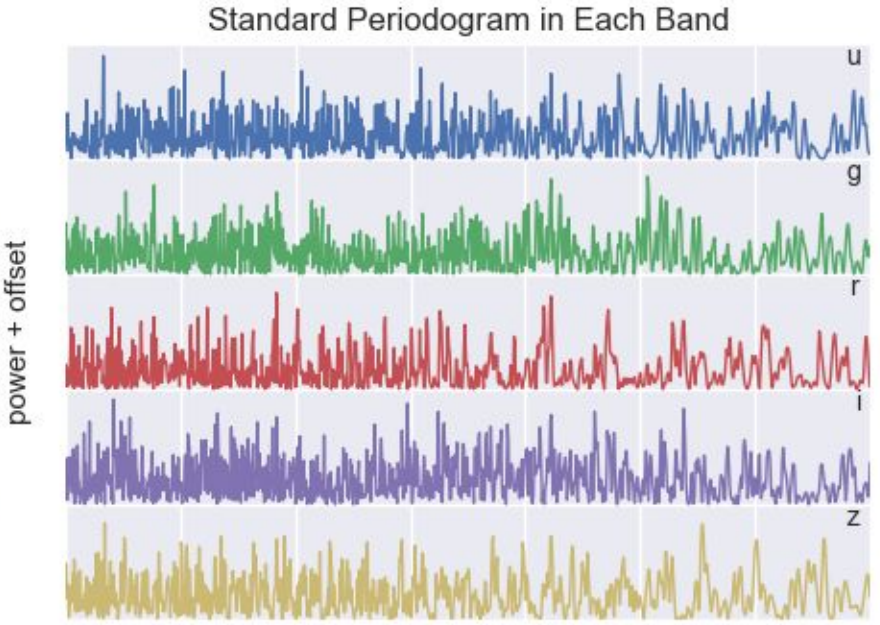


Quick Test:

(LSST-like data)

- Standard Lomb-Scargle Periodogram
- 60 visits over 6 months; single band each visit





Standard single-band methods fail for sparse LSST-type data!

Let's think about a periodogram which can utilize multiple bands simultaneously . . .

Period Detection in Multi-band Photometry . . .

- **Welch & Stetson 1993** – *variability index* for two simultaneous bands
- **Sesar 2010** – Supersmoother on g-band primarily; on r & i to evaluate; skip u & z
- **Suveges 2012** – use PCA to combine info from *simultaneous* measurements
- **Oluseyi 2012** – single-band SuperSmoother analysis in g-r-i, look for ⅔ agreement

The Lomb-Scargle Periodogram

If you've ever come across the Lomb-Scargle Periodogram, you've probably seen something like this...

$$P_N(\omega) = \frac{1}{2\,V_y}\left[\frac{\left[\sum_k (y_k - \bar{y})\cos\omega(t_k-\tau)\right]^2}{\sum_k \cos^2\omega(t_k-\tau)} + \frac{\left[\sum_k (y_k - \bar{y})\sin\omega(t_k-\tau)\right]^2}{\sum_k \sin^2\omega(t_k-\tau)}\right]$$

But this obfuscates the beauty of the algorithm: the classical periodogram is essentially the χ^2 of a single sinusoidal model-fit to the data:

$$y(t|\omega,\theta) = \theta_1\sin(\omega t) + \theta_2\cos(\omega t).$$

$$\chi^2_{min}(\omega) = \chi^2_0[1 - P_N(\omega)]$$

Standard Lomb-Scargle

Periodogram peaks are frequencies where a sinusoid fits the data well:

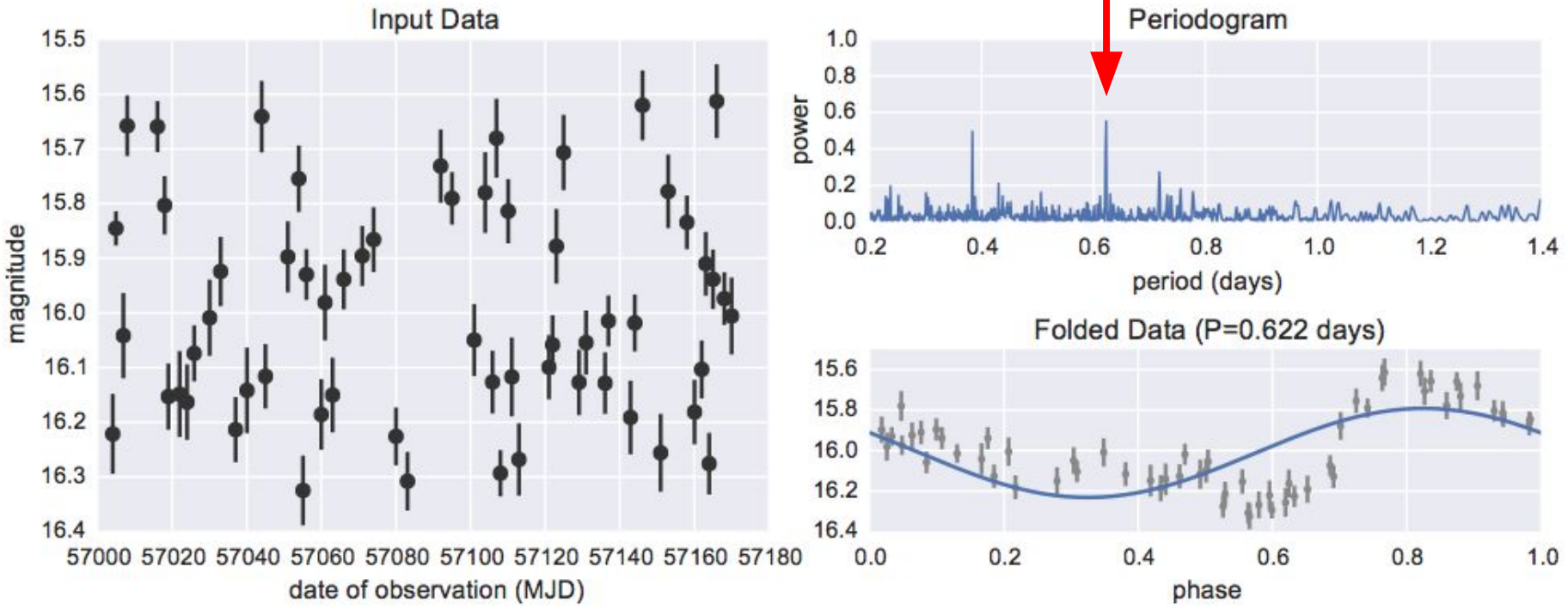
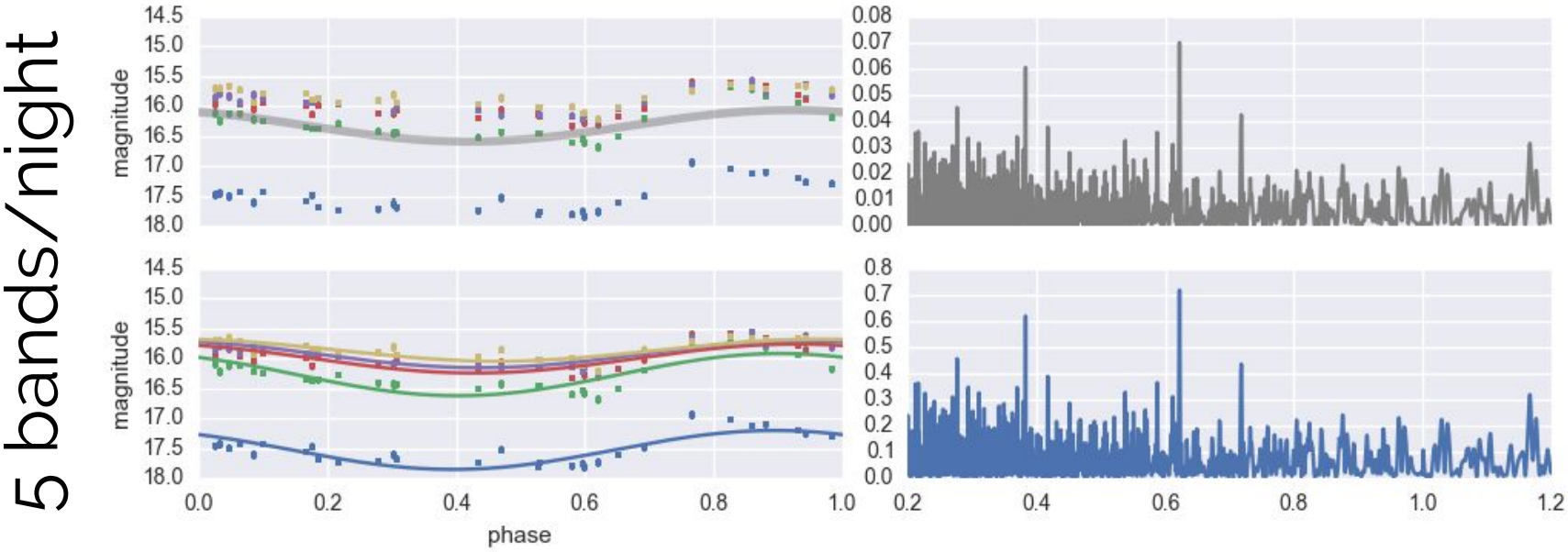


Figure: VanderPlas & Ivezic 2015

cf. Lomb (1976), Scargle (1982)

Two Naive Multiband Approaches

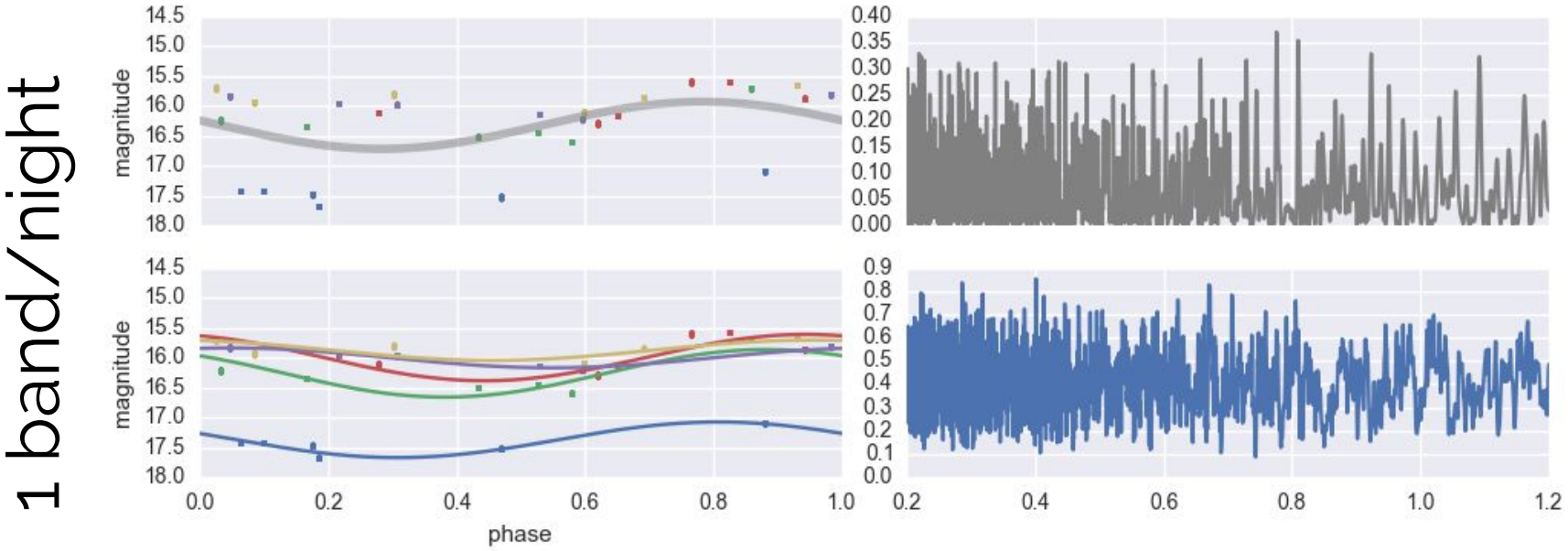
1. Ignore band distinction and fit a single periodogram to all bands.
   (model is highly biased: under-fits the data)



2. Fit an independent periodogram within each band; combine the χ^2 of all K bands
   (model is too flexible: over-fits the data)

Two Naive Multiband Approaches

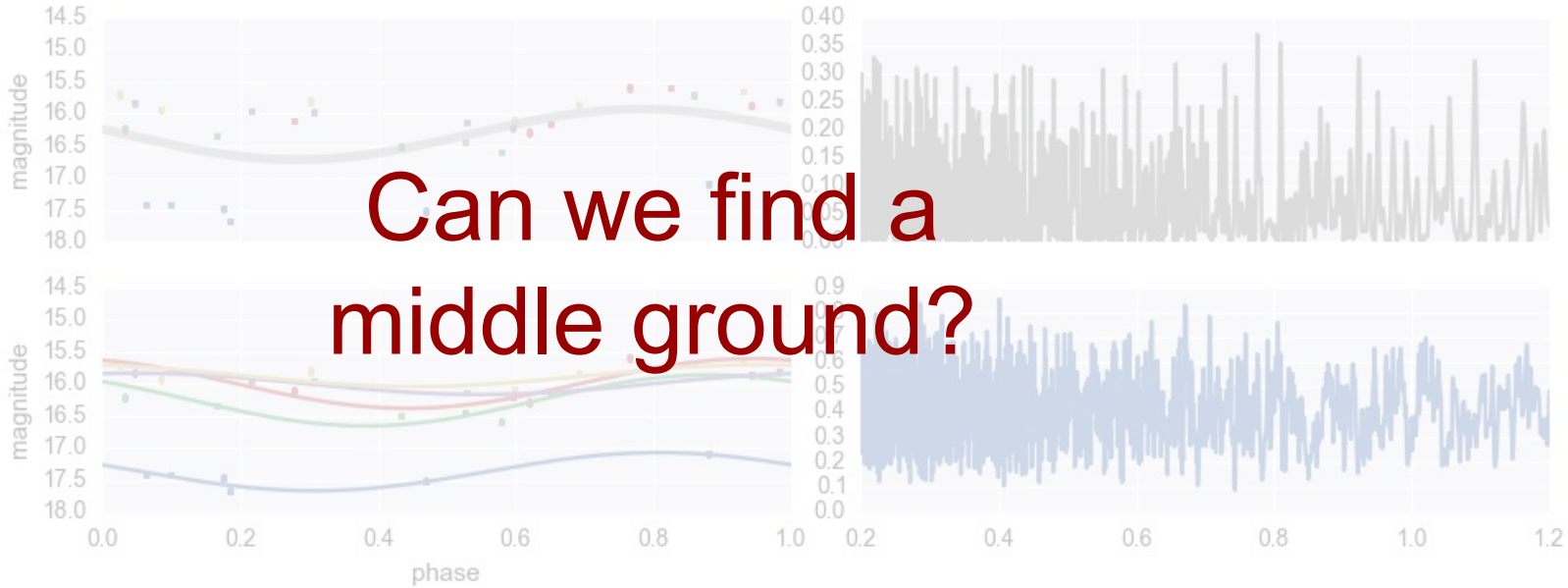
1. Ignore band distinction and fit a single periodogram to all bands.
   (model is highly biased: under-fits the data)



2. Fit an independent periodogram within each band; combine the χ^2 of all K bands
   (model is too flexible: over-fits the data)

Two Naive Multiband Approaches

1. Ignore band distinction and fit a single periodogram to all bands.
   (model is highly biased: under-fits the data)

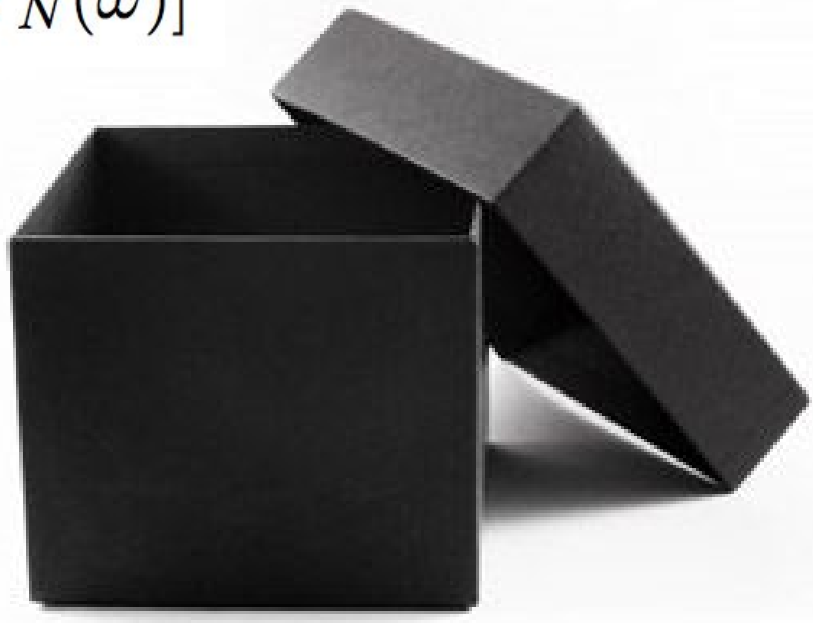


2. Fit an independent periodogram within each band; combine the χ^2 of all K bands
   (model is too flexible: over-fits the data)

Connection between Fourier periodogram and least squares allows us begin generalizing the periodogram . . .

$$y(t|\omega, \theta) = \theta_1 \sin(\omega t) + \theta_2 \cos(\omega t).$$

$$\chi^2_{min}(\omega) = \chi^2_0[1 - P_N(\omega)]$$



Floating Mean Model

. . . in which we simultaneously fit the mean

$$y(t \mid \omega, \theta) = \theta_0 + \theta_1 \sin \omega t + \theta_2 \cos \omega t$$

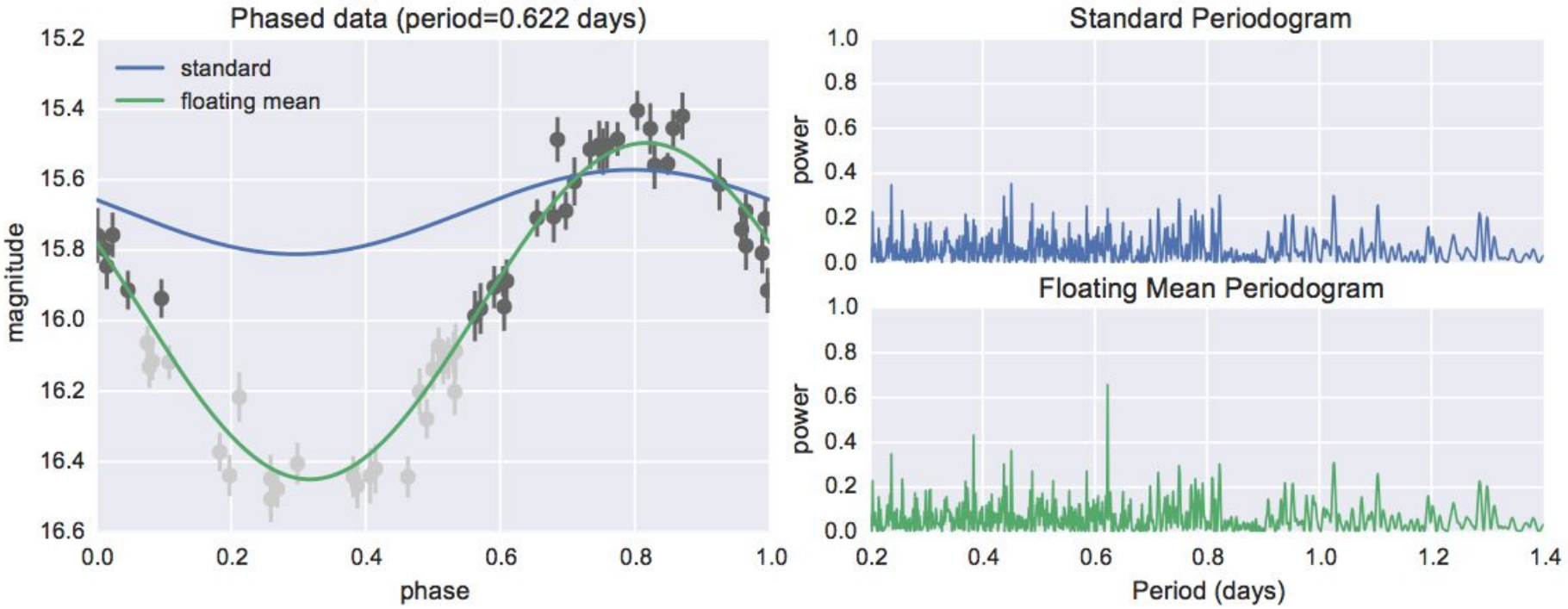


Figure: VanderPlas & Ivezic 2015

cf. Ferraz-Mello (1981); Cumming *et al* (1999); Zechmeister & Kurster (2009)

Truncated Fourier Model

. . . in which we fit for higher-order periodicity

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

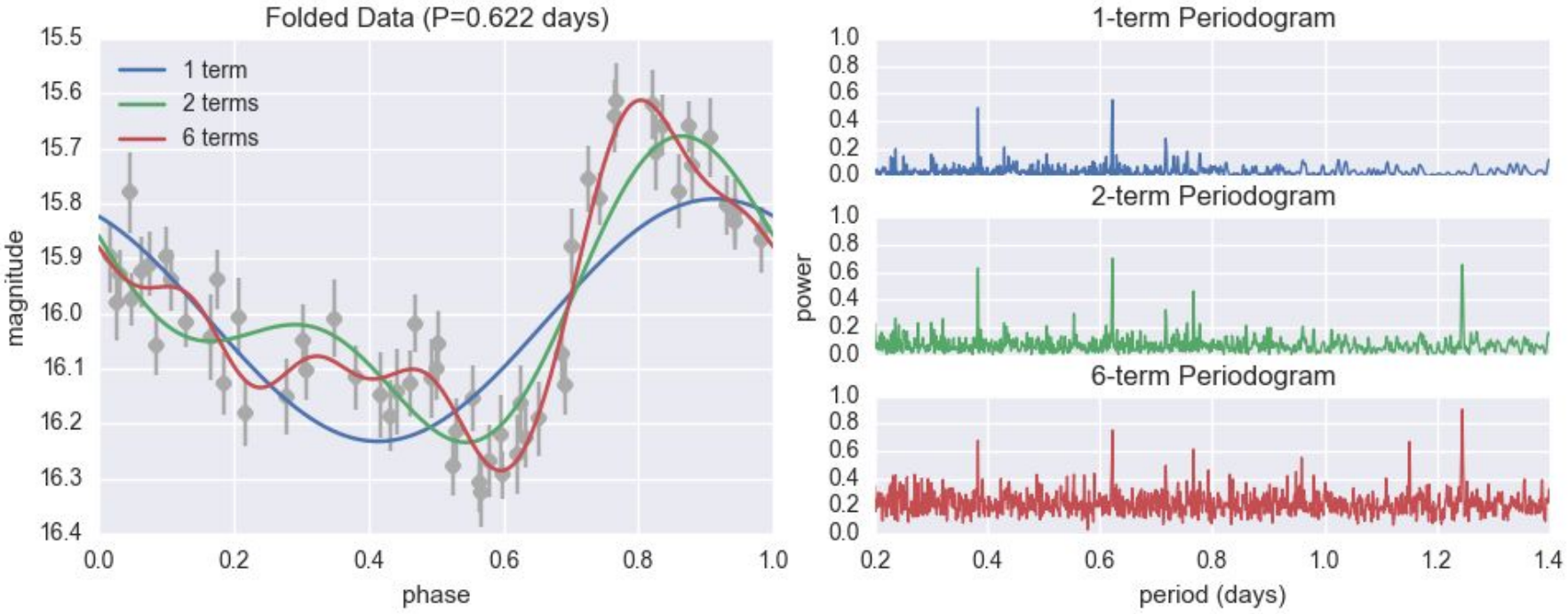


Figure: VanderPlas & Ivezic 2015

cf. Bretthorst (1988)

Truncated Fourier Model

. . . in which we fit for higher-order periodicity

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

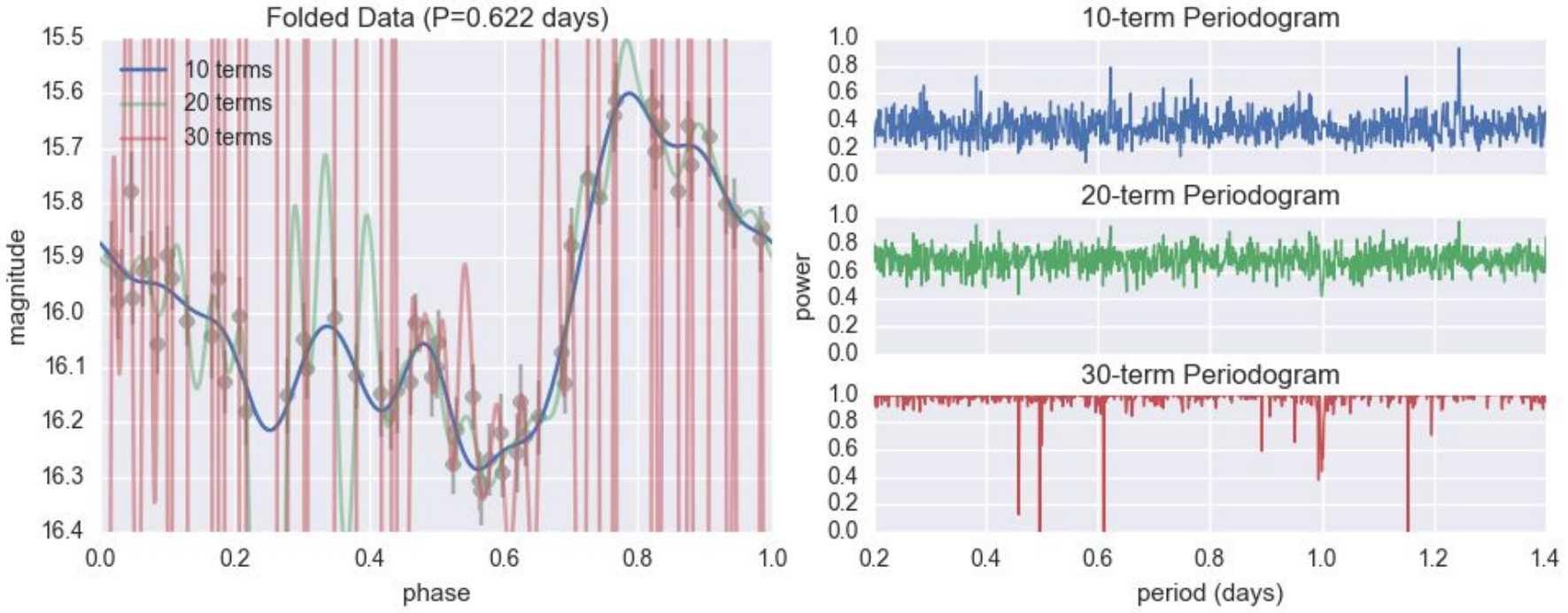
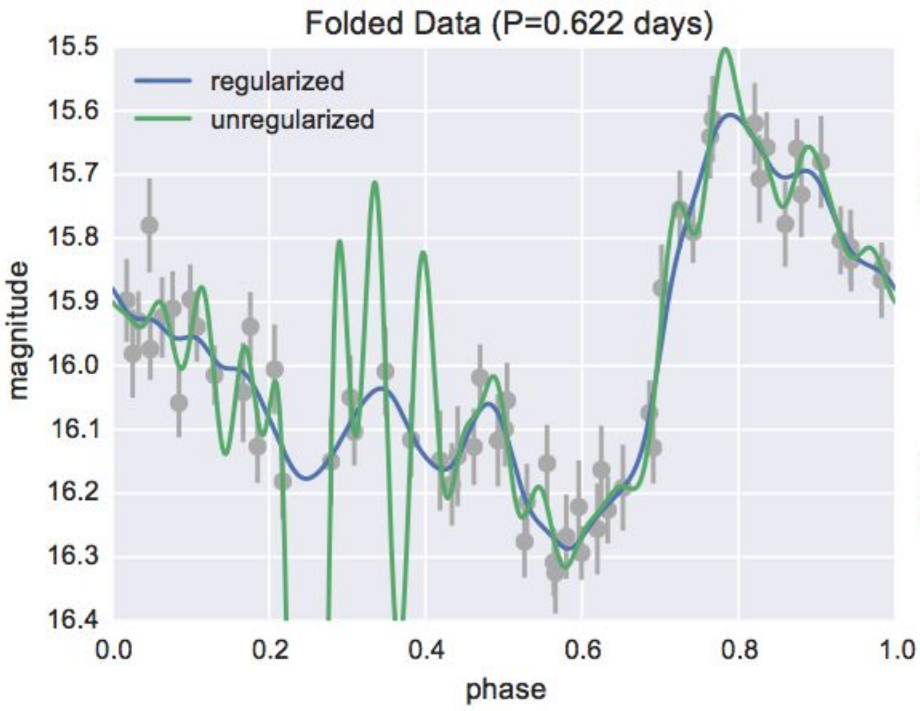


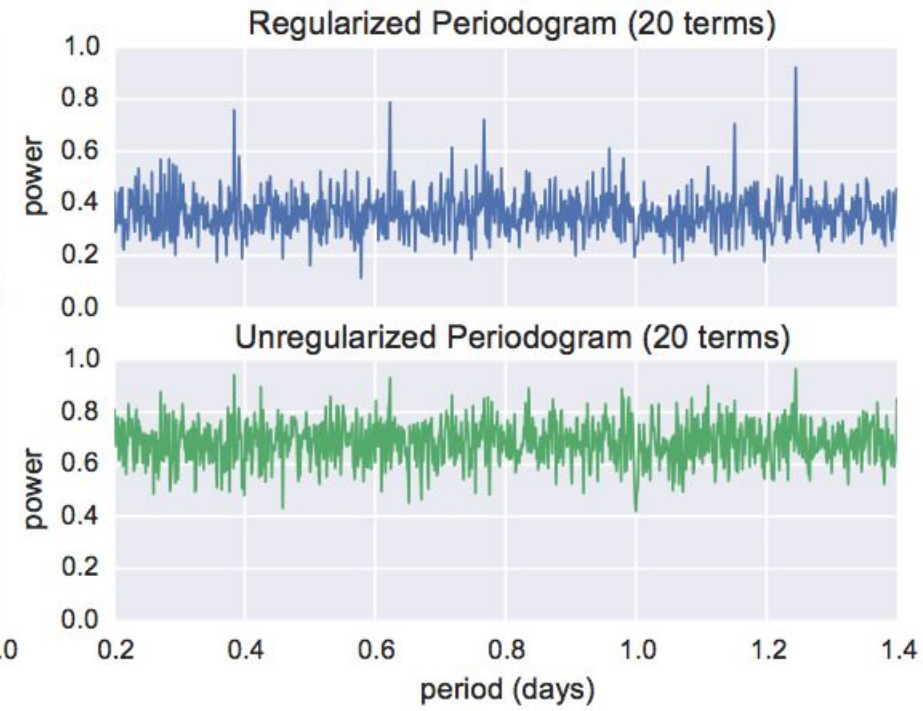
Figure: VanderPlas & Ivezic 2015

cf. Bretthorst (1988)

Regularized Model

. . . in which we penalize regression coefficients to simplify an overly-complex model.





The "trick" is adding a strong prior which pushes coefficients to zero: higher terms are only used if actually needed!

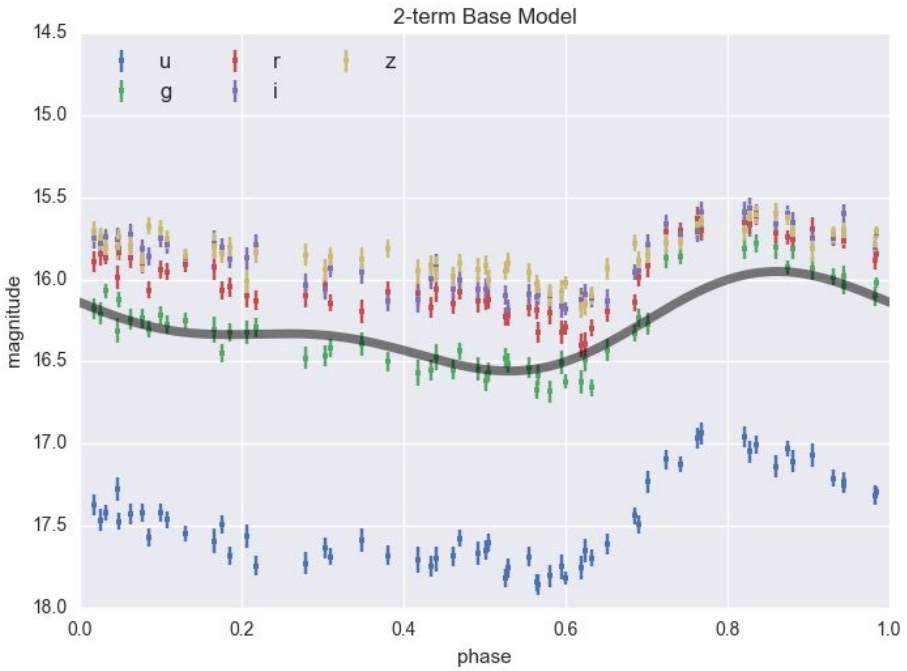
Putting it all together: The Multiband Periodogram

$$y_k(t|\omega,\theta) = \theta_0 + \sum_{n=1}^{N_{base}} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right] + \theta_0^{(k)} + \sum_{n=1}^{N_{band}} \left[\theta_{2n-1}^{(k)}\sin(n\omega t) + \theta_{2n}^{(k)}\cos(n\omega t)\right]. \quad (18)$$

Putting it all together: The Multiband Periodogram

$$y_k(t|\omega,\theta) = \theta_0 + \sum_{n=1}^{N_{base}} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right] + \theta_0^{(k)} + \sum_{n=1}^{N_{band}} \left[\theta_{2n-1}^{(k)}\sin(n\omega t) + \theta_{2n}^{(k)}\cos(n\omega t)\right]. \quad (18)$$

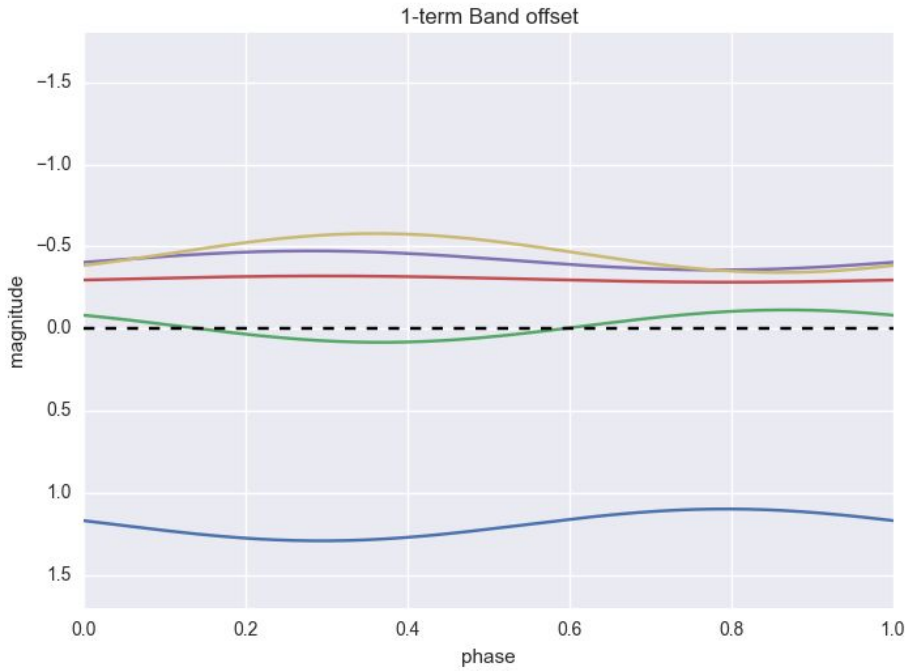
- define a truncated Fourier ***base component*** which contributes equally to all bands.



Putting it all together: The Multiband Periodogram

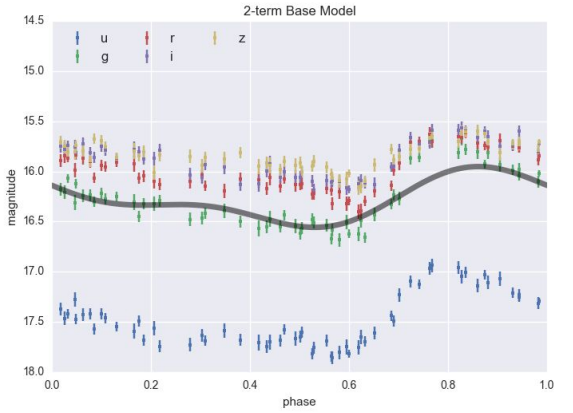
$$y_k(t|\omega,\theta) = \theta_0 + \sum_{n=1}^{N_{base}} \left[\theta_{2n-1}\sin(n\omega t) + \theta_{2n}\cos(n\omega t)\right] + \theta_0^{(k)} + \sum_{n=1}^{N_{band}} \left[\theta_{2n-1}^{(k)}\sin(n\omega t) + \theta_{2n}^{(k)}\cos(n\omega t)\right] . \quad (18)$$

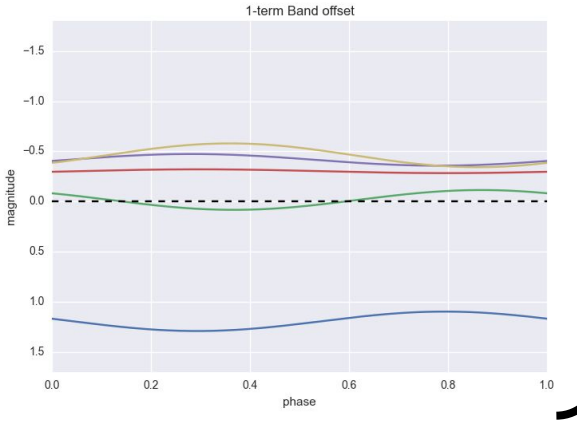
- for each band, add a truncated Fourier ***band component*** to describe deviation from base model



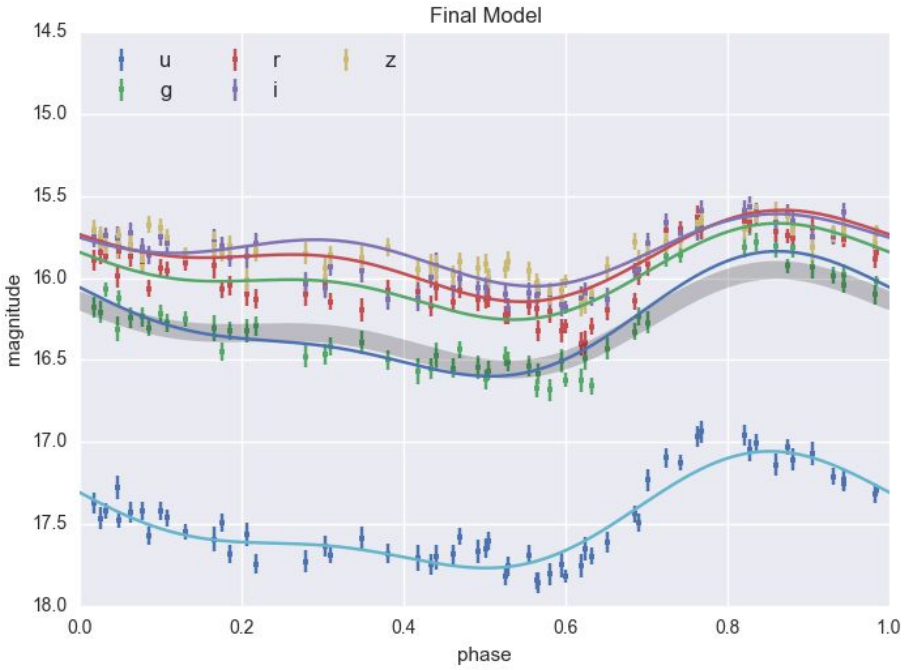
Putting it all together: The Multiband Periodogram

Regularize the band component to drive common variation to the base model.





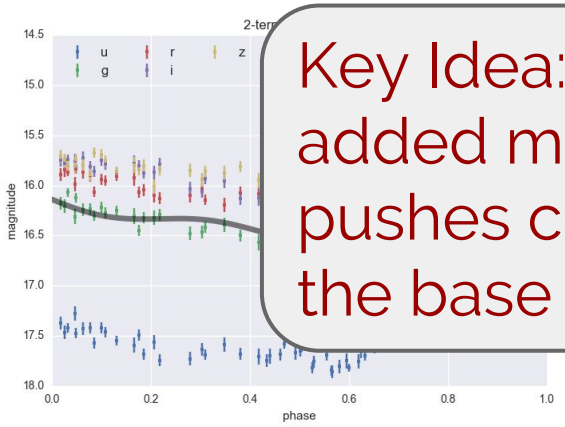
=

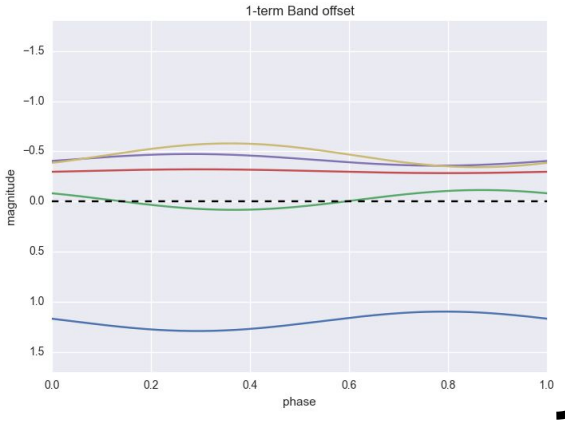


Putting it all together: The Multiband Periodogram

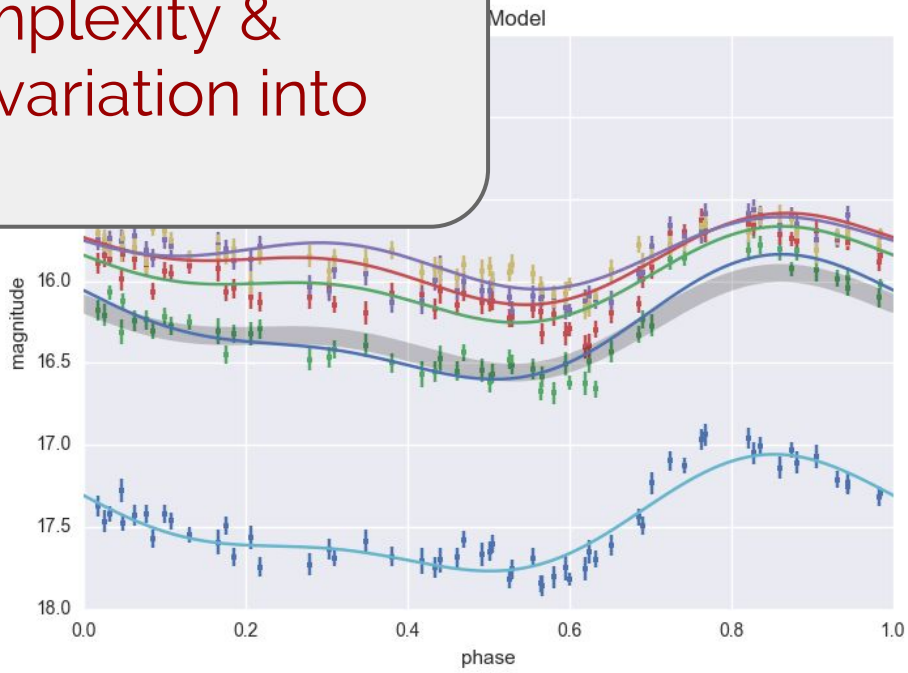
Regularize the band component to drive common variation to the base model.

Key Idea: Regularization reduces added model complexity & pushes common variation into the base model.



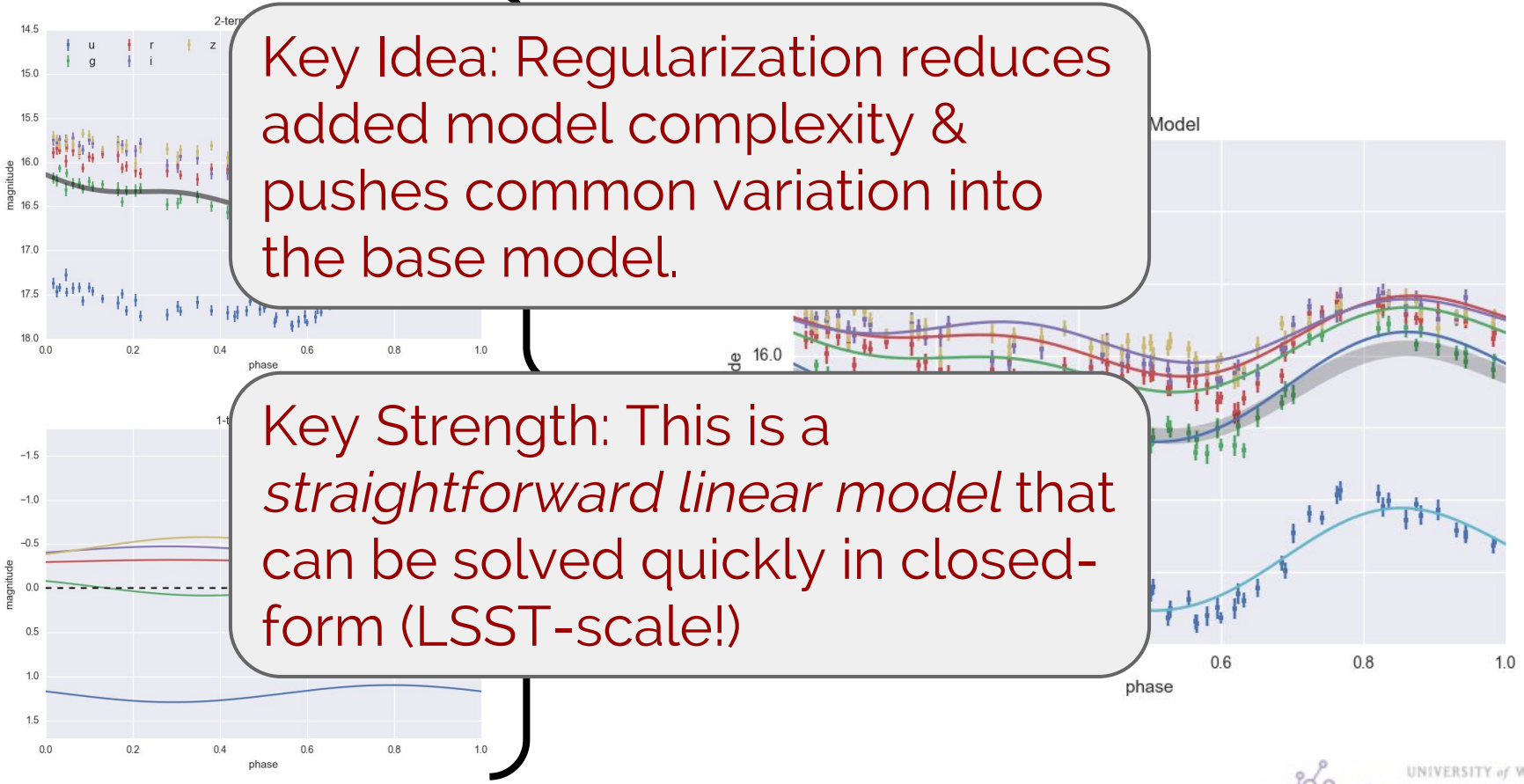


=



Putting it all together: The Multiband Periodogram

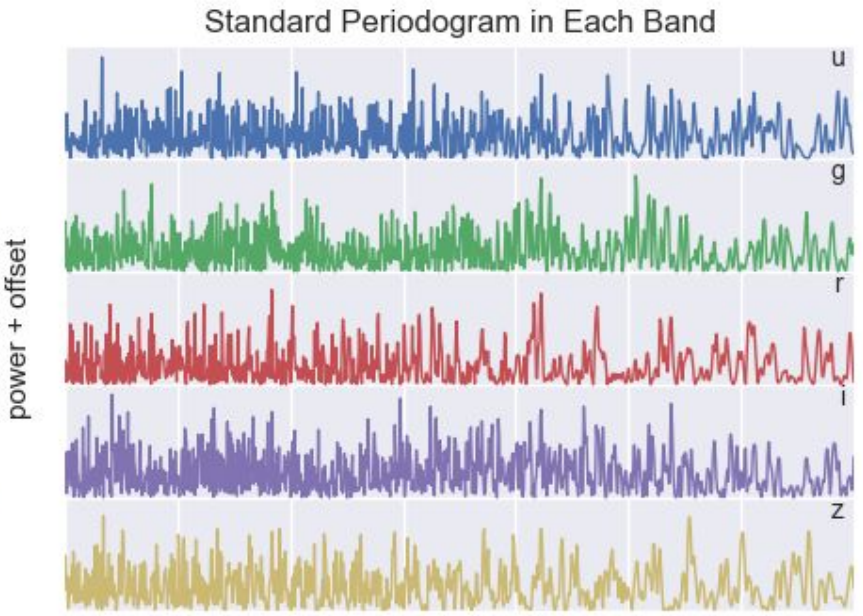
Regularize the band component to drive common variation to the base model.



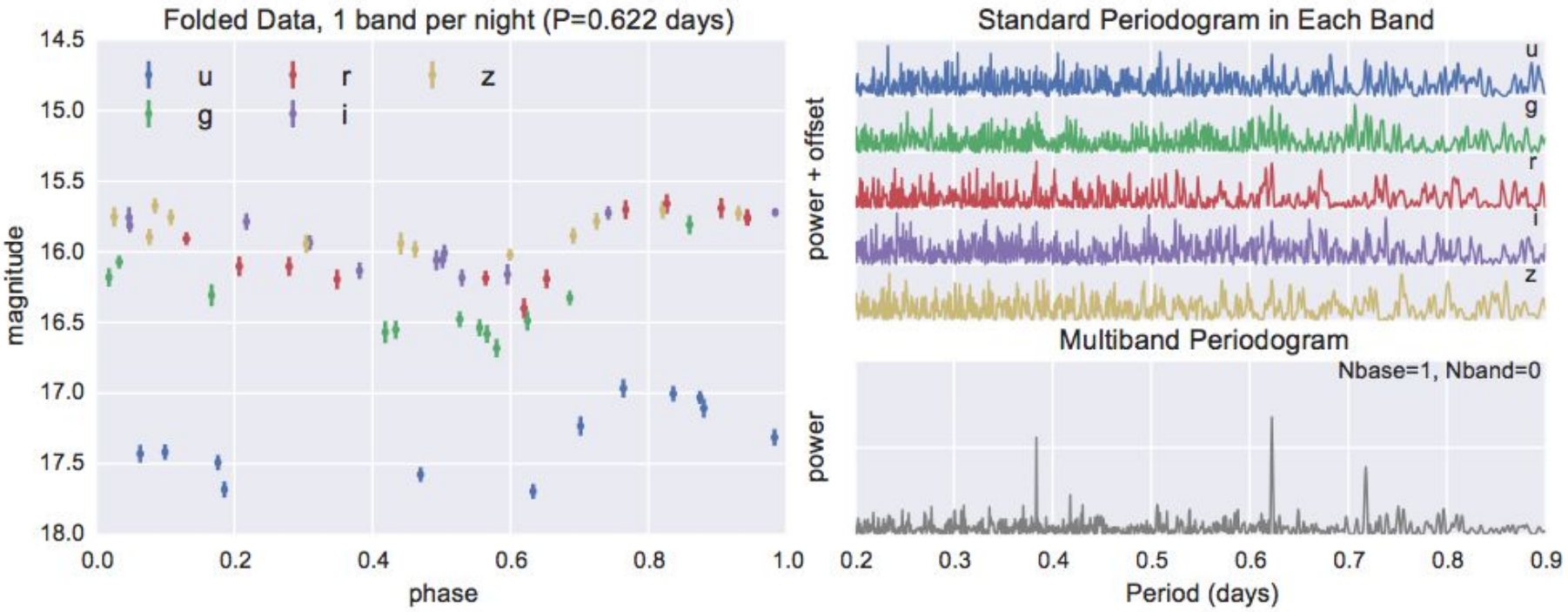
Key Idea: Regularization reduces added model complexity & pushes common variation into the base model.

Key Strength: This is a *straightforward linear model* that can be solved quickly in closed-form (LSST-scale!)

Multiband Periodogram on sparse, LSST-style data . . .

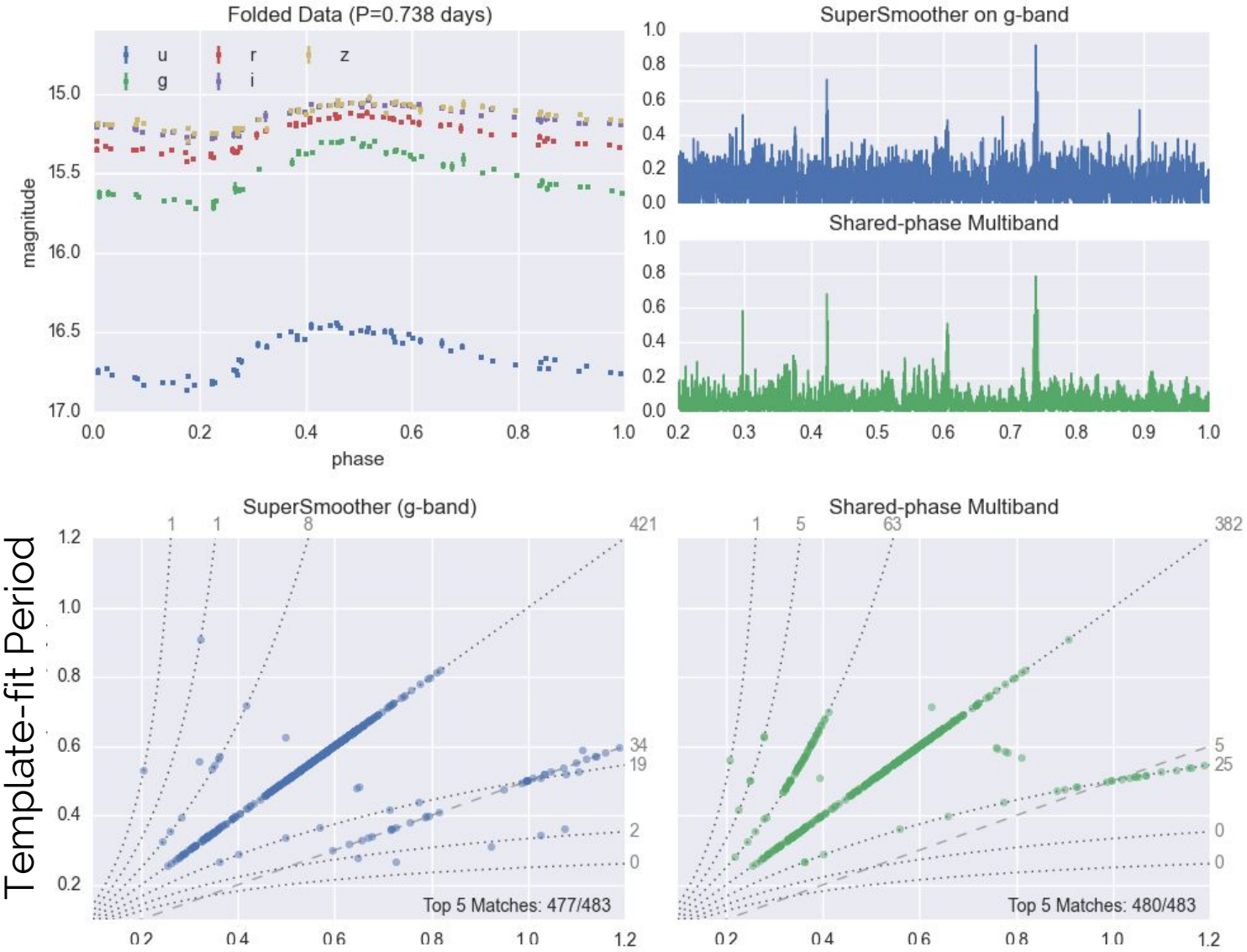


Multiband Periodogram on sparse, LSST-style data . . .

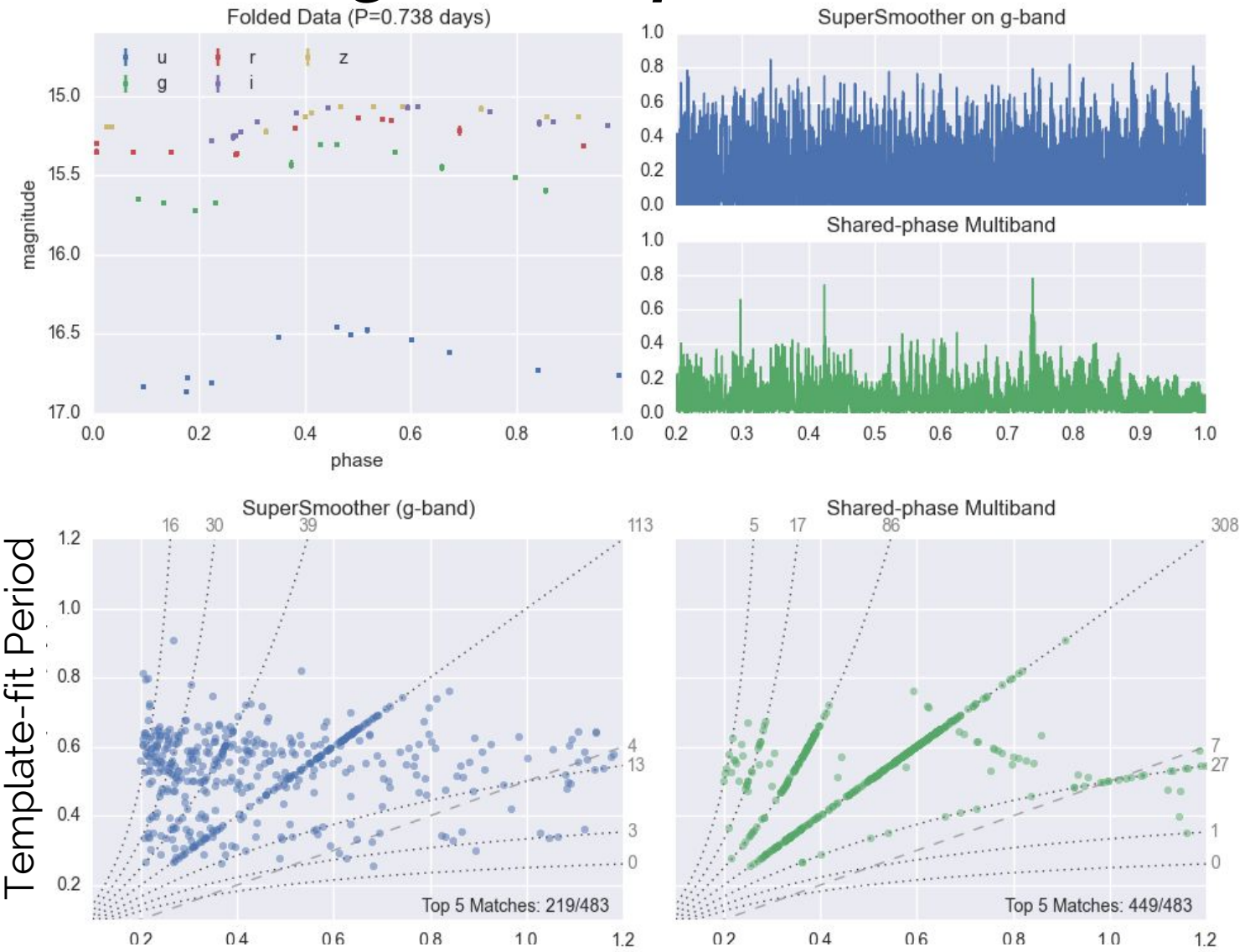


Detects period with high significance when single-band approaches fail!

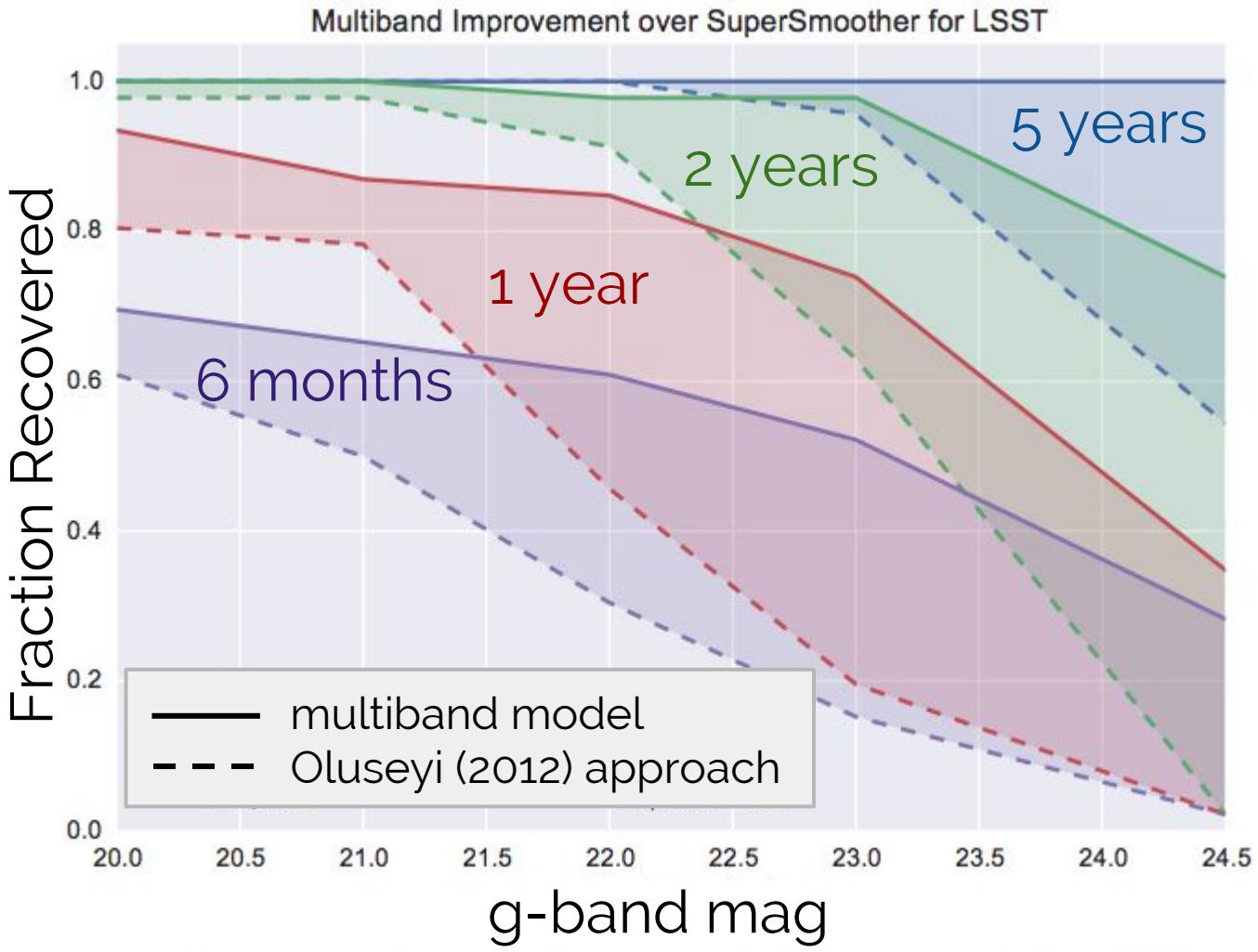
Comparing Approaches:
Stripe-82 Data *(5 bands per visit)*



Comparing Approaches:
Stripe-82 Data *(single band per visit)*

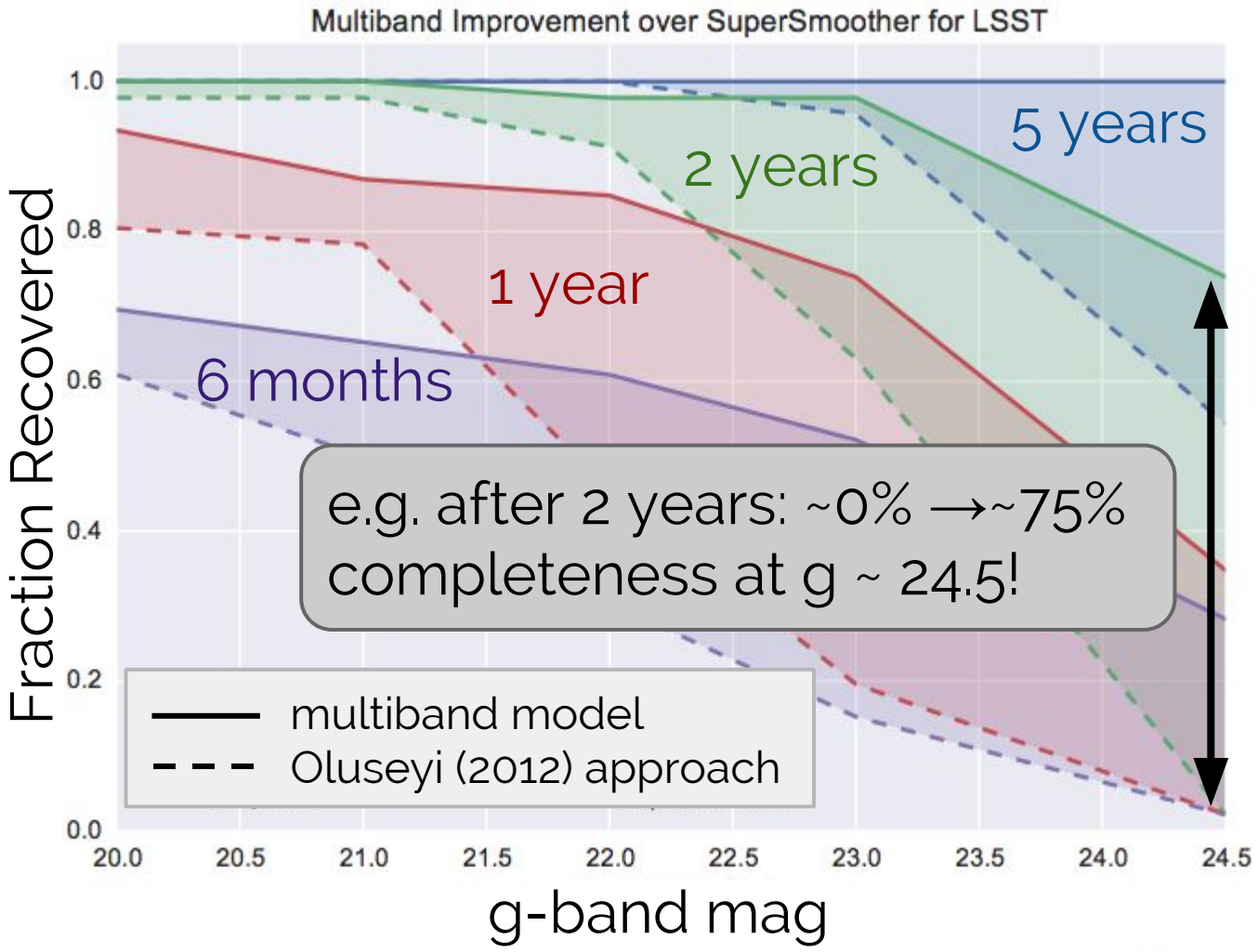


Prospects for LSST



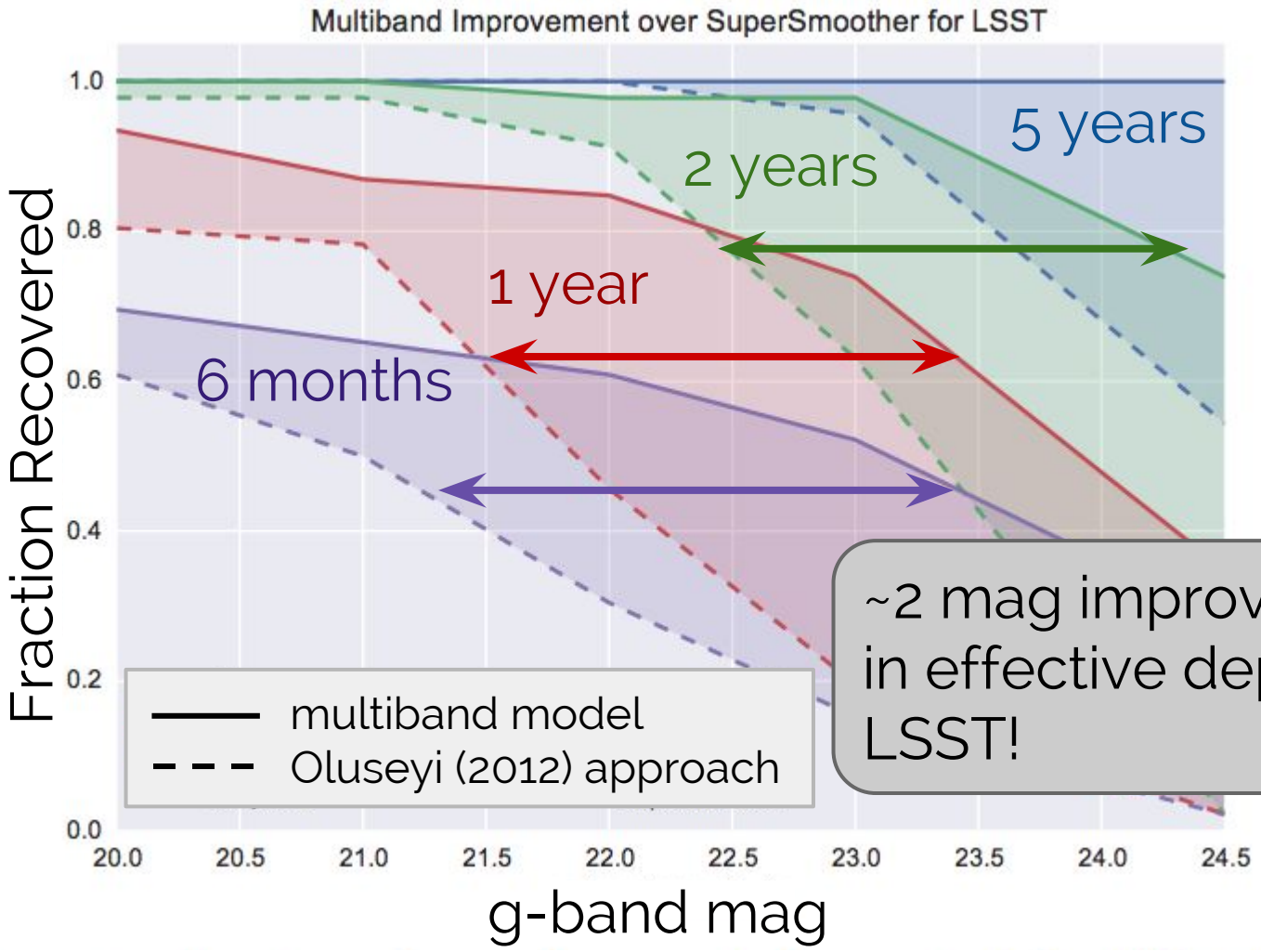
Based on simulated LSST cadence & photometric errors;
see VanderPlas & Ivezic (2015)

Prospects for LSST



Based on simulated LSST cadence & photometric errors; see VanderPlas & Ivezic (2015)

Prospects for LSST



~2 mag improvement in effective depth of LSST!

Based on simulated LSST cadence & photometric errors; see VanderPlas & Ivezic (2015)

Code to reproduce the study & figures (including all figures in these slides):

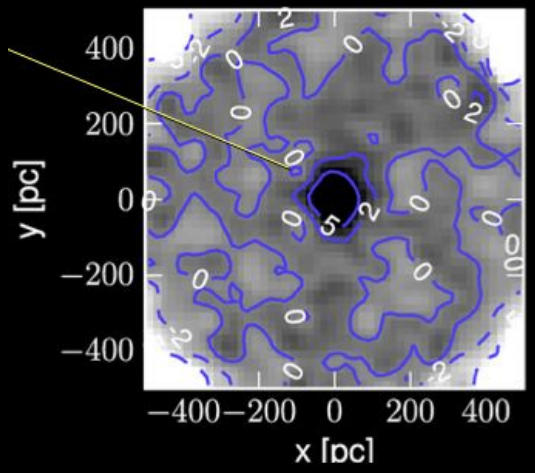
http://github.com/jakevdp/multiband_LS/

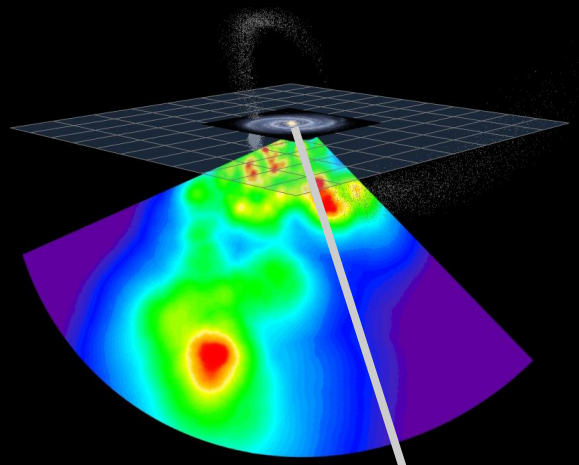
Python multiband implementation:

http://github.com/jakevdp/gatspy/

"If it's not reproducible, it's not science."

Back to our Motivation:





? ? ?

~100kpc

Other Recent Progress:

- **Long, Chi, & Baraniuk (2015)**

  Multiband extension of Lomb-Scargle — uses a *nonlinear* regularization on amplitude & phase offset. Better physical motivation, but more computationally intensive.

- **Mondrik, Long, & Marshall (2015)**

  Multiband extension of Analysis of Variance periodogram — also explores dependence of multiband detections on survey cadence.

Interesting Pre-LSST Datasets

- **Pan-STARRS**

  Natural testing ground for multiband methods, though data is *very* sparse; currently some RR Lyrae studies underway (B. Sesar; in prep).

- **SDSS Stripe 82 Reprised**

  LSST's analysis pipeline is capable of going much deeper via "forced photometry". Stripe 82 reanalysis is leading to interesting progress in QSO science (Y. AlSayyad; in prep) Could we find deeper RR Lyrae in re-processed SDSS data?

Astrostatistics: Opening the Black Box

- **Realize** that future surveys will likely not be optimized for your particular science interests
- **Identify** where standard algorithms & statistical methods will fail ..
- **Understand** the methods you want to apply & the assumptions behind them.
- **Adapt** the methods for use with sparse, heterogeneous, noisy, large datasets.

Thank You!

Email: jakevdp@uw.edu

Twitter: @jakevdp

Github: jakevdp

Web: http://vanderplas.com/

Blog: http://jakevdp.github.io/

Astrostatistics: Opening the Black Box

abstract: The large datasets being generated by current and future astronomical surveys give us the ability to answer questions at a breadth and depth that was previously unimaginable. Yet datasets which strive to be generally useful are rarely ideal for any particular science case: measurements are often sparser, noisier, or more heterogeneous than one might hope. To adapt tried-and-true statistical methods to this new milieu of large-scale, noisy, heterogeneous data often requires us to re-examine these methods: to pry off the lid of the black box and consider the assumptions they are built on, and how these assumptions can be relaxed for use in this new context. In this talk I'll explore a case study of such an exercise: our extension of the Lomb-Scargle Periodogram for use with the sparse, multi-color photometry expected from LSST. For studies involving RR-Lyrae-type variable stars, we expect this multiband algorithm to push the effective depth of LSST two magnitudes deeper than for previously used methods.