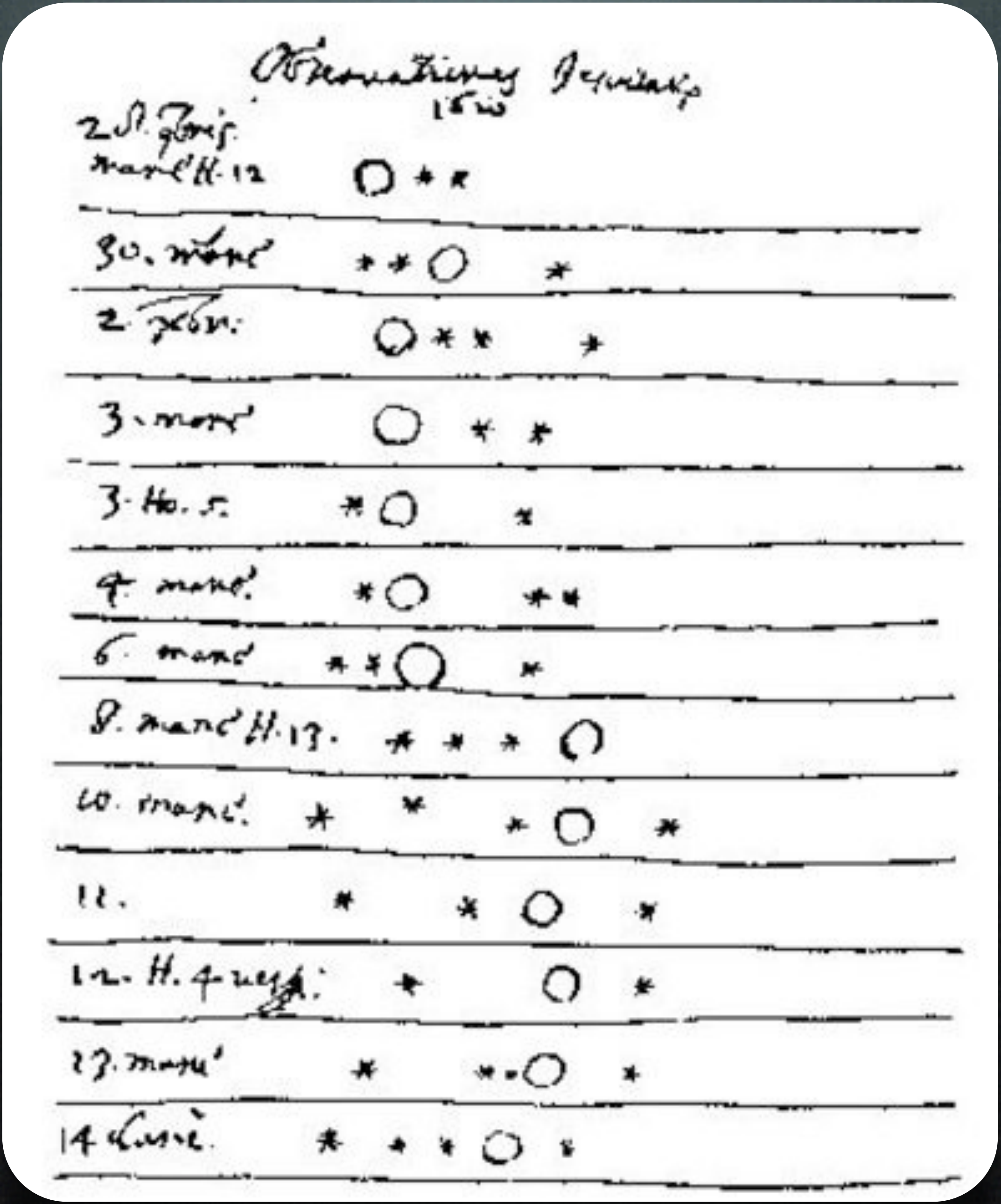


Observationes Jesuilary
1610

2 S. Jovis.
mane H. 12 ○ * *

30. mane * * ○ *

2. xbris: ○ * * *

3. mane ○ * *

3. Ho. 5. * ○ *

4. mane * ○ * *

6. mane * * ○ *

8. mane H. 13. * * * ○

10. mane. * * * ○ *

11. * * ○ *

12. H. 4 negg: * ○ *

13. mane * * ○ *

14 Cane. * * * ○ *

January 7 - March 2, 1610

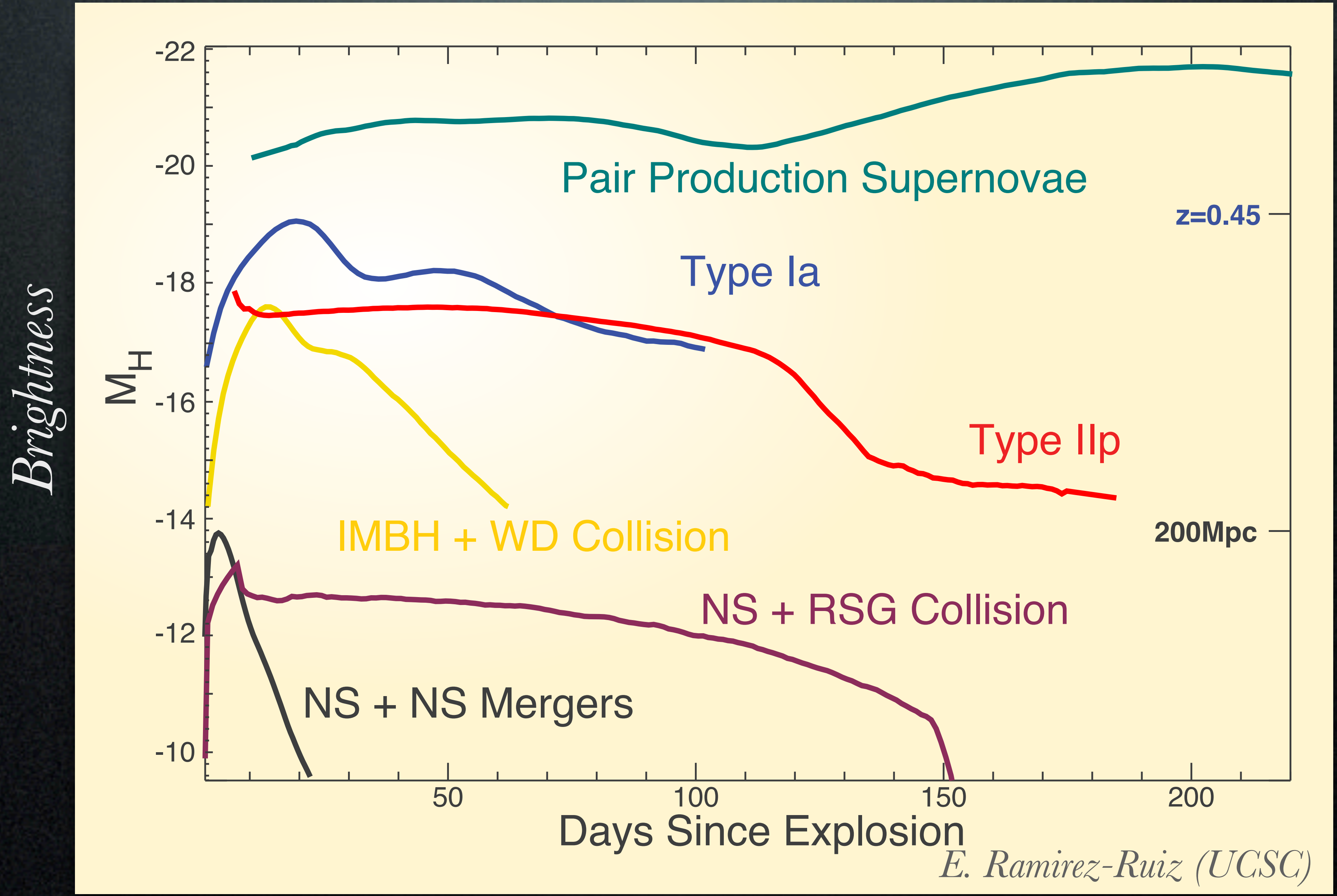
Real-Time Knowledge Extraction from Massive Time-Series Datastreams

Josh Bloom
Astronomy Department

jbloom@astro.berkeley.edu

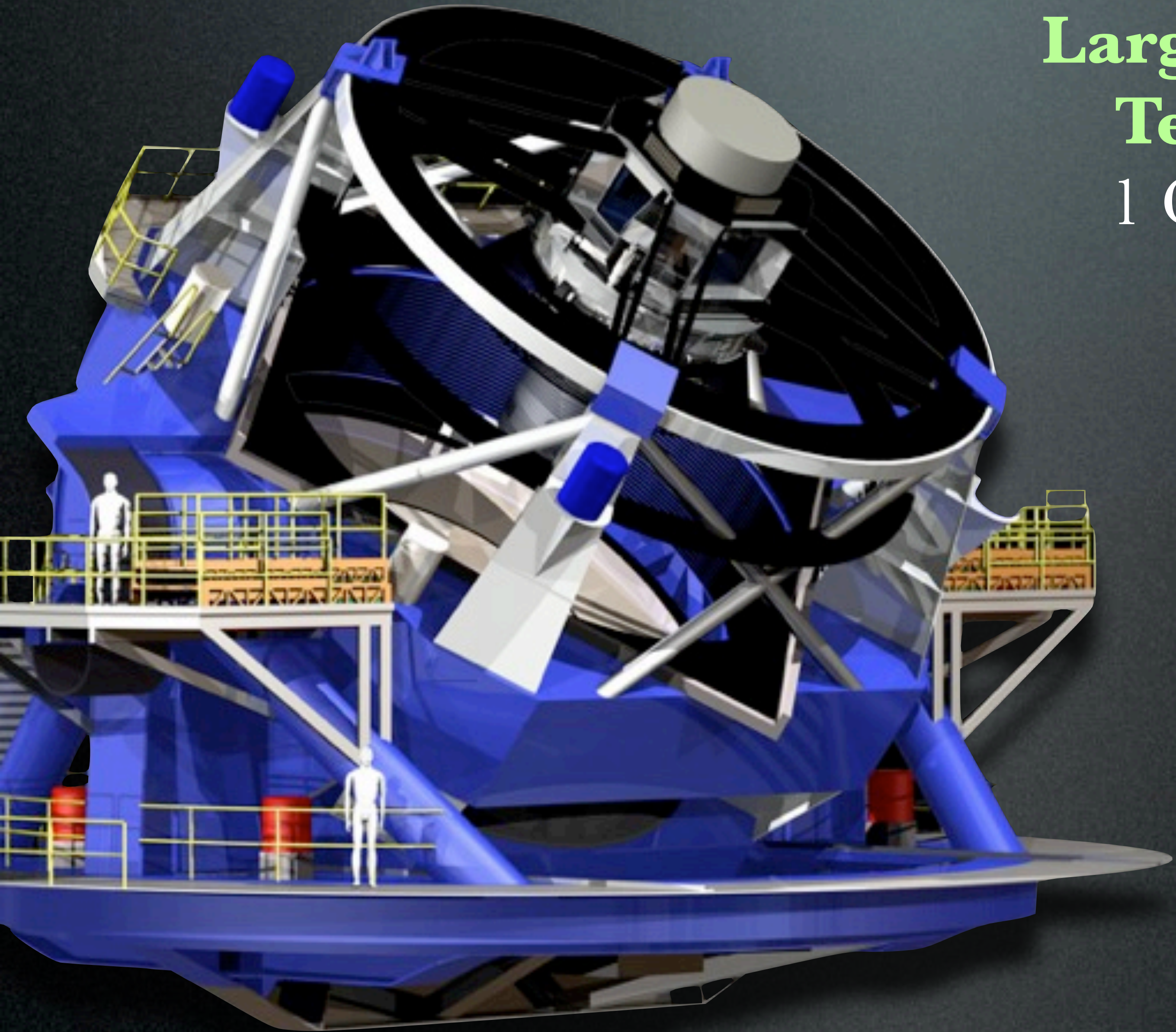
CS267: Feb 16, 2010

Extragalactic Transient Universe: Explosive Systems



E. Ramirez-Ruiz (UCSC)

“Bad” News:
Discoveries Swamp Followup Resources



Large Synoptic Survey Telescope (LSST):

1 Gb every 2 seconds

10^6 supernovae/yr
10^5 eclipsing systems
10^7 asteroids...

light curves of 800 million sources every 3 days

Transients Classification Project

***Berkeley Astronomy*:**

Dan Starr, Dovi Poznanski, Maxime Rischard, Nat Butler, Chris Klein, Rachel Kennedy, Justin Huggins, Adam Morgan, Adam Miller, JSB

San Francisco State University:

John M. Brewer

***Berkeley Statistics*:**

Noureddine El Karoui, John Rice

***Berkeley CS*:**

Martin Wainwright, Masoud Nikravesh

***Lawrence Berkeley Lab*:**

Peter Nugent, Horst Simon

***Los Alamos Nat. Lab. / UC Santa Cruz*:**

Damian Eads

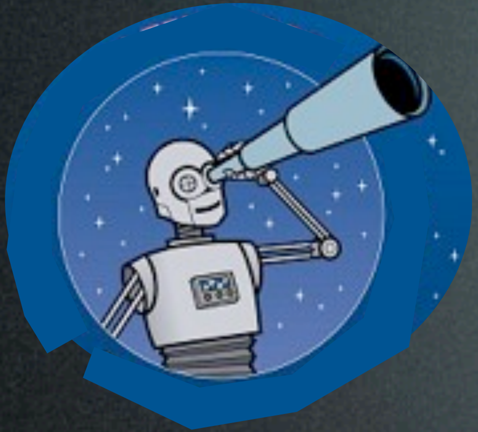
Google

YAHOO!

NSF

SciDAC
Scientific Discovery through
Advanced Computing

LCOGT.net



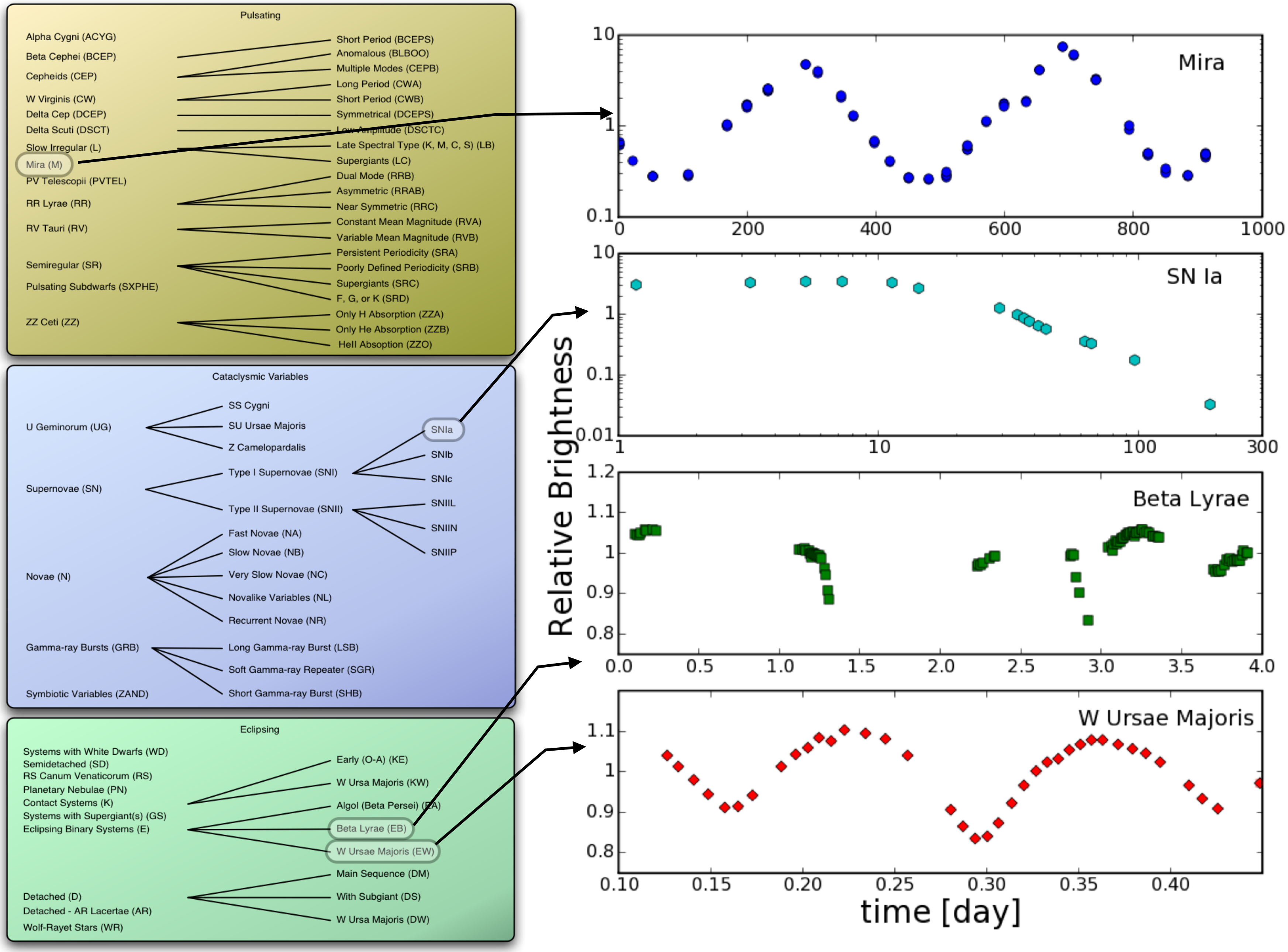
Goal: Autonomous creation of new knowledge, that itself spurs further resource allocation & inquiry

- Generate **probabilistic statements** about the nature of events (ie. classification)
- Provide push/pull **access** to current & past events
- (bootstrap) Learning from feedback
- Operate at sufficient & **scalable** rates

Pulsating Stars

Cataclysmic Variables

Eclipsing Systems



Pulsating

Alpha Cygni (ACYG)
Beta Cephei (BCEP)
Cepheids (CEP)
W Virginis (CW)
Delta Cep (DCEP)
Delta Scuti (DSCT)
Slow Irregular (L)
Mira (M)
PV Telescopii (PVTEL)
RR Lyrae (RR)
RV Tauri (RV)
Semiregular (SR)
Pulsating Subdwarfs (SXPHE)
ZZ Ceti (ZZ)

Short Period (BCEPS)
Anomalous (BLBOO)
Multiple Modes (CEPB)
Long Period (CWA)
Short Period (CWB)
Symmetrical (DCEPS)
Low Amplitude (DSCTC)
Late Spectral Type (K, M, C, S) (LB)
Supergiants (LC)
Dual Mode (RRB)
Asymmetric (RRAB)
Near Symmetric (RRC)
Constant Mean Magnitude (RVA)
Variable Mean Magnitude (RVB)
Persistent Periodicity (SRA)
Poorly Defined Periodicity (SRB)
Supergiants (SRC)
F, G, or K (SRD)
Only H Absorption (ZZA)
Only He Absorption (ZZB)
HeII Absoption (ZZO)

Cataclysmic Variables

U Geminorum (UG)
Supernovae (SN)
Novae (N)
Gamma-ray Bursts (GRB)
Symbiotic Variables (ZAND)

SS Cygni
SU Ursae Majoris
Z Camelopardalis
Type I Supernovae (SNI)
Type II Supernovae (SNII)
Fast Novae (NA)
Slow Novae (NB)
Very Slow Novae (NC)
Novalike Variables (NL)
Recurrent Novae (NR)
Long Gamma-ray Burst (LSB)
Soft Gamma-ray Repeater (SGR)
Short Gamma-ray Burst (SHB)

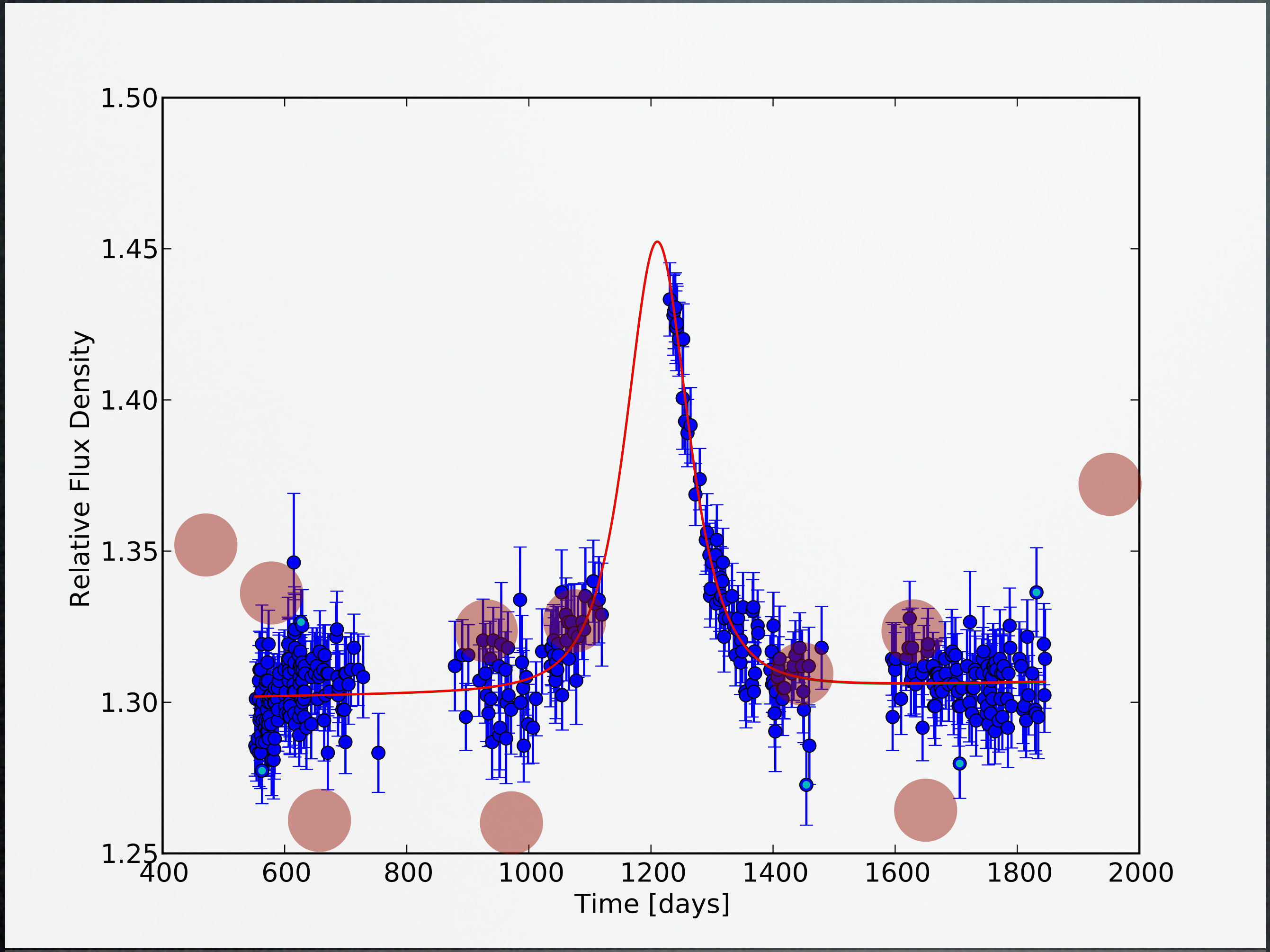
SNIa
SNIb
SNIc
SNIIL
SNIIN
SNIIP

Eclipsing

Systems with White Dwarfs (WD)
Semidetached (SD)
RS Canum Venaticorum (RS)
Planetary Nebulae (PN)
Contact Systems (K)
Systems with Supergiant(s) (GS)
Eclipsing Binary Systems (E)
Detached (D)
Detached - AR Lacertae (AR)
Wolf-Rayet Stars (WR)

Early (O-A) (KE)
W Ursa Majoris (KW)
Algol (Beta Persei) (EA)
Beta Lyrae (EB)
W Ursae Majoris (EW)
Main Sequence (DM)
With Subgiant (DS)
W Ursa Majoris (DW)

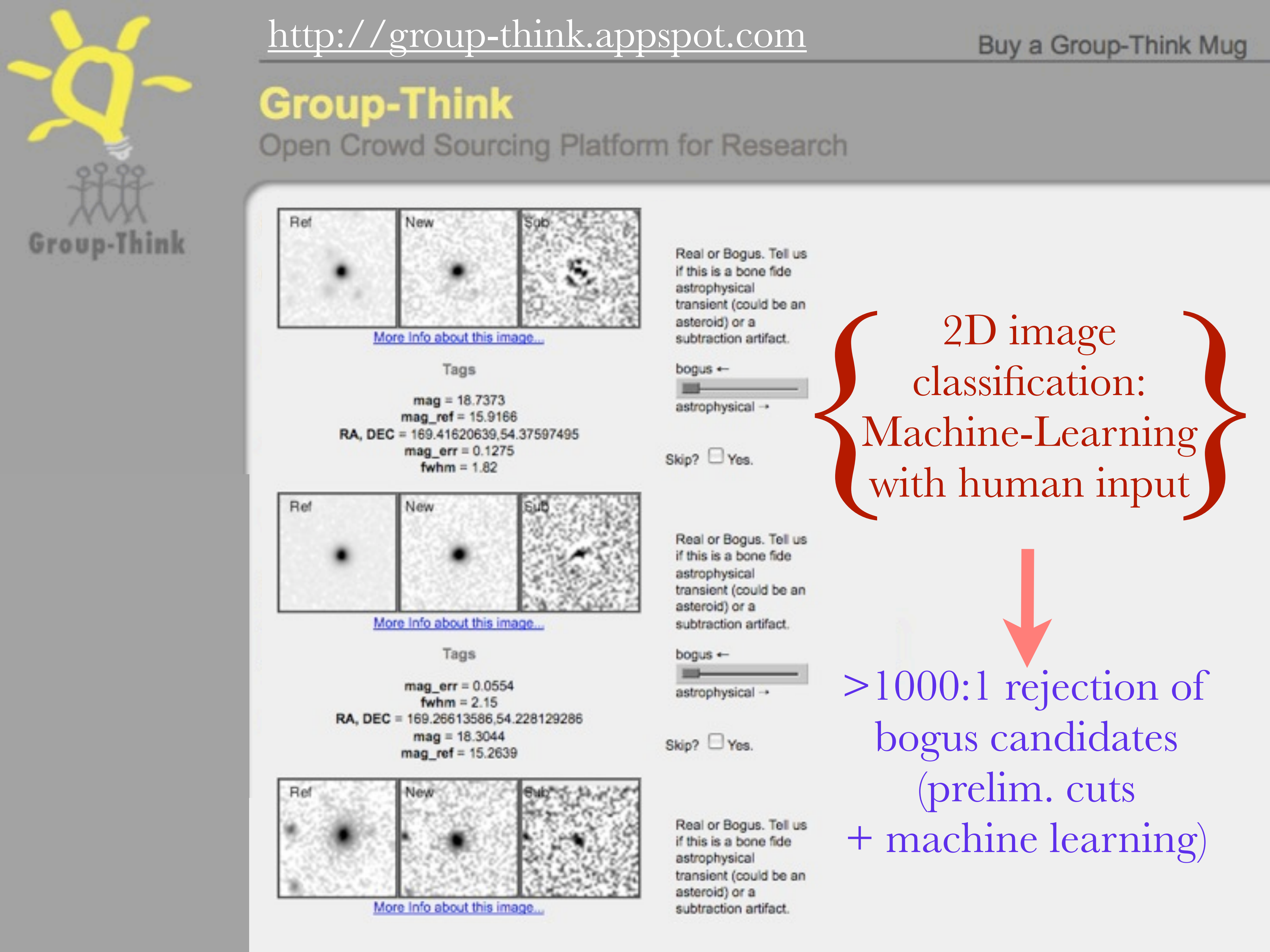
Considerable Complications with Time Series Data



- noisy, irregularly sampled
- spurious data
- telltale signature event may not have happened yet

class: *microlensing*

http://group-think.appspot.com



Group-Think

Open Crowd Sourcing Platform for Research

Tags

mag = 18.7373
mag_ref = 15.9166
RA, DEC = 169.41620639,54.37597495
mag_err = 0.1275
fwhm = 1.82

Real or Bogus. Tell us if this is a bone fide astrophysical transient (could be an asteroid) or a subtraction artifact.

Skip? □ Yes.

Tags

mag_err = 0.0554
fwhm = 2.15
RA, DEC = 169.26613586,54.228129286
mag = 18.3044
mag_ref = 15.2639

Real or Bogus. Tell us if this is a bone fide astrophysical transient (could be an asteroid) or a subtraction artifact.

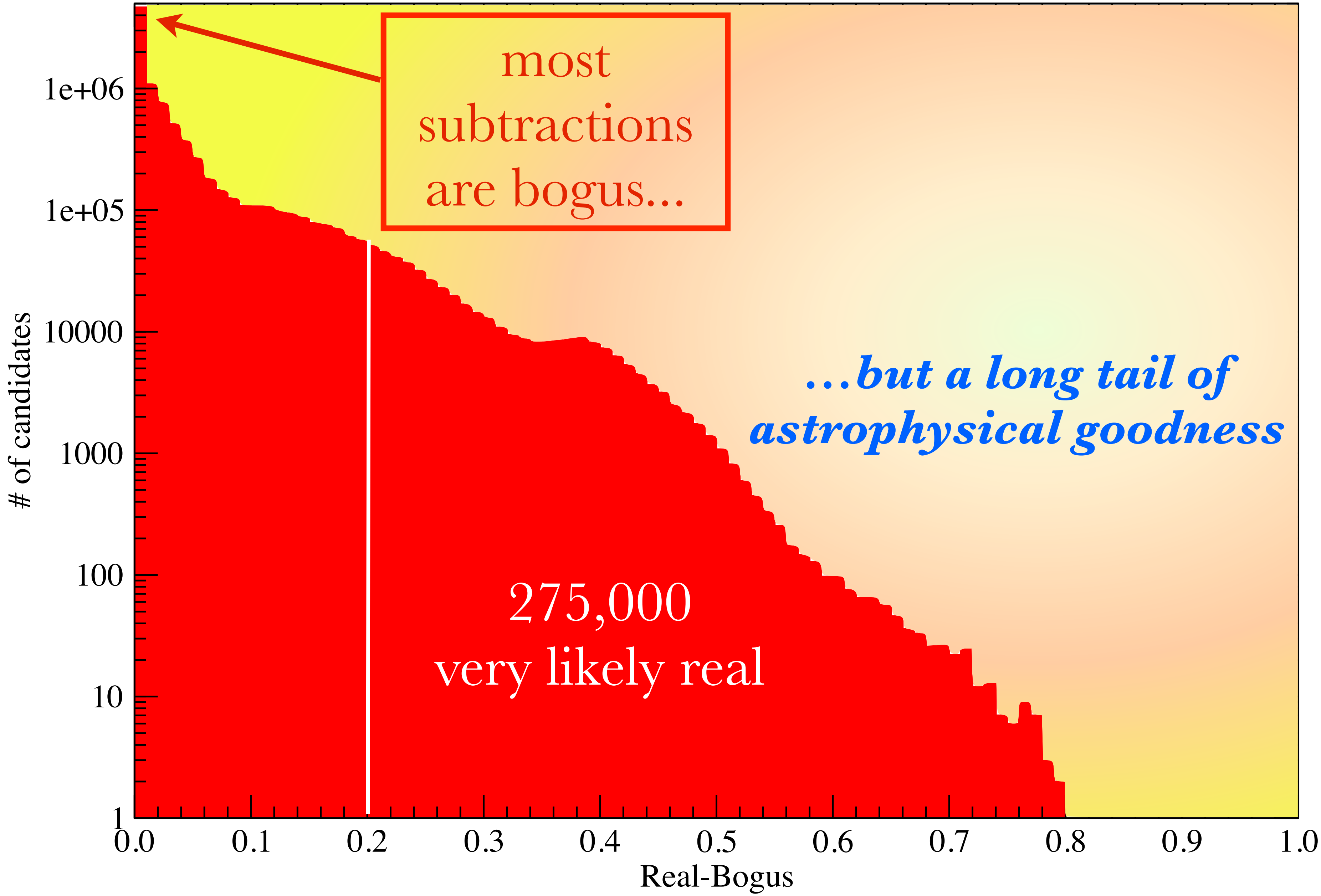
Skip? □ Yes.

Real or Bogus. Tell us if this is a bone fide astrophysical transient (could be an asteroid) or a subtraction artifact.

{ 2D image classification: Machine-Learning with human input }

↓

>1000:1 rejection of bogus candidates (prelim. cuts + machine learning)



most subtractions are bogus...

...but a long tail of astrophysical goodness

275,000
very likely real

10M PTF subtractions (1 month of data)

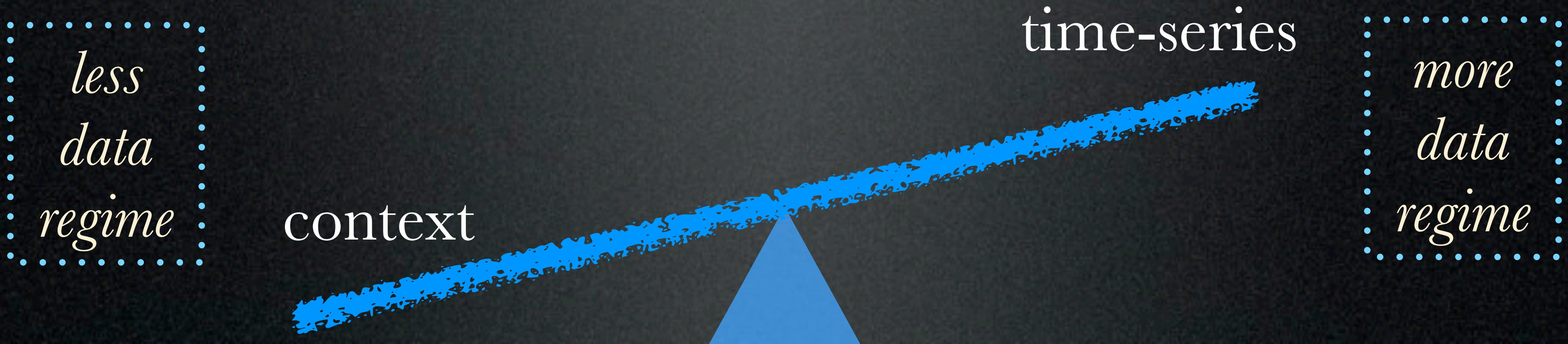
Major Challenge:

how do we use ***domain knowledge*** & ***known ("labelled") instances*** to create a classifier?

traditional fitting, machine learning, ...

Machine-Learning Approach to Classification

Data	Utility for Classification
Time Series (e.g. color, brightness change, etc.)	• comparison to previously observed sources, & theoretical/ numerical models •historical images: extend time baseline
Context (e.g. sky location, nearest galaxy type)	situational awareness: expectations of different classes



less data regime

time-series

context

more data regime

Feature Extraction: Homogenizing Heterogenous Data

“Features”: real-number metrics that describe the time-domain characteristics & context of a source.

variability metrics:

e.g. Stetson indices, χ^2/dof (constant hypothesis)

periodic metrics:

e.g. dominant frequencies in Lomb-Scargle, phase offsets between periods

shape analysis

e.g. skewness, kurtosis, Gaussianity

context metrics

e.g. distance to nearest galaxy, type of nearest galaxy, location in the ecliptic plane

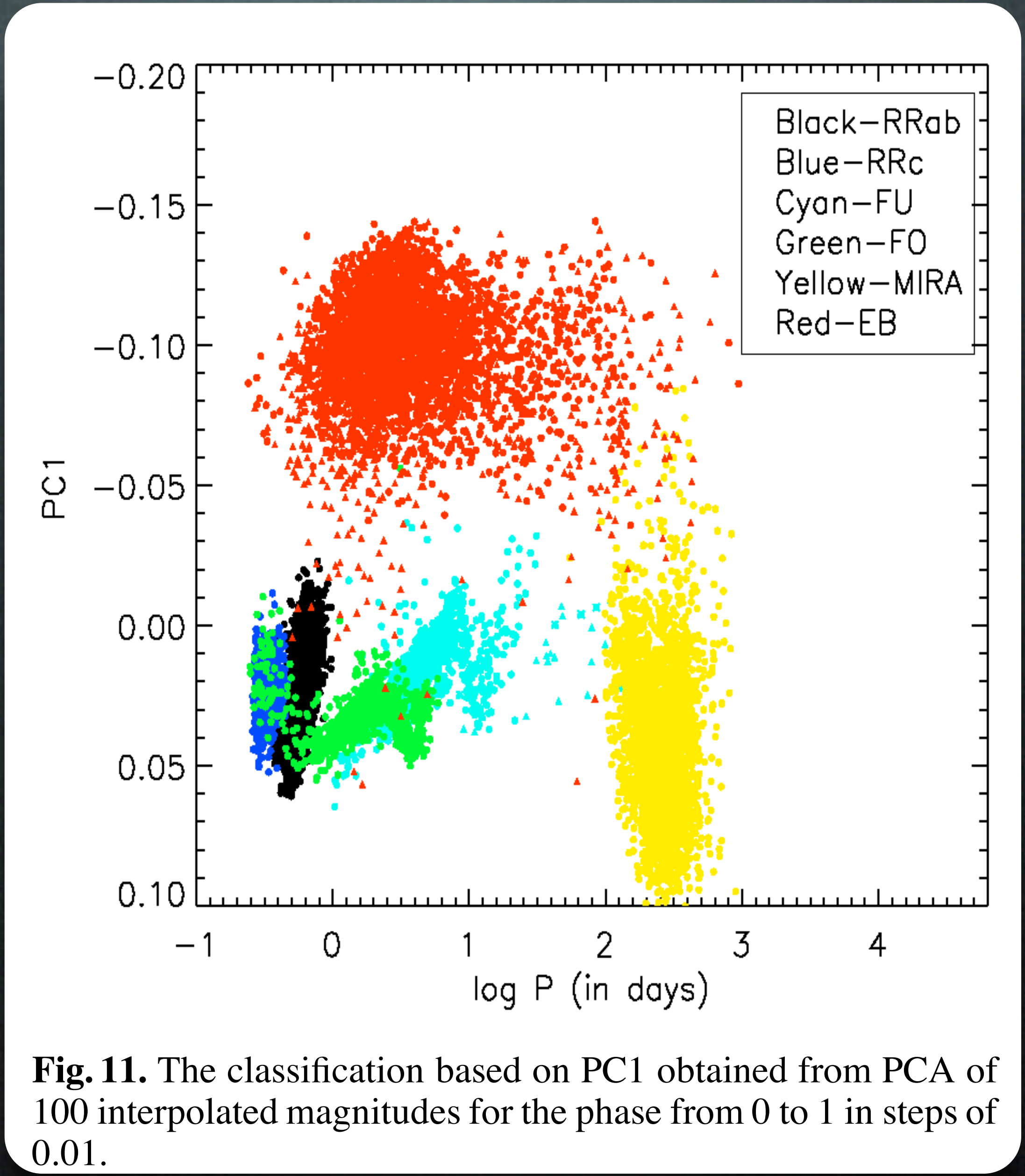


Fig. 11. The classification based on PC1 obtained from PCA of 100 interpolated magnitudes for the phase from 0 to 1 in steps of 0.01.

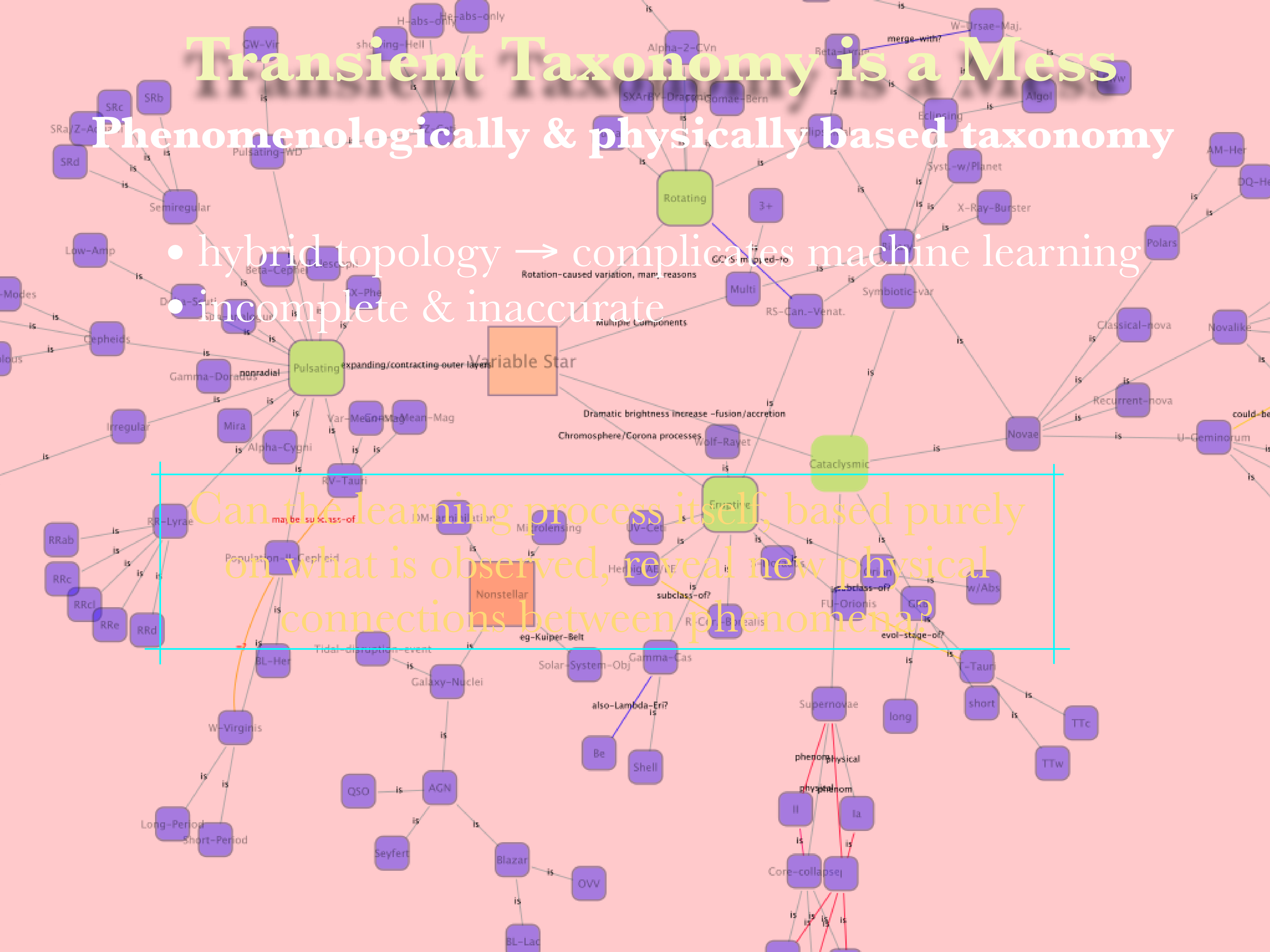
Deb & Singh+09

Transient Taxonomy is a Mess

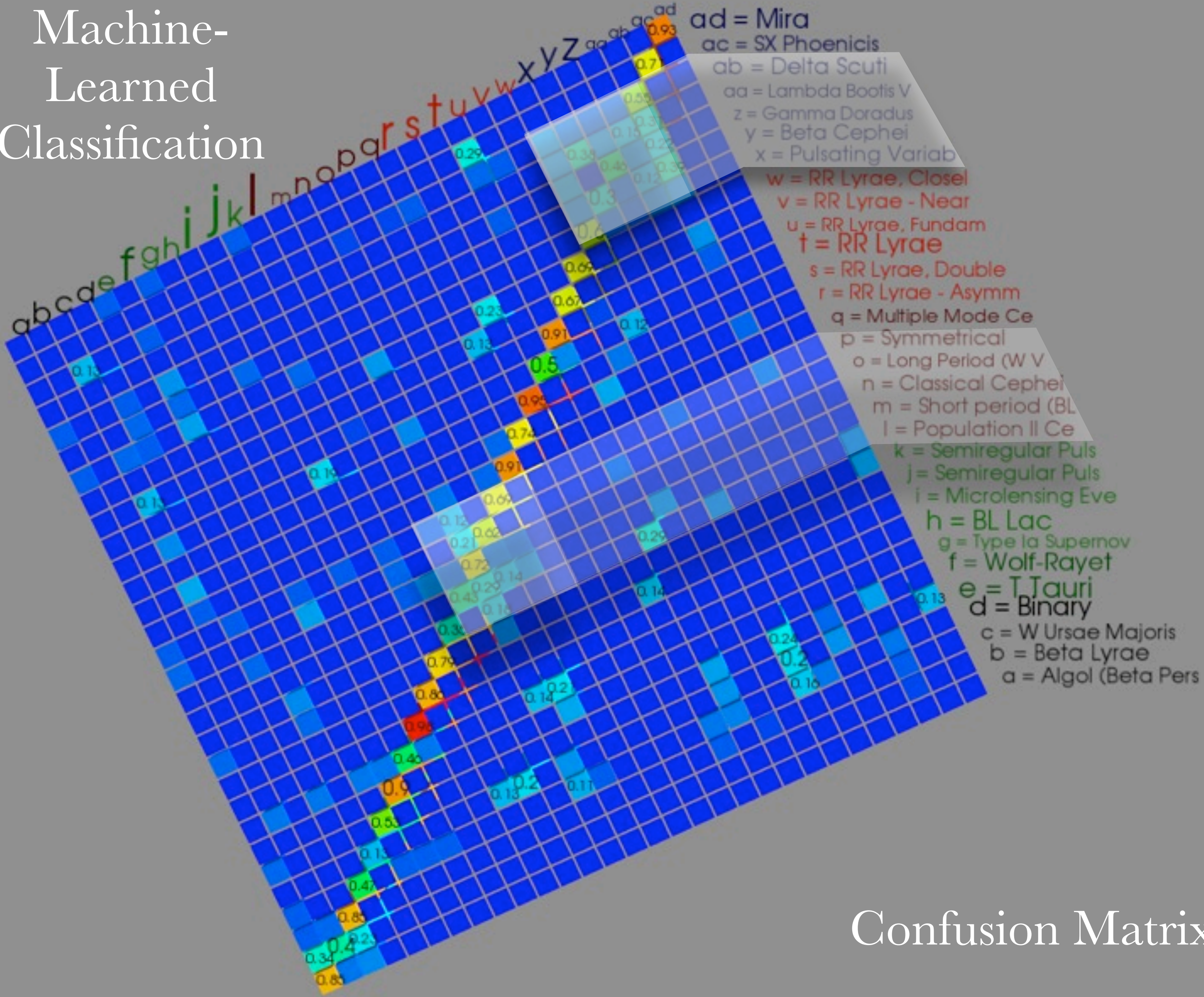
Phenomenologically & physically based taxonomy

- hybrid topology → complicates machine learning
- incomplete & inaccurate

Can the learning process itself, based purely on what is observed, reveal new physical connections between phenomena?



Machine-Learned Classification



Confusion Matrix

1. Parallelize the Learning Phase of Machine Learning

Problem:

frameworks like Weka (`http://www.cs.waikato.ac.nz/ml/weka/`) are not natively parallel. We will need to burst out training requests on specific time/observation vectors & classify quickly with the results

Solution:

build a parallel platform for weka

(GridWeka, Weka-parallel etc. are out of date & probably not elegant)

- develop/adapt Mahout (http://lucene.apache.org/mahout/), ML for Hadoop

http://userweb.port.ac.uk/~khusainr/weka

1. Parallelize the Learning Phase of Machine Learning

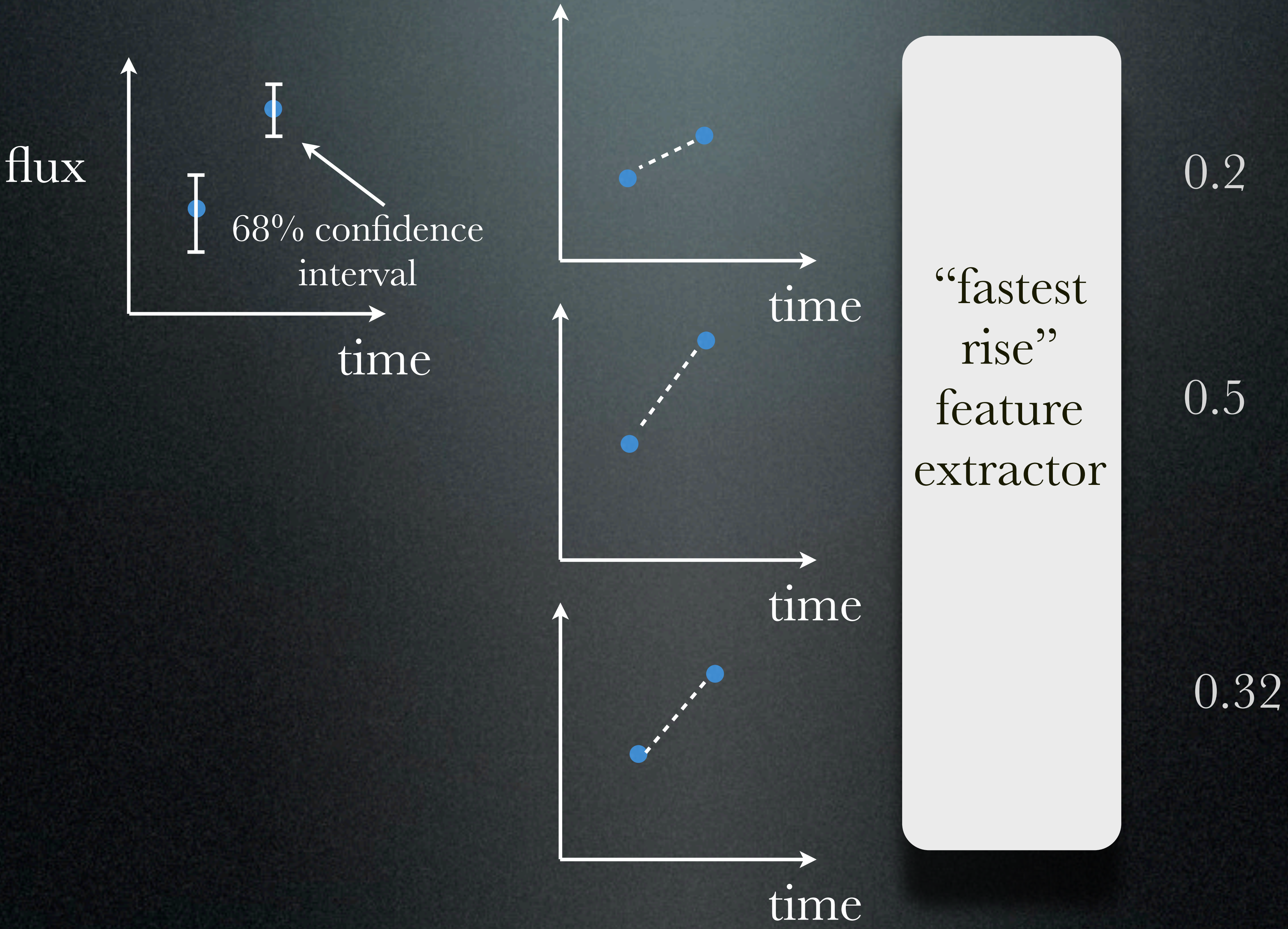
Problem:

we have errors on our data (both training sets and instances) & we dont know how to deal with them

Sledgehammer Solution:

use a parallel platform to generate distribution of trained models & apply to distribution of instance-based sets

1. Parallelize the Learning Phase of Machine Learning



68% confidence interval

“fastest rise” feature extractor

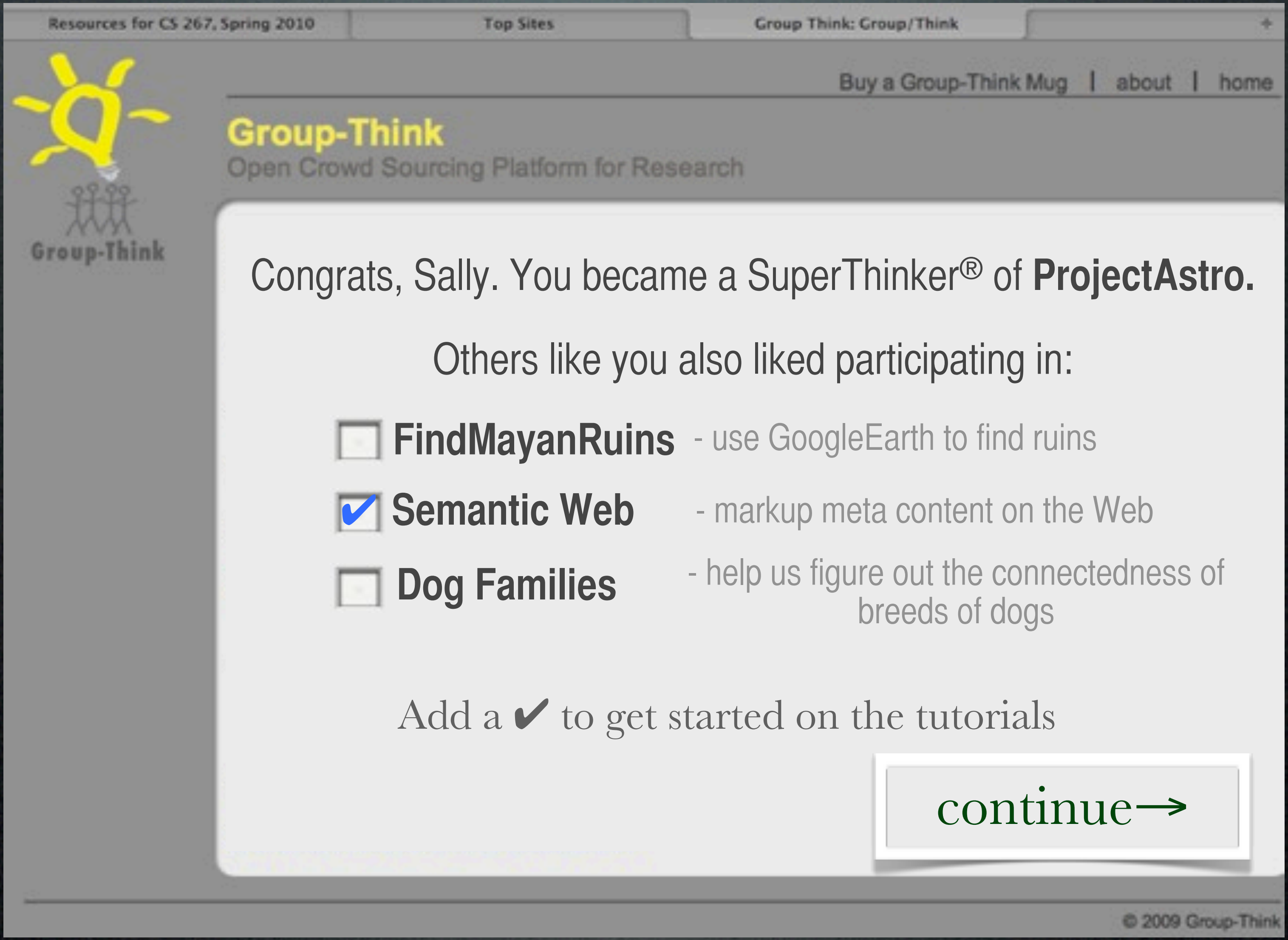
0.2

0.5

0.32

2. Build a General Crowdsourcing Platform (GroupThink2.0)

- production scale site (GoogleAppEngine or elsewhere), allowing interconnection of projects


Resources for CS 267, Spring 2010
Top Sites
Group Think: Group/Think
Buy a Group-Think Mug | about | home
Group-Think
Group-Think
Open Crowd Sourcing Platform for Research

Congrats, Sally. You became a SuperThinker® of **ProjectAstro.**

Others like you also liked participating in:

□ **FindMayanRuins** - use GoogleEarth to find ruins

☑ **Semantic Web** - markup meta content on the Web

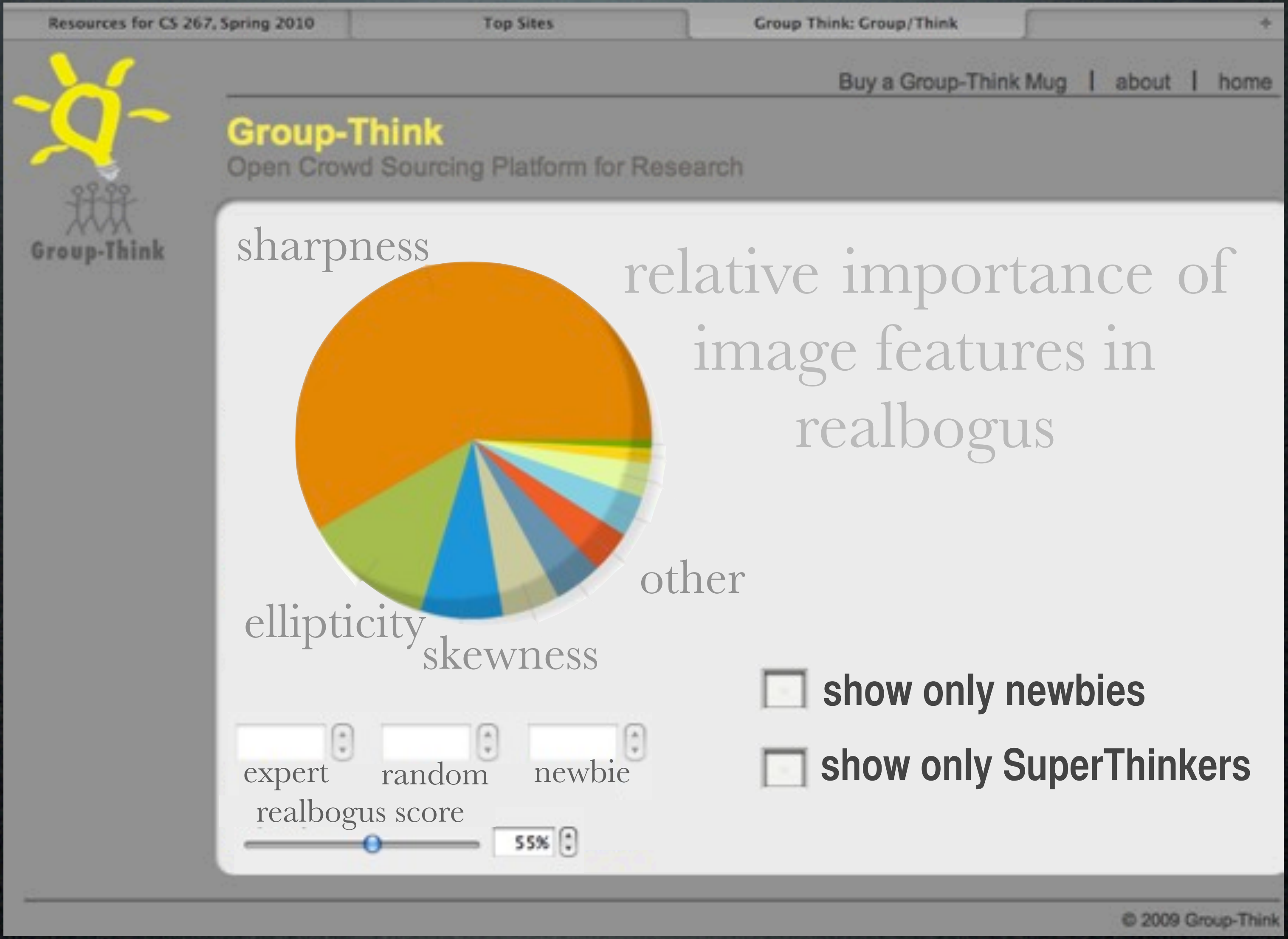
□ **Dog Families** - help us figure out the connectedness of breeds of dogs

Add a ✔ to get started on the tutorials

continue→
© 2009 Group-Think

2. Build a General Crowdsourcing Platform (GroupThink2.0)

- build innovative analytics plugins for projects;
- could require grid/cloud-based analysis for on-the-fly results

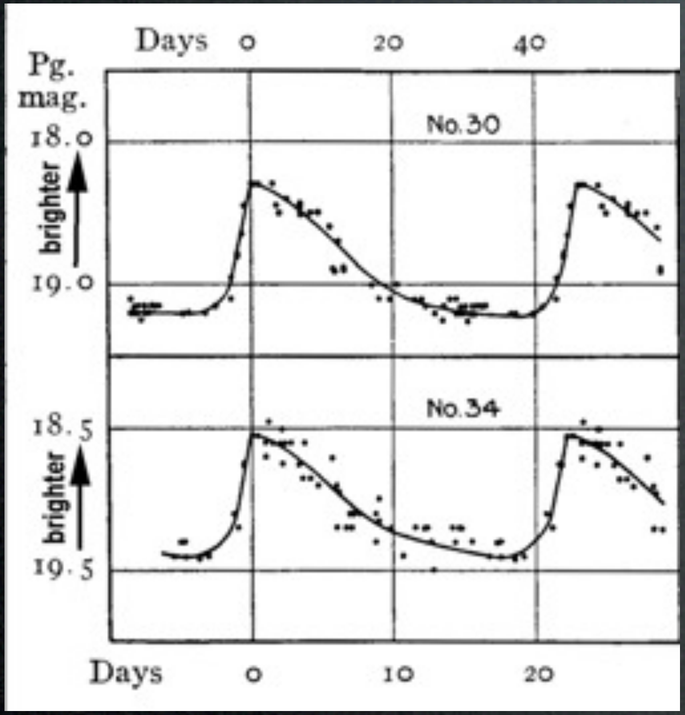

Resources for CS 267, Spring 2010
Top Sites
Group Think: Group/Think
Buy a Group-Think Mug | about | home
Group-Think
Group-Think
Open Crowd Sourcing Platform for Research
sharpness
relative importance of image features in realbogus
other
ellipticity
skewness
expert
random
newbie
realbogus score
55%
© 2009 Group-Think

☐ show only newbies

☐ show only SuperThinkers

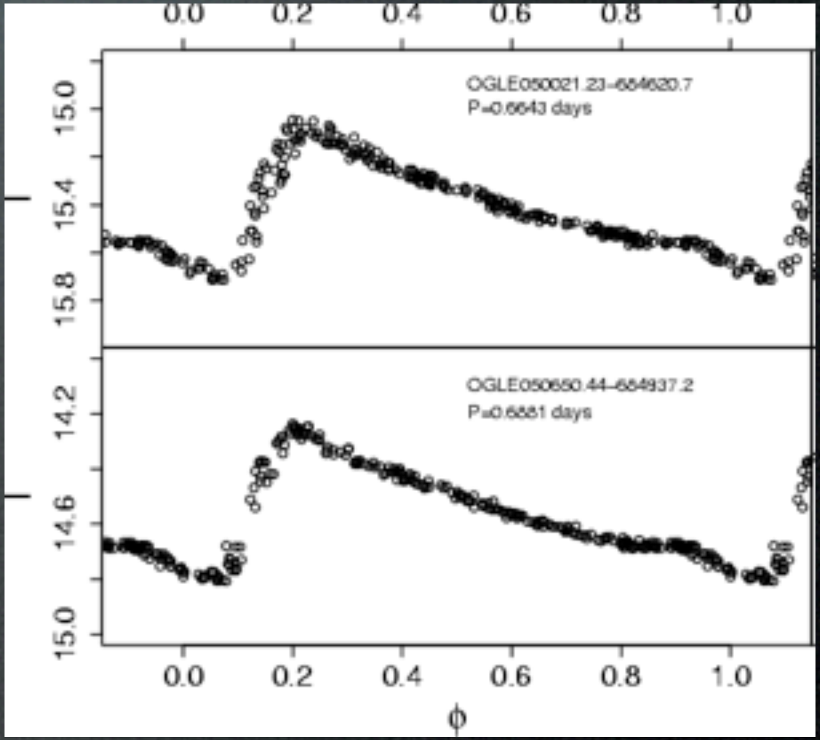
3. Parallelized Genetic Programming for Feature Discovery

Instead of handcoding "features" for ML, using GP (in parallelized environment) to ***discover features*** which give the best classification



Cepheid

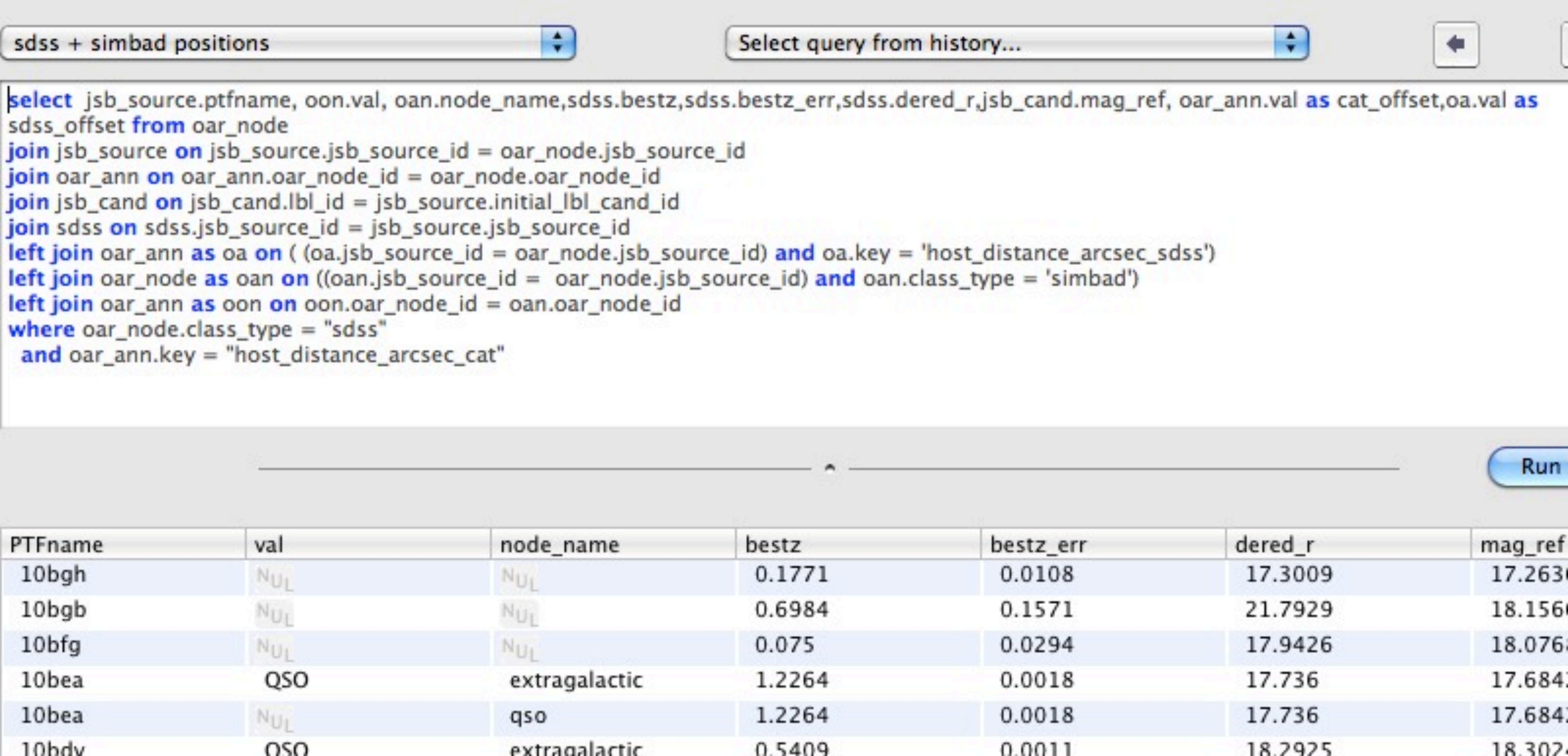
vs.



RR Lyrae

4. Parallelized Visual Exploration Tool

allow the armchair astronomer to ask complex questions of the databases & visualize and interact with the results (100M+ rows)



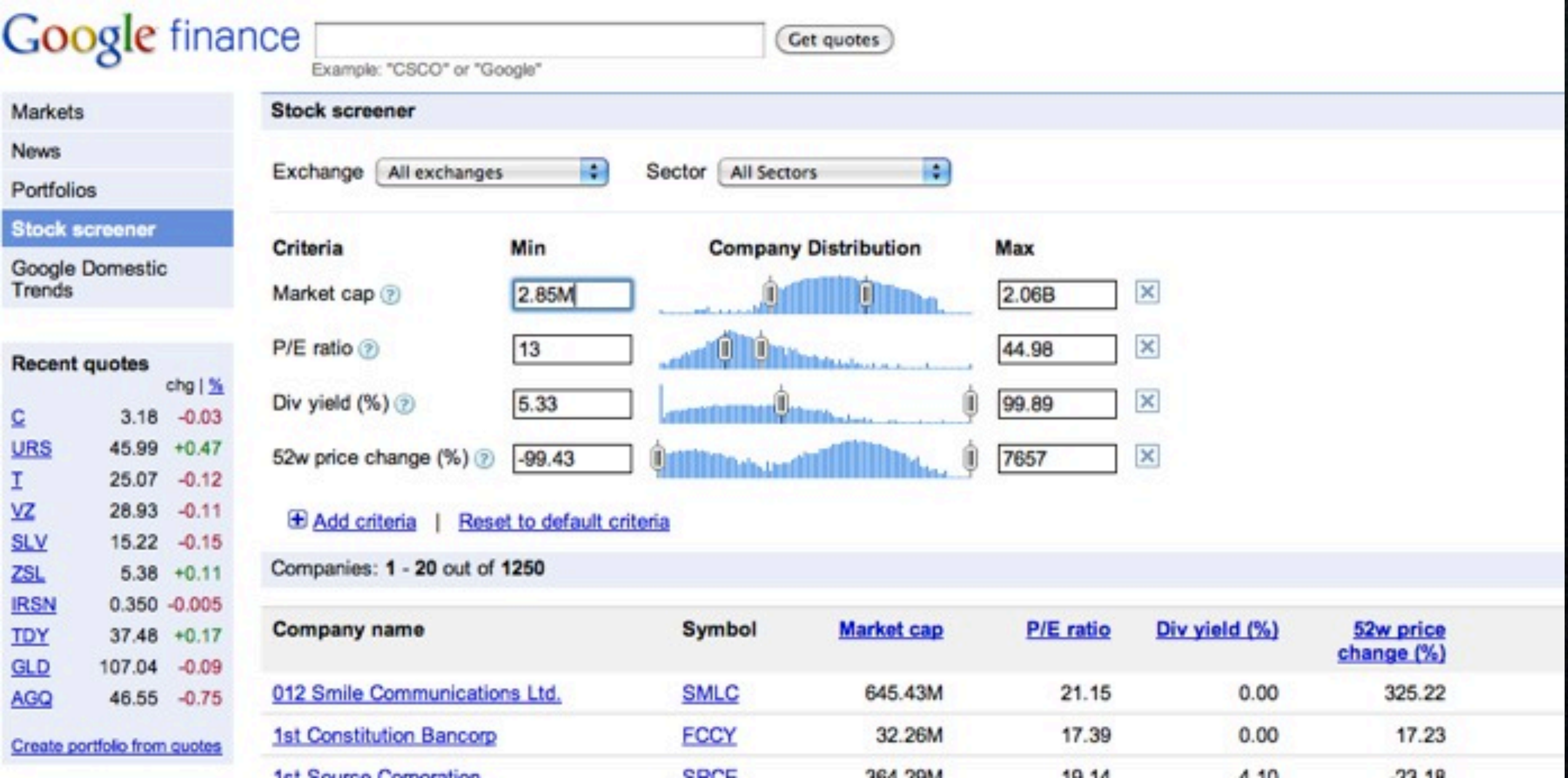
```
select  jsb_source.ptfname, oon.val, oan.node_name,sdss.bestz,sdss.bestz_err,sdss.dered_r,jsb_cand.mag_ref, oar_ann.val as cat_offset,oa.val as
sdss_offset from oar_node
join jsb_source on jsb_source.jsb_source_id = oar_node.jsb_source_id
join oar_ann on oar_ann.oar_node_id = oar_node.oar_node_id
join jsb_cand on jsb_cand.lbl_id = jsb_source.initial_lbl_cand_id
join sdss on sdss.jsb_source_id = jsb_source.jsb_source_id
left join oar_ann as oa on ( (oa.jsb_source_id = oar_node.jsb_source_id) and oa.key = 'host_distance_arcsec_sdss')
left join oar_node as oan on ((oan.jsb_source_id =  oar_node.jsb_source_id) and oan.class_type = 'simbad')
left join oar_ann as oon on oon.oar_node_id = oan.oar_node_id
where oar_node.class_type = "sdss"
  and oar_ann.key = "host_distance_arcsec_cat"
```

PTFname	val	node_name	bestz	bestz_err	dered_r	mag_ref
10bgh	NUL	NUL	0.1771	0.0108	17.3009	17.263
10bgb	NUL	NUL	0.6984	0.1571	21.7929	18.156
10bfg	NUL	NUL	0.075	0.0294	17.9426	18.076
10bea	QSO	extragalactic	1.2264	0.0018	17.736	17.684
10bea	NUL	qso	1.2264	0.0018	17.736	17.684
10bdy	QSO	extragalactic	0.5409	0.0011	18.2925	18.302

4. Parallelized Visual Exploration Tool

allow the armchair astronomer to ask complex questions of the databases & visualize and interact with the results

- parallel database calls with embedded custom code (e.g. Hadoop SQL “hive”)



Recent quotes

	chg | %
C	3.18	-0.03
URS	45.99	+0.47
T	25.07	-0.12
VZ	28.93	-0.11
SLV	15.22	-0.15
ZSL	5.38	+0.11
IRSN	0.350	-0.005
TDY	37.48	+0.17
GLD	107.04	-0.09
AGQ	46.55	-0.75

Create portfolio from quotes

Stock screener

Exchange: All exchanges Sector: All Sectors

Criteria	Min	Company Distribution	Max
Market cap	2.85M		2.06B
P/E ratio	13		44.98
Div yield (%)	5.33		99.89
52w price change (%)	-99.43		7657

Companies: 1 - 20 out of 1250

Company name	Symbol	Market cap	P/E ratio	Div yield (%)	52w price change (%)
012 Smile Communications Ltd.	SMLC	645.43M	21.15	0.00	325.22
1st Constitution Bancorp	FCCY	32.26M	17.39	0.00	17.23
1st Source Corporation	SRCE	364.29M	19.14	4.10	23.18

Resources

1. dotastro.org

2. Harvard TimeSeries Center:

   http://timemachine.iic.harvard.edu/

3. “The Fourth Paradigm: Data-Intensive Scientific Discovery”

   http://research.microsoft.com/en-us/collaboration/fourthparadigm/