Automated Classification of Transient and Variable Sources

Ciro Donalek
(S.G. Djorgovski, A. Mahabal, M. Graham, A. Drake, M. Yang, G. Longo)
California Institute of Technology

AstroInformatics 2012
Redmond, WA, Sep 10-13

Summary

- Time Domain Astronomy
  - Transient and Variable Source Classification
- Ensemble of Ensembles
  - Our data: CRTS, archival
  - Dimensionality Reduction
  - DT, NN, kNN, “Sup” SOM, dm/dt histograms, BN
  - how to combine different models?
- Results and Conclusions

How to Help?

- Can’t learn the discipline before you start (takes 4 years.)
- Can’t go native – you are a CS person not a bio,… person
- Have to learn how to communicate
  Have to learn the language
- Have to form a working relationship with domain expert(s)
- Have to find problems that leverage your skills

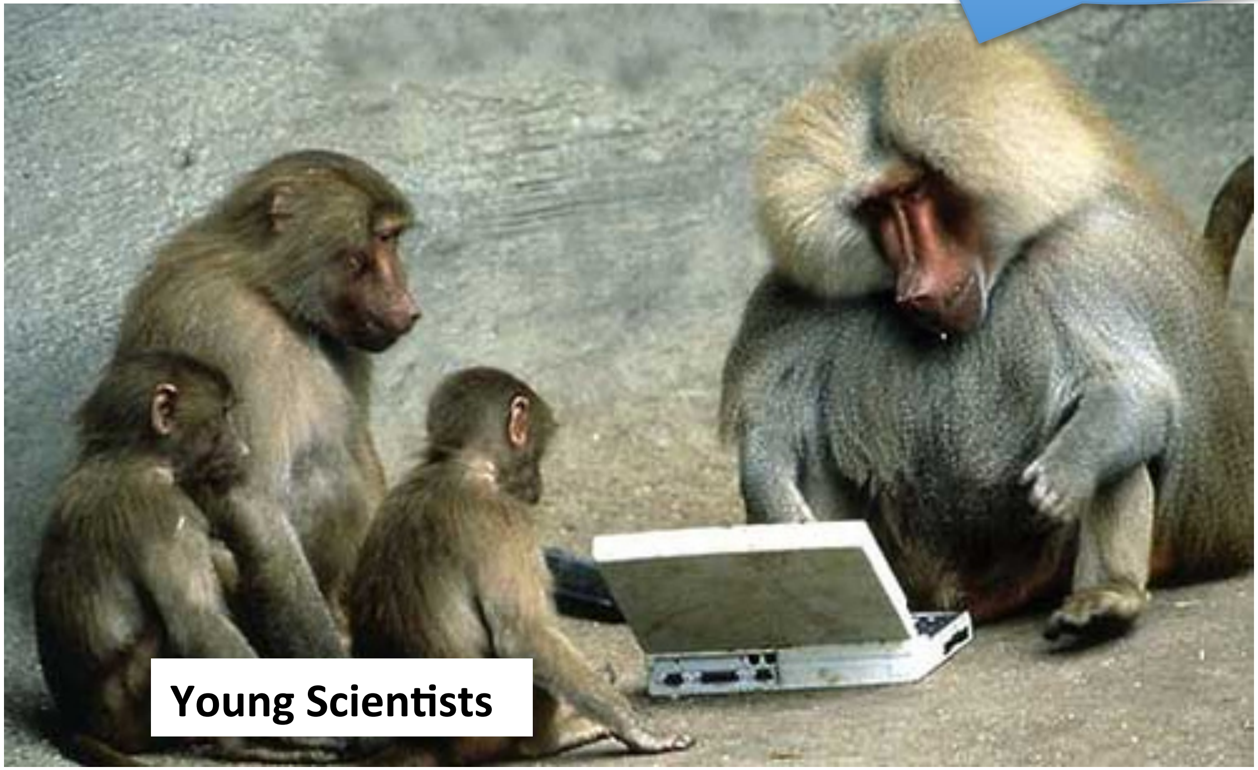
J. Gray, 2004

Why we need Data Mining in Science

- Better and faster technology is increasing the amount of collected data in many scientific fields.
  - billions of G C A T in the genome
  - synoptic sky surveys
  - climate data
  - path that millions of users take through a website
  - changing in nature
- Goal: **Extract Knowledge**
  - as rapidly and efficiently as possible
  - classification, clustering, regression, path analysis, visualizatiom
- **The main reason why we need data mining tools is...**

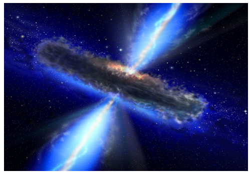
...to have a **future** better than this!

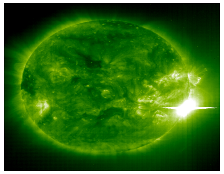
Wish I had some
DATA MINING
tools!



Time Domain Astronomy

- Has rapidly become one of the most exciting new research frontiers in Astronomy.
- Enabled by Telescope Systems dedicated to discovery of moving objects, transient or explosive astrophysical phenomena, GRBs, detection of extrasolar planets
- Each of them requiring rapid alerts and follow-up observations.

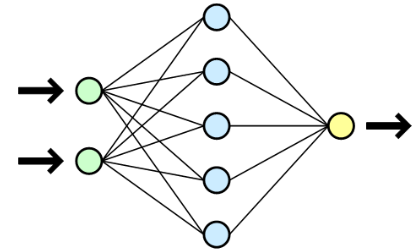


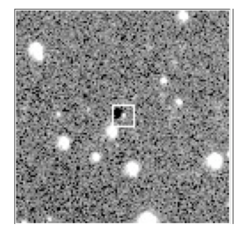


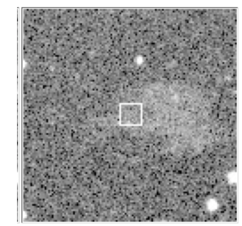
Catalina Sky Survey 0.7m Schmidt

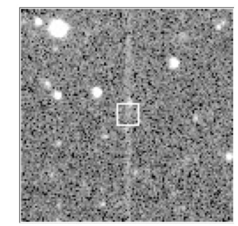
TDA: Classification Challenges

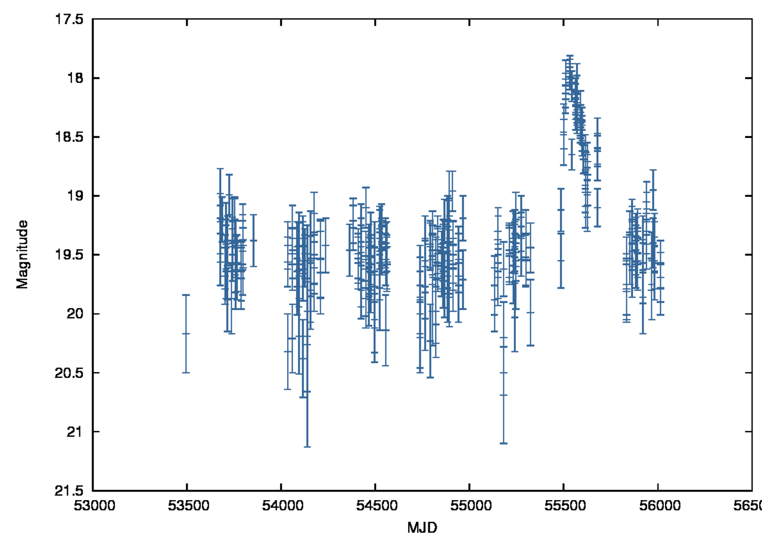
- From a CS perspective, Synoptic Sky Surveys are also posing several new objects classification challenges:
  - Real-Time artifact removal
  - Real-Time transient classification
  - Next-Day transient classification
  - Decision Making
    - choose the objects worthy of follow-ups with expensive facilities
    - resources allocation
- Most systems today rely on a delayed human judgment. This “manual” approach will simply not scale to the next generation of surveys.



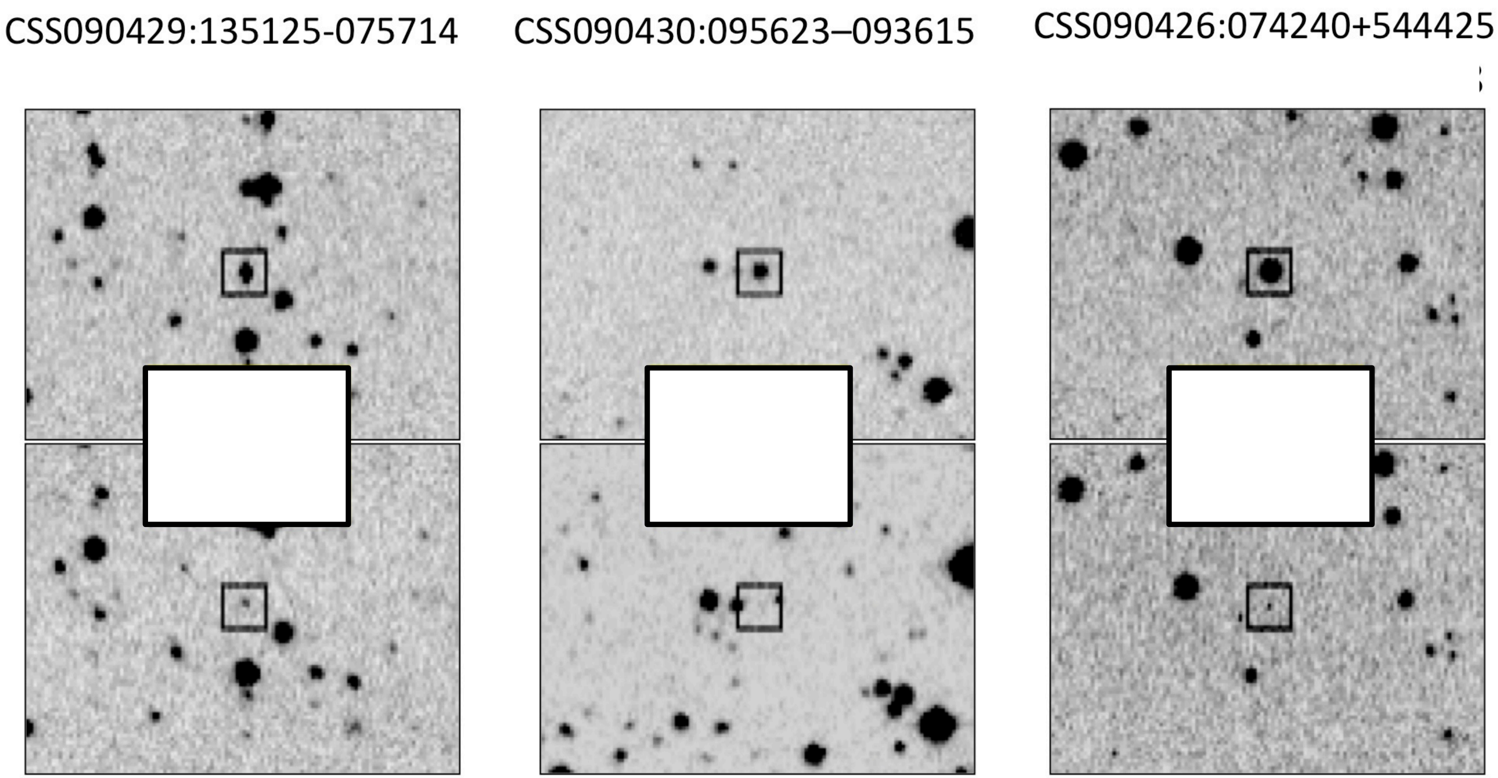








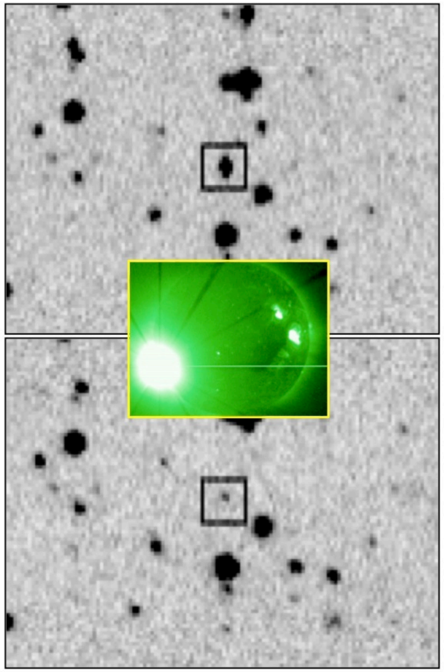
Discovering Transients



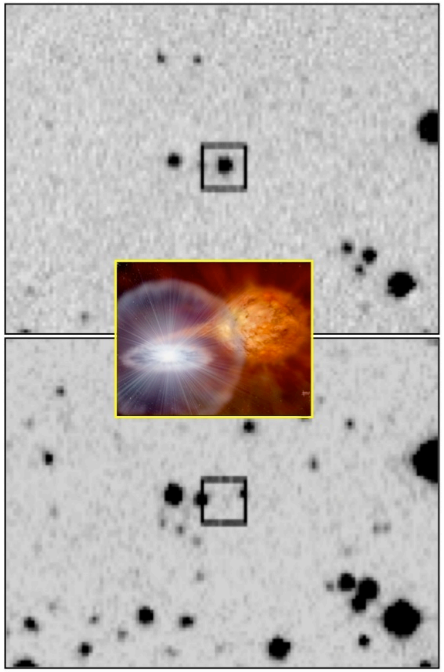
Top row show objects which appear much brighter that night, relative to the baseline images obtained earlier (bottom row).
On this basis alone, the three transients are physically indistinguishable.

Discovering Transients

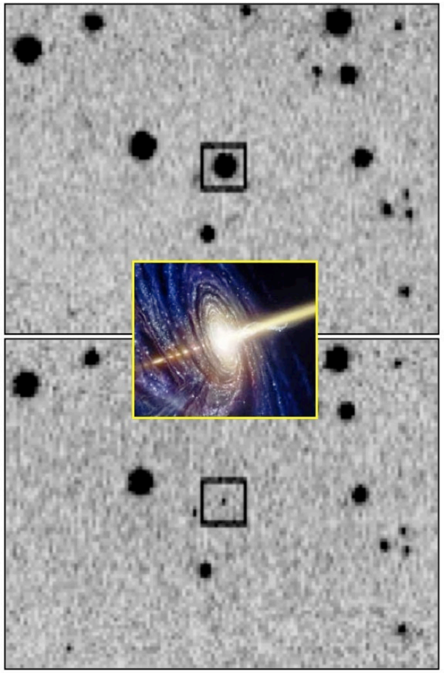
CSS090429:135125-075714
Flare star



CSS090430:095623–093615
Dwarf Nova



CSS090426:074240+544425
Blazar, 2EG J0744+5438

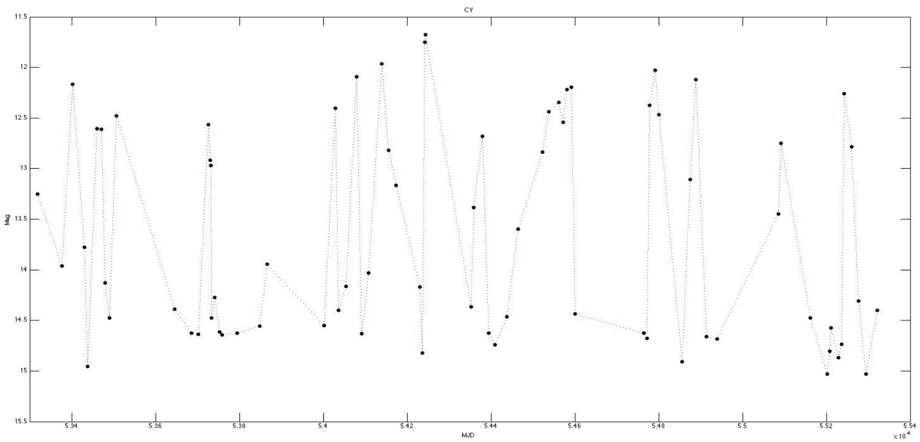


Top row show objects which appear much brighter that night, relative to the baseline images obtained earlier (bottom row).

On this basis alone, the three transients are physically indistinguishable.

Subsequent follow-ups show them to be three vastly different types of phenomena.

From a CS perspective...



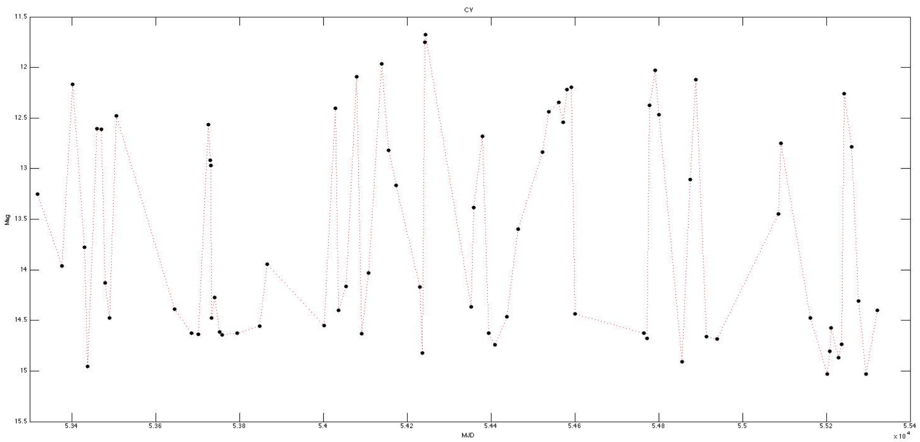
Many people interested in different classes/shares

On Demand => optimize the classifiers for a given class

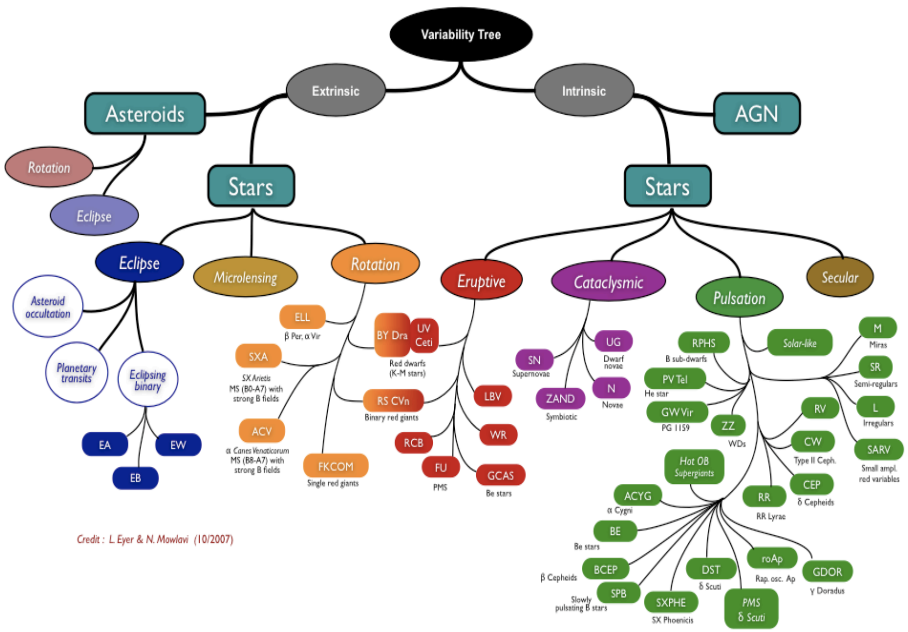
High Completeness => Maximizie Gain

Low Contamination => Minimize Losses

Challenges

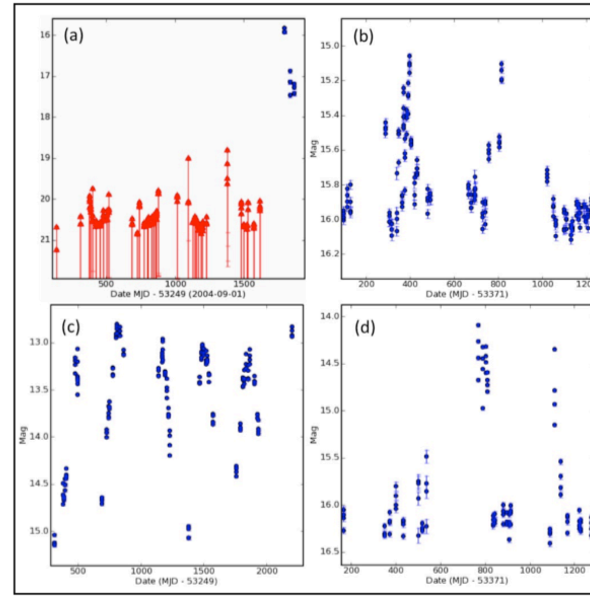


- Massive multiparametric dataset
  - Petascale ready
  - Sparse and Heterogeneous Data
  - High number of Features and Classes
- Classification
  - Real Time
  - Reliable
  - High completeness
  - Low contamination
  - Use minimum amount of points
  - *Learn* from the past experience
  - As automated as possible
- Include External Knowledge



Our Data: transients from CRTS

- Parameters
  - Discovery
    - mag, delta-mag
  - Contextual
    - distance from the nearest star and galaxy, distance to nearest radio source, gb, etc.
  - Follow up Colors
  - Lightcurve Characterization
- Class
- Data
  - Heterogeneous
  - Unbalanced
  - Sparse
  - Missing Data



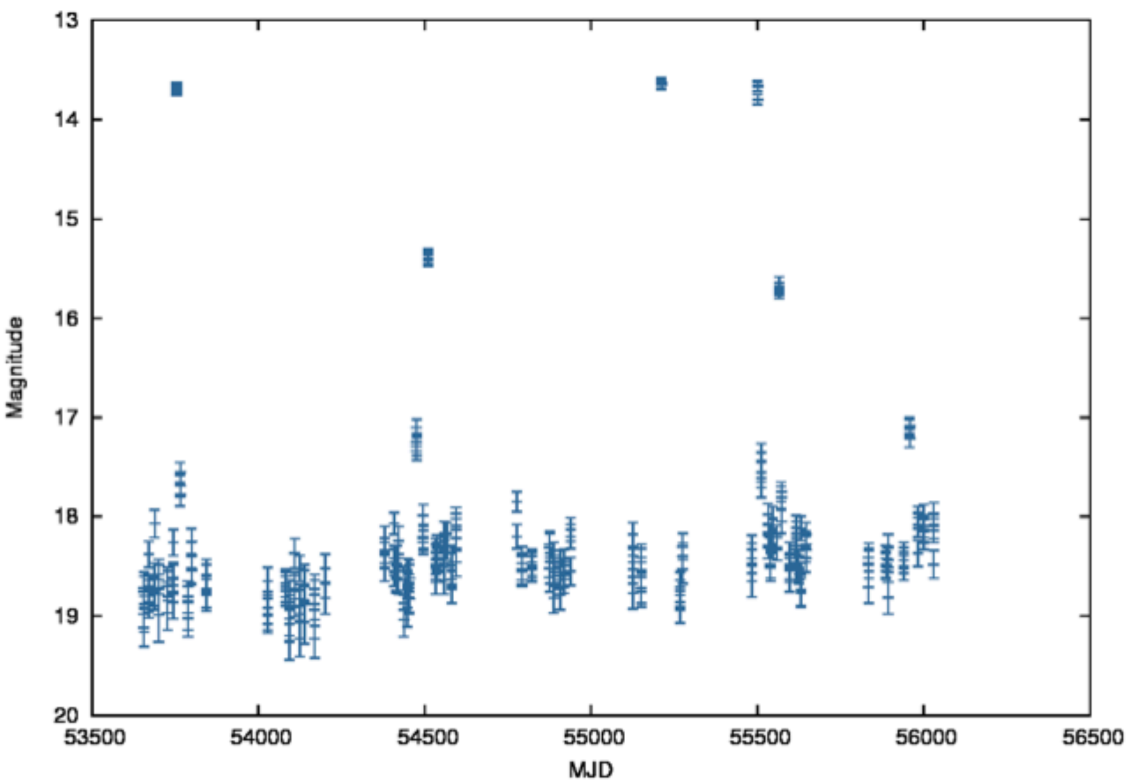
CRTS: (a) Supernova; (b) Blazar; (c) Pulsating Variable Star; (d) CV.

transient_dataset.dat

```
mjd g r i z g-r r-i i-z GB galdist galmag stardist starmag radiodist radioflux c1 c2 p1 p2
81090804133432 54907.3 16.50 15.95 15.60 15.21 0.55 0.36 0.38 54.580612 226.1041400 10.4942200 20.3368856087
81090804133432 54894.3 16.66 16.11 15.67 15.24 0.55 0.44 0.42 54.580612 226.1041400 10.4942200 20.3368856087
81090804133432 54901.5 16.77 16.23 15.76 15.41 0.54 0.46 0.36 54.580612 226.1041400 10.4942200 20.3368856087
81090804133432 54903.3 16.78 16.27 15.83 15.51 0.51 0.44 0.32 54.580612 226.1041400 10.4942200 20.3368856087
11180834110558 54894.3 18.35 18.42 18.77 18.72 -0.07 -0.35 0.05 44.261376 241.4694400 17.4126500 0.621239291
51090614128796 54860.4 19.74 19.64 19.75 99.00 0.10 -0.12 -99.00 63.273174 170.8197900 10.2948700 0.75858099
51090614128796 54858.5 20.66 20.70 20.72 99.00 -0.04 -0.02 -99.00 63.273174 170.8197900 10.2948700 0.7585809
51090614128796 54863.4 20.18 19.37 99.00 99.00 0.80 -99.00 -99.00 63.273174 170.8197900 10.2948700 0.7585809
51090614128796 54863.5 99.00 99.00 19.36 99.00 -99.00 -99.00 -99.00 63.273174 170.8197900 10.2948700 0.75858
91180704127993 54863.5 24.86 99.00 23.40 99.00 -74.14 75.60 -99.00 76.760843 203.8707000 18.6033100 5.777447
91210644131318 54863.5 22.51 22.26 22.46 99.00 0.25 -0.20 -99.00 83.112788 188.4711500 21.5189900 0.31676106
01150724101062 54862.3 17.26 16.78 16.67 99.00 0.48 0.11 -99.00 70.926227 207.8404900 14.1437700 2.322211871
31090464114479 54882.1 17.91 17.74 18.18 19.03 0.17 -0.44 -0.86 27.554329 129.1736700 9.2348200 1.2161416998
40010554111683 54882.2 18.73 18.33 18.20 18.30 0.40 0.14 -0.11 41.602274 152.7154700 -2.0055200 9.2066406014
41180544115201 54881.5 19.86 20.08 19.86 18.99 -0.22 0.22 0.87 55.122105 156.2640000 17.9725900 9.0104617404
61070544137952 54882.2 18.75 18.55 19.31 22.72 0.20 -0.76 -3.40 47.086963 152.2921000 7.9095600 1.449553391
90010534124257 54886.2 18.57 18.31 18.19 18.52 0.26 0.12 -0.33 38.921028 148.8571100 -1.4725800 0.552522075
```

Lightcurve Characterization

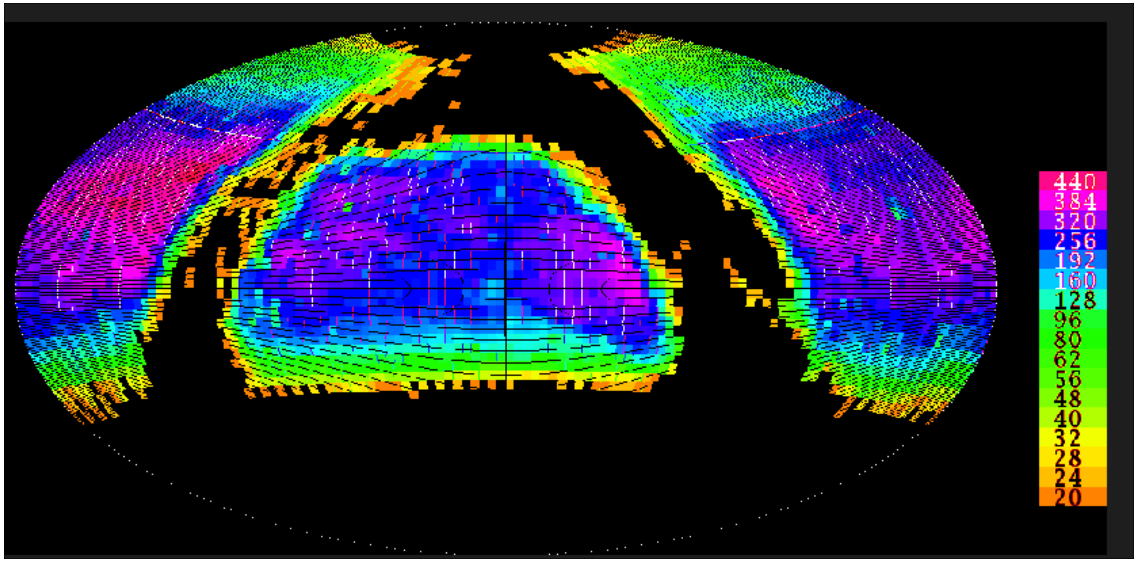
- Data: lightcurves from Catalina Real Time Sky Survey.
  - **CV**, Blazars, SN, AGN, RR Lyrae, Mira, WUMa, Flare Stars, ...
- Caltech Time Series Characterization Service (Graham M., 2011), ~60 periodic and non-periodic features can be extracted.



CRTSID	1112047000737
RA	133.838254
DEC	11.304235
amplitude	2.815000
beyond1std	0.073077
fpr_mid20	0.008531
fpr_mid35	0.018134
fpr_mid50	0.027501
fpr_mid65	0.042434
fpr_mid80	0.130268
linear_trend	-0.000221
max_slope	108.655617
med_abs_dev	0.230000
med_buf_range_per	0.926923
pair_slope_trend	0.533333
percent_amplitude	86.096359
pdfp	16.432213
skew	-2.956801
kurtosis	8.378567
std	1.130028
ls	0.313022
rcorbor	0.000000
magratio	0.488462
data_num	260

The Catalina Surveys DR1

- The CRTS survey searches ¾ of sky for highly varying astronomical sources ("optical transients").
- Transients are mainly found based on catalog comparisons
  - CSS coadds, USNO-B, PQ, SDSS.
  - Remove artifacts and asteroids by detection coincidence and motion.
- All data is fully processed within minutes of observation.
- **All discoveries are made public instantly to enable rapid follow-up.**
- CSDR1 consists of all photometry from seven years of photometry taken with the CSS Schmidt telescope. This data release encompasses the photometry for **198 million objects** with V magnitudes between 12 and 20 from an area of **24,000** square degrees.

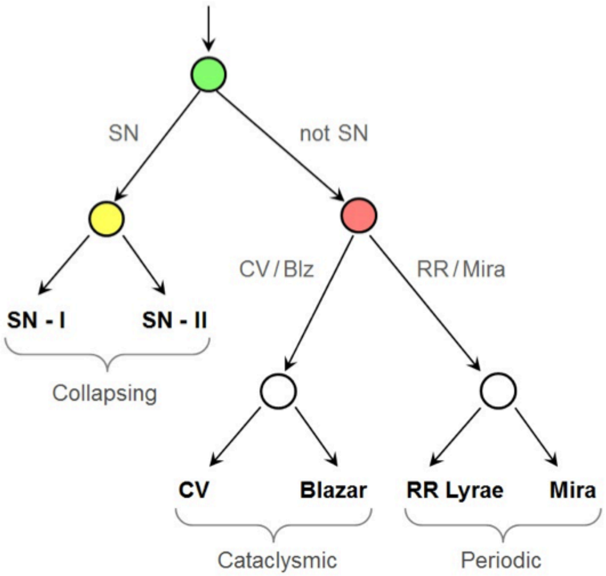


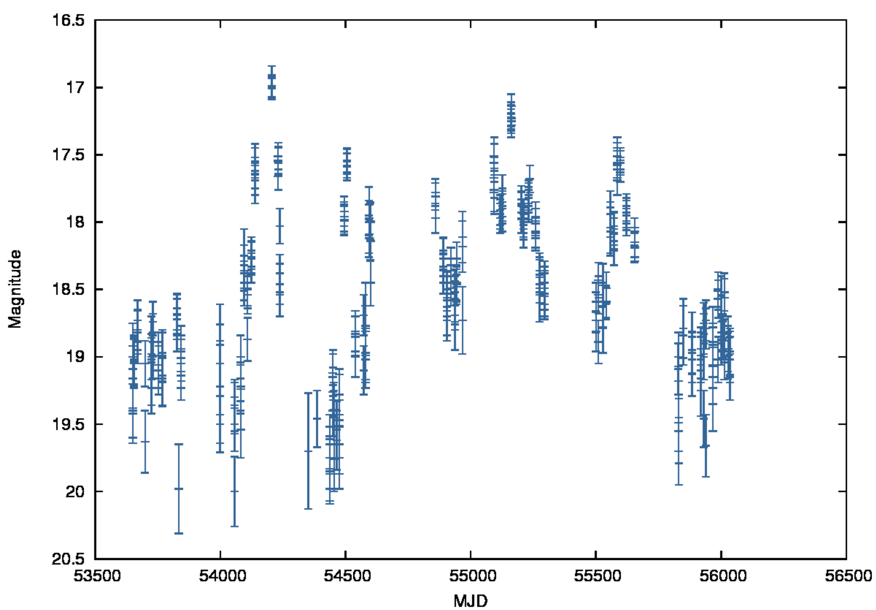
The colour key gives the number of epochs at each location.

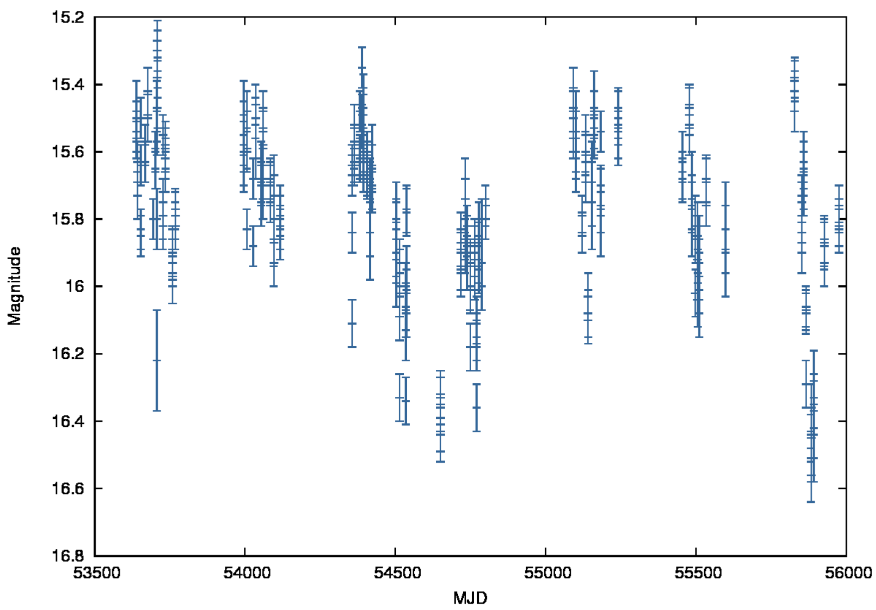
http://crts.caltech.edu/

Problems Attacked

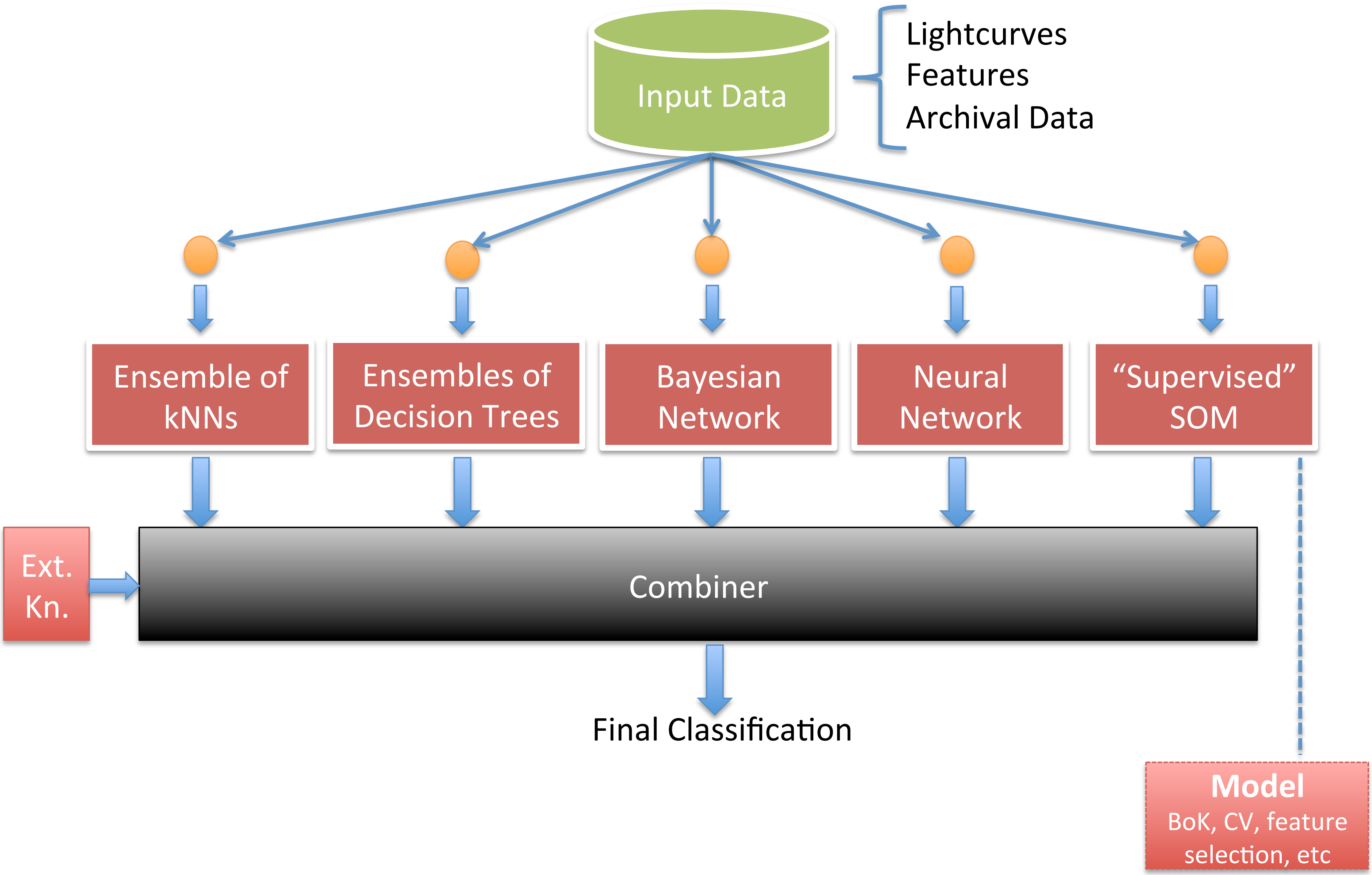
- Classification
  - different types of classifiers perform better for some event classes than for the others.
  - some astrophysically motivated major features are used separate different groups of classes
- Systematic Search for CV
  - help us to know the formation and distribution of the cataclysmic variables
- RR Lyrae vs Eclipsing Binaries
  - W U Ma are the main contaminant in studies using RR Ly as tracers of Galactic structures





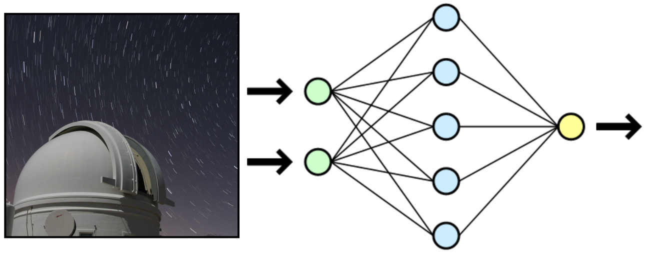


Classification Schema

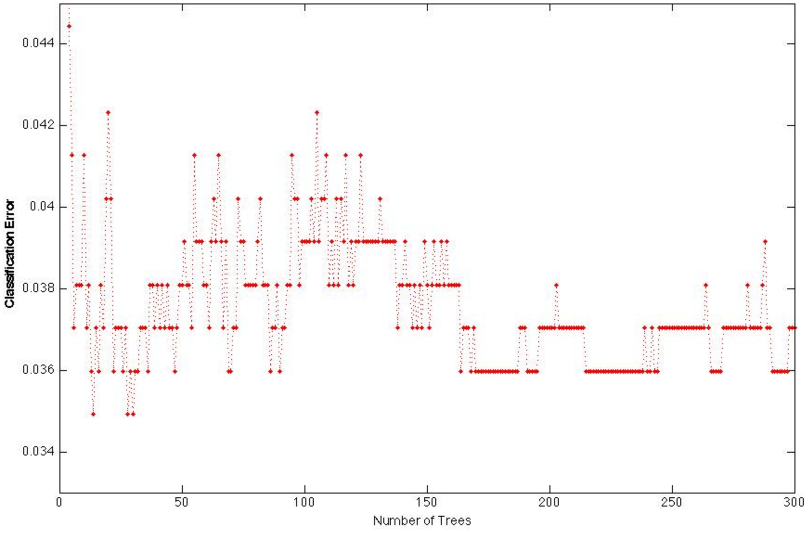


Experiments Framework

- Construct the BoK (preprocessing)
- Ensemble Algorithms: Gentle AdaBoost (2 classes), AdaBoost (3 or more classes), Bag



- Test Ensemble Quality
  - Stratified 10 fold Cross Validation
  - Set the appropriate number of Ensemble Members
  - Compute Completeness and Contamination



- *Derive "classifier weights"*

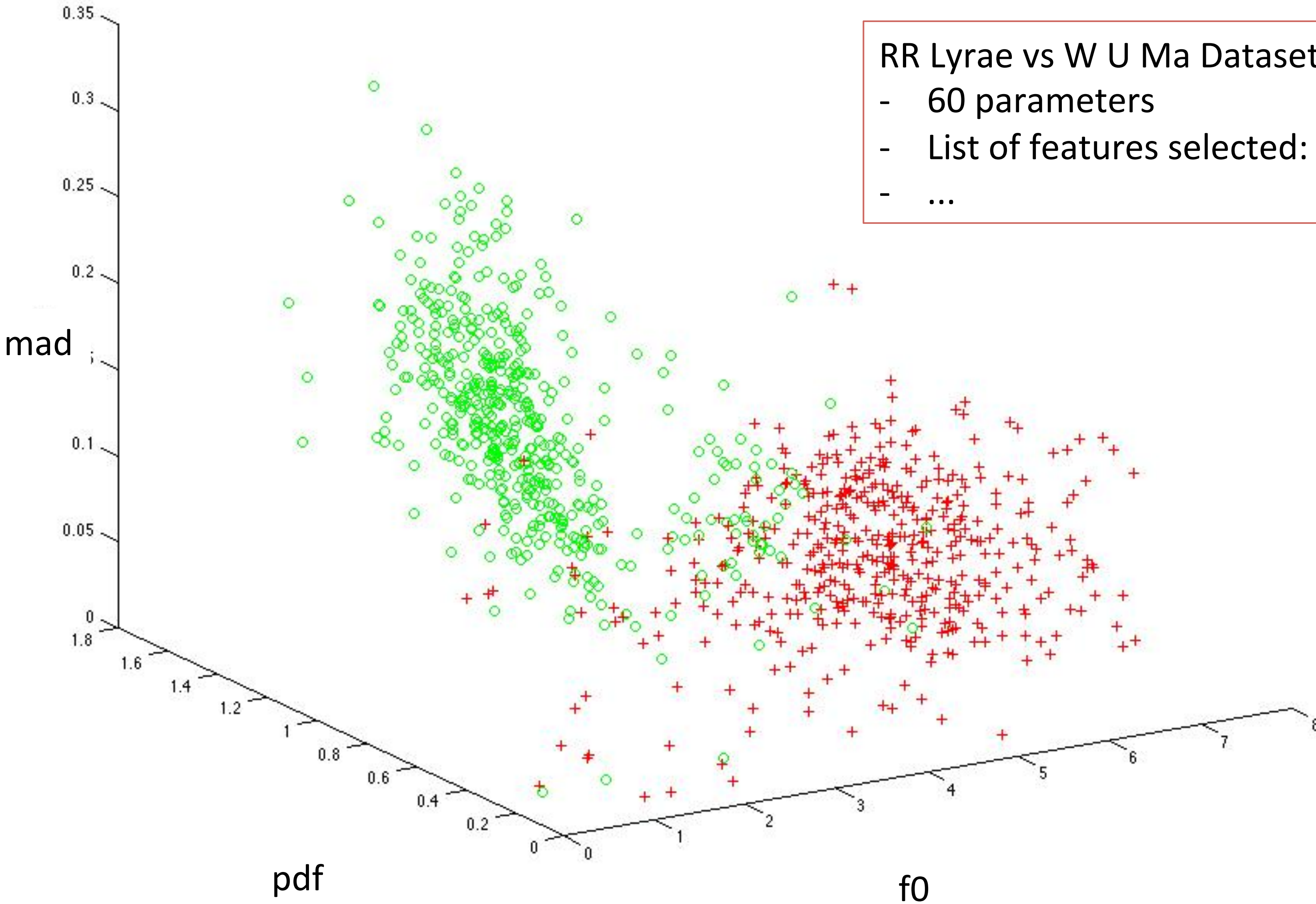
Pre-Processing: Feature Selection

- Over 100 features
  - address the curse of dimensionality selecting only a subset of features;
  - some features may be misleading;
  - preferable to PCA when the meaning of features are important and one of the goal is to identify an influential subset
- Interested in finding *k* of the *d* dimensions that give us the most information
- Analyze these sets with the domain scientist.

Estimating Feature Importance

- Model Dependent Approaches
  - Sequential Feature Selection
    - start with no variables and add at each step the one that decreases the error the most, until any further addition does not significantly decrease the error.
  - Backward Feature Selection
- Attribute estimators
  - detect conditional dependencies between attributes in order to provide a unified view on the attribute estimation in regression and classification (eg, relief, t-test)
- Correlation Hunting
  - scatterplots
  - Self Organizing Maps

rank features - ttest

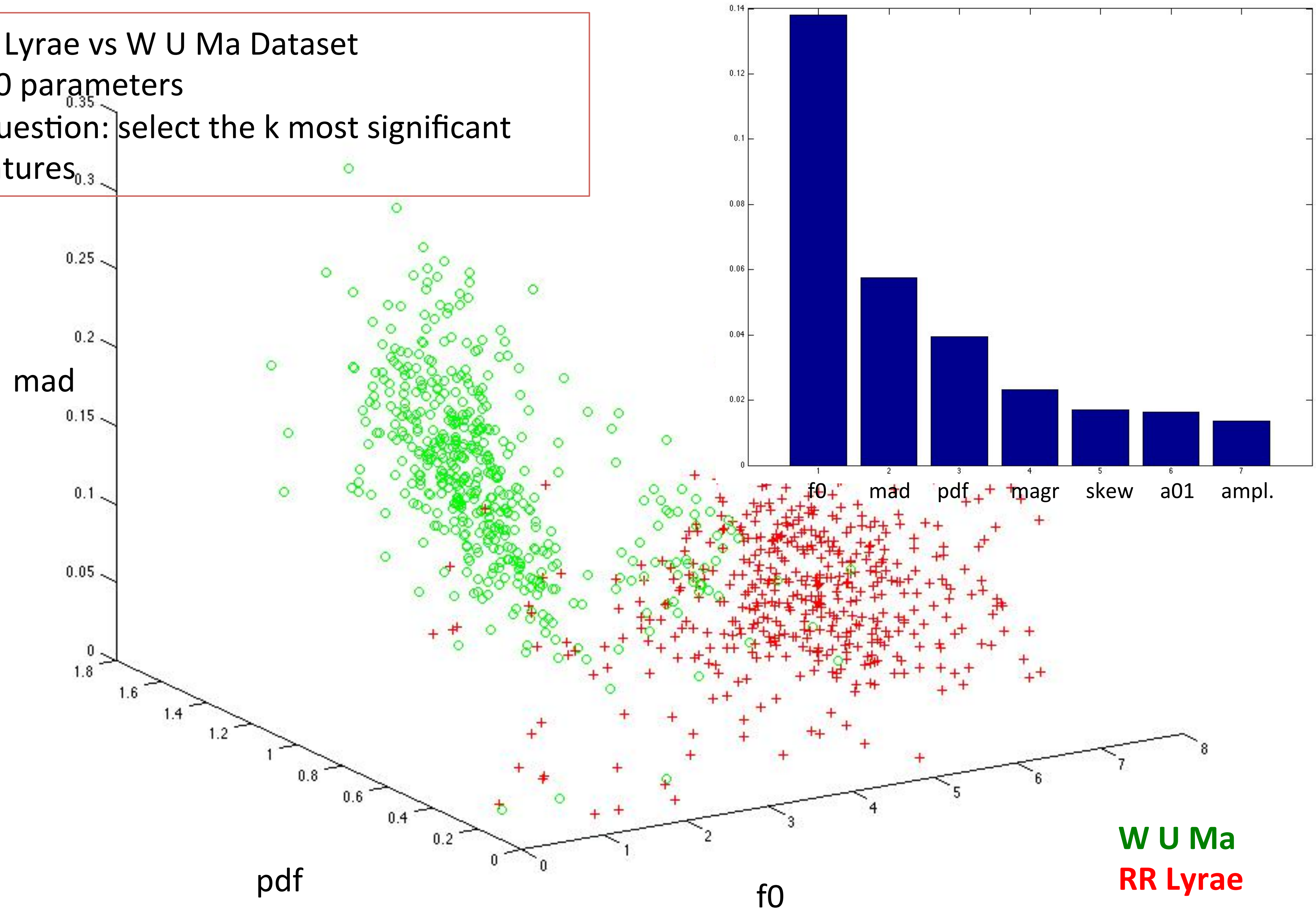


RR Lyrae vs W U Ma Dataset

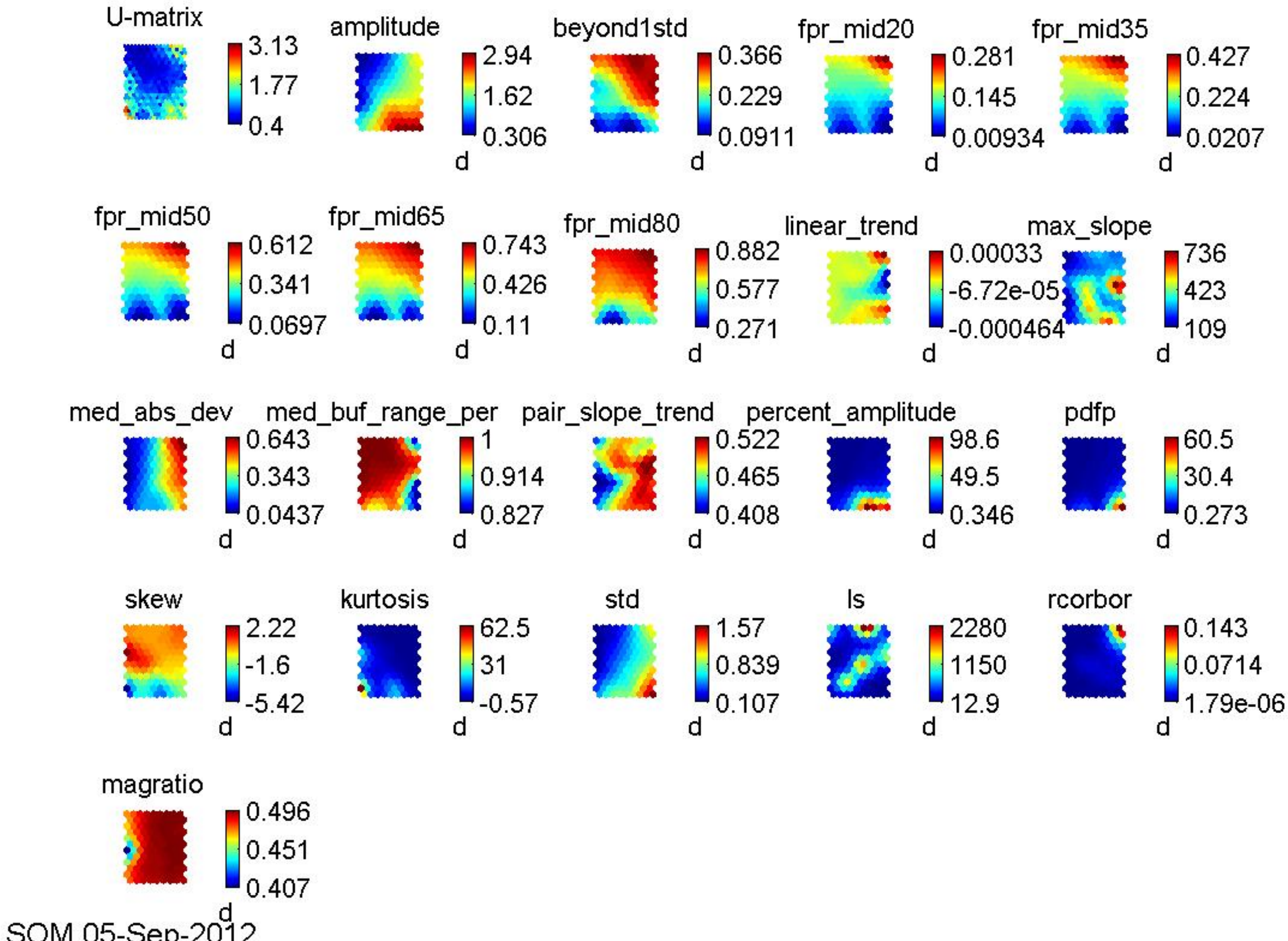
- 60 parameters
- List of features selected:
- ...

Rank features - relieff

RR Lyrae vs W U Ma Dataset
- 60 parameters
- question: select the k most significant features



Self-Organizing Maps



Results: W U Ma vs RR Lyrae

RR Lyrae vs Eclipsing Binaries

- W U Ma are the main contaminant in studies using RR Lyrae as tracers of Galactic structures
- ~60 periodic/non periodic features extracted from CRTS lightcurves
- “IRIS” for Astronomy

	Completeness MLP	Contamination MLP
WUMa (463)	94%	7%
RR Lyrae (482)	97%	3%

SEQ: max slope, pair slope trend, skew, small kurtosis, stetsonj

	Completeness kNN	Contamination kNN
WUMa (463)	96%	4%
RR Lyrae (482)	95%	5%

TTEST: f0, pdf, mad, stetson j, std, magratio, vf1v

	Completeness DT	Contamination DT
WUMa (463)	97%	3%
RR Lyrae (482)	96%	4%

BACKWARD FS: f0, mad, beyond1st, fpr_mid50, ...20 more (out of ~60)

	Completeness kNN	Contamination kNN	Completeness DT	Contamination DT
WUMa (463)	92%	8%	93%	5%
RR Lyrae (482)	93%	7%	95%	6%

RRELIEF
f0, vf1v, magratio, a02, skew, a01 amplitude

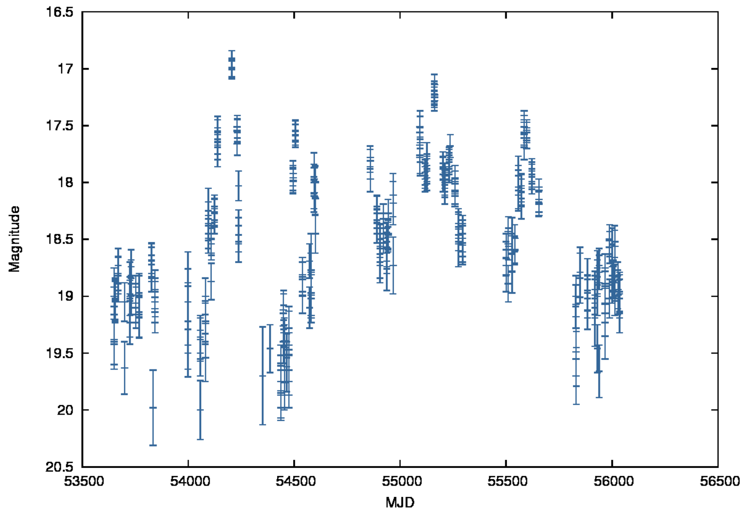
Results: CV vs (Blazars, all)

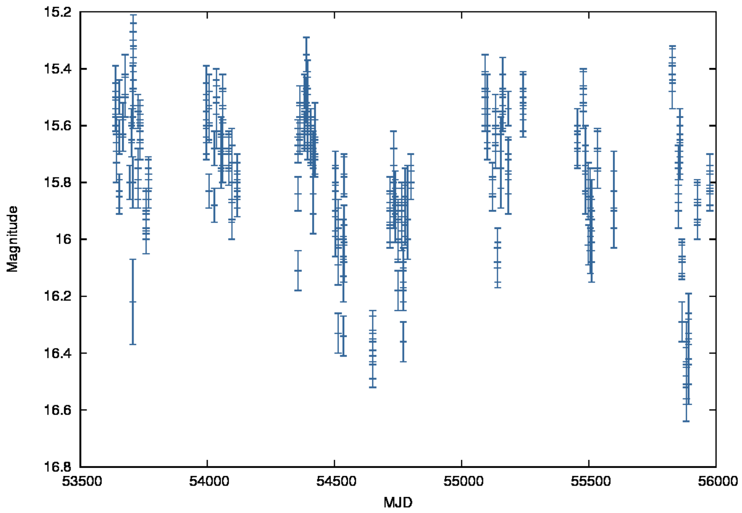
Systematic Search for CV (with M. Yang)

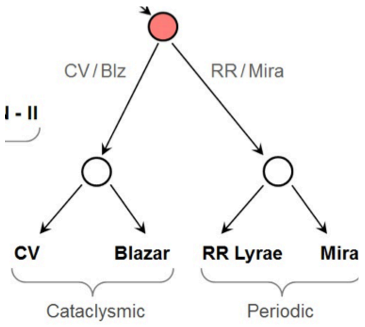
- systematic study will help us to know the formation and distribution of the cataclysmic variables (CVs)
- CRTS provide observations with ~ 7 years baseline, which is good for the cataclysmic variable search
- Light curve features are calculated
- Cataclysmic variable candidates are classified using ML
- Spectroscopic follow-up

	Compl. DT	Cont. DT	Compl. kNN	Cont. KNN
Blazar (100)	83%	13%	52%	35%
CV (376)	92%	8%	91%	12%

	Compl. DT	Cont. DT	Compl. Som	Cont. Som
Other (264)	90%	8%	84%	14%
CV (376)	93%	5%	89%	11%







kNN - Distances

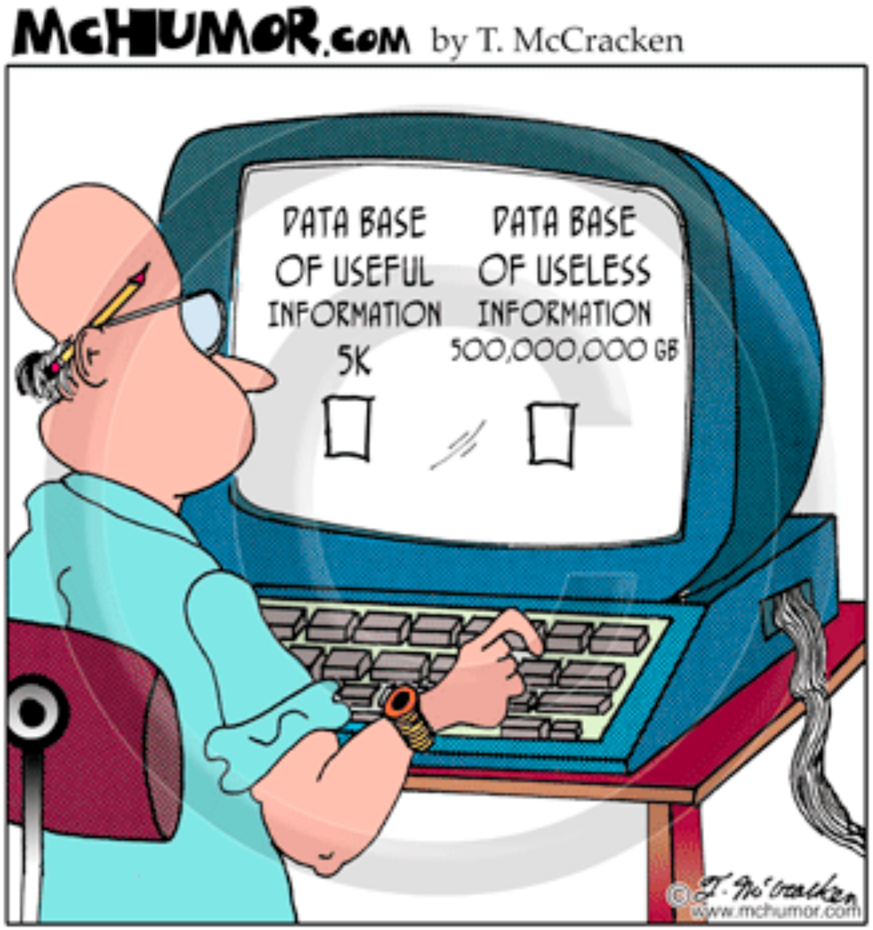
- W uma vs RR Lyrae

	Overall Classification Rate
Cityblock	96%
Chebychev	95%
Cosine	95%
Euclidean	96%
Hamming	71%
Jaccard	71%
Mahalanobis	93%
Minkowski	96%
Spearman	92%

`'cityblock'`	City block distance.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors).
`'euclidean'`	Euclidean distance.
`'hamming'`	Hamming distance, percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix `C`. The default value of `C` is the sample covariance matrix of `X`, as computed by `nancov(X)`. To specify a different value for `C`, use the `'Cov'` name-value pair.
`'minkowski'`	Minkowski distance. The default exponent is `2`. To specify a different exponent, use the `'P'` name-value pair.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between `X` and a query point is scaled, meaning divided by a scale value `S`. The default value of `S` is the standard deviation computed from `X`, `S = nanstd(X)`. To specify another value for `S`, use the `Scale` name-value pair.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

Future Work

- Better way to combine outputs from the single classifiers (crispy and probabilistic classification)
- Gather more data
- Refine existing models and strategies
- Add/investigate more models
- Any suggestion is appreciated!



EXTREME DATA MINING!