Machine-learned classification of variable stars detected by (NEO)WISE

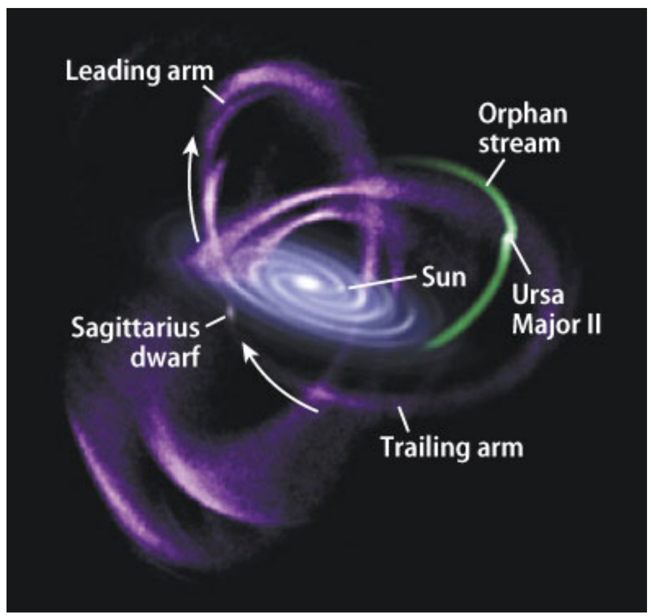
Frank Masci, Jeff Rich, Roc Cutri, Carl Grillmair & Doug Hoffman

WISE at 5: Legacy & Prospects – February 10, 2015

Goals

- We were awarded a NASA--ADAP grant in March 2013 to construct a generic WISE Variable Source Catalog (P.I. Roc Cutri) from first 13 months of data (~ 2.16 full sky coverages)
- Primary science driver: discover as many RR-Lyrae variable stars as possible in an attempt to associate with stellar debris streams around Milky Way (from disrupted satellite galaxies)
  - ➢ RR Lyrae in mid-IR provide excellent distance indicators (standard candles)
  - ➢ Accurate distances to just a few locations in streams + kinematic information

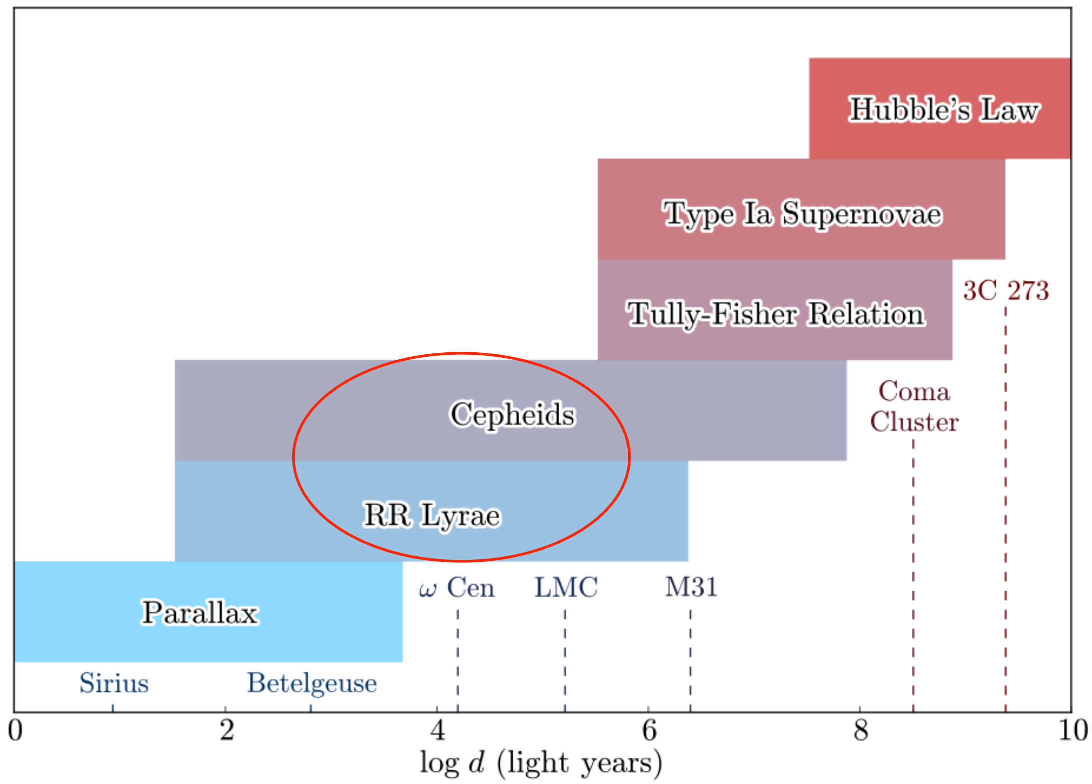
    => constrain gravitational potential, distribution of dark matter, …
- This catalog will be a valuable resource for the community



Belokurov et al. (2006)
Grillmair et al. (2006)
Also, see poster by Carl Grillmair

Anchors to the size-scale of the Universe

- RR Lyrae and Cepheid variables are used to establish the size-scale of the Universe
- Distance ladder with viable ranges of some common calibrators:

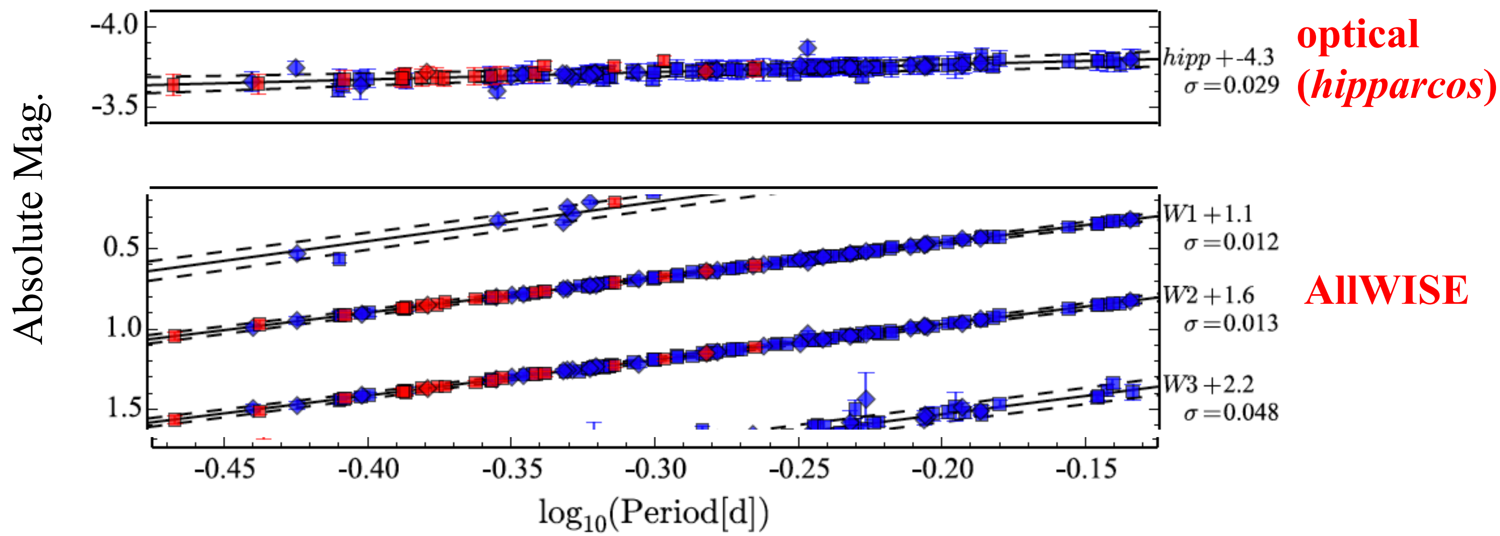


courtesy: *Zaritsky, Zabludo, & Gonzalez (2013)*

Mid-IR *Period-Luminosity* Relations

- Studies with WISE (& Spitzer) show that mid-IR provides a more accurate calibration (<~ 2%)
- *WISE RR Lyrae studies:* Madore et al. 2013; Klein et al. 2014; Dambis et al. 2014;
  - relatively immune to dust extinction: photometric scatter down by >50% cf. to optical!
  - SED ~ Rayleigh Jeans: surface brightness changes are less sensitive to temperature variations
  - leads to more homogeneous samples

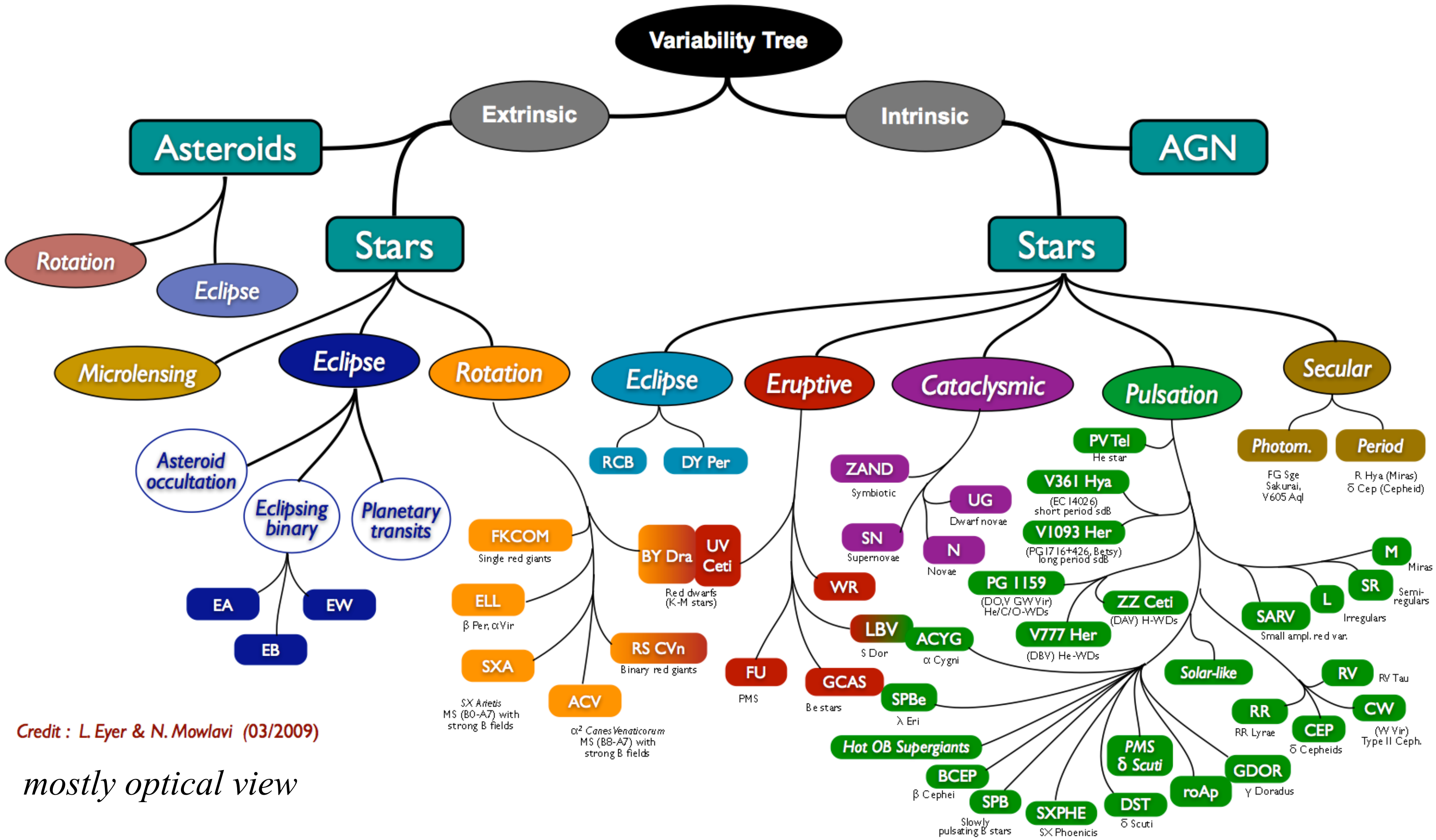
Klein & Bloom (2014)



Notice difference in slopes and scatter (sigmas) between optical and mid-IR

The ever growing tree…

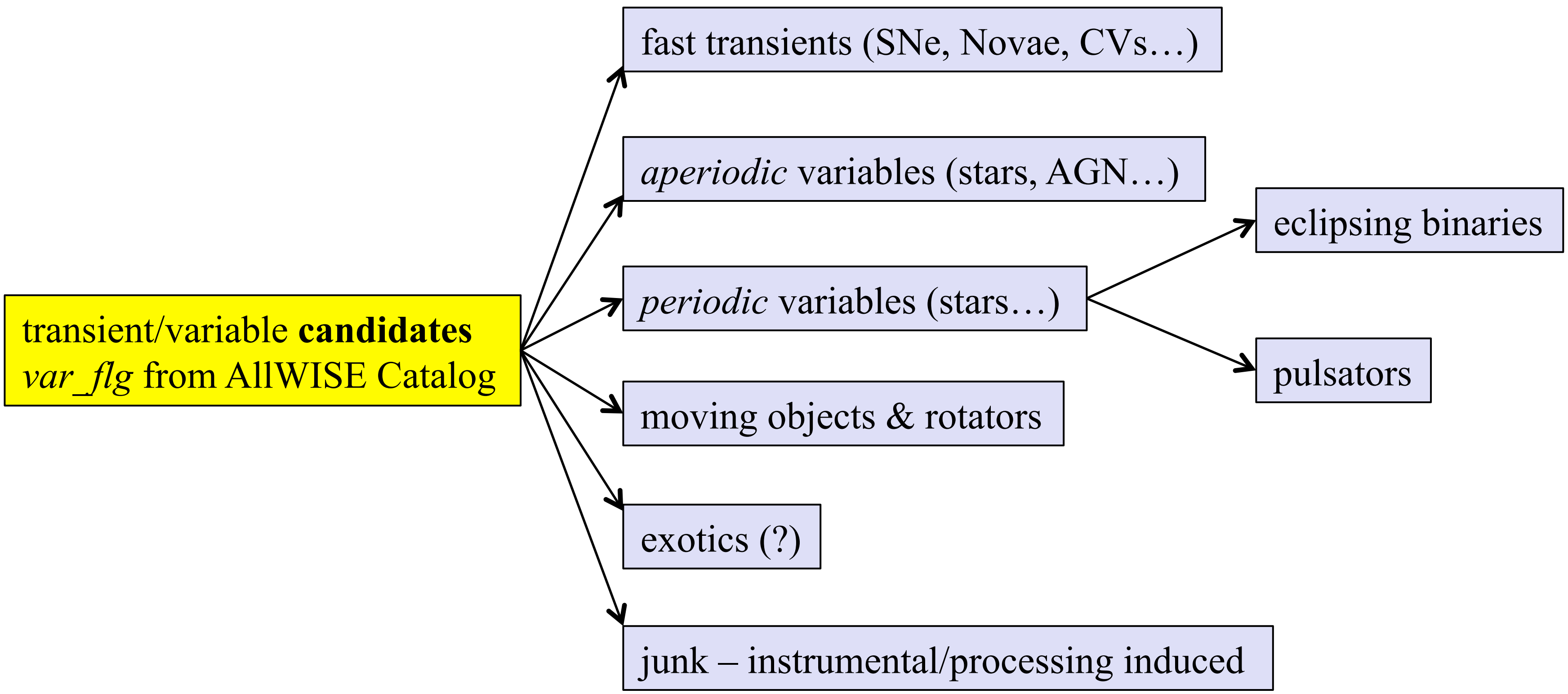
An attempt to classify the transient/variable sky (as of 2009)



mostly optical view

Constructing the WVSC

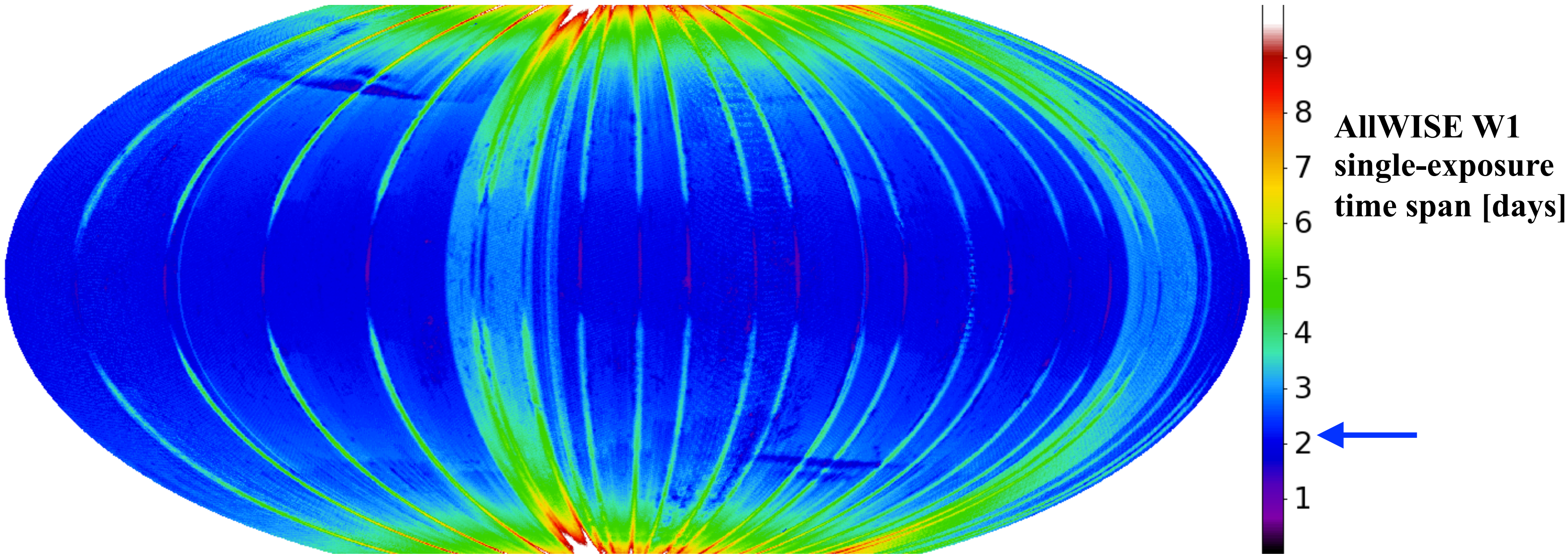
- The WISE Variable Source Catalog will potentially contain many transients/variables from previous slide, classified or not. Some will simply be one-off events from single-exposures.



- Goal is to classify (label) as much as possible according to available taxonomy
- But WISE's survey constraints and limitations presents a challenge

What is possible with ~1yr of (NEO)WISE?

- To *characterize and classify new* variables requires good-quality, well sampled light-curves
- The types of variables observed by (NEO)WISE that *best* lend themselves to classification depends on available single-exposure observing cadence and baseline

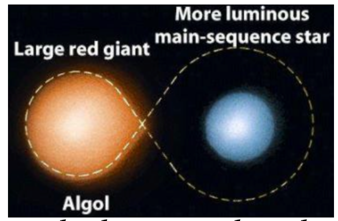


- 1 year survey => 2-sky passes => time-span per position near ecliptic is ~ 2 days (minus 6 month gap)
  - ➢ Two spliced quasi-continuous 1-day spans over most of sky: provides good phase sampling
  - ➢ Longer baselines near the ecliptic poles: see poster by **Jeffrey Rich:** "Ecliptic Pole Sources…"
- **Cadence:** same positions near the ecliptic visited ~ every 3 hours

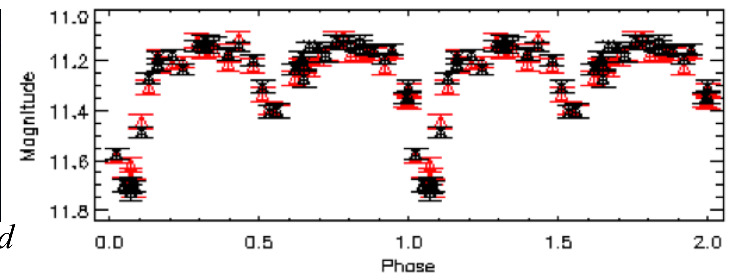
What is possible with ~1yr of (NEO)WISE?

Given survey constraints, the most common variables we expect to encounter from ~1 year of data are:

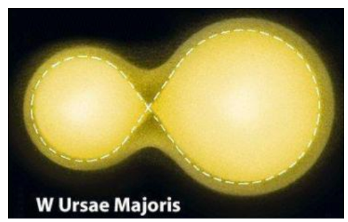
- Algol & β Lyrae type eclipsing binaries
  (*periods* <~ 2.5 days)



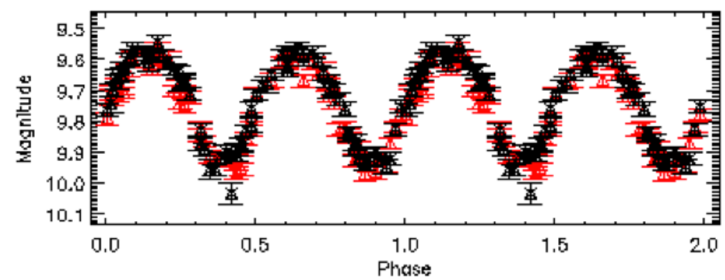
detached + semi-detached



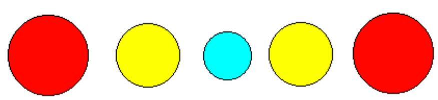
- W Ursae Majoris eclipsing binaries
  (*periods* <~ 2.5 days)

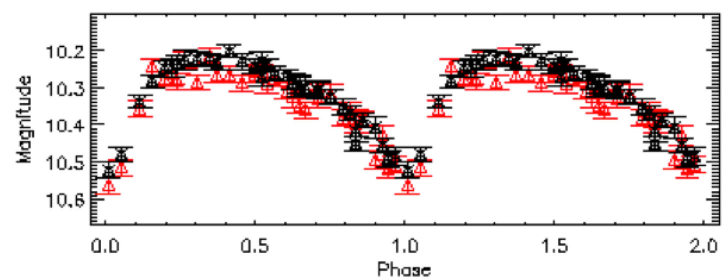


contact binaries



- RR Lyrae pulsators
  (*periods* <~ 1 day)



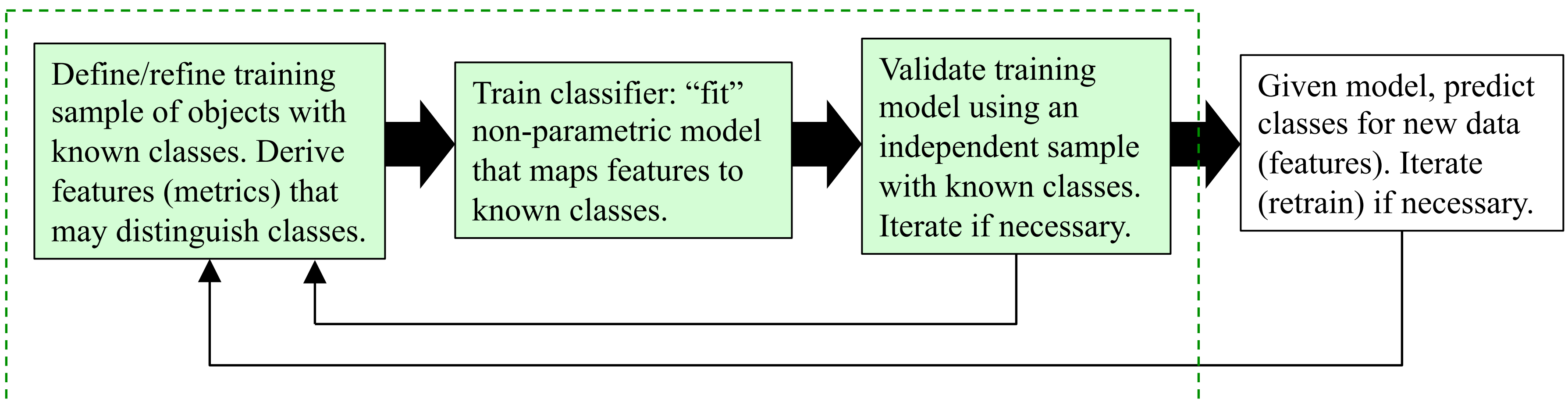


- Some short-period Cepheid variables (*periods* ~> 2 days), mostly at higher ecliptic latitude

Classification via Machine Learning

- Human based classification can be subjective, inconsistent, and is usually not reproducible
- ML is deterministic (i.e., consistently right or wrong), given same training model
- Can quantify class membership probabilistically instead of a simple binary yes/no decision

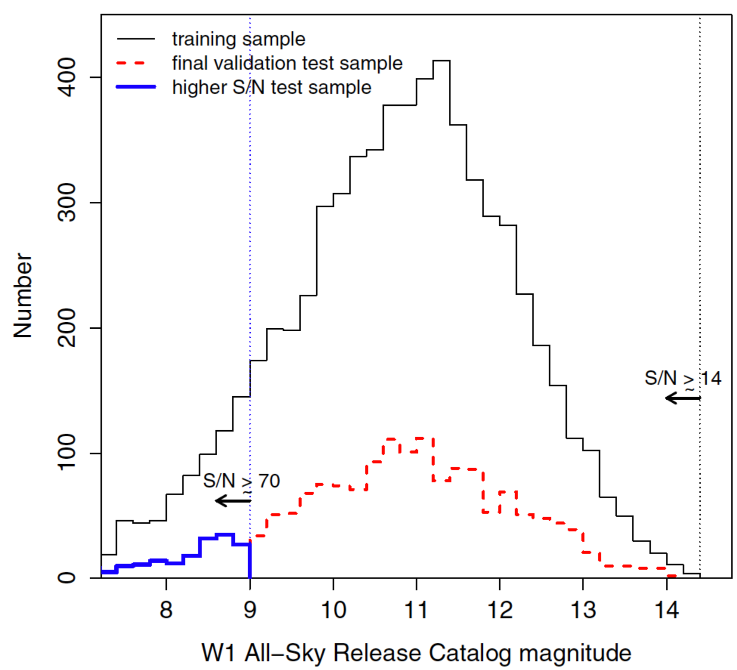
ML classification life-cycle



- **Green boxes:** what we’ve accomplished so far: proof-of-concept study for a subset of variables
- For details, see Masci et al. 2014, AJ, 148, 21

Training ("truth") sample

- First phase of study: explored classification performance for specific classes
- We focused on the 3 (most abundant) classes: **RR-Lyrae**, **Algols** (+β Lyrae), and **W Uma** variables
- First step was to construct a "training" (truth) sample of variables with known classifications.
  - selected from three optical variability surveys: GCVS, MACHO, ASAS.
- After matching to the WISE AllSky Catalog and other quality filtering, 8273 variables were retained
  - **Breakdown:** 1736 RR Lyrae, 3598 Algols, 2939 W Uma
  - more than 90% have an average single-exposure S/N > 20



(NEO)WISE light-curve features/metrics

- Extracted W1,W2 light-curves from the single-exposure source DB.
- Computed the following 7 features per light-curve => a point in our 7-D "feature space".

1. **Periods:** using periodograms computed using the Generalized Lomb-Scargle (GLS) method.

2. **Stetson-*L* variability index:** quantifies both degree of correlation between W1,W2 and the kurtosis of the time-collapsed magnitude distribution.

3. **Magnitude Ratio:** quantifies fraction of time a variable spends above or below its median mag:

$$0 \leq \frac{\max(m_i) - median(m_i)}{\max(m_i) - \min(m_i)} \leq 1$$

4. **Coefficient $|A_2|$** from Fourier decomposition (light-curve fitting). Quantifies light-curve shape.

$$m(t) = A_0 + \sum_{j=1}^{5} A_j \cos\left[2\pi j \Phi(t) + \phi_j\right], \qquad \Phi(t) = \frac{t - t_0}{P} - \mathrm{int}\left(\frac{t - t_0}{P}\right)$$

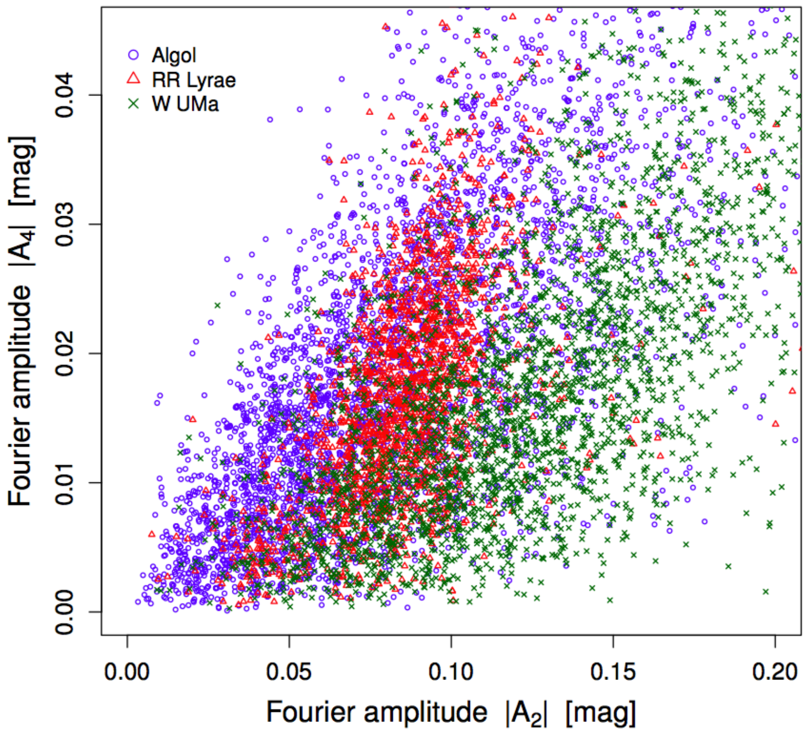
5. **Coefficient $|A_4|$**

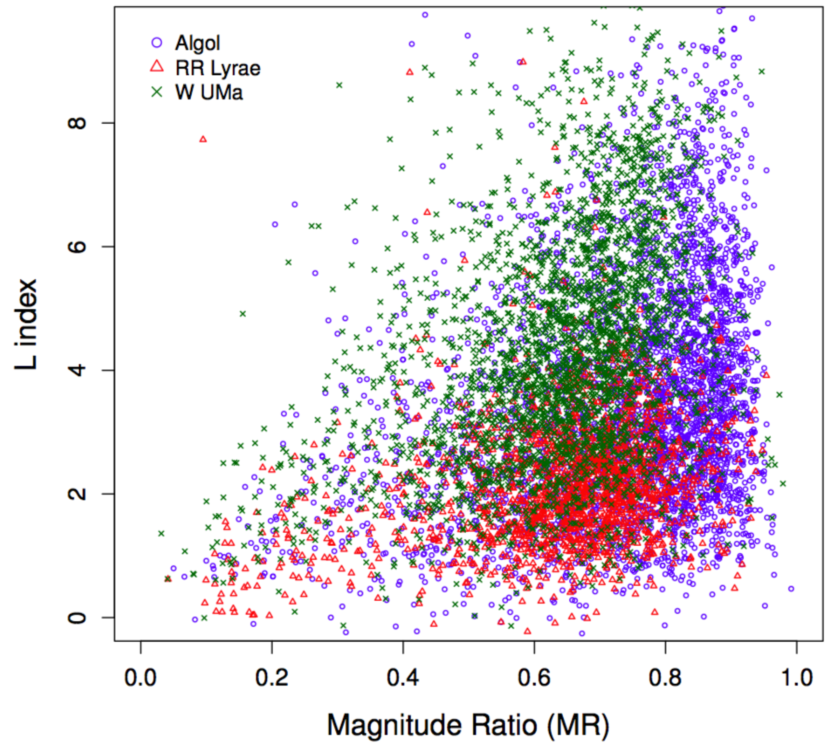
6. **Relative phase ϕ_{21}** from Fourier decomposition: $\phi_{21} = \phi_2 - 2\phi_1$

7. **Relative phase ϕ_{31}** from Fourier decomposition: $\phi_{31} = \phi_3 - 3\phi_1$

Some 2-D projections of 7-D feature space

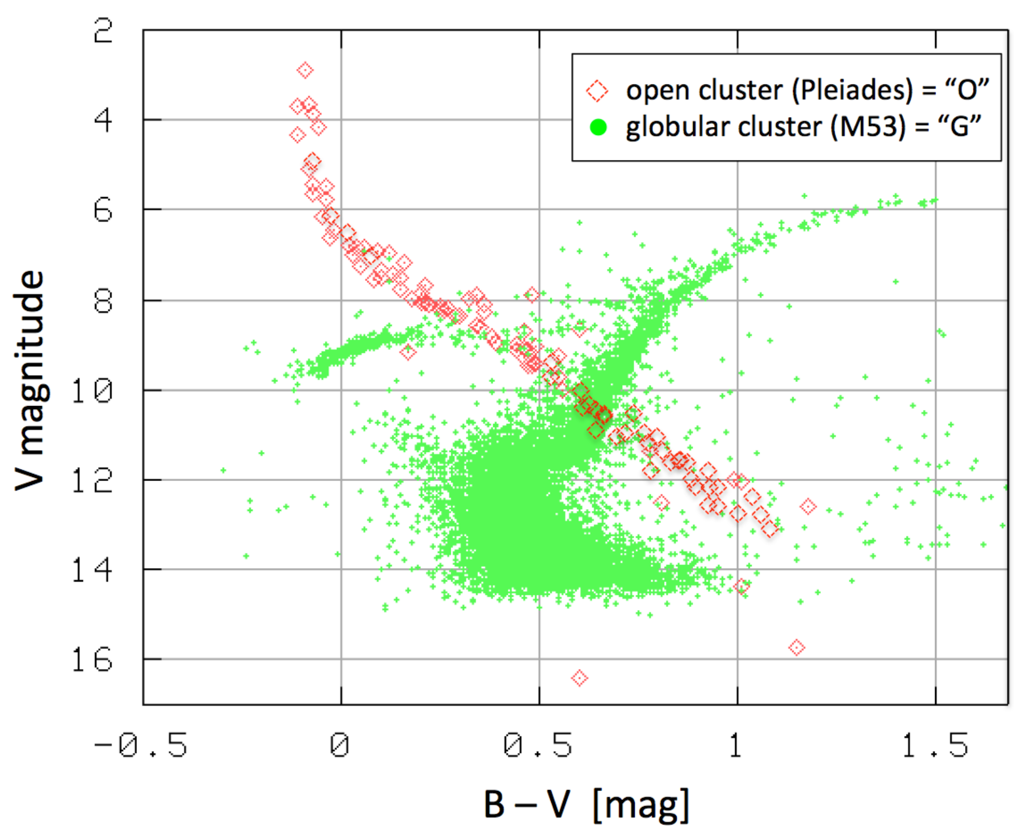
- Overlap (ambiguous) regions separable in higher dimensions. More features the better.
- Fourier decomposition works well in mid-IR. Just like in optical variability studies.





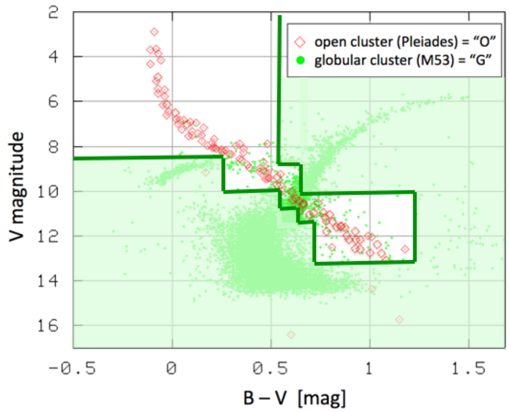
Classification using Random Forests™

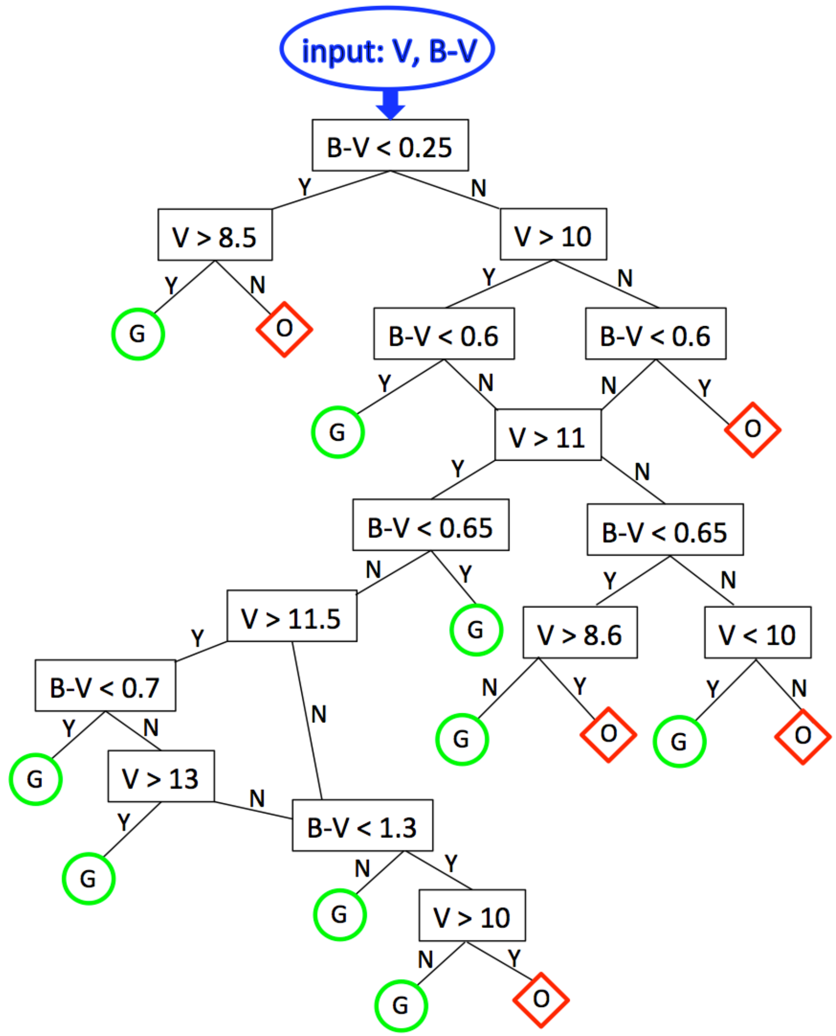
- Random Forests are based on "decision trees". Popularized by Breiman & Cutler ~ 2001.
- Here's an example of a classification problem involving 2-classes: stars in young open clusters (e.g., Pleiades) versus those in globular clusters, using only 2 features: color and magnitude
- A simple hypothetical example. In practical ML applications, can have >100 features (dimensions)



Classification using Random Forests™

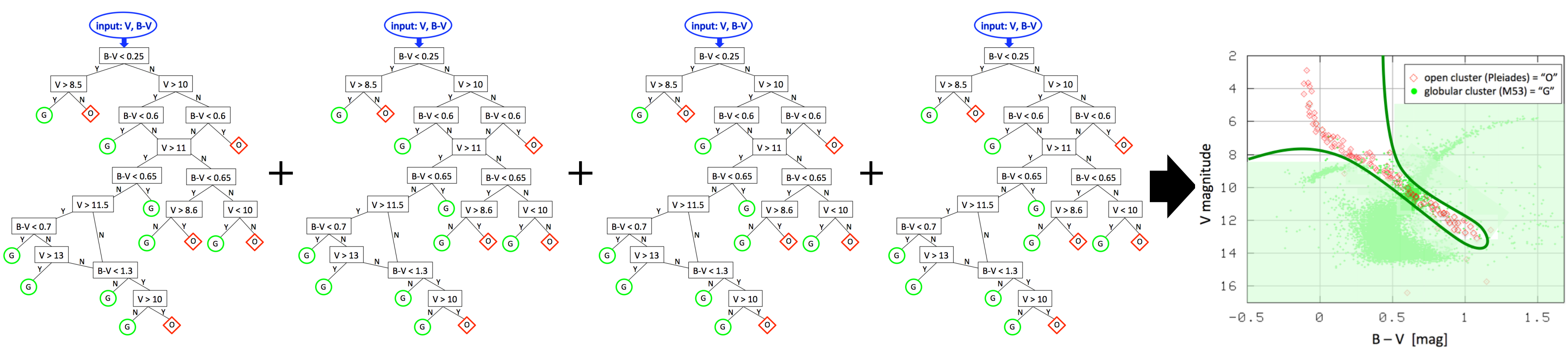
- Here's **one "decision tree"** for classifying "Open" vs "Globular" stars. Many other trees are possible
- Create a decision tree from pre-labeled cases to train a classifier, then use to predict future outcomes





Forest = lots of random trees

- However, the results from a single tree are prone to a high variance (i.e., sharp class boundaries)
- Instead, we grow lots of trees (e.g., >~ 1000) from:
  1. bootstrapped replicates of the training data-set (random sampling with replacement)
  2. randomly sample from set of N features at each "decision-node" of tree to find best split
- **The key is randomness!** Make the same number of (unbiased) mistakes everywhere in feature space
- Combine outcomes from all trees by averaging: boundaries become sharper; prediction error reduced
- Relative class probability of a future candidate = fraction of votes for each class across all trees
  - ➢ Can then threshold this probability to assign most probable class



Decorrelated random decision-trees (replicated here for simplicity)

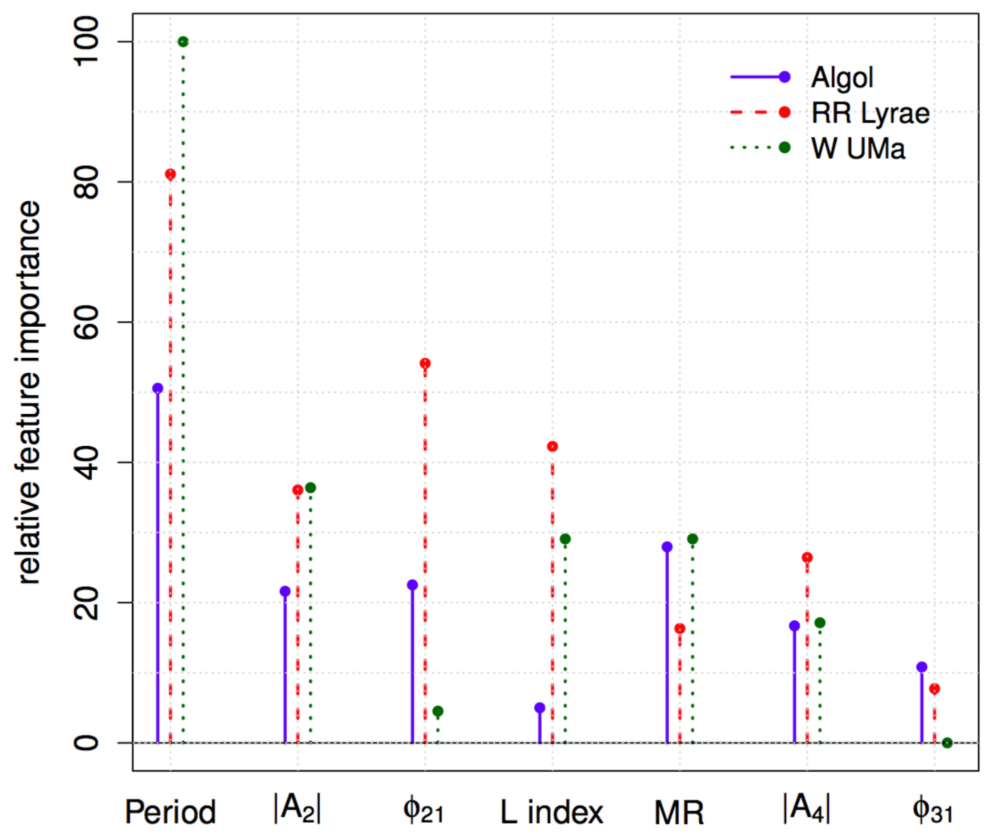
Why Random Forests?

- Intuitive & interpretable
- Can deal with complex non-linear patterns in N-D feature space where N can be >1000
- Can have more features than actual data points (objects to be classified)
- Training model “fitting” is parameter free (non-parametric) and distribution free
- Robust against over-fitting and outliers
- Relatively immune to irrelevant and correlated (redundant) features
- Can handle missing data for features
- Automatic optimal feature selection and node-splitting when creating decision trees
- Includes a framework to support active learning (iterative training & reclassification)
- Ability to assess the relative importance of each feature (more later)
- The following companies use some variant of RFs. Do a pretty good job at predicting what I like!

- Previous optical-variability classification studies successfully used RFs, e.g., Richards et al. 2011
- We explored other ML methods and Random Forests came out on top (see Masci et al. 2014)

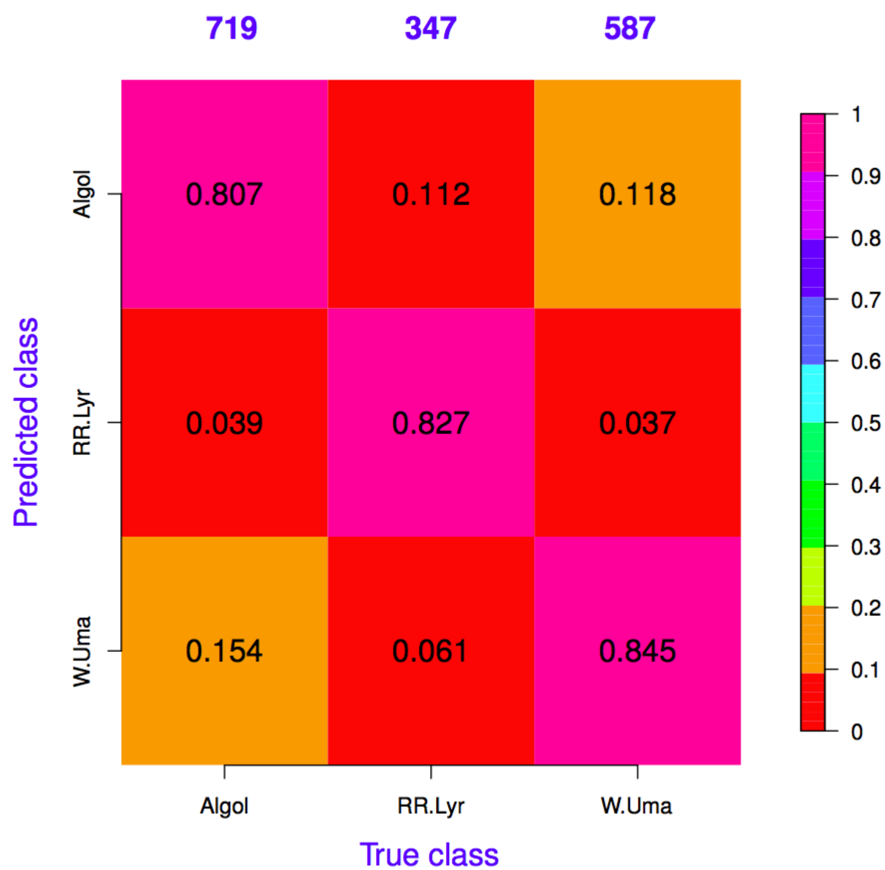
Feature Importance Evaluation

- Easy with Random Forests!
- Based on examining drop/increase in classification accuracy (ability to predict known outcomes) with and without specific feature(s) included during training



Classification performance for most common periodic variable stars

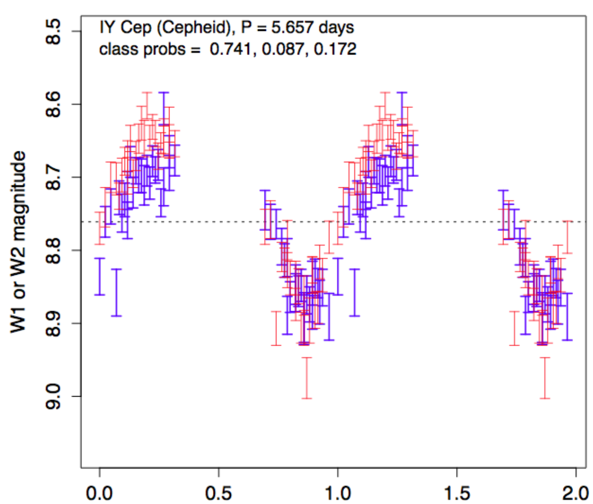
- **Confusion matrix:** summary of classification efficiency & purity (contamination) level of each class
- Obtain classification accuracies (efficiencies) of 80 – 85% across the three classes
- And purity levels of >~ 90% ("1 – false-positive-rate" from cross-class contamination)
- Consistent with previous automated classification studies for variable stars from optical surveys



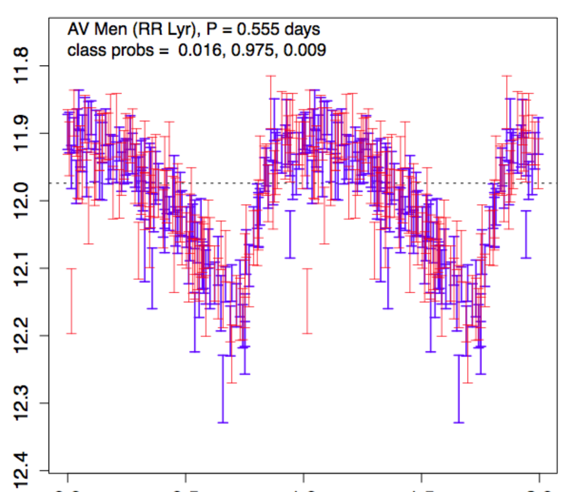
Example light-curve classifications

- Cepheids were not in our initial training sample due to low statistics; can only assign to 3 classes
- This is also at the period-recoverability limit (~6 days) given (NEO)WISE cadence and baseline
- Goal is to introduce more classes by identifying clusters in full feature space as statistics improve

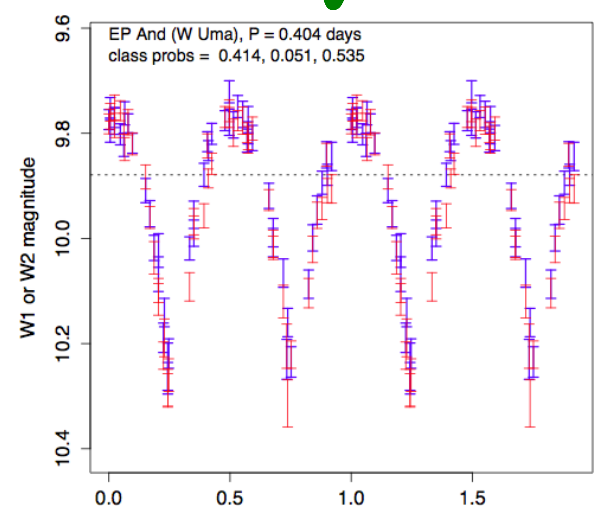
Truth: Cepheid variable
Classifier: Algol (!)



Truth: RR Lyrae
Classifier: RR Lyrae



Truth: W Uma
Classifier: W Uma



Phase (using estimated period)

All made possible with "R"

- Freely available at *http://cran.r-project.org*
- A powerful statistics software environment/toolbox. Not another blackbox. Lots of tutorials/examples
- **Warning:** R is addictive!

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2014-10-31, Pumpkin Helmet) R-3.1.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.

Summary and closing thoughts

- Explored feasibility of automatic classification of periodic variable stars from ~1yr of (NEO)WISE
  - All looks very promising for at least the most common variables
  - Consistent with (and sometimes exceeding) performance of previous optical surveys
  - Provides a crucial first step towards constructing the WISE Variable Source Catalog (WVSC)

- **Challenges:**
  - "Feature engineering" step – which features best separate known classes?
  - Validation of ML classifier – only as good as the data it was trained on – is it generic enough?

- **Near-future:**
  - Narrow down list of variable candidates that "best" lend themselves to classification
  - Retrain using AllWISE Multi-Epoch Photometry (MEP) DB, then construct WVSC

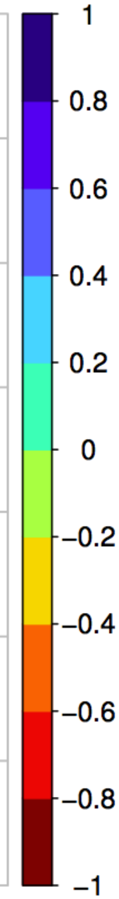
- Encourage everyone to dabble in machine learning when working with large datasets with lots of metrics (e.g., WISE Source Catalogs)
  - A rich software-base is *freely* available.
  - Power of probabilistic classification: results are more open to scientific interpretation.

Back up slides

Correlation Matrix

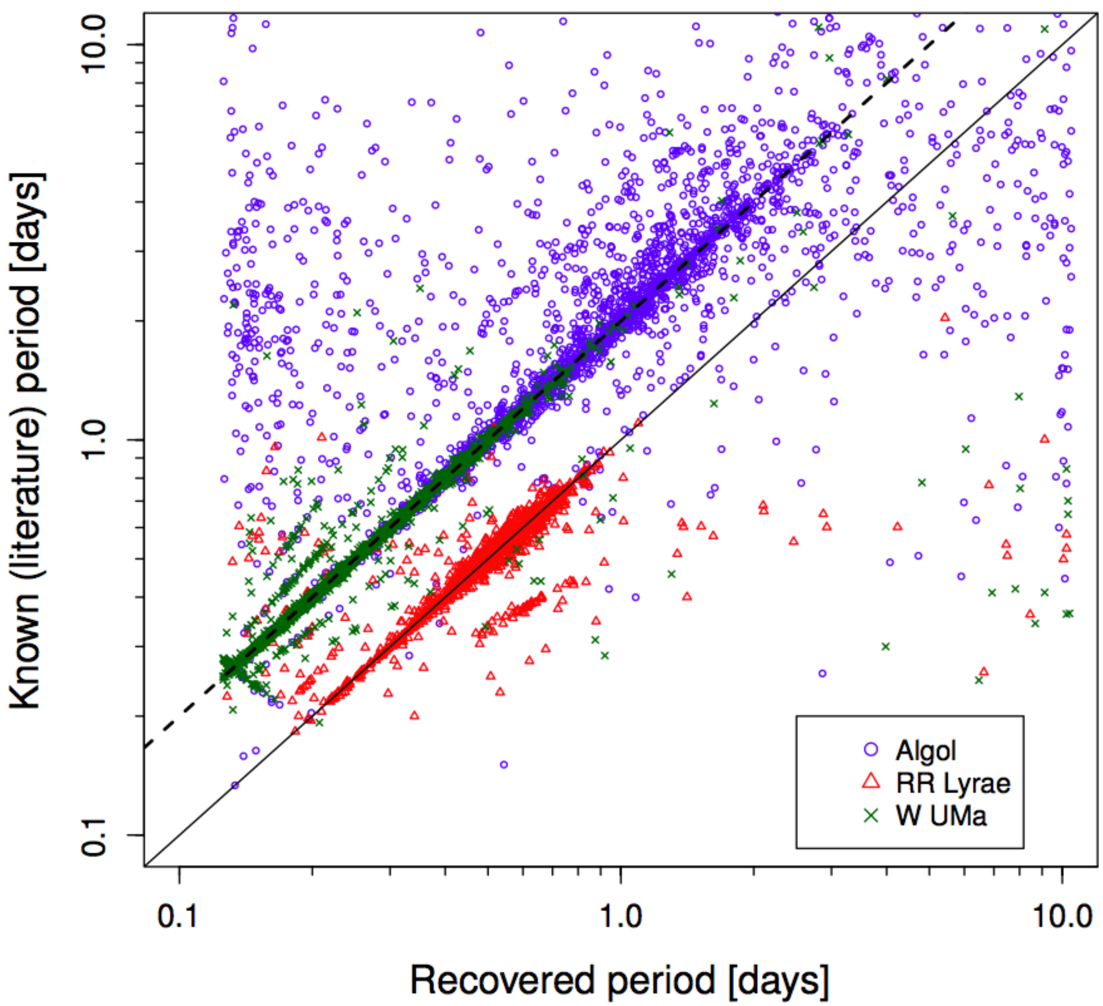
- Check degree of correlation ("redundancy") amongst 7 features for possible feature reduction
- Random Forests however are relatively immune to moderately correlated features

	$\|A_2\|$	$\|A_4\|$	ϕ_{21}	ϕ_{31}	Period	MR	L index
$\|A_2\|$	1	0.62	0.06	−0.01	−0.29	0.02	0.71
$\|A_4\|$	0.62	1	0.01	0	−0.08	0.21	0.5
ϕ_{21}	0.06	0.01	1	0.02	0.05	0.04	0.13
ϕ_{31}	−0.01	0	0.02	1	0	−0.01	0
Period	−0.29	−0.08	0.05	0	1	0.01	−0.09
MR	0.02	0.21	0.04	−0.01	0.01	1	0.27
L index	0.71	0.5	0.13	0	−0.09	0.27	1



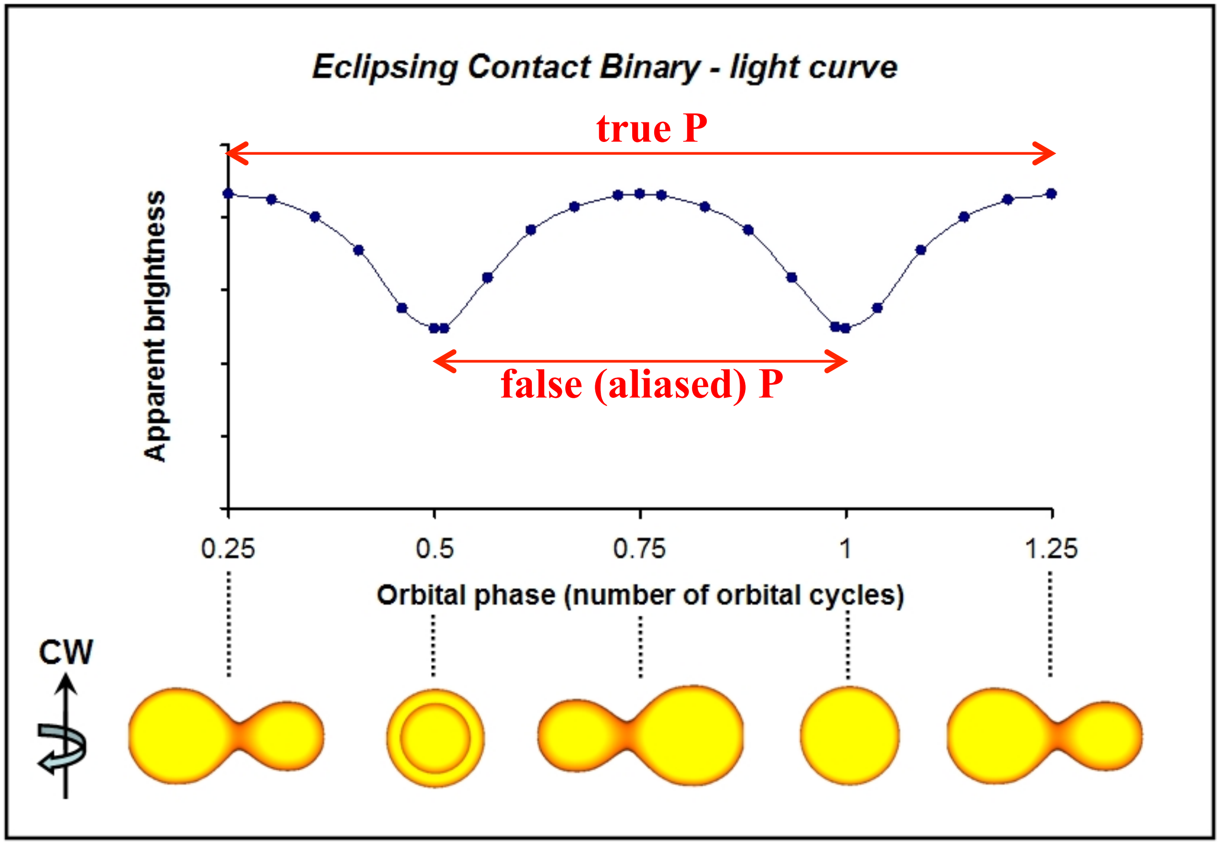
Period recoverability from W1 light curves

New periods from Generalized Lomb Scargle (GLS) periodograms versus literature:



Biased periods for eclipsing binaries?

- GLS sometimes returns half the true (literature) period, assuming latter is correct
- Typically occurs when primary and secondary eclipses in light-curve have similar depths
- Need other features to mitigate this aliasing/ambiguity, i.e., to classify into different types



Random Forests: the originators

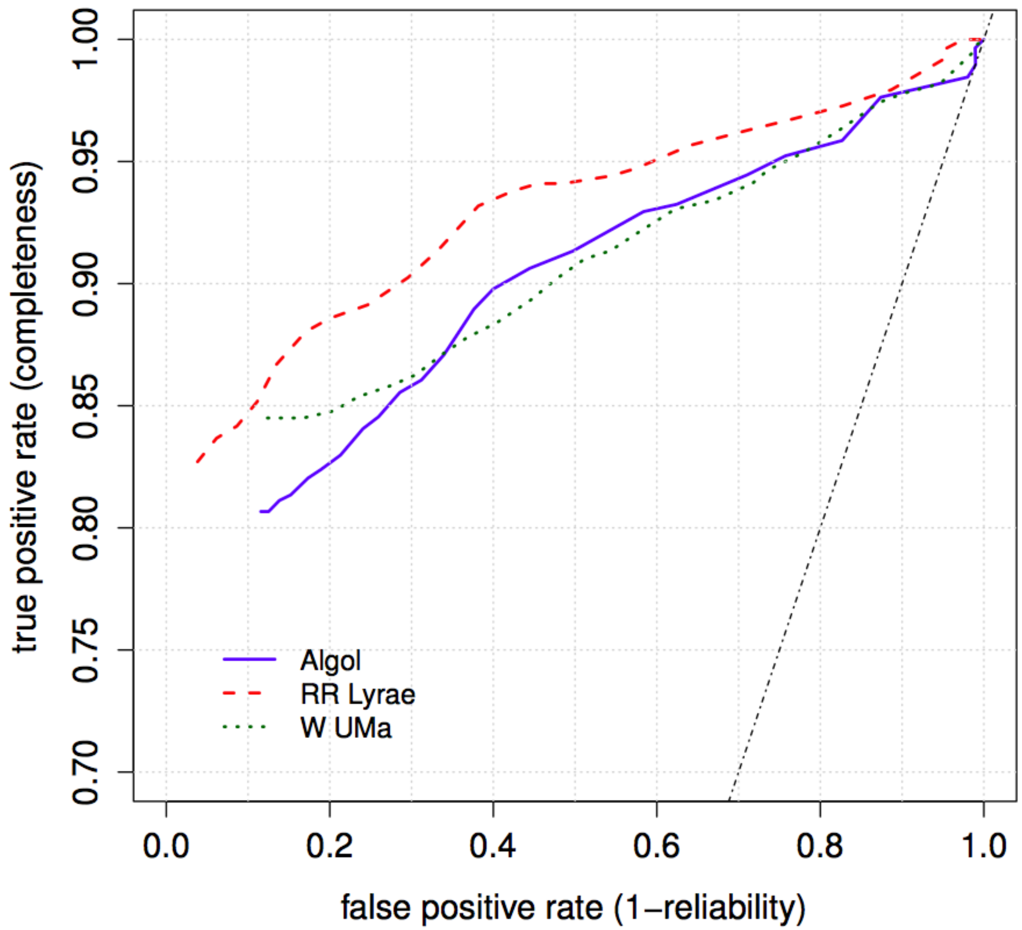
- Classification and regression trees (CART) methods have been around since mid 1980s
- The averaging results from lots of random decision trees is known as bagging (bootstrap aggregation)
- Idea popularized by Leo Breiman & Adele Cutler in ~ 2001 (UC Berkeley)

Leo Breiman
1928 – 2005

Adele Cutler
now at Utah State

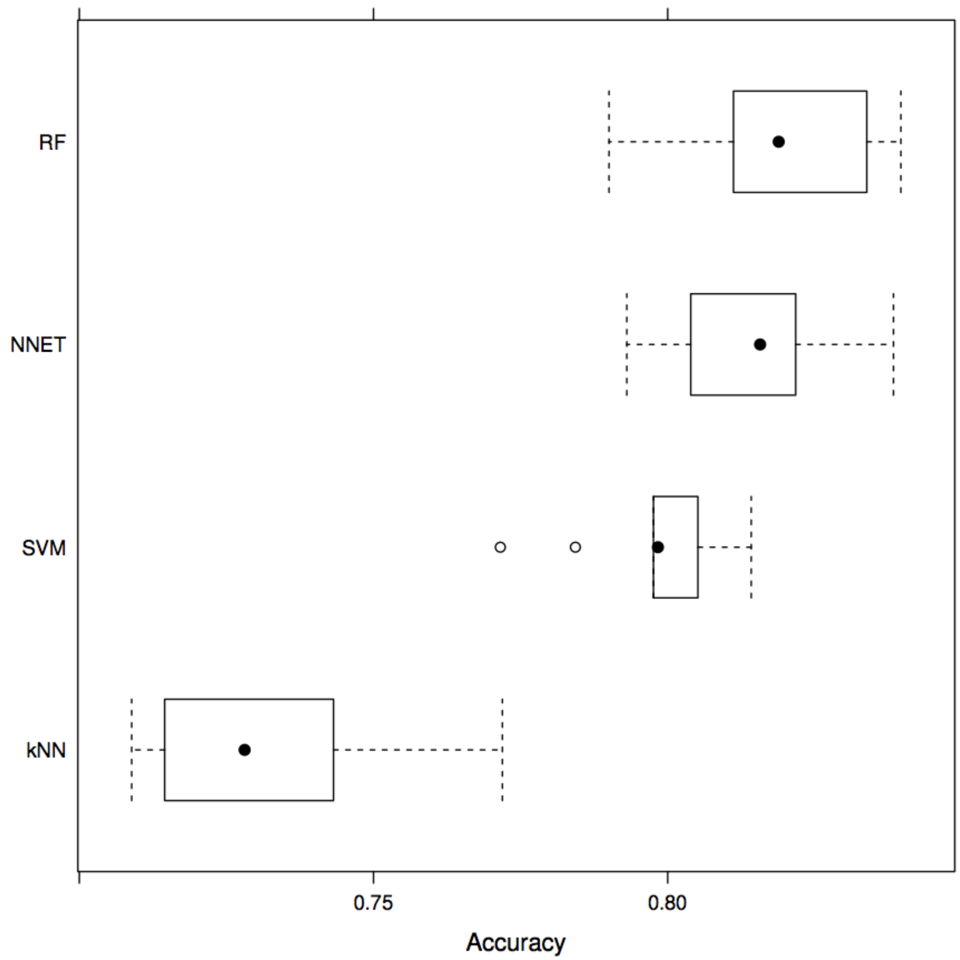
ROC curves
(Receiver Operating Characteristic)

- **aka:** "Completeness" versus "1 – Reliability"
- Thresholded on classification probability for each class (increases from right to left)



Performance of other classifiers?

- Also explored Support Vector Machines (SVM), Neural Networks (NNET), *k*-Nearest Neighbors (kNN) and compared to Random Forests (RF)
- RFs have the edge! Masci et al. 2014, AJ, 148, 21.



Performance of other classifiers?

Performance metrics (Masci et al. 2014, AJ, 148, 21)

Table 1. Classifier comparison

Method	Med. Accuracy [a]	Max. Accuracy [a]	Training time [b] (sec)	Pred. time [c] (sec)	p-value [d] (%)
NNET	0.815	0.830	375.32	0.78	99.99
kNN	0.728	0.772	6.42	0.55	< 0.01
RF	0.819	0.840	86.75	0.77	$\cdots$
SVM	0.798	0.814	75.66	1.77	3.11

[a] Median and maximum achieved accuracies from a 10-fold cross-validation on the training sample.

[b] Average runtime to fit training model using parallel processing on a 12-core 2.4 GHz/core Macintosh with 60 GB of RAM.

[c] Average runtime to predict classes and compute probabilities for 1653 feature vectors in our final validation *test sample* (Section 5.3).

[d] Probability value for H0: difference in mean accuracy relative to RF is zero.

Recommended book on ML with R