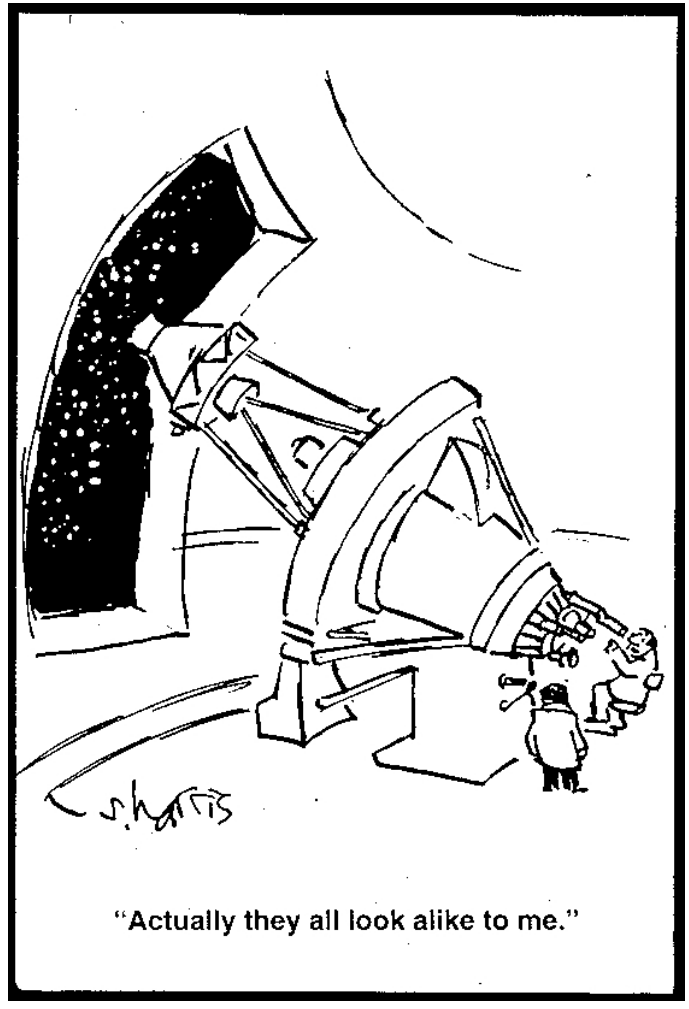
Lightcurve classification for periodically varying stars [Lightcurves]

David Jones

SAMSI

May 12, 2017

Variable sources



http://cartoonsmix.com

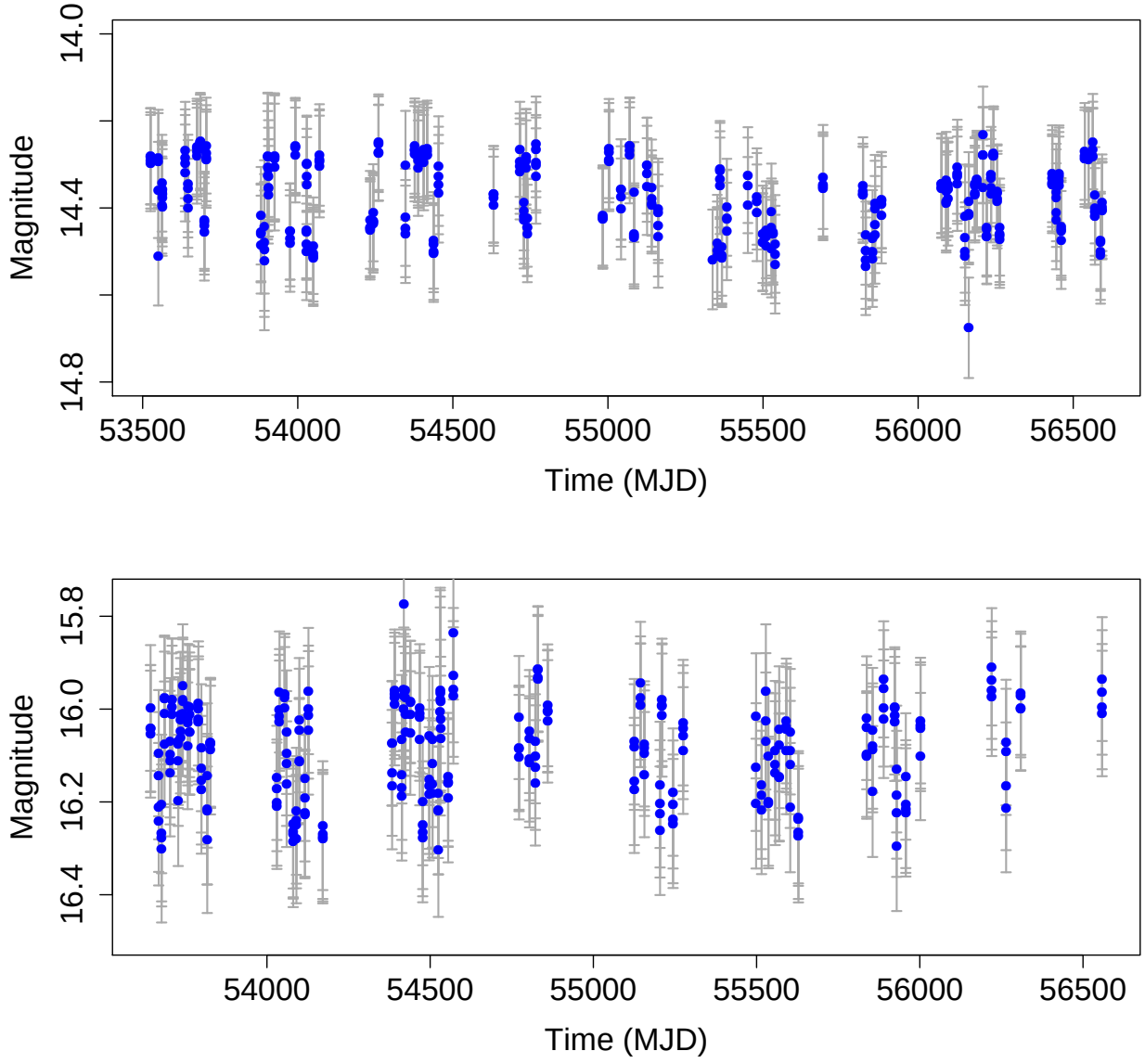
Eclipsing binaries

https://www.eso.org

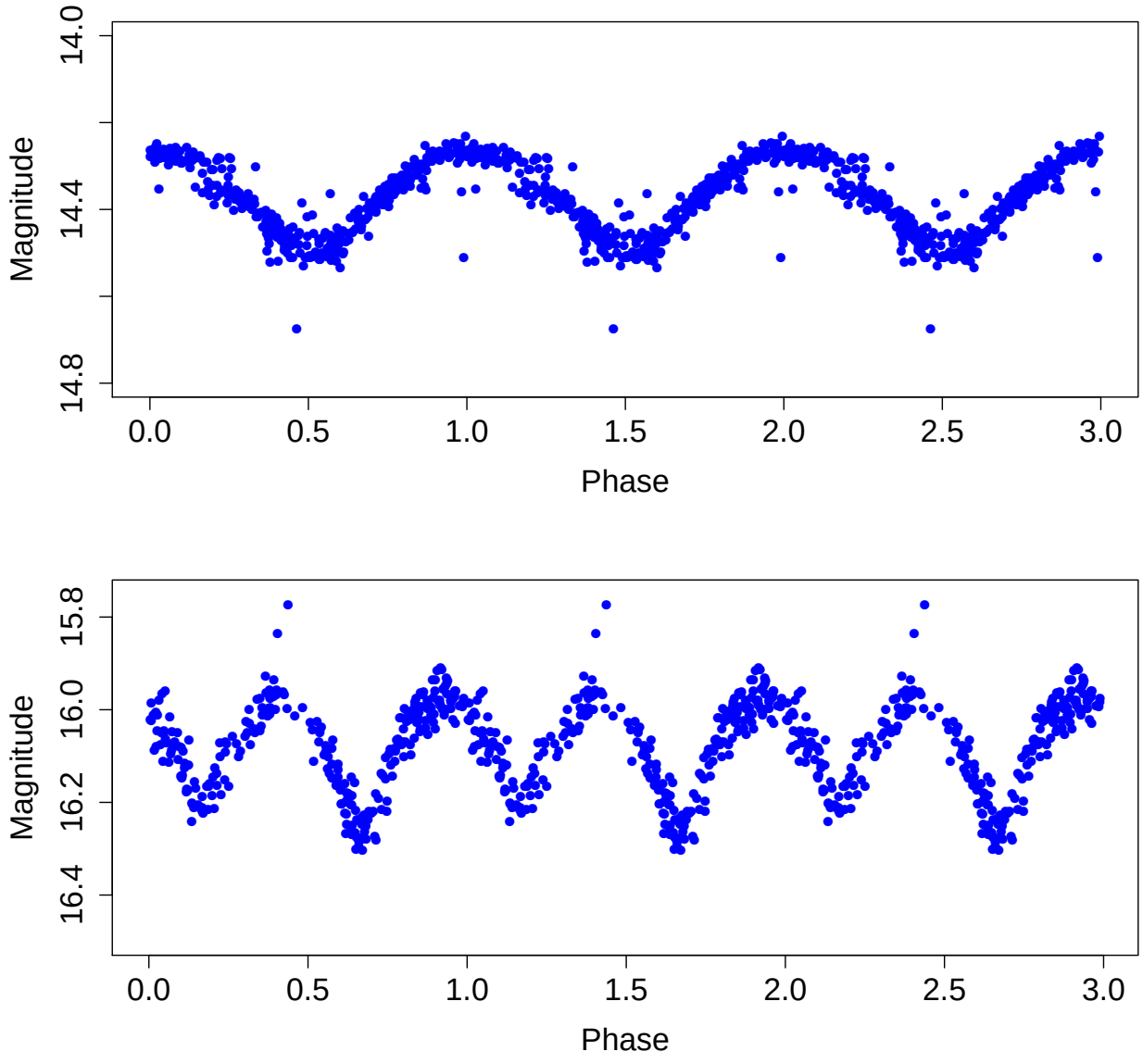
Pulsating stars

https://www.spacetelescope.org

Raw lightcurves



Folded lightcurves



Catalina Real-Time Transient Survey (CRTS)

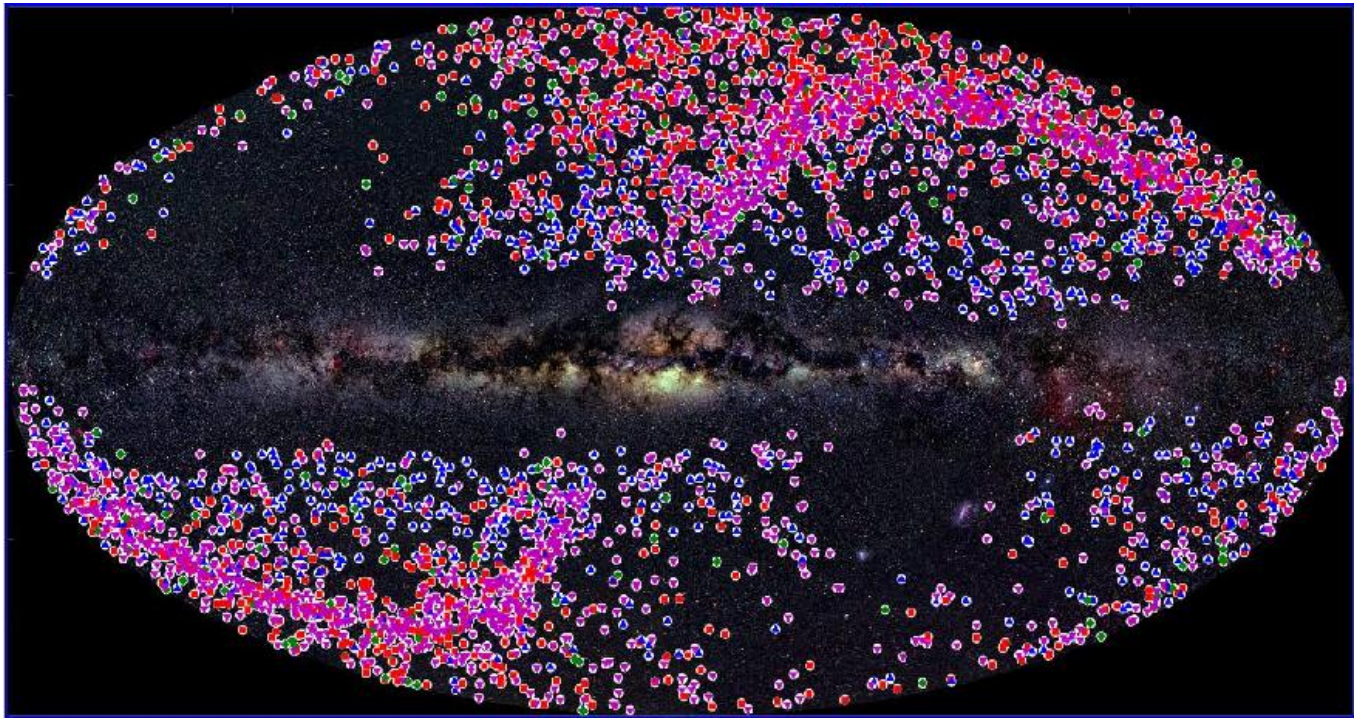
http://crts.caltech.edu/

Mt. Lemmon Survey, Arizona

Catalina Sky Survey, Mt. Bigelow, Arizona

Siding Spring Survey, NSW, Australia

Catalina Real-Time Transient Survey (CRTS)



CRTS Transient Discoveries, `http://crts.caltech.edu/`

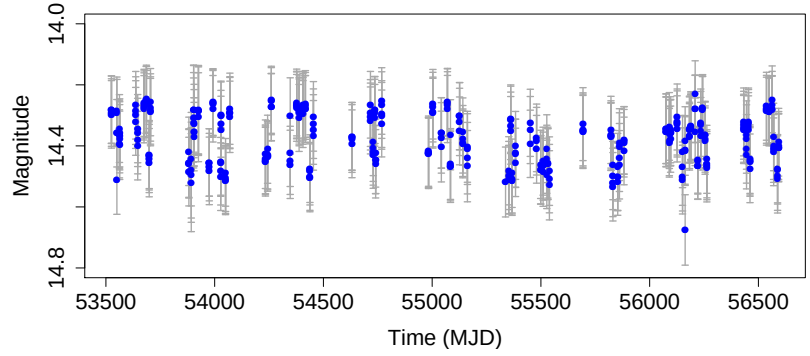
- Covers 33k square degress of the sky
- Goal: "discover rare and interesting transient phenomena"
- Aims to automate discovery and classification as far as possible
- Publishes discoveries within minutes
- About 50k periodic variables discovered so far

Data

Table 2
Types of Periodic Variables

Type	F (%)	N	Class
EW	49.93	30743	1
EA	7.61	4683	2
β Lyrae	0.45	279	3
RRab	27.28	16797[a]	4
RRc	8.88	5469	5
RRd	0.82	502	6
Blazhko	0.36	223[a]	7
RS CVn	2.47	1522	8
ACEP	0.10	64	9
Cep-II	0.20	124	10
HADS	0.39	242	11
LADS	0.01	7	12
LPV	0.83	512	13
ELL	0.23	143	14
Hump	0.04	25	15
PCEB	0.14	85	16
EA_{UP}	0.25	155	17

CRTS data summary, Drake et al. 2014



1) Raw lightcurves

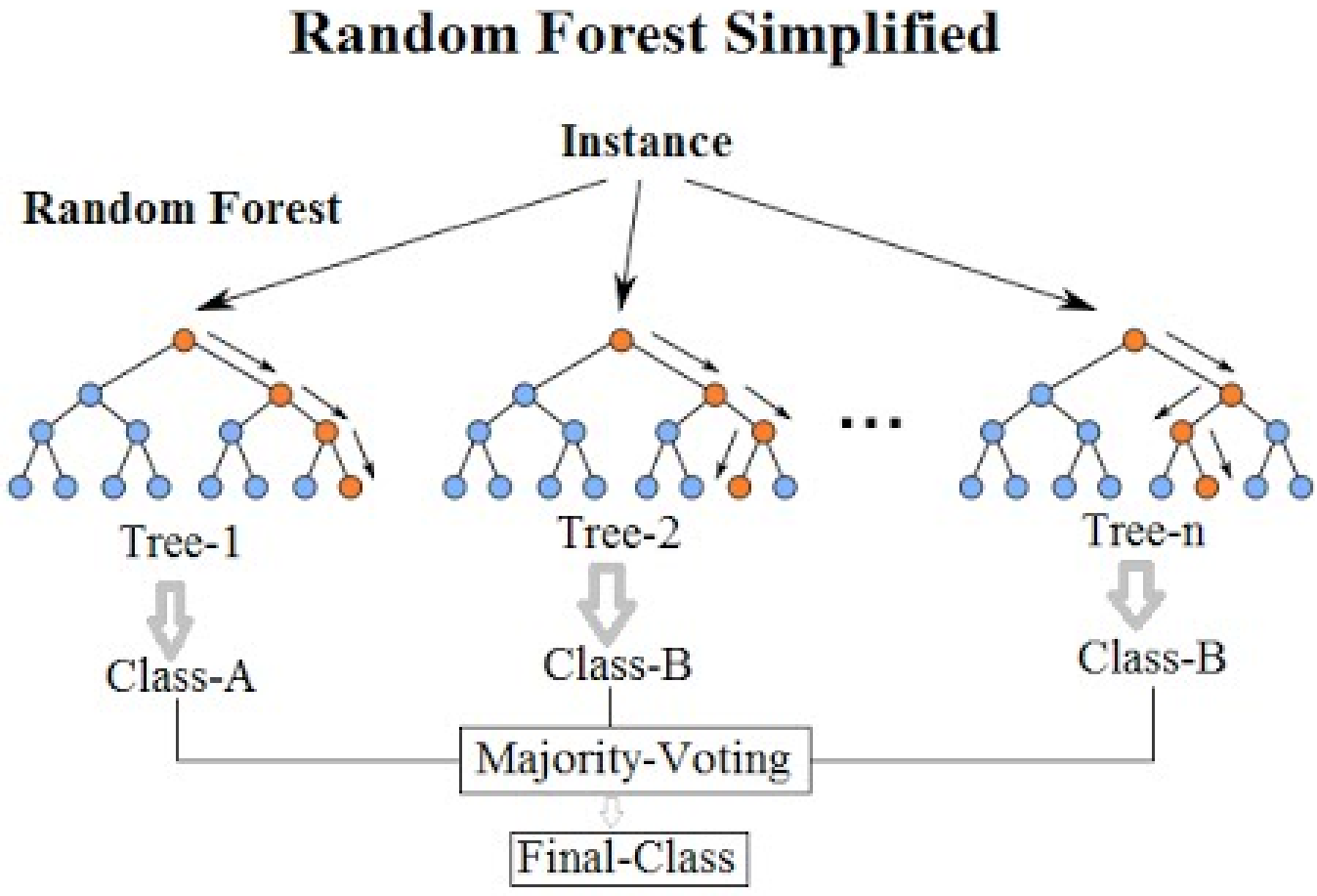
```
> head(training_set_features)
             ID V_mag      period range num_obs  class
1 1.109080e+12 17.71   0.3984844  0.48     265    RRd
2 1.009115e+12 14.37   0.3054930  0.23     220    RRc
3 1.007077e+12 15.90 513.4235200  0.24     213 RS_CVn
4 1.163018e+12 14.17   0.5256680  0.29      93   RRab
5 1.104029e+12 16.10   0.3657230  0.32     239     EW
6 1.146063e+12 16.28   0.3519940  0.54     180    RRd
```

2) Basic features

Classification trees

-

Random Forest classifiers



Niraj Kumar, `https://i.ytimg.com/vi/ajTc5y3OqSQ/hqdefault.jpg`

Random Forest classifiers

-

Training data and test data

- Training data:
- Test data:

Project goal

Goal: construct new features and train a classifier that is as accurate as possible.

The following two criteria will be used to assess the performance of the classifier (higher values are better):

- Primary: the percentage of test data sources correctly classified
- Secondary: minimize .....

Target to beat

-

Training data confusion matrix

Assessment on test data (not shown):

- Primary value =
- Secondary value =

Questions to consider

1. Which classes are likely to get mixed up?
2. What aspects of the lightcurves do the three basic features not capture?
3. What new features could be computed quickly?
4. How do we decide which features are most useful?
5. What classification method should we use? If we use a random forest classifier, are there options/inputs we can tune?
6. **What do we do if a new dataset includes only features for each lightcurve (not the raw lightcurves) and some features are missing?

References

1. Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. (2015). A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3), 2269-2291.
2. Dumusque, X., Boisse, I., & Santos, N. C. (2014). SOAP 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar sports and plages. *The Astrophysical Journal*, 796(2), 132.
3. Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D., & Ford, E. B. (2017). Insights on the spectral signatures of RV jitter from PCA. In *American Astronomical Society Meeting Abstracts*, 229.
4. Loredo, T. J., Berger, J. O., Chernoff, D. F., Clyde, M. A., & Liu, B. (2012). Bayesian methods for analysis and adaptive scheduling of exoplanet observations. *Statistical Methodology*, 9(1), 101-114.
5. Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning. The MIT Press.