Overcoming Sample Selection Bias in Variable Star Classification

Joseph Richards

UC Berkeley
Department of Astronomy
Department of Statistics

jwrichar@stat.berkeley.edu

GREAT Workshop on Astrostatistics and Data Mining in Astronomical Databases

June 1, 2011

Center for Time-Domain Informatics

UC Berkeley (UCB):

Faculty/Staff

Josh Bloom, Dan Starr (Astro), John Rice, Noureddine El Karoui (Stats), Martin Wainwright, Masoud Nikravesh (CS)

Post-Docs

Dovi Poznanski, Brad Cenko, Nat Butler, Berian James, JWR

Grad Students

Dan Perley, Adam Miller, Adam Morgan, Chris Klein, **James Long**, Tamara Broderick, Sahand Negahban, John Brewer, Henrik Brink

Undergrads

Maxime Rischard, Justin Higgins, Rachel Kennedy, Jason Chu, Arien Crellin-Quick

Lawrence Berkeley National Laboratory (LBNL):

Peter Nugent, David Schlegel, Nic Ross, Horst Simon

Visit our website: `http://cftd.info/`

CDI Seminar Series

NSF-sponsored Cyber-Enabled Discovery and Innovation (CDI) seminars on interface of Statistics, Astronomy, and CS

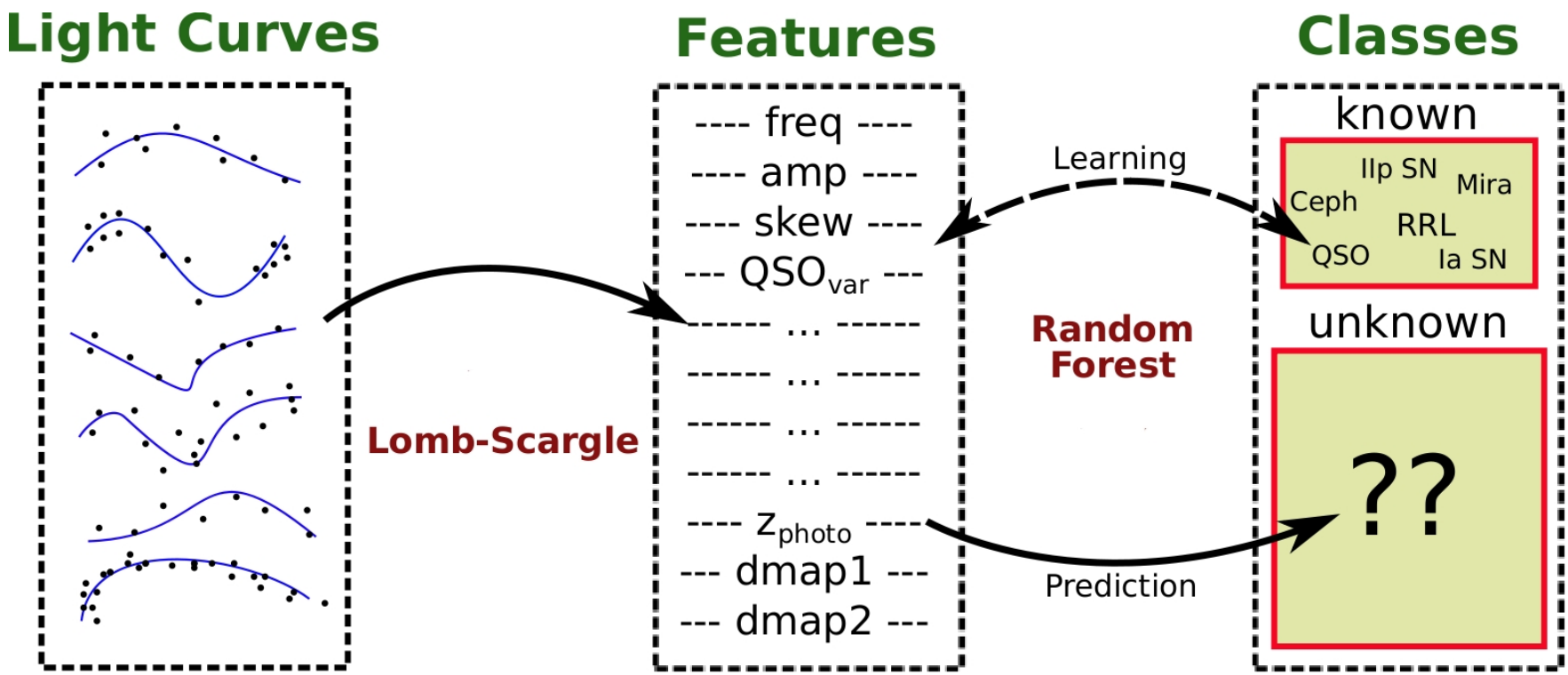
All talks hosted by the Center for Time-Domain Informatics: http://cftd.info/

Past Speakers:

- Andy Connolly, UW Astro
- Brad Efron, Stanford Stats
- Jogesh Babu, Penn St. Stats
- Jim Berger, Duke Stats
- Chad Schafer, CMU Stats
- David van Dyk, UCI Stats
- Tamas Budavari, JHU Astro

Motivation

A road map for light curve classification:

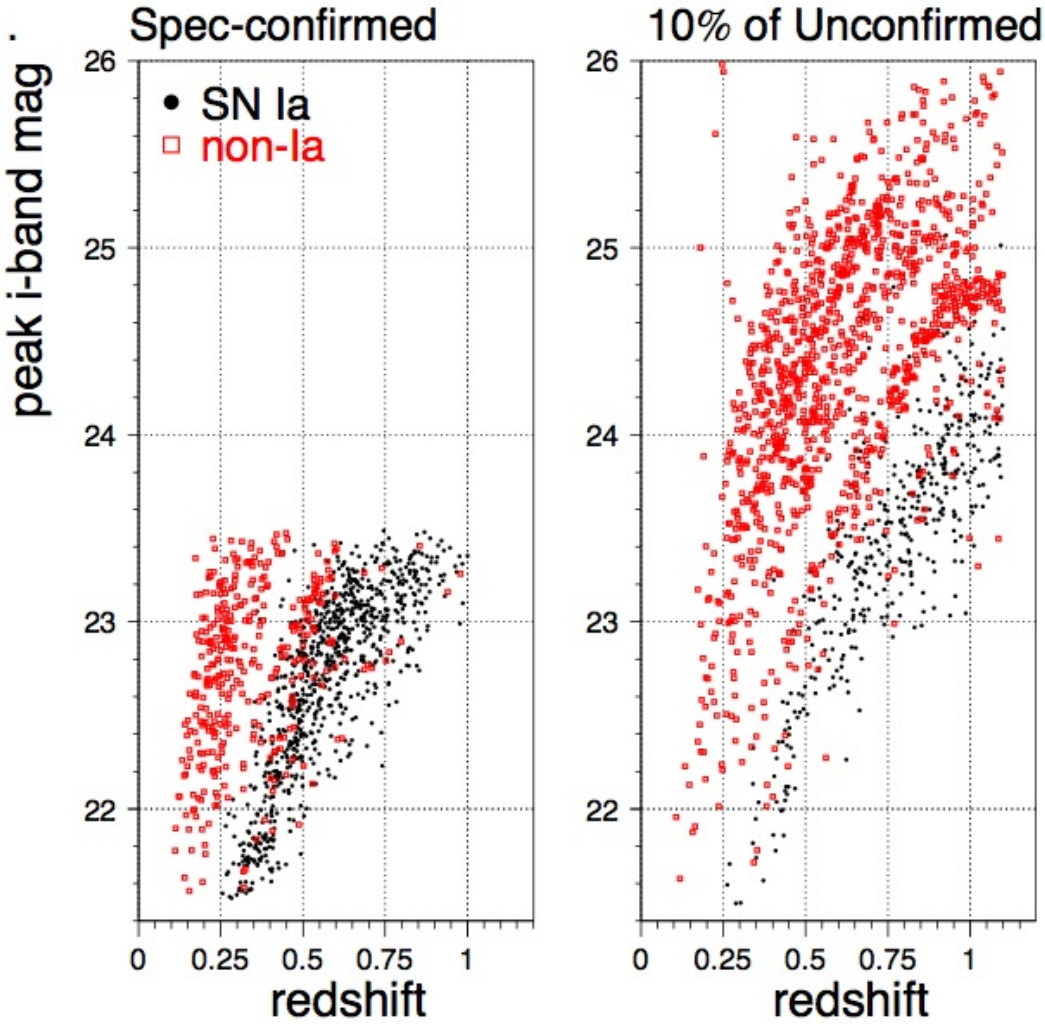


See: Richards et al. (2011) arXiv:1101.1959

Bloom & Richards (2011) arXiv:1104.3142

Sample Selection Bias

In astronomical problems, the training (labeled) and testing (unlabeled) sets are often generated from different distributions.



Left: Training set
Right: Testing set

This problem is referred to as Sample Selection Bias or Covariate Shift.

SN Challenge Data

Kessler et al. (2010)

arXiv:1008.1024

Sample Selection Bias

For SN Ia typing, it is **better to use deeper spectroscopic training samples**, even though they produce data from fewer SNe

$S_{m,25}$ - 25th mag-limited spec survey is optimal (23.5th mag was used in SN Challenge)

From Richards et al. (2011) arXiv:1103.6034

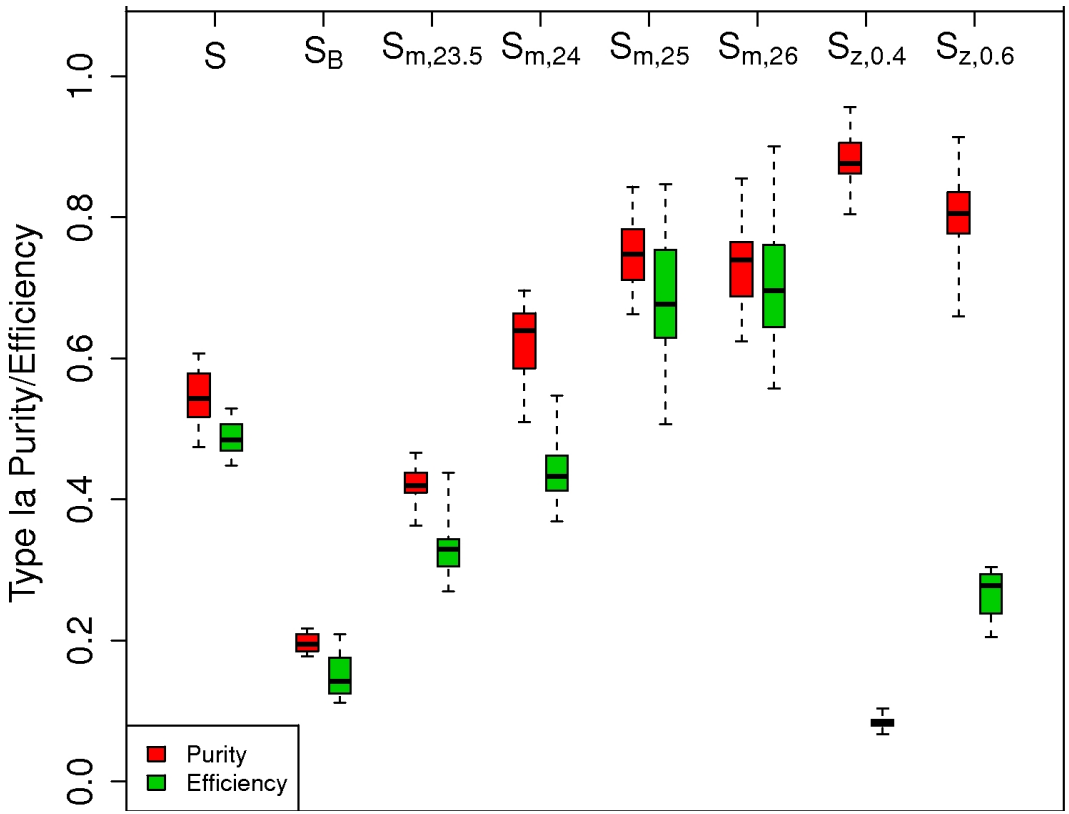


Figure: Type Ia SN Purity and Efficiency of Random Forest classifier on SN Challenge testing data

Sample Selection Bias in Variable Star Classification

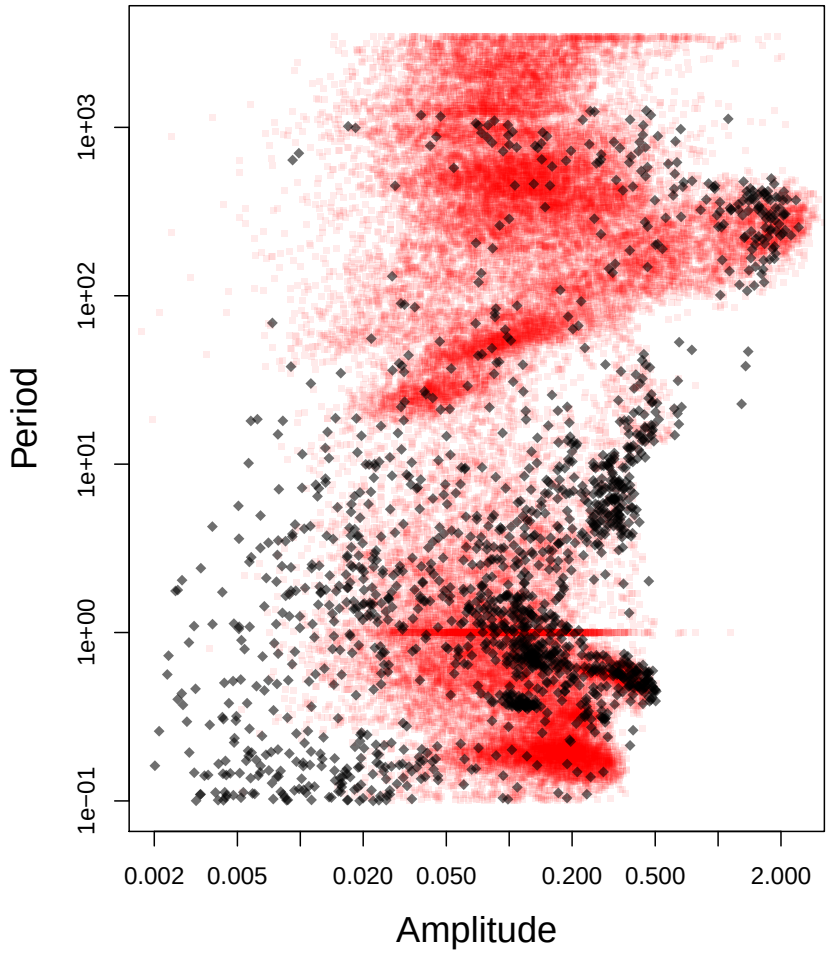
with Dan Starr, Adam Miller, Nat Butler, James Long, John Rice, Josh Bloom (UC Berkeley), Henrik Brink & Berian James (DARK)

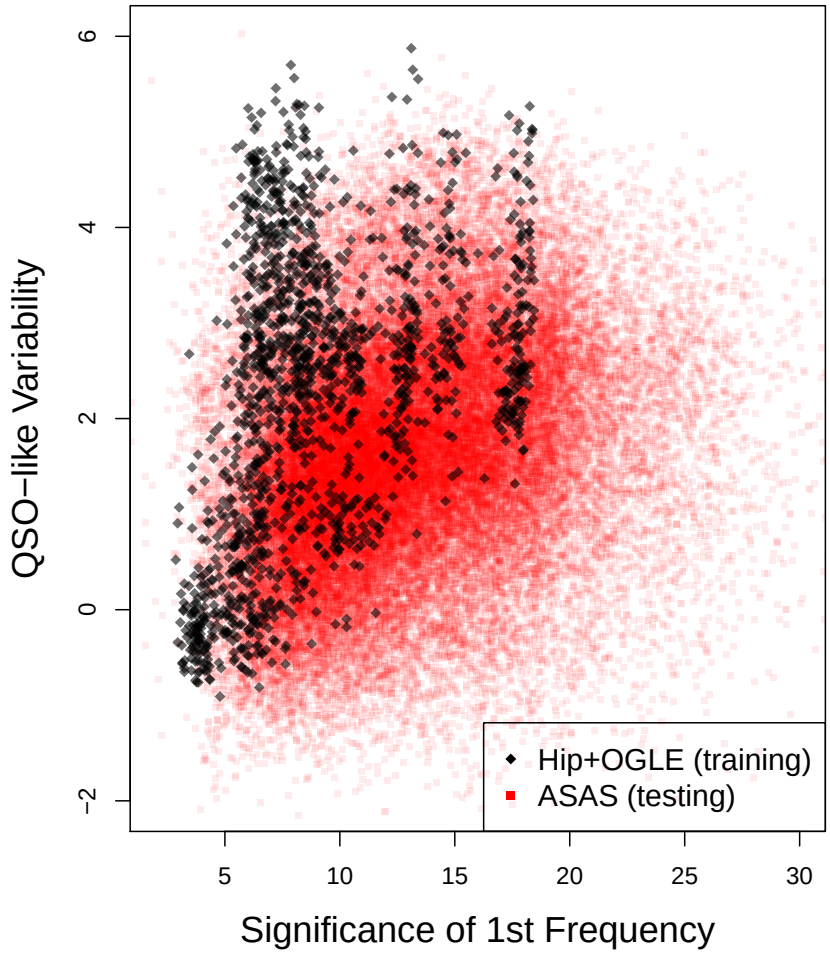
Richards et al. (2011), in prep.

Sample Selection Bias in VarStar Classification

Black: Training set (OGLE+Hipparcos, see Debosscher et al. 2007)

Red: Testing set (All Sky Automated Survey, ASAS; Pojmanski 2002)





Sample Selection Bias in VarStar Classification

Training sets in variable star studies are biased:

1. Populations of well-studied objects are inherently biased toward brighter/closer sources with better quality data **See James Long's talk!**
2. Available training data are typically from older, lower quality detectors
3. Each survey has different characteristics, aims, cadences...
4. Training data are often generated from idealized models

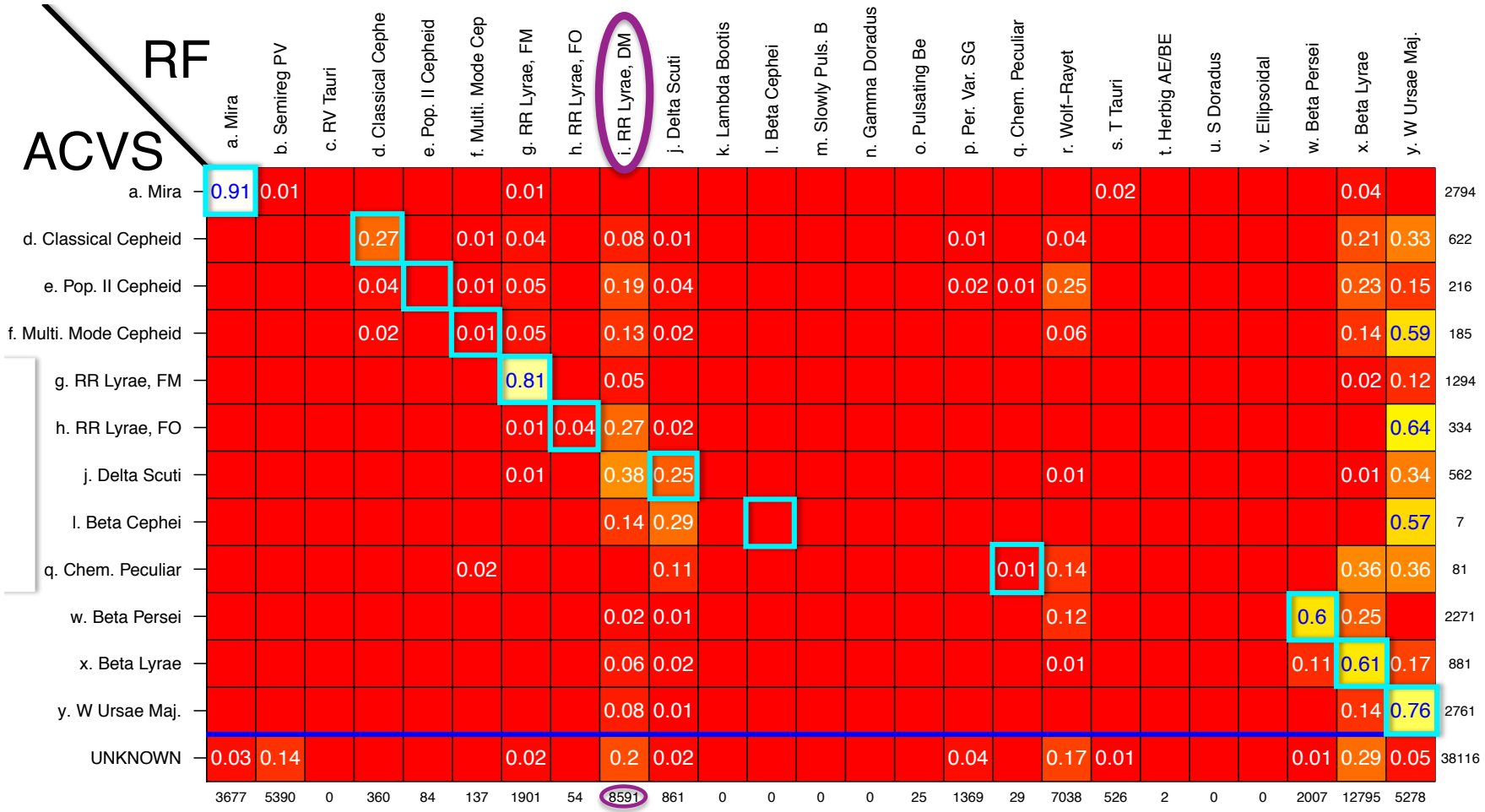
This can cause significant problems for off-the-shelf supervised methods:

1. Poor model selection – risk minimization (e.g., by cross-validation) is performed with respect to $\mathbf{P}_{\mathrm{Train}}(\mathbf{x}, y)$
2. Regions of feature space ignored by the training data – catastrophically bad extrapolation

Sample Selection Bias in VarStar Classification

Example: ASAS varstar classification (50,124 stars in ACVS)

Results of off-the-shelf Random Forest classifier:



Some Methods

Methods: Importance Weighting (IW)

Idea: Choose classifier that minimizes statistical risk over distribution of the *testing* set.

Methods: Importance Weighting (IW)

Idea: Choose classifier that minimizes statistical risk over distribution of the *testing* set.

Use importance weights on *training* set:

$$w_i = \frac{\mathbf{P}_{\text{Test}}(\mathbf{x}_i, y_i)}{\mathbf{P}_{\text{Train}}(\mathbf{x}_i, y_i)} = \frac{\mathbf{P}_{\text{Test}}(\mathbf{x}_i)\mathbf{P}_{\text{Test}}(y_i|\mathbf{x}_i)}{\mathbf{P}_{\text{Train}}(\mathbf{x}_i)\mathbf{P}_{\text{Train}}(y_i|\mathbf{x}_i)} = \frac{\mathbf{P}_{\text{Test}}(\mathbf{x}_i)}{\mathbf{P}_{\text{Train}}(\mathbf{x}_i)}$$

Issues:

1. Difficult to estimate high-dimensional feature densities.
2. IW is asymptotically sub-optimal when the statistical model is correctly specified (Shimodaira 2000)
3. Requires the support of the testing distribution be a subset of the support of the training distribution

Methods: Co-training (CT)

Idea: Iteratively add to the training set the most confidently classified testing data

Methods: Co-training (CT)

Idea: Iteratively add to the training set the most confidently classified testing data

CT approach (Blum & Mitchell 1998)

Iterate until all data are in training set:

1. Build two separate classifiers, h_1 & h_2 on disjoint feature sets $\mathbf{x}_1$ & $\mathbf{x}_2$
2. Add the most confidently classified testing instances to the training set of the other classifier

Final classifier: $p(y|\mathbf{x}) = h_1(y|\mathbf{x}_1)h_2(y|\mathbf{x}_2)$

Self-training (ST) performs iterations on a single classifier

Drawback: CT & ST are greedy: dominant classes in the training data gain undue influence

Methods: Active Learning (AL)

Idea: Manually label the testing data that would most help future iterations of the classifier

Methods: Active Learning (AL)

Idea: Manually label the testing data that would most help future iterations of the classifier

Key: In astronomy, we often have the ability to selectively follow up on sources:

- Spectroscopic study
- Query other databases; cross-match
- "Look at" the data

On each AL iteration, select a batch of objects from the entire testing set for manual labeling via a **query function** (pool-based, batch-mode AL)

Heuristic: Query data in regions of feature space that are densely populated with testing data and sparsely populated with training data.

Methods: Active Learning (AL)

Proposed RF AL query functions; Richards et al. (2011)

AL1. Select testing data point ($\mathbf{x}' \in \mathcal{U}$) that is **most under-sampled by the training data** ($\mathcal{L}$):

$$S_1(\mathbf{x}') = \frac{\mathbf{P}_{\text{Test}}(\mathbf{x}')}{\mathbf{P}_{\text{Train}}(\mathbf{x}')} \approx \frac{\sum_{\mathbf{x}\in\mathcal{U}} \rho(\mathbf{x}', \mathbf{x})/N_{\text{Test}}}{\sum_{\mathbf{z}\in\mathcal{L}} \rho(\mathbf{x}', \mathbf{z})/N_{\text{Train}}} \tag{1}$$

AL2. Select testing data point that **maximizes the total change in the RF probabilities over the testing data**:

$$S_2(\mathbf{x}') = \frac{\sum_{\mathbf{x}\in\mathcal{U}} \rho(\mathbf{x}', \mathbf{x})(1 - \max_y \widehat{P}_{\text{RF}}(y|\mathbf{x}))}{\sum_{\mathbf{z}\in\mathcal{L}} \rho(\mathbf{x}', \mathbf{z}) + 1} \tag{2}$$

Here, $\widehat{P}_{\text{RF}}(y|\mathbf{x})$ is the estimated RF prob and $\rho(\mathbf{x}', \mathbf{x})$ is the RF proximity measure

Experiment: OGLE+Hipparcos

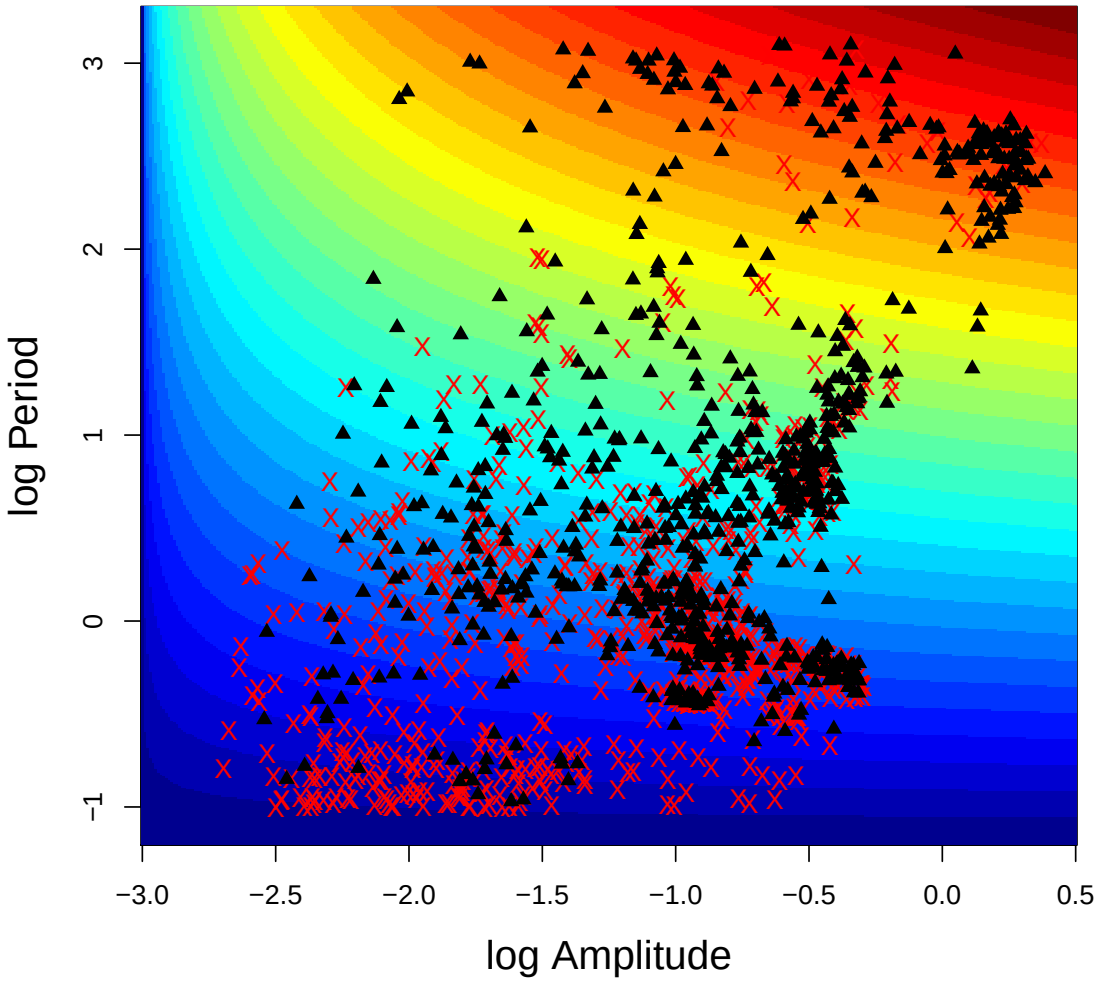
Experiment

1542 OGLE+Hip sources (25 classes) randomly split into:
Training (black ▲)
Testing (red x)

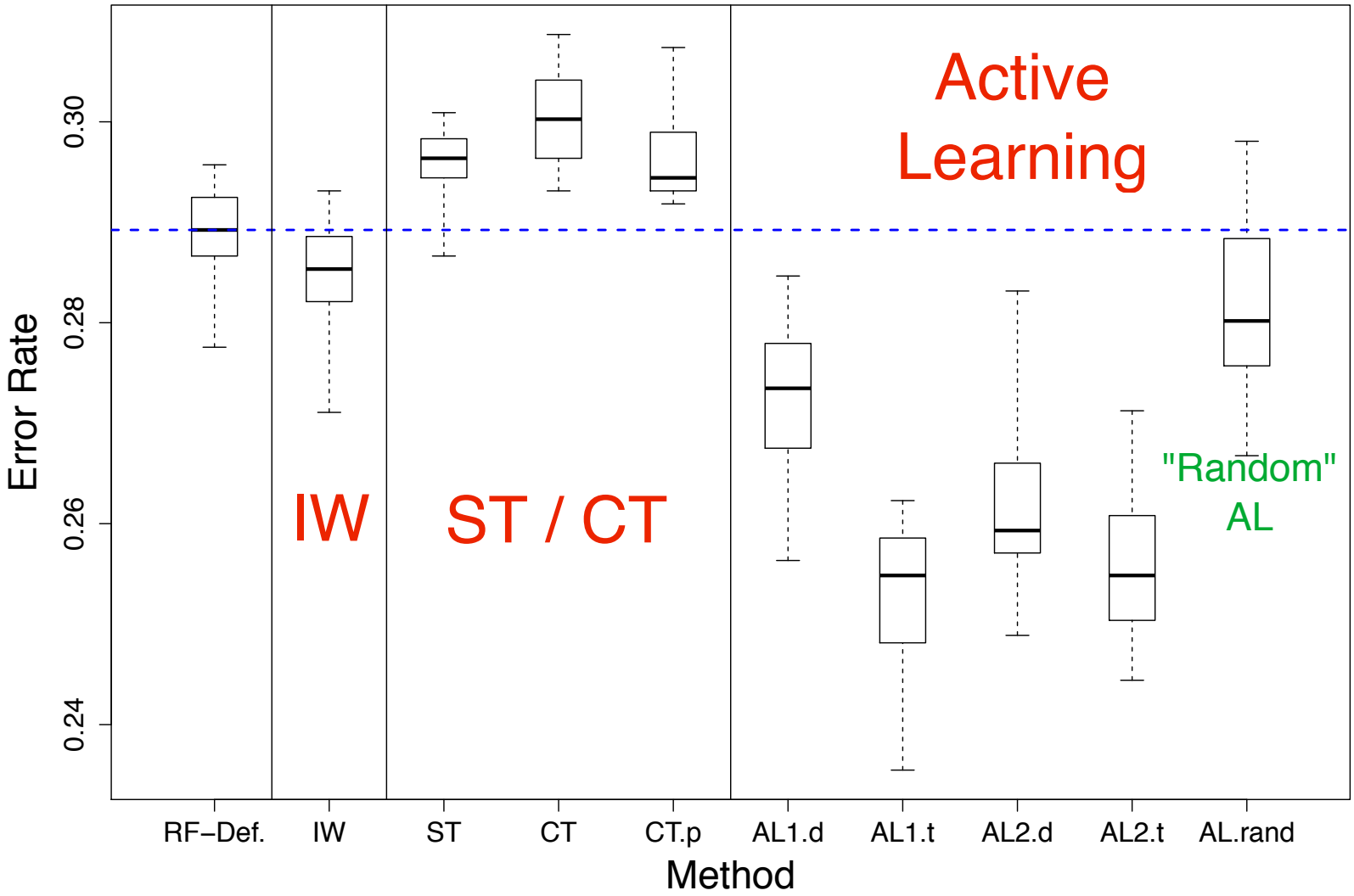
Selection function:

$$\Gamma \propto \log(P)\log(A)^{1/4}$$

We compare methods based on error rate on testing data



Experiment: Results



Note: AL routines evaluated only on non-selected testing data

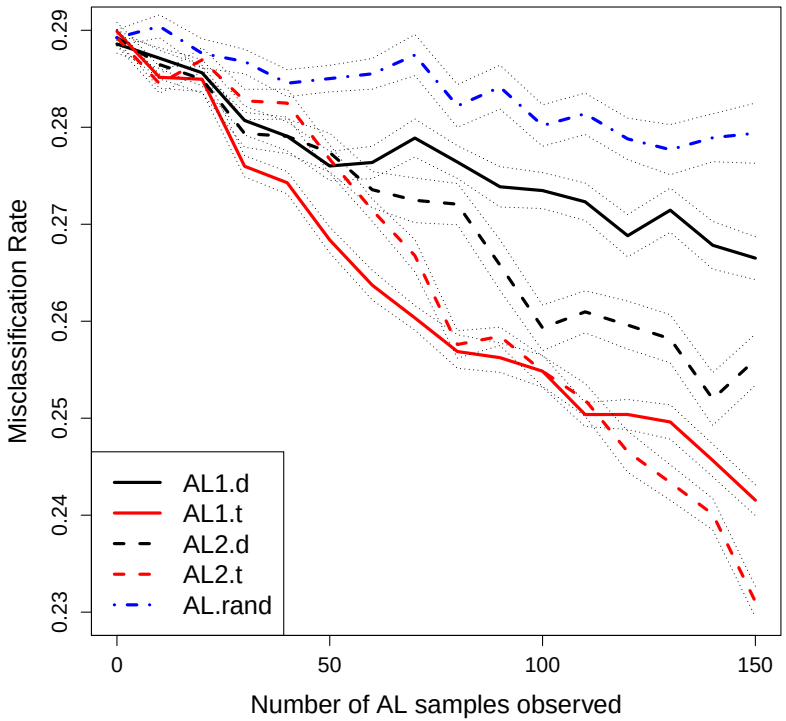
Experiment: Results

AL performs better than the default RF for under-represented classes

Class	N_{Train}	N_{Test}	RF	IW	ST	CT	AL1	AL2
All	771	771	28.9	28.5	29.6	30.0	27.3	25.9
Delta Scuti	25	89	15.7	15.7	15.7	15.7	15.4	15.6
W Ursa Maj.	16	43	40.7	36.0	51.2	60.5	**27.0**	**27.1**
Mira	121	23	8.7	8.7	8.7	8.7	9.1	8.7
Class. Cepheid	122	68	2.9	2.9	1.5	1.5	3.1	1.6

Top: % error on testing subsets
Right: AL testing error vs. # samples

Error rates decrease more quickly using AL1 (solid) and AL2 (dashed) than random selection (dot-dashed)



Application: ASAS

Application: ASAS

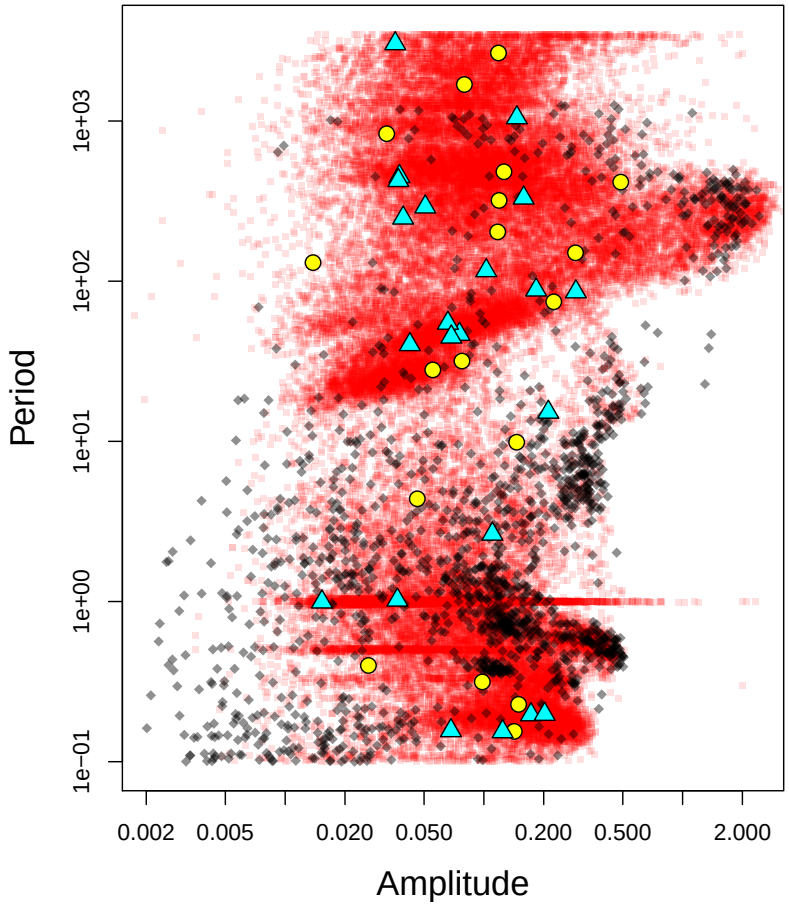
We use AL to classify all 50124 sources in the ACVS catalog

Training set: 1524 well-understood stars (in 25 classes) from OGLE+Hipparcos

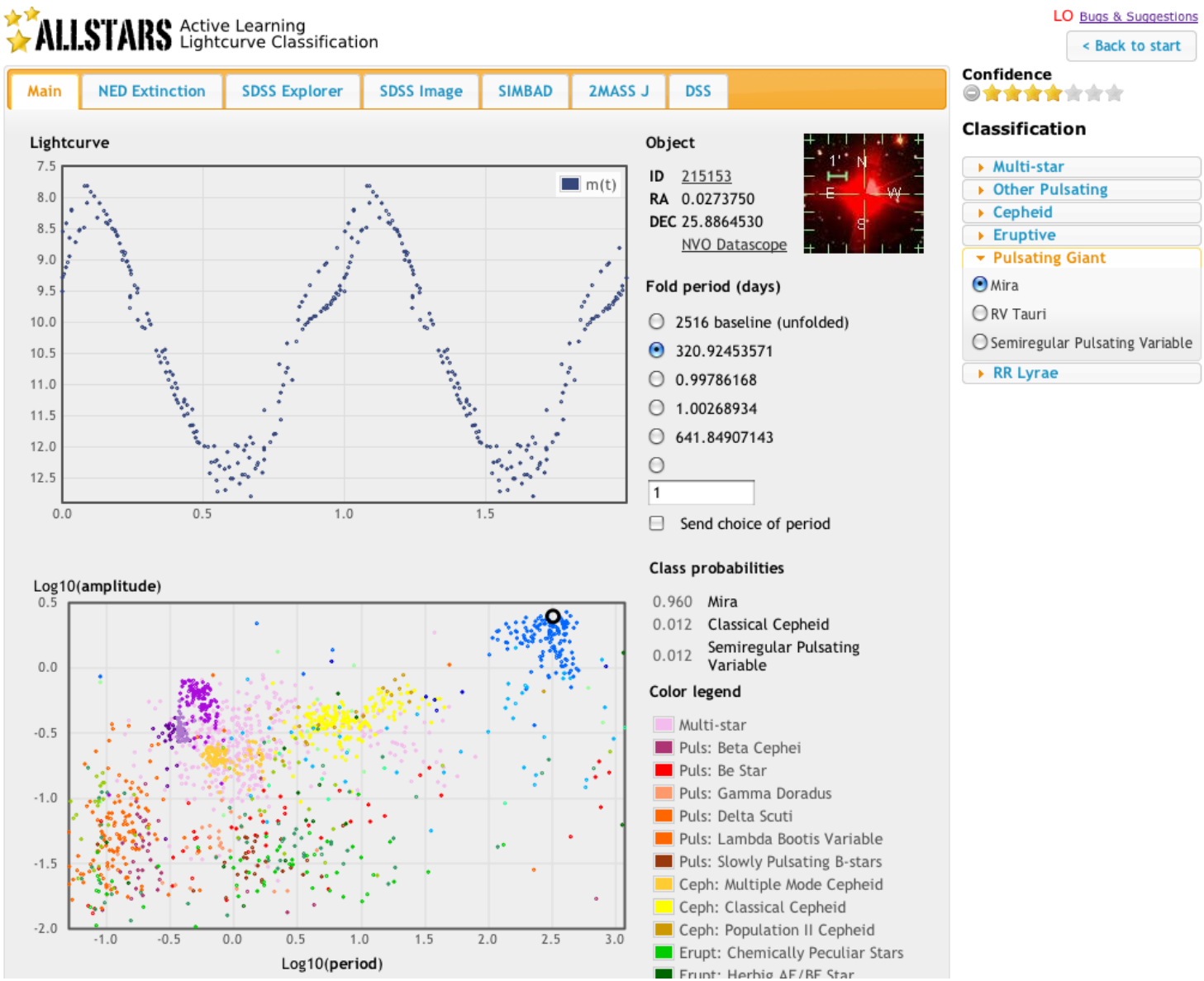
Perform 9 AL iterations of 50 sources each selected by sampling from S

Incorporate labeling "cost": $S(\mathbf{x}) = S_2(\mathbf{x})(1 - C(\mathbf{x}))$

11 users labeled sources. Use `IEThresh` crowd-sourcing of Donmez et al. (2009)

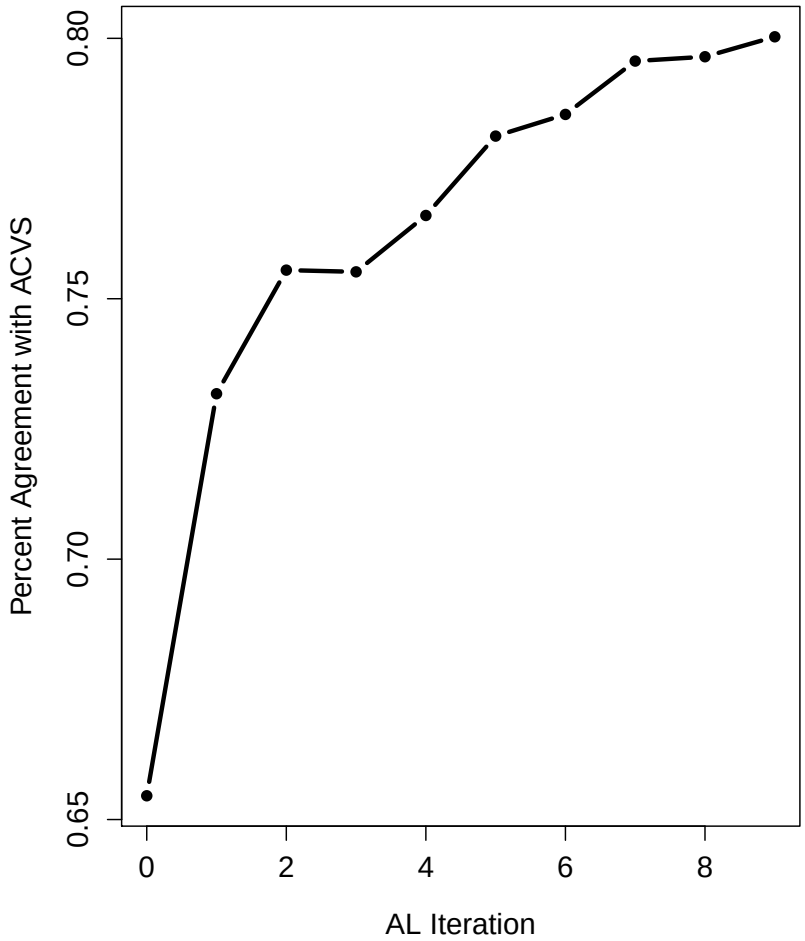


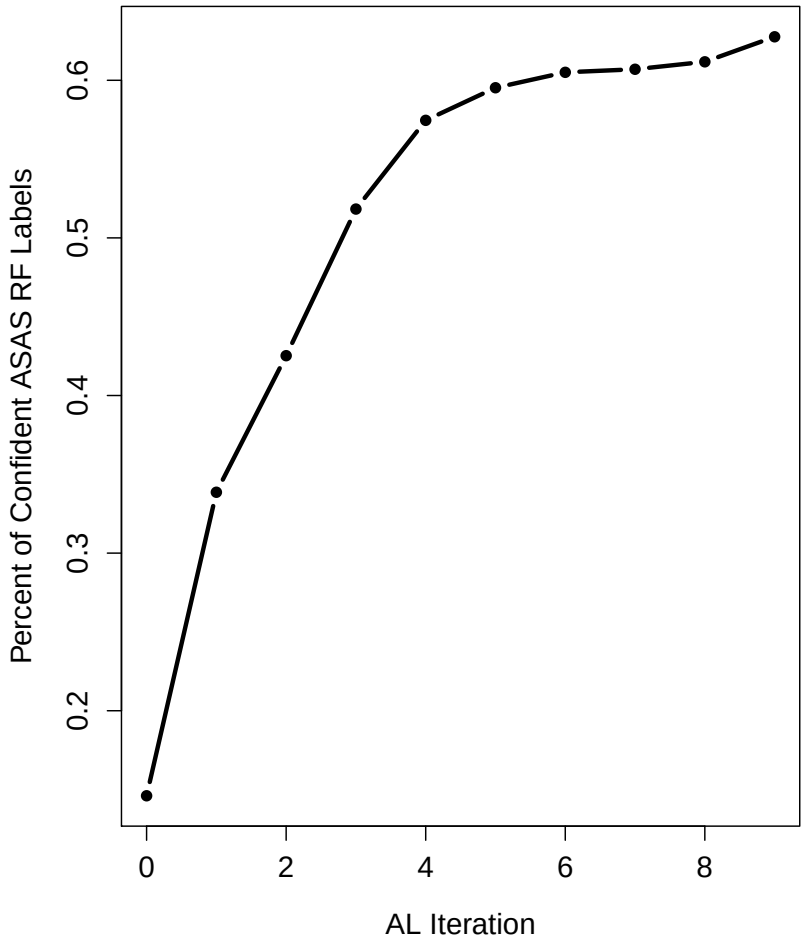
ALLSTARS AL Web Interface



Results: ASAS

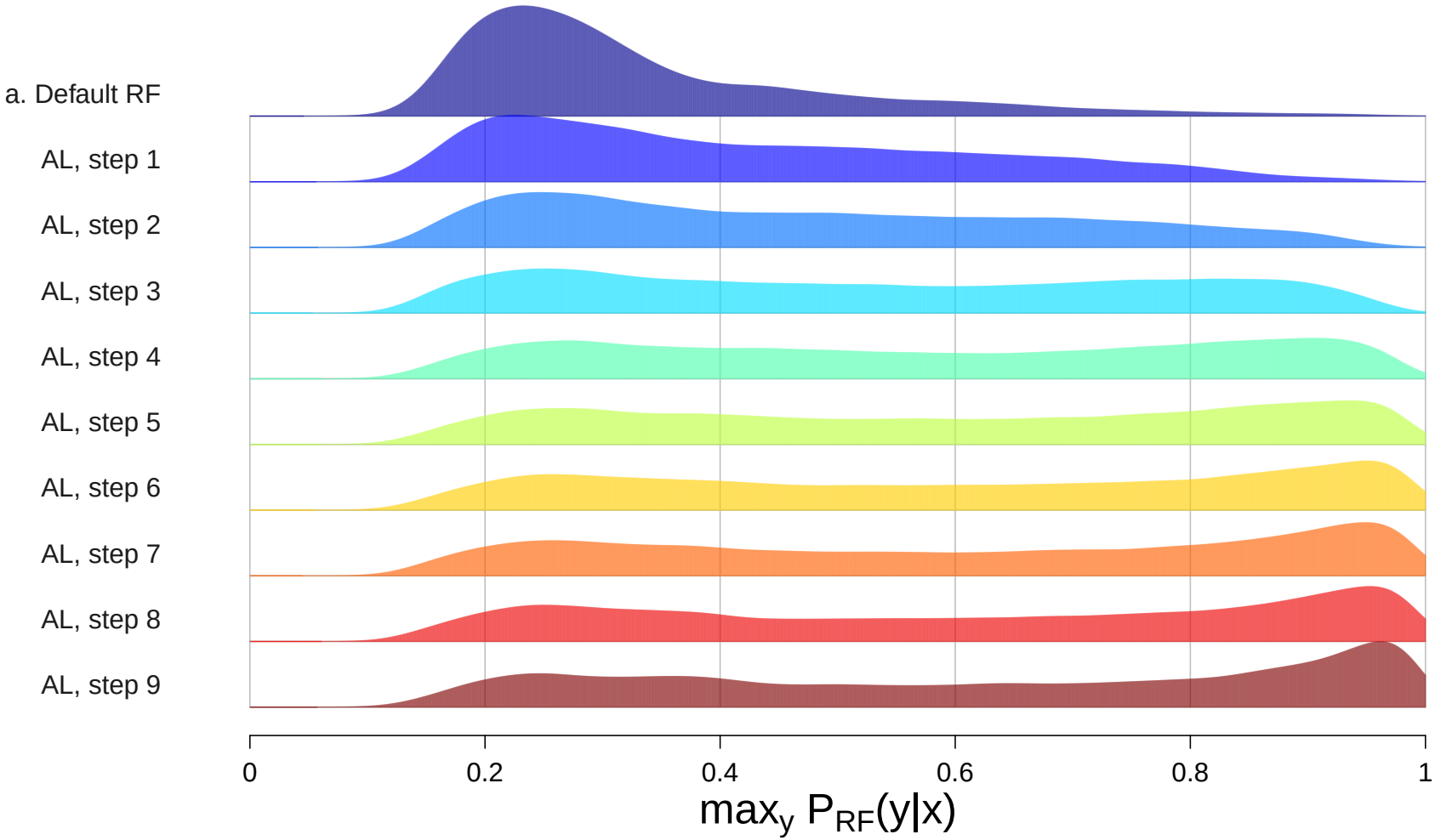
Performance metrics of classifier vs. AL iteration:





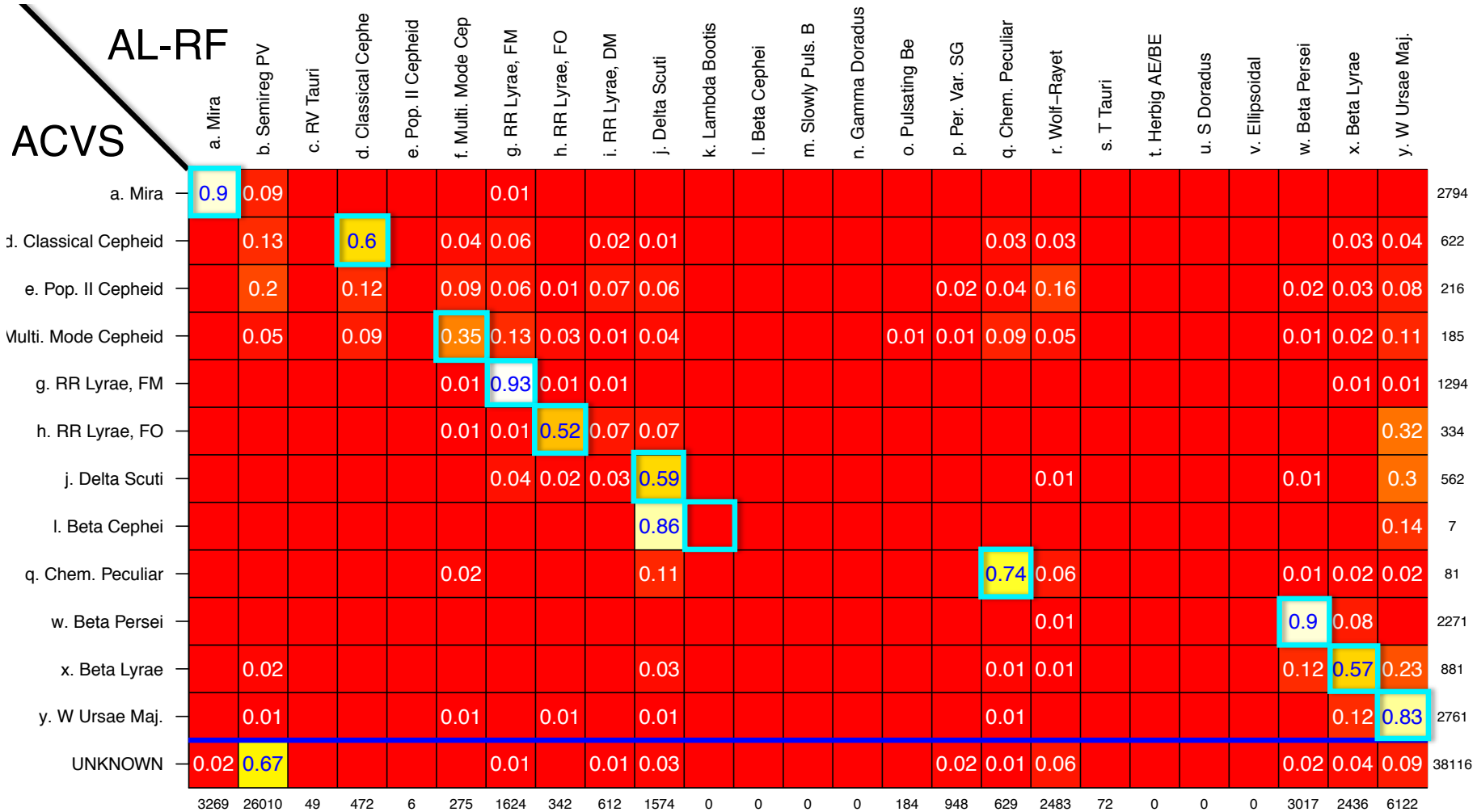
Results: ASAS

Distribution of max RF probabilities over 50124 ASAS sources:



Results: ASAS

AL classifications, compared to ACVS:



Summary

1. Sample selection bias is a debilitating problem in many areas of astronomy
2. Biases in training samples cause (1) poor model selection, and (2) catastrophic extrapolation
3. To mitigate its effects, can (1) use weights on training set, or (2) selectively add data to the training set
4. We find that Active Learning is a viable & effective approach to the problem of sample selection bias
5. In both an experiment & application to a real survey, we find that AL outperforms other approaches

Center for Time-Domain Informatics Publications

Bloom, Joshua S., Starr, D. L., Butler, N. R., Poznanski, D., Rischard, M., Kennedy, R., Brewer, J. **Rapid and Automated Classification of Events from the Palomar Transient Factory** (2009, AAS, 41, 419)

Starr, D. L., Bloom, J. S., Brewer, J. M., Butler, N. R., Poznanski, D., Rischard, M., Klein, C. **The Berkeley Transient Classification Pipeline: Deriving Real-time Knowledge from Time-domain Surveys** (2009, ASPC, 411, 493)

Butler, Nathaniel R., Bloom, Joshua S. **Optimal Time-Series Selection of Quasars** (2011, AJ, 147, 93)

Richards, Joseph W., et al. **On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data** (2011, ApJ, 733, 1)

Bloom, Joshua S. & Richards, Joseph W. **Data Mining and Machine-Learning in Time-Domain Discovery & Classification** (2011, Chapter in the forthcoming book "Advances in Machine Learning and Data Mining for Astronomy")

Richards, Joseph W., Homrighausen, Darren, Freeman, Peter E., Schafer, Chad M. & Poznanski, Dovi **Semi-supervised Learning for Photometric Supernova Classification** (2011, submitted, MNRAS)

Richards, Joseph W. et al. **Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification** (2011, in preparation)