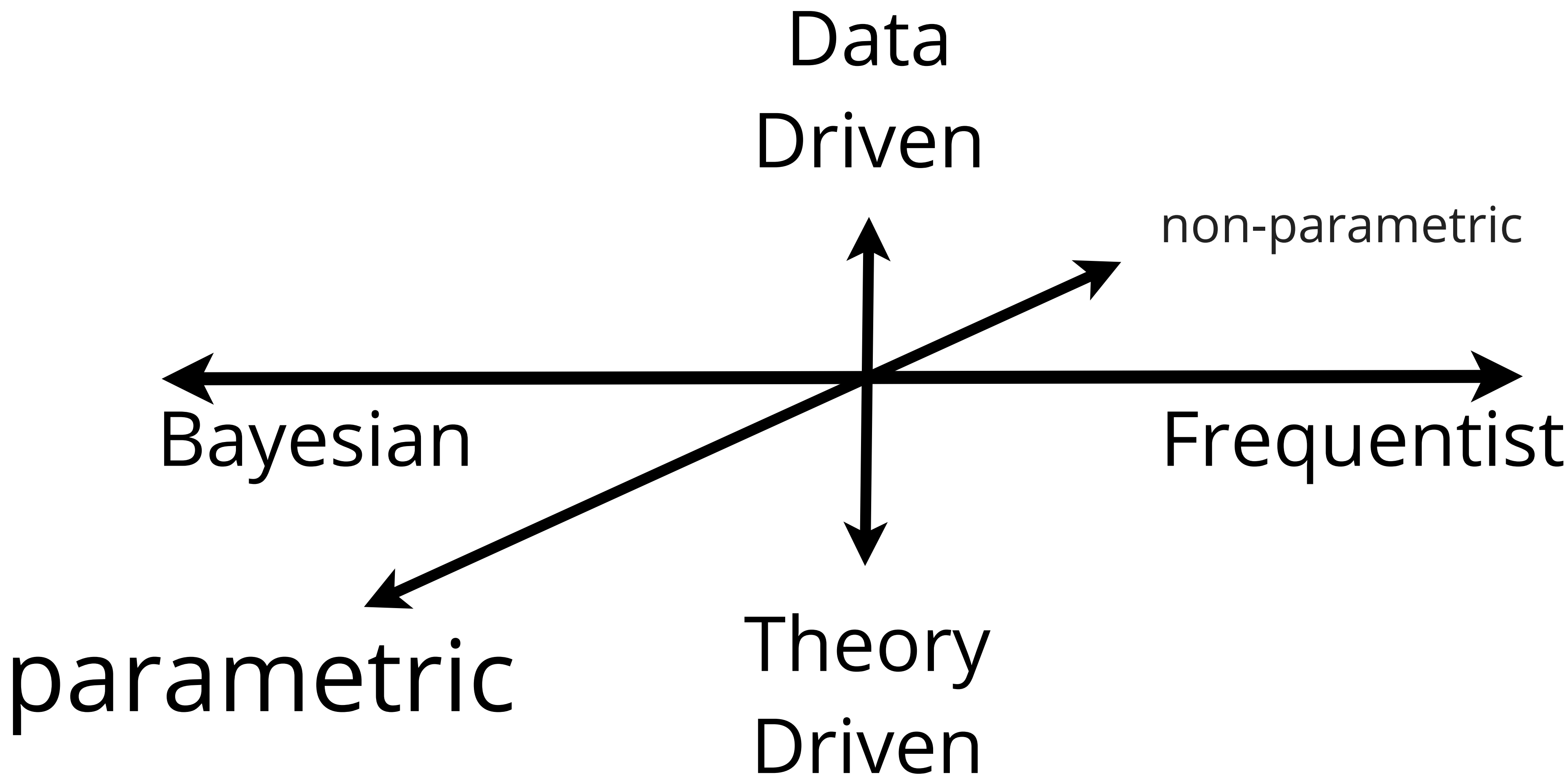
Data-Driven Astronomical Inference

Josh Bloom, UC Berkeley

@profjsb

Computational Astrophysics 2014–2020: Approaching Exascale   LBNL 21 March 2014

Inference Space



Hardware	laptops → NERSC
Software	Python/Scipy, R, ...
Carbonware	astro grad students, postdocs

Bayesian Distance Ladder

Pulsational Variables: Period-Luminosity Relation

$$\begin{aligned} m_{ij} = \mu_i &+ M_{0j} \\ &+ \alpha_j \log_{10}(P_i/P_0) \\ &+ E(B-V)_i \times [R_V \times a(1/\lambda_i) + b(1/\lambda_i)] \\ &+ \epsilon_{ij} \end{aligned}$$

i indexes over individual stars
j indexes over wavebands
a and *b* are fixed constants at each waveband

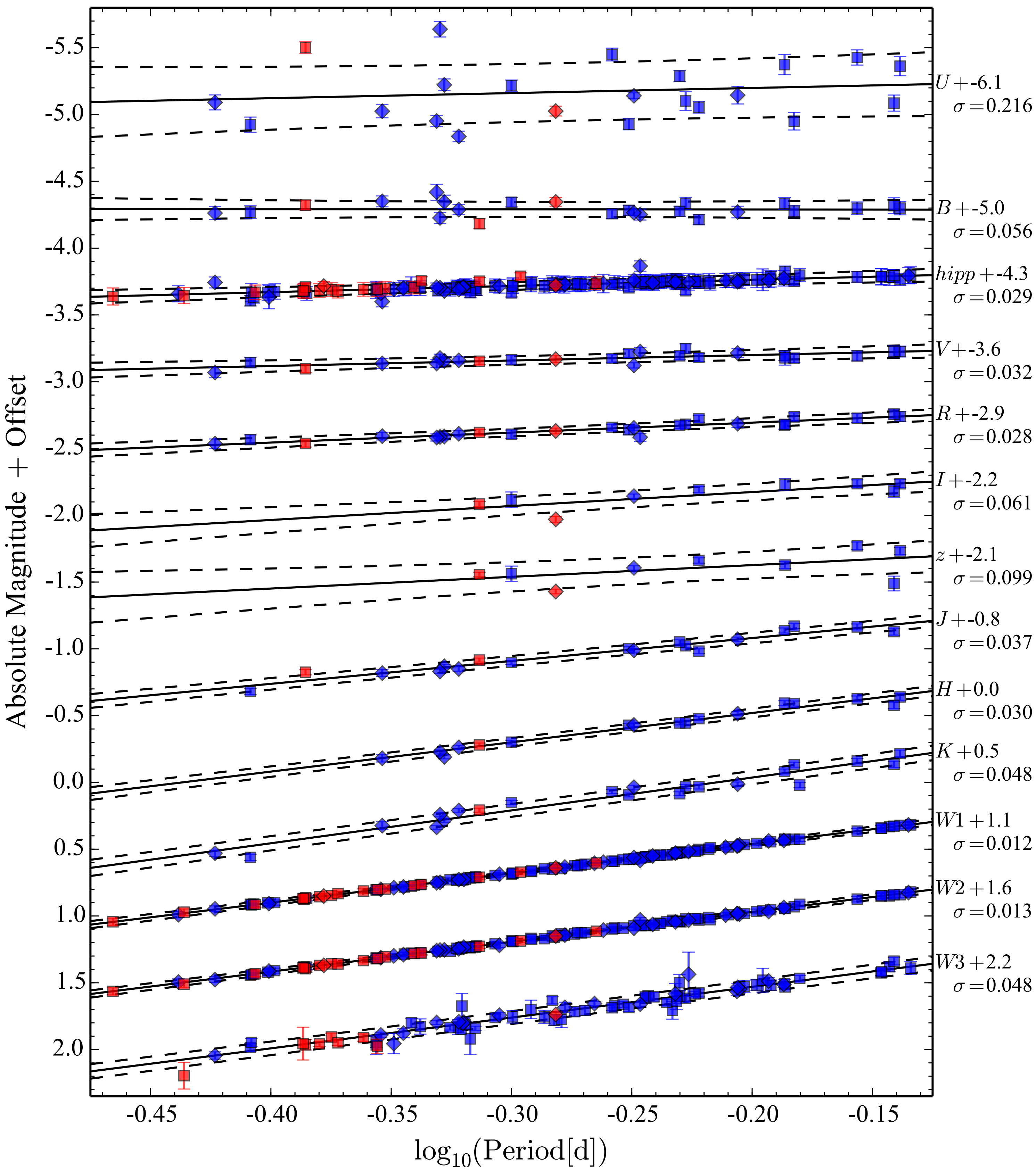
Data 134 RR Lyrae (WISE, Hipparcos, UVIORJHK)

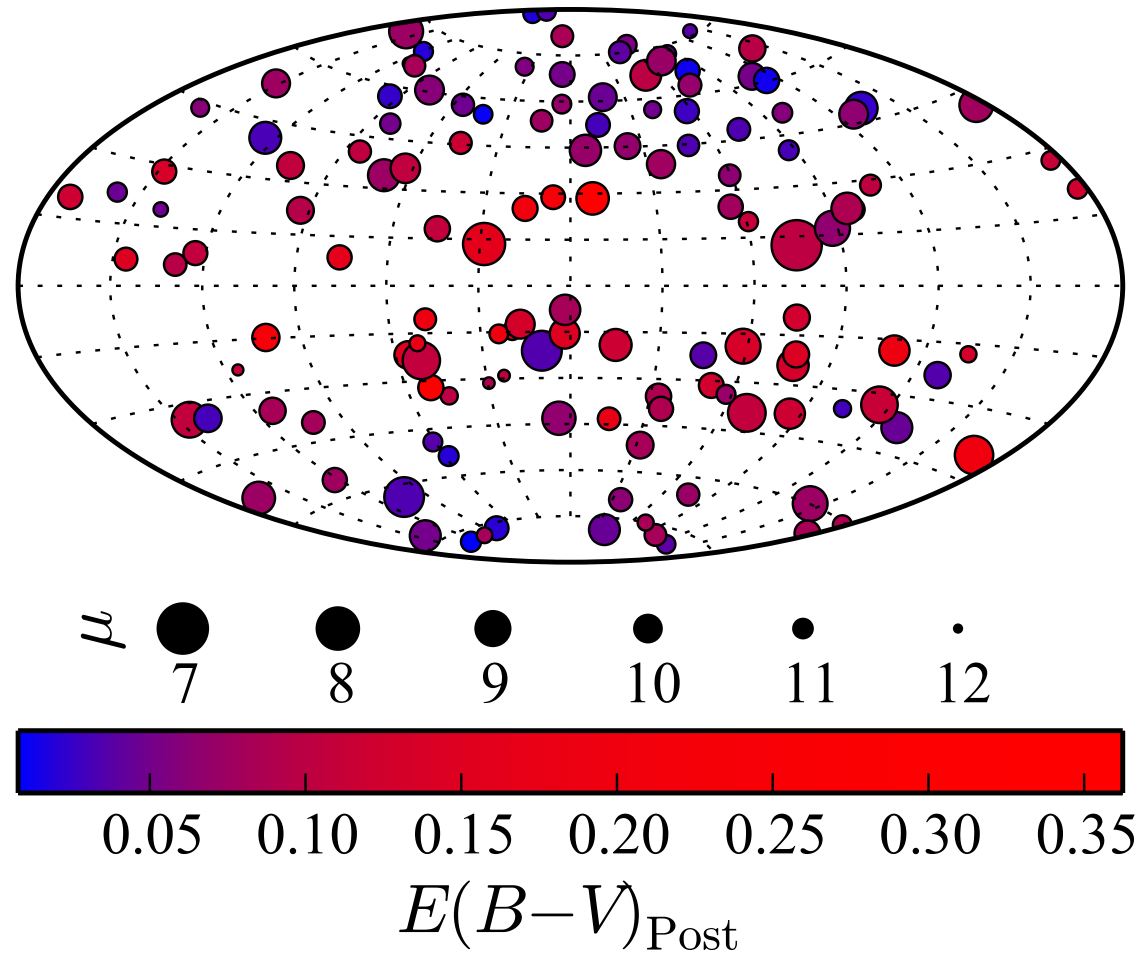
Fit 307 dimensional model parameter inference

- deterministic MCMC model
- ~6 days for a single run (one core)
- parallelism for convergence tests

Klein+12; Klein,JSB+14,

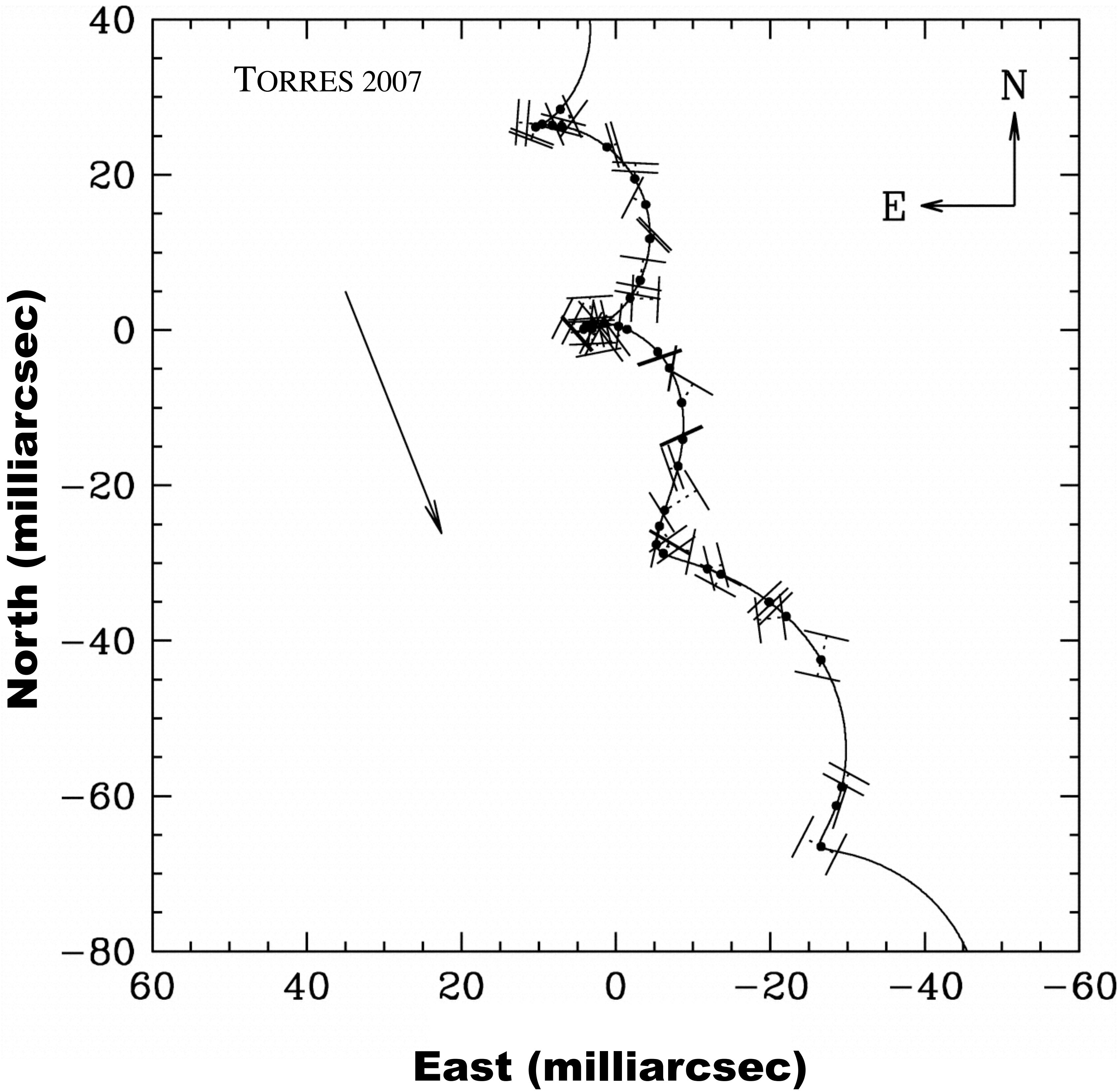
Bayesian Distance Ladder





- Approaching 1% distance uncertainty
- Precision 3D dust measurements

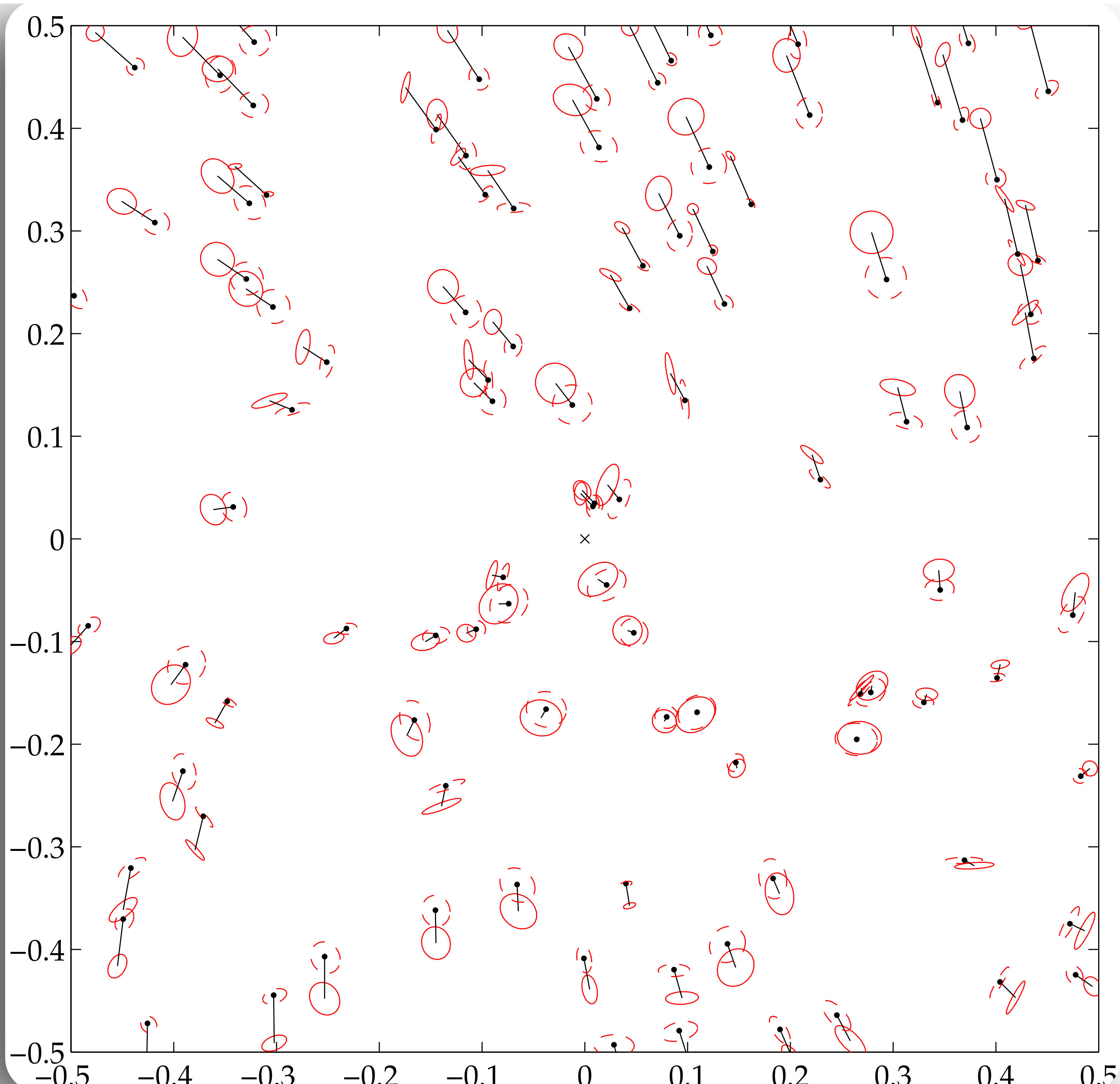
Bayesian Astrometry



Fitting for: Parallax, Proper Motion, Binary Parameters, Microlensing...

Hipparcos: 10^6 stars
Gaia: $\sim 10^9$ stars

Bayesian Astrometry



Step 1:

Regress 7-d parametric affine transformation (scale, rotation, shear, etc.)

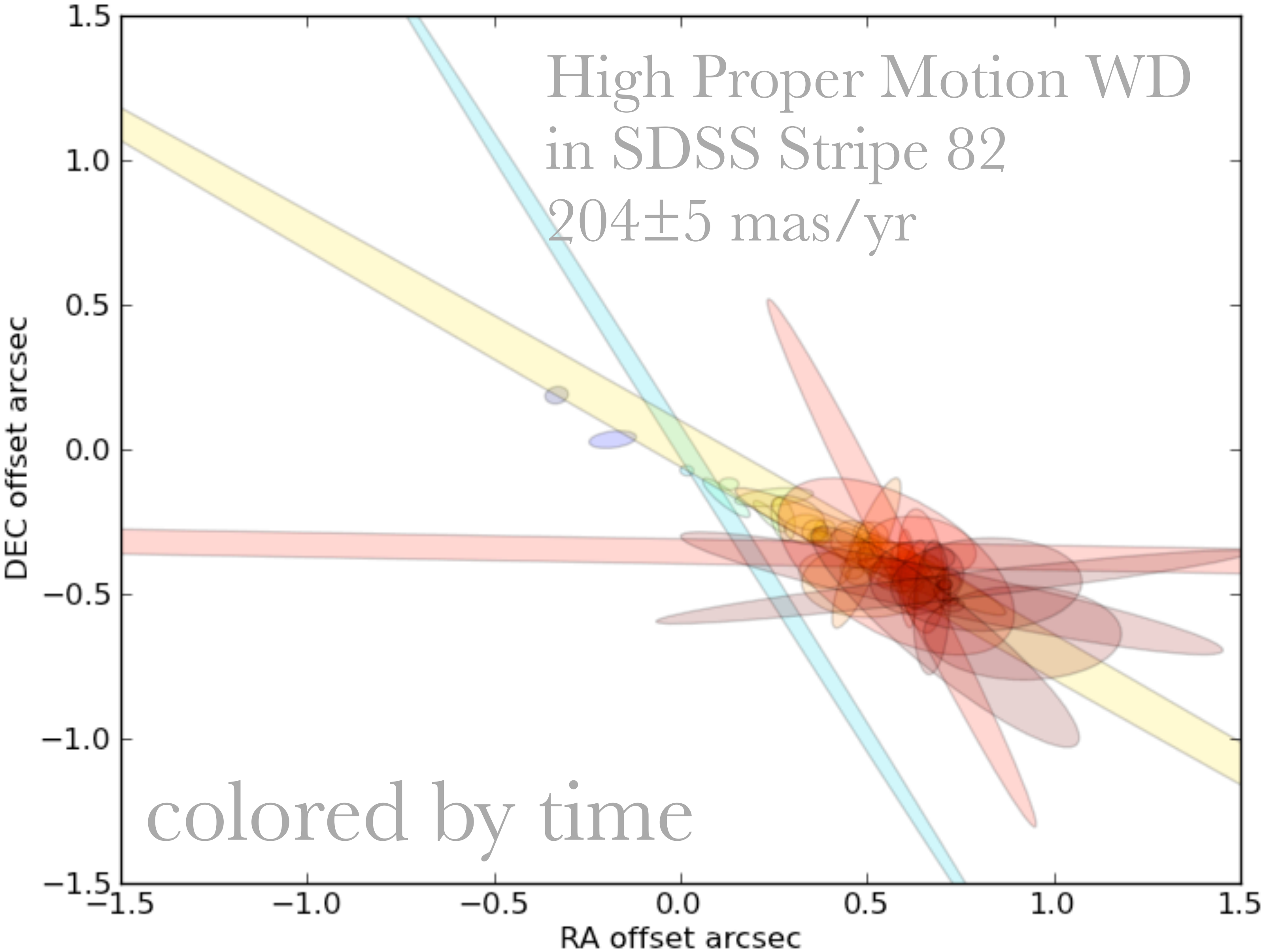
Step 2:

Learn a non-parametric distortion map with Gaussian processes

http://berianjames.github.com/pyBAST/

James,JSB+14

Bayesian Astrometry



Some Clear Benefits

- covariate uncertainties in celestial coordinates
- mapping observed points can incorporate variance throughout image, extending even to highly non-trivial distortion effects
- astrometry can be treated as *Bayesian updating*, allowing incorporation of prior knowledge about proper motion & parallax

non-parallel Cholesky + MCMC: ~hour for 71 observations

http://berianjames.github.com/pyBAST/

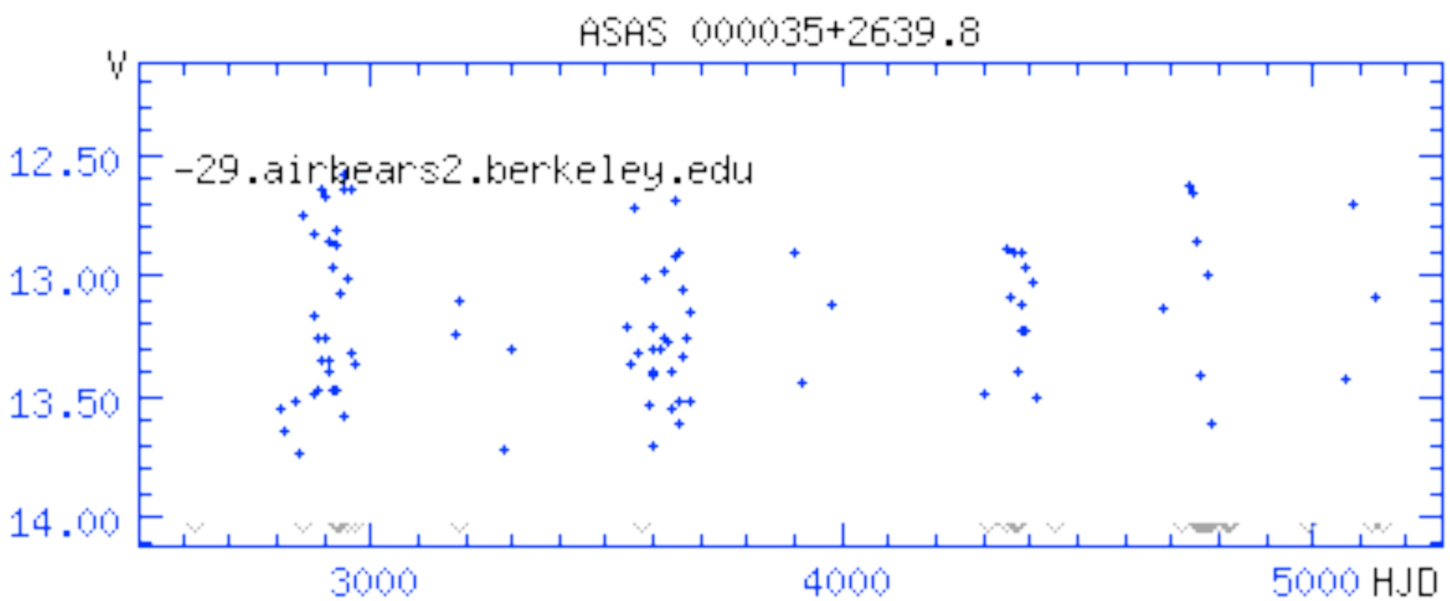
Fork me on GitHub

Machine Learned Classification

25-class variable star

Data: 50k from ASAS, 810 with known labels

(timeseries, colors)



$\longrightarrow$ $P_{RRL} = 0.94$

Richards+12

Machine Learned Classification

True Class

Predicted Class

	a	b1	b2	b3	b4	c	d	e	f	g	h	i	j	j1	l	o	p	q	r1	s1	s2	s3	t	u	v	w	x	y
a. Mira	0.923	0.015				0.042											0.057		0.176									
b1. Semireg PV	0.066	0.824		0.276	0.111	0.125											0.086			0.25		0.059	0.375					
b2. SARG A			0.6	0.034	0.019																	0.059						
b3. SARG B		0.015	0.267	0.586																0.25	0.05			1				
b4. LSP	0.011	0.074	0.067	0.069	0.87											0.2	0.171					0.059						
c. RV Tauri		0.029				0.75	0.011												0.059				0.125					0.015
d. Classical Cepheid							0.955	0.538	0.25											0.25	0.1							
e. Pop. II Cepheid						0.042		0.308	0.25																			
f. Multi. Mode Cepheid							0.011	0.077	0.5	0.012								0.043			0.1	0.059			1			
g. RR Lyrae, FM							0.011	0.077		0.965		0.5		1														
h. RR Lyrae, FO										0.012	0.897		0.036															0.015
i. RR Lyrae, DM												0.5																
j. Delta Scuti											0.034		0.786		0.278													0.015
j1. SX Phe																												
l. Beta Cephei													0.107		0.722													0.015
o. Pulsating Be			0.067													0.4						0.059						
p. RSG		0.044															0.686						0.125					
q. Chem. Peculiar													0.036			0.2		0.913										0.015
r1. RCB																			0.706									
s1. Class. T Tauri																												
s2. Weak–line T Tauri				0.034			0.011											0.043			0.7	0.118					0.061	
s3. RS CVn																					0.05	0.059				0.037		
t. Herbig AE/BE																0.2							0.375					
u. S Doradus																												
v. Ellipsoidal																												
w. Beta Persei																						0.353				0.889	0.182	
x. Beta Lyrae																			0.059			0.118				0.074	0.667	0.044
y. W Ursae Maj.						0.042				0.012	0.069		0.036							0.25		0.059					0.091	0.882
	91	68	15	29	54	24	89	13	4	86	29	2	28	1	18	5	35	23	17	4	20	17	8	1	1	27	33	68

74 dimensional feature set for learning

featurization is the bottleneck (but embarrassingly parallel)

Richards+12

SELECT CLASS:

All

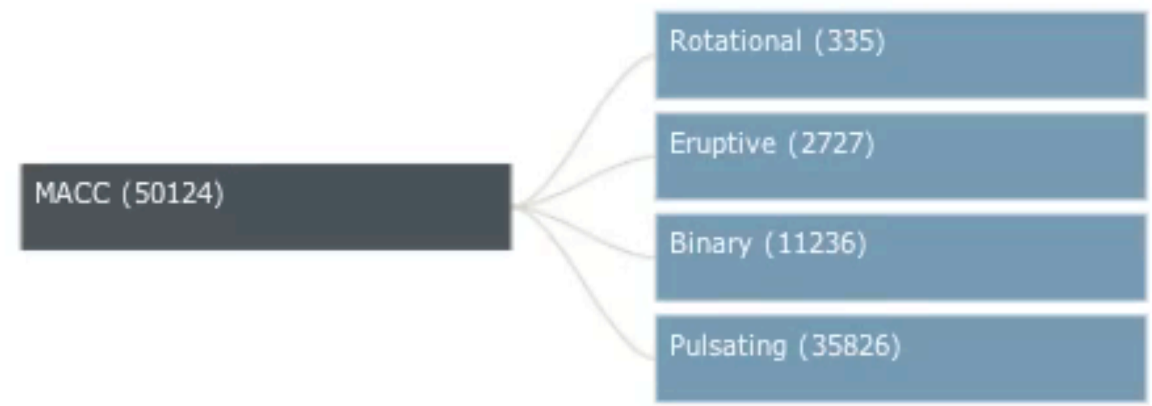
Specify RA/Dec

Customize Table

Export Query Results

Displaying sources 1 to 500

Previous 500 | Next 500



ASAS_ID	dotAstro_ID	RA	DEC	Class	P_Class	Anomaly	ACVS_Class	Train_Class	P	P_signif	N_epochs	V
080940-3810.5	227867	122.415885	-38.174595	Mira		0.040	MIRA=SR	Mira	328.386	20.859	456	7.8
115501-5915.2	236087	178.762155	-59.258671	Mira		0.075	MIRA		200.167	21.082	339	8.2
132500-6439.8	238210	201.24495	-64.663232	Mira		0.066	MIRA		350.507	23.531	503	10.34
161441-3223.5	244080	243.671565	-32.391181	Mira		0.037	MIRA		358.451	16.717	563	9.99
165413-5615.9	245810	253.55541	-56.265033	Mira		0.033	MIRA	Mira	286.697	16.654	497	9.46
165538-4506.2	245884	253.907535	-45.102913	Mira		0.040	MIRA		316.996	23.126	402	7.82
194952+0923.8	258863	297.46893	9.401204	Mira		0.042	MIRA		287.103	12.820	470	10.06
221800-2936.2	263989	334.501365	-29.604124	Mira		0.055	MIRA		293.459	16.501	415	9.17
235627-4947.2	265240	359.121555	-49.786453	Mira		0.066	MIRA		266.197	16.322	408	8.69
044030-3814.2	218251	70.125405	-38.235209	Mira		0.029	MIRA	Mira	390.997	20.138	382	8.98
070729+0459.1	224085	106.87182	4.986432	Mira		0.050	MIRA		262.462	13.700	480	10.02
091646-0435.2	230899	139.19409	-4.585538	Mira		0.050	MIRA		268.569	20.205	907	9.56
094755-6726.9	232018	146.98116	[illegible]	Mira		0.092	MIRA		[illegible]	18.072	290	10.62
103823-8046.8	233741	[illegible]	[illegible]	Mira		0.082	MIRA		[illegible]	23.139	823	10.1
120517-5511.2	236361	[illegible]	[illegible]	Mira		[illegible]	MIRA		[illegible]	18.337	574	9.75
121938-1915.3	236646	[illegible]	[illegible]	Mira		0.075	MIRA		[illegible]	18.677	615	7.83

MACC ABOUT CONTACT FAQ

Machine-learned varstar catalog: **http://bigmacc.info**

Doing Science with Probabilistic Catalogs

<u>Demographics</u> (with little followup):

trading high purity at the cost of lower efficiency

e.g., using RRL to find new Galactic structure

<u>Novelty Discovery</u> (with lots of followup):

trading high efficiency for lower purity

e.g., discovering new instances of rare classes

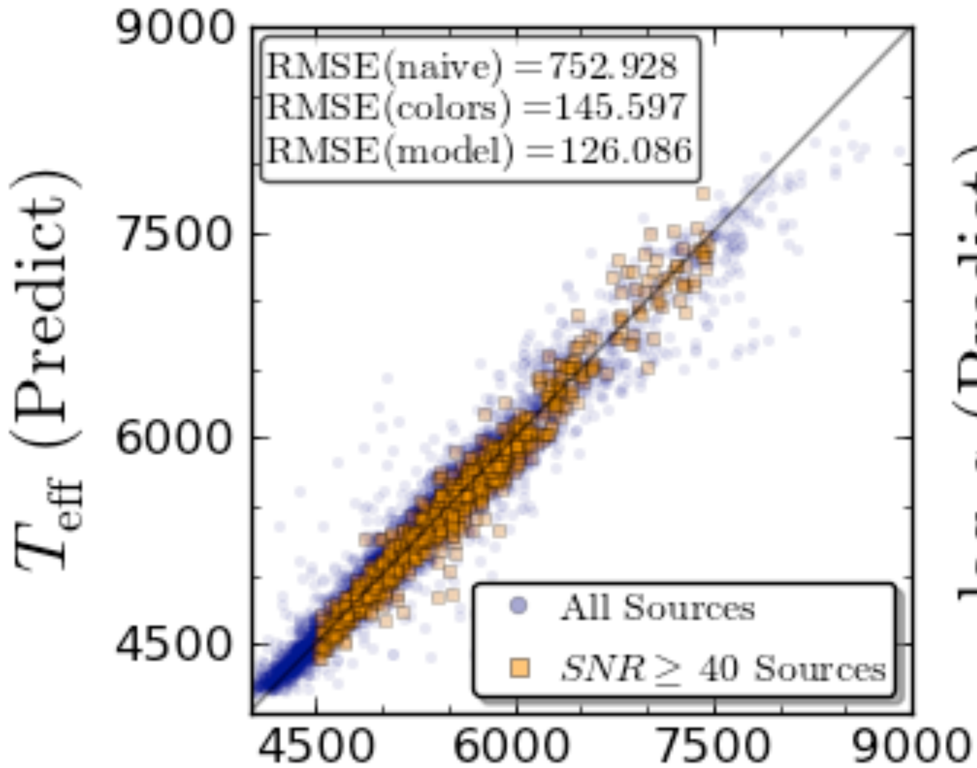
Discovery of Bright Galactic R Coronae Borealis and DY Persei Variables: Rare Gems Mined from ASAS

A. A. Miller[1,*], J. W. Richards[1,2], J. S. Bloom[1], S. B. Cenko[1], J. M. Silverman[1], D. L. Starr[1], and K. G. Stassun[3,4]

arXiv.org > astro-ph > arXiv:1204.4181

Turning Imagers into Spectrographs

Time variability + colors → fundamental stellar parameters



Data: 5000 variables in SDSS Stripe 82 with spectra

~80 dimensional regression with Random Forest

Miller, JSB+14

Big Data Challenge: Time & Resources

Large Synoptic Survey Telescope (LSST) - 2018

Light curves for 800M sources every 3 days

10^6 supernovae/yr, 10^5 eclipsing binaries

3.2 gigapixel camera, 20 TB/night

LOFAR & SKA

150 Gps (27 Tflops) → 20 Pps (~100 Pflops)

Gaia space astrometry mission - 2013

1 billion stars observed ~70 times over 5 years

Will observe 20K supernovae

<u>Many other astronomical surveys are already producing data:</u>

SDSS, i**PTF**, CRTS, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES etc.,

Big Data Challenge: Time & Resources

Large Synoptic Survey Telescope (LSST) - 2018

Light curves for 800M sources every 3 days

10^6 supernovae/yr, 10^5 eclipsing binaries

3.2 gigapixel camera, 20 TB/night

LOFAR & SKA

150 Gps

Gaia space

1 billion stars

Will observe 20K supernovae

How do we do discovery, follow-up, and inference when the data rates (& requisite timescales) preclude human involvement?

Many other astronomical surveys are already producing data:
SDSS, iPTF, CRTS, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES etc.,

Towards a Fully Automated Scientific Stack for Transients

papers

inference

current state-of-the-art stack

typing

followup

classification

discovery

finding

reduction

observing

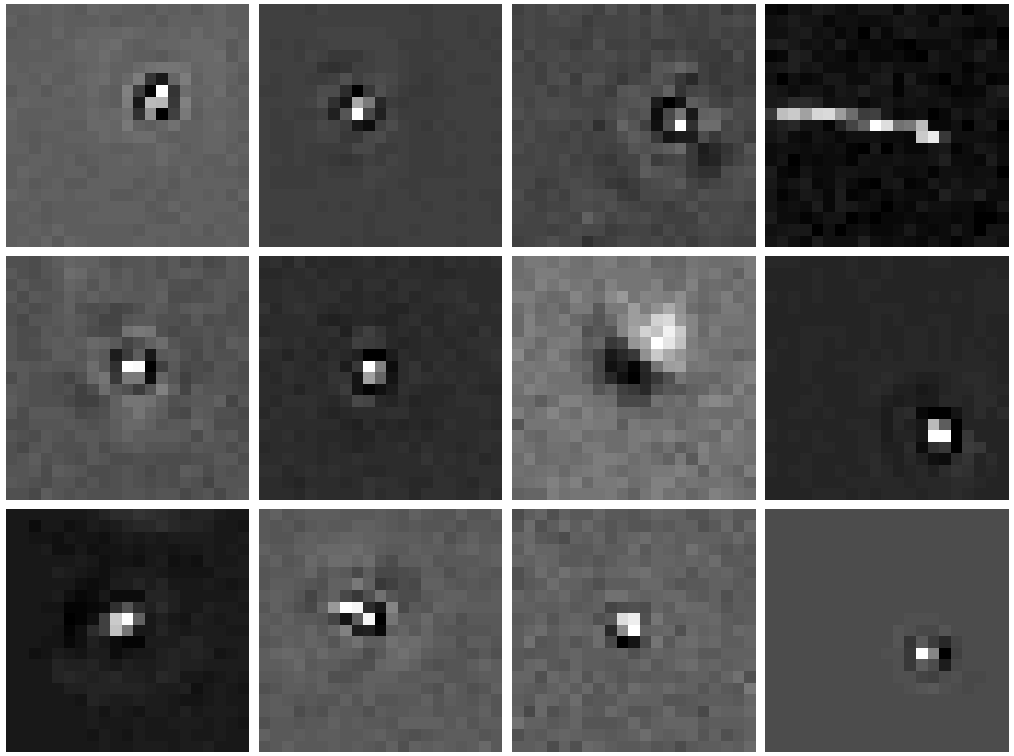
scheduling

strategy

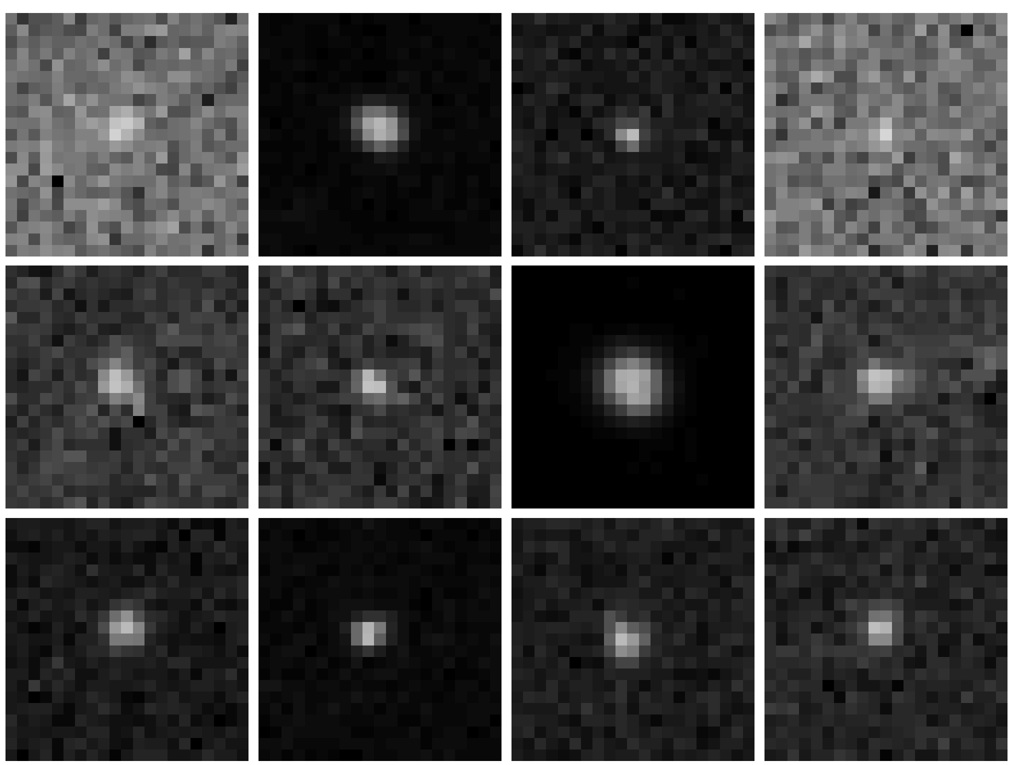
automated (e.g. iPTF)

not (yet) automated

“bogus”



“real”



PTF subtractions

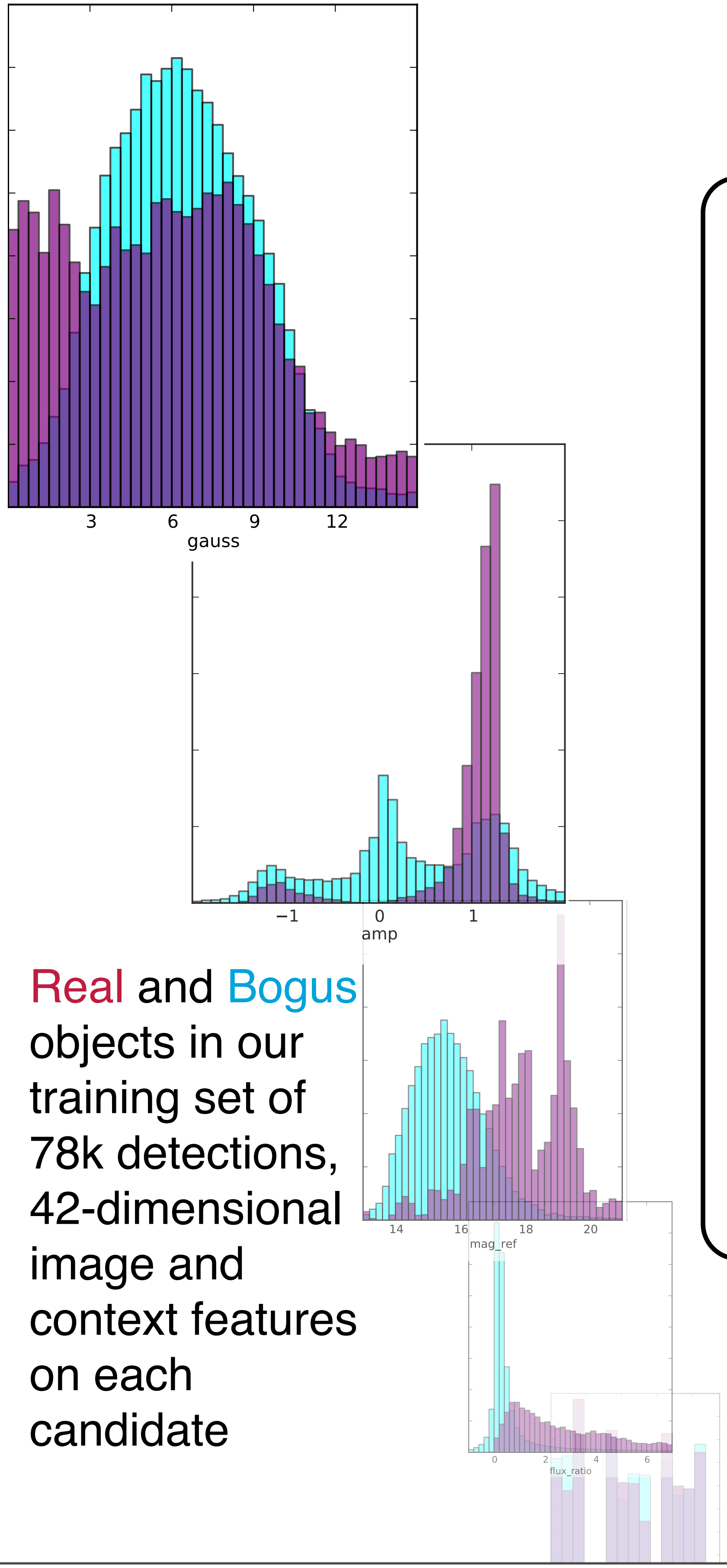
<u>Goal:</u>
build a framework to discover variable/ transient sources without people

- fast (compared to people)
- parallelizeable
- transparent
- deterministic
- versionable

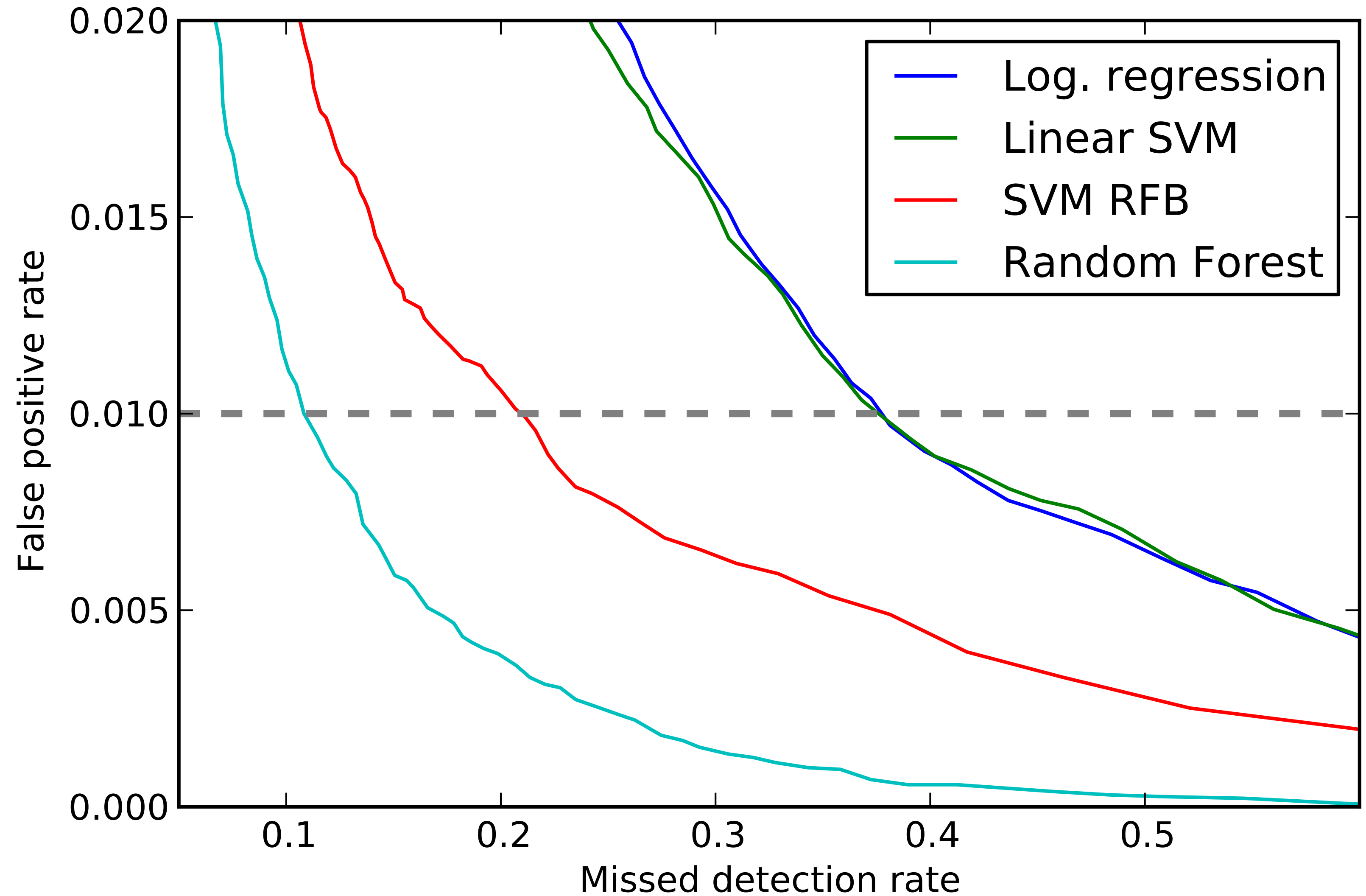
1000 to 1 needle in the haystack problem

“Discovery” is Imperfect

but some classifiers work better than others



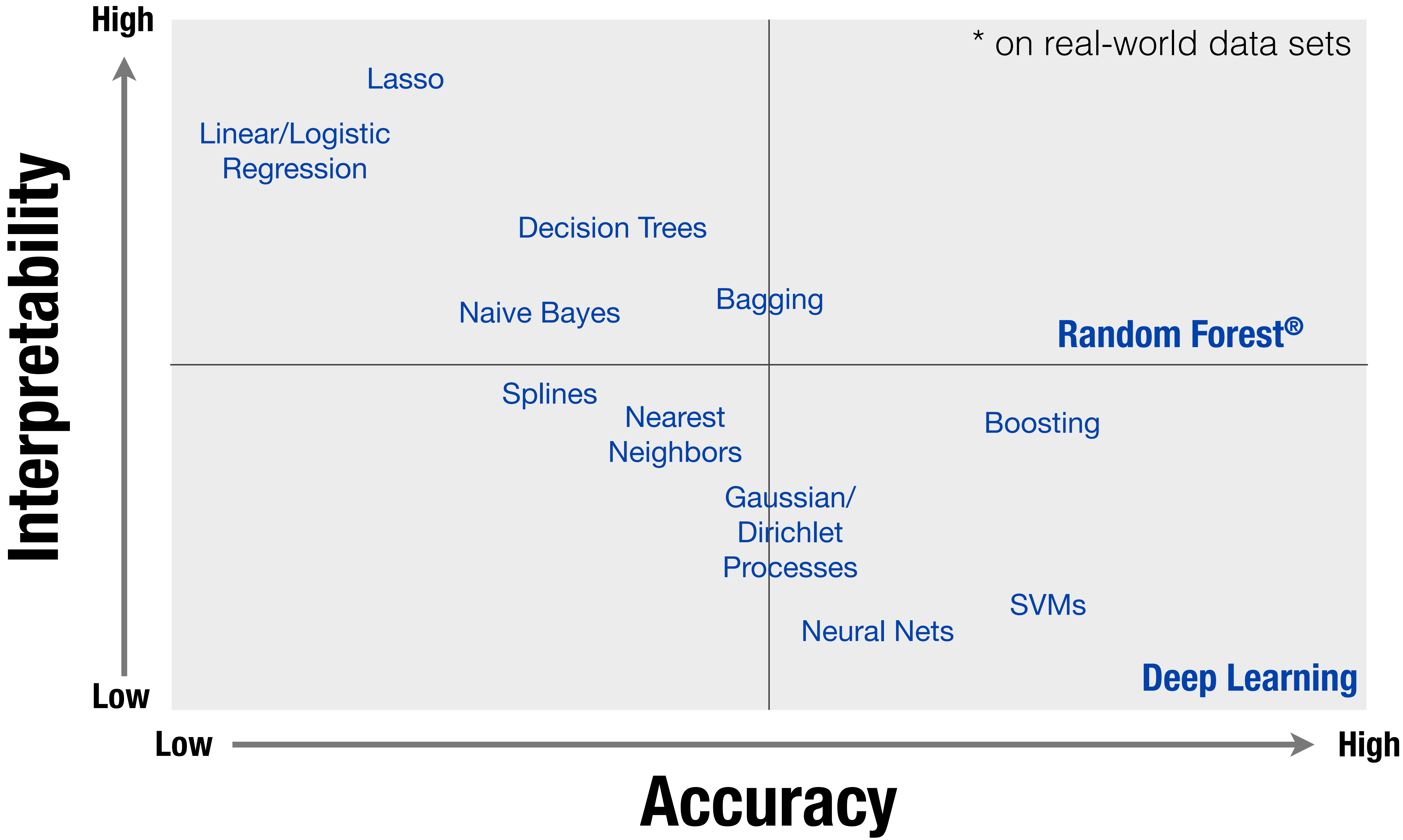
Real and Bogus objects in our training set of 78k detections, 42-dimensional image and context features on each candidate



Brink+2012

arXiv.org > astro-ph > arXiv:1209.3775

ML Algorithmic Trade-Off

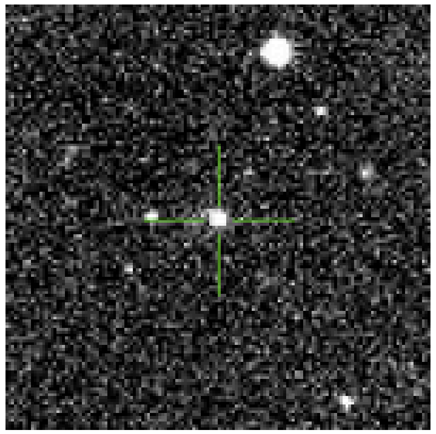


Random Forest is a trademark of Salford Systems, Inc.

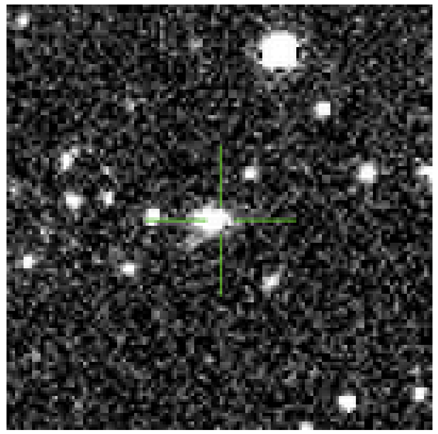
10tel SN IIn Real-time Classifications...

OVERVIEW PHOTOMETRY SPECTROSCOPY FOLLOWUP OBSERVABILITY FINDING CHART SCANNING

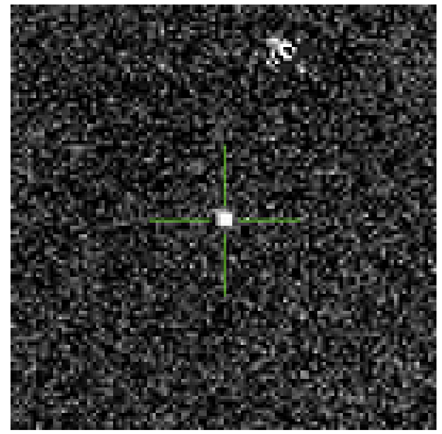
NEW



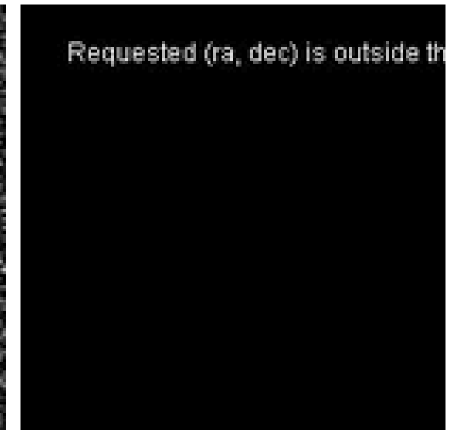
REF

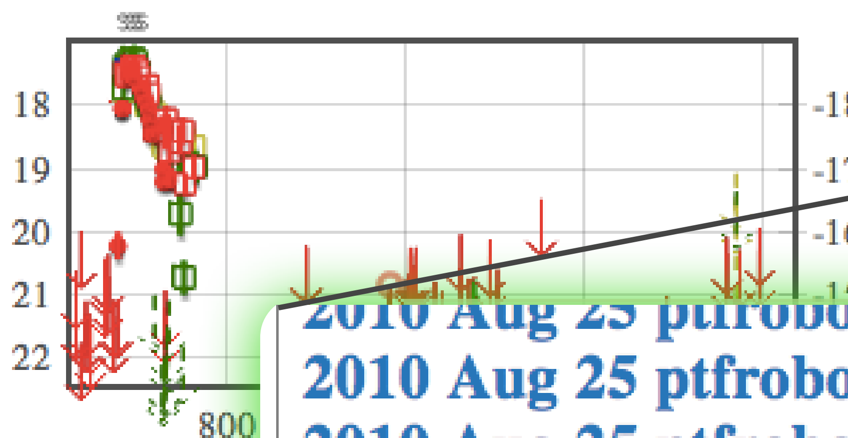


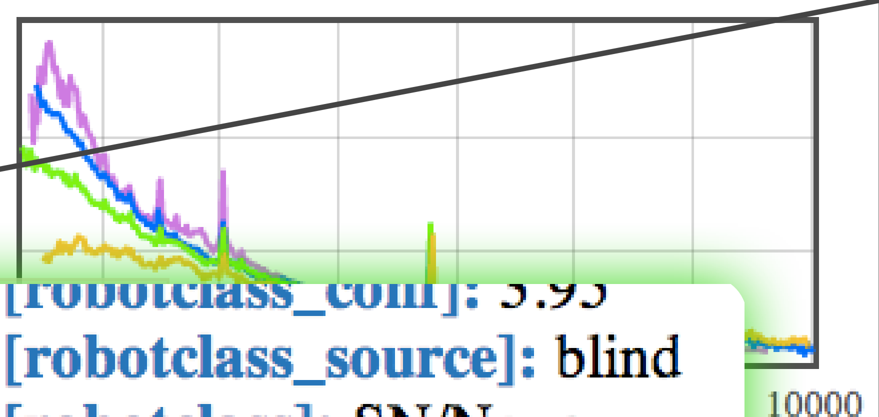
SUB



SDSS







r >19.9 (204.6 d)

2010 Aug 25 ptfrobot [robotclass_conf]: 3.95
2010 Aug 25 ptfrobot [robotclass_source]: blind
2010 Aug 25 ptfrobot [robotclass]: SN/Nova
2010 Aug 25 ptfrobot [type]: Transient

Mentioned in 7 Email(s)

ADDITIONAL IN...

IPAC | NED | DSS | SIMBAD | HEASARC | WISE | SkyView | PyMP | PTFInfoBot | Extinction

Add to Cart

COMMENTS

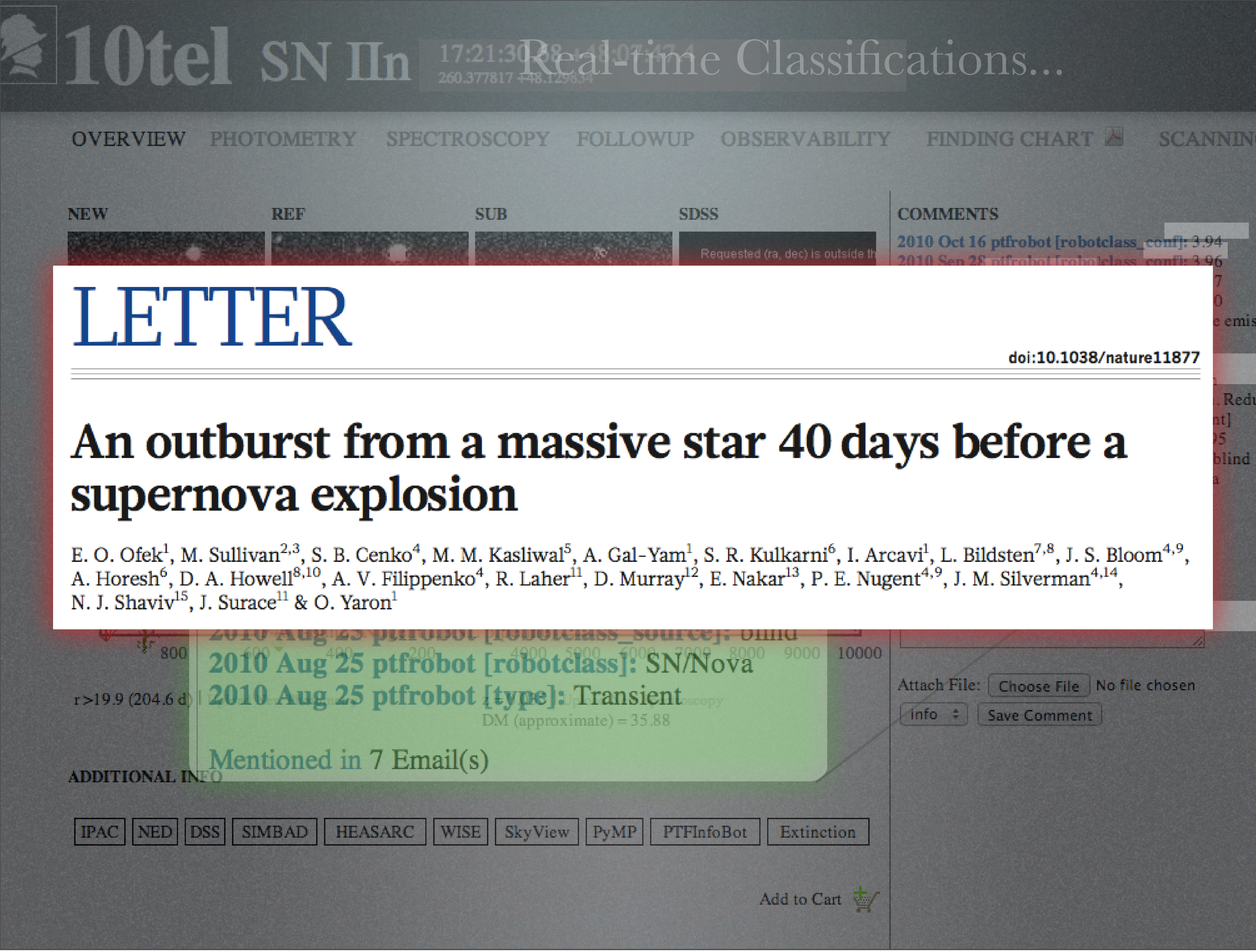
2010 Oct 16 ptfrobot [robotclass_conf]: 3.94
2010 Sep 28 ptfrobot [robotclass_conf]: 3.96
2010 Sep 21 ptfrobot [robotclass_conf]: 3.97
2010 Sep 18 ptfrobot [robotclass_conf]: 4.00
2010 Aug 26 ahowell [info]: Redshift. Notice emiss
Na I D. [view attachment]
2010 Aug 26 ahowell [redshift]: 0.035
2010 Aug 26 ahowell [classification]: SN IIn
2010 Aug 26 ahowell [comment]: A type IIn. Reduc
Daniel Murray's pyraf script! [view attachment]
2010 Aug 25 ptfrobot [robotclass_conf]: 3.95
2010 Aug 25 ptfrobot [robotclass_source]: blind
2010 Aug 25 ptfrobot [robotclass]: SN/Nova
2010 Aug 25 ptfrobot [type]: Transient

Mentioned in 7 Email(s)

Add a Comment:

Attach File: Choose File No file chosen

info Save Comment

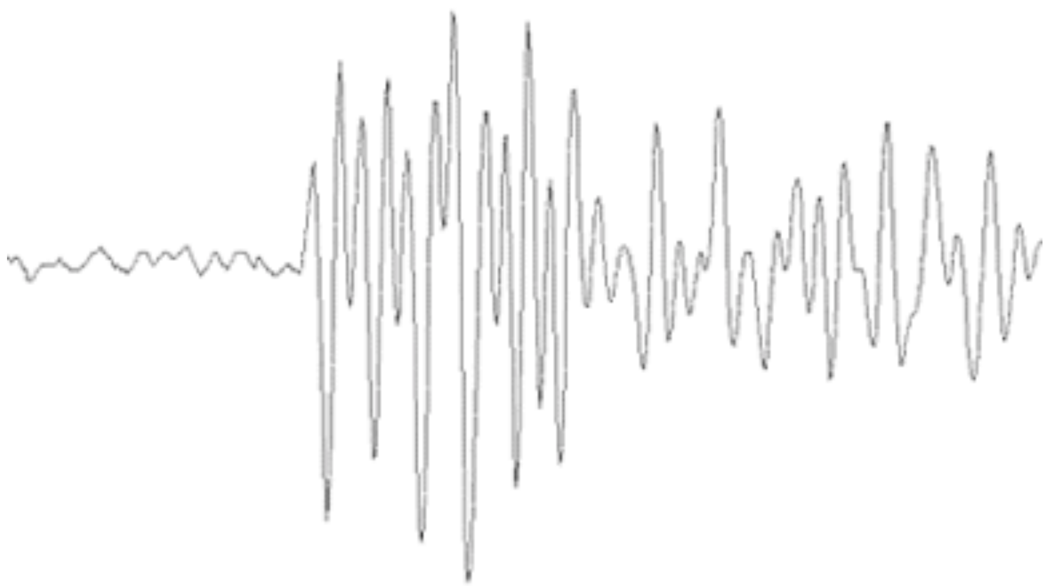


LETTER

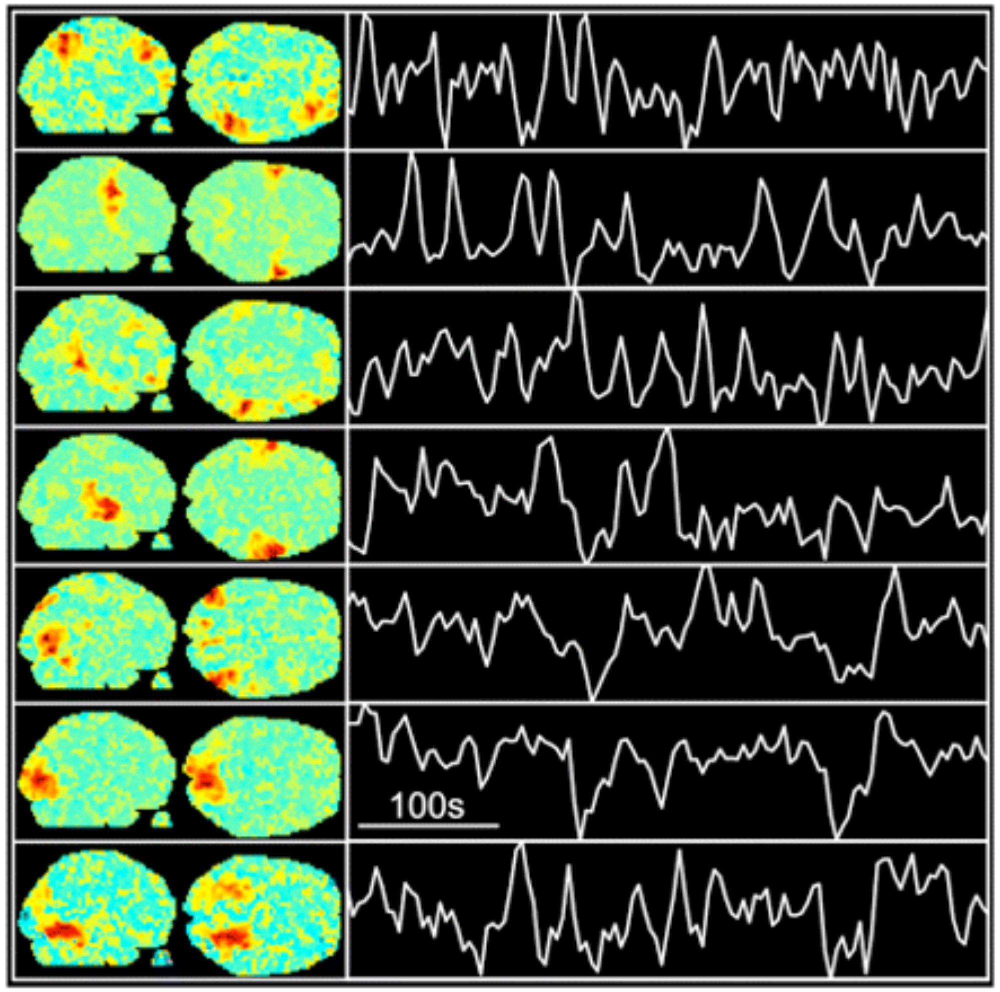
doi:10.1038/nature11877

An outburst from a massive star 40 days before a supernova explosion

E. O. Ofek[1], M. Sullivan[2,3], S. B. Cenko[4], M. M. Kasliwal[5], A. Gal-Yam[1], S. R. Kulkarni[6], I. Arcavi[1], L. Bildsten[7,8], J. S. Bloom[4,9], A. Horesh[6], D. A. Howell[8,10], A. V. Filippenko[4], R. Laher[11], D. Murray[12], E. Nakar[13], P. E. Nugent[4,9], J. M. Silverman[4,14], N. J. Shaviv[15], J. Surace[11] & O. Yaron[1]



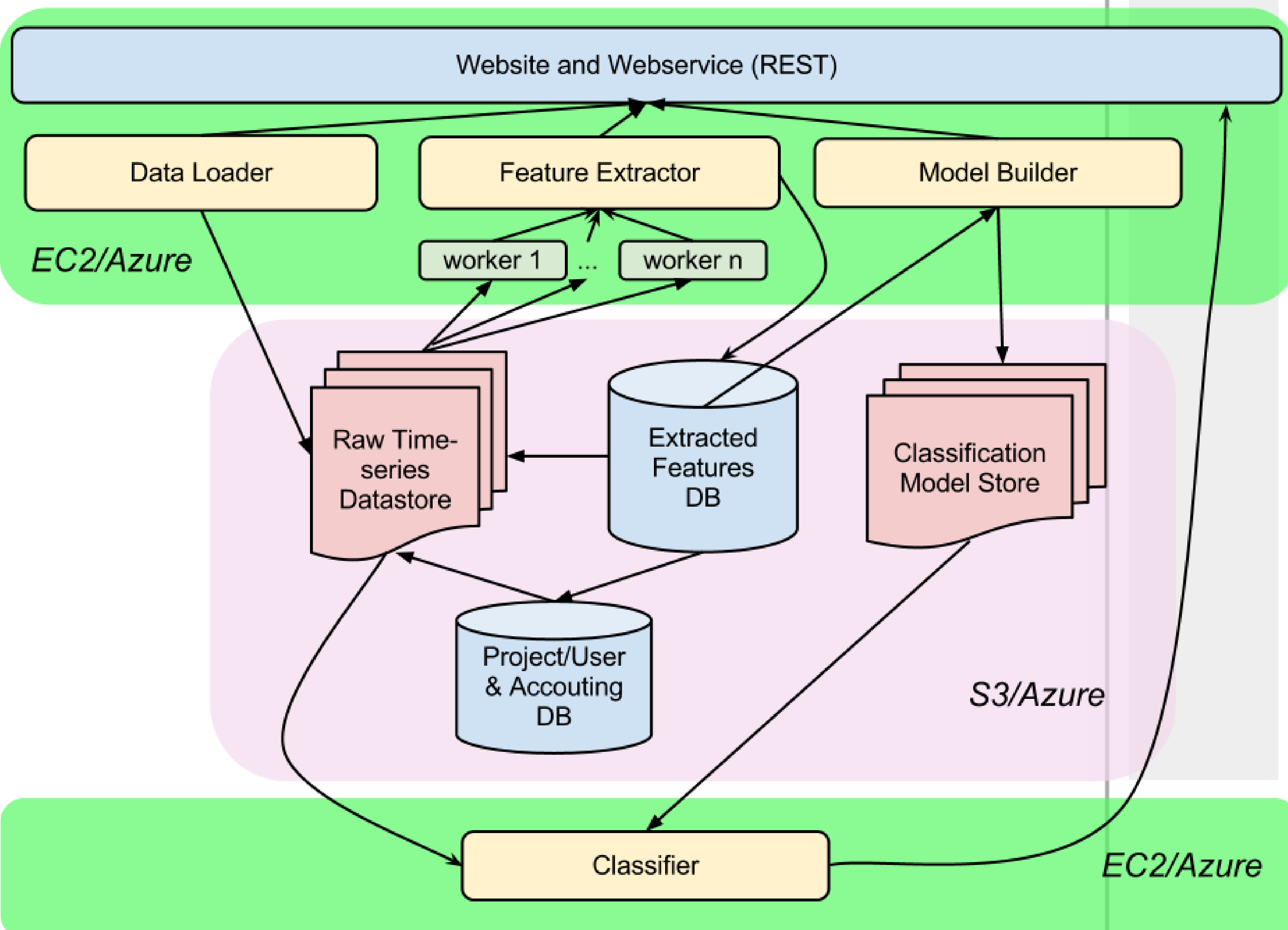
Seismology

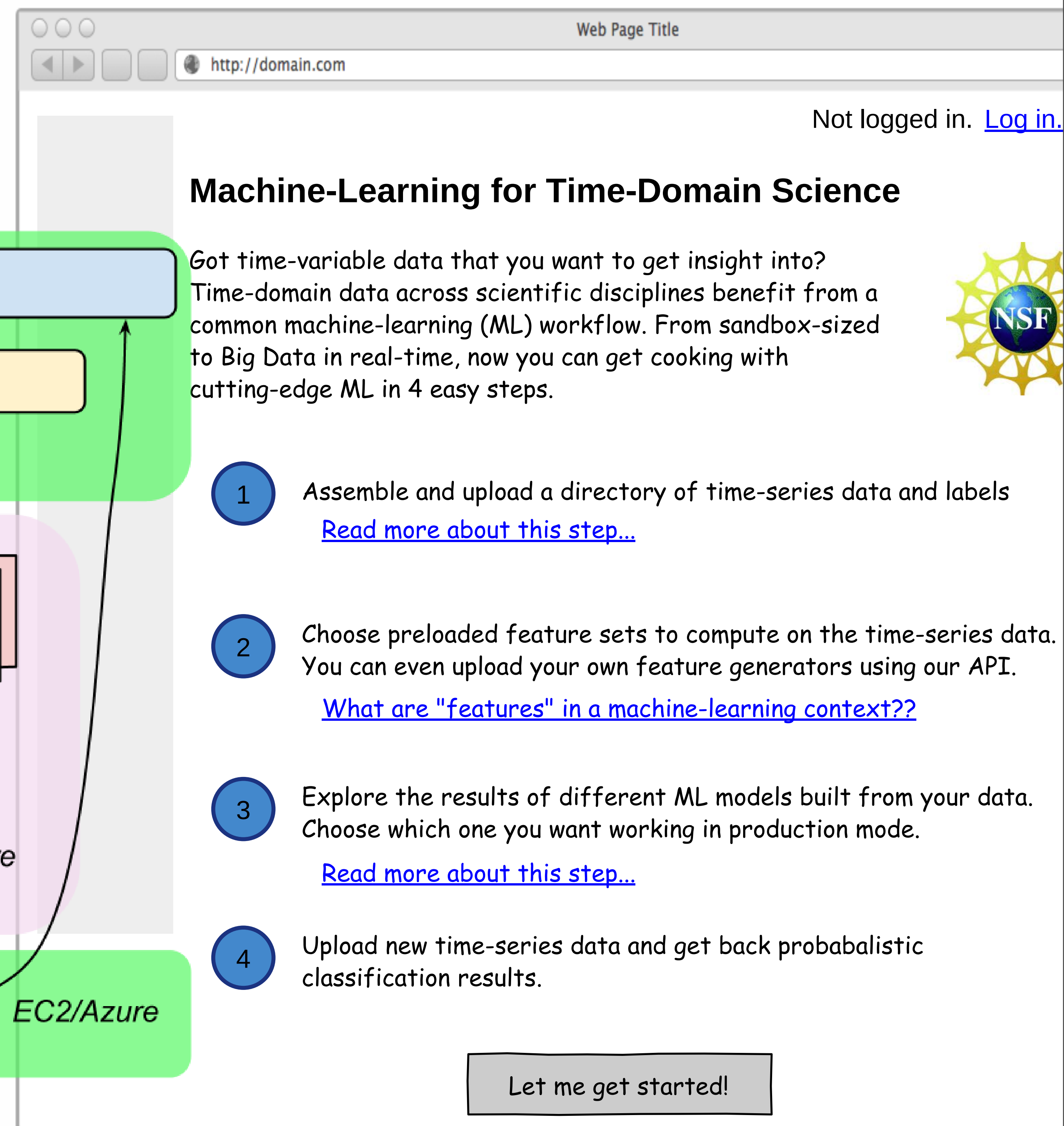


Neuroscience

Classification Platform for Novel Scientific Insight on Time-Series Data

NSF/BIGDATA project





Parting Thoughts

- Astronomy's data deluge demands an abstraction of the traditional roles in the scientific process. Despite automation, crucial (insightful) roles remain for people
- Machine learning is an emerging & useful tool
- Deterministic prediction with verifiable uncertainties is crucial to maximize scientific impact under real-world resource constraints
- Major Challenge: Training & access to great learning frameworks
- In the time-domain, machine-learning ***prediction <u>is</u> the gateway to understanding***