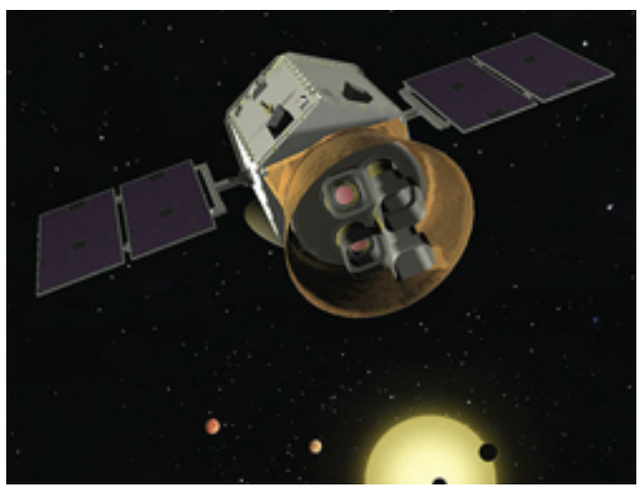
Finding Needles in the Haystack: Outlier Detection in Astronomical Datasets

Rafael Martínez-Galarza

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Collaborators: **Dennis Crake(CfA/Southampton),** Federica Bianco (NYU), Matthew Graham (Caltech), Ashish Mahabal (Caltech), Francesca Civano (CfA), Raffaele D'Abrusco (CfA), Ian Evans (CfA),



Fermilab seminar - July 10, 2019

This talk

- Motivation :
  - The exploration approach
  - The upcoming golden age of large astronomical datasets
- What is an outlier? - Feature extraction with and without Deep Learning
- Methods: Unsupervised Random Forest, t-SNE, Persistent Homology
- Applications 1: Time domain
- Applications 2: High Energy Astrophysics
- Results: the weirdest Kepler/TESS light curves and Chandra sources

1. Motivation

The question-driven approach

Well defined, pre-established research questions, e.g.:

- What is the nature of gravitational wave sources progenitors?

- What is the value of Ω_0 to a precision below 1% percent?

- How common are terrestrial planets around main-sequence stars?

Known unknowns. Based on present knowledge

Design targeted observations/surveys. Model specific objects

The exploration approach

Gain the capability to pose **new** questions

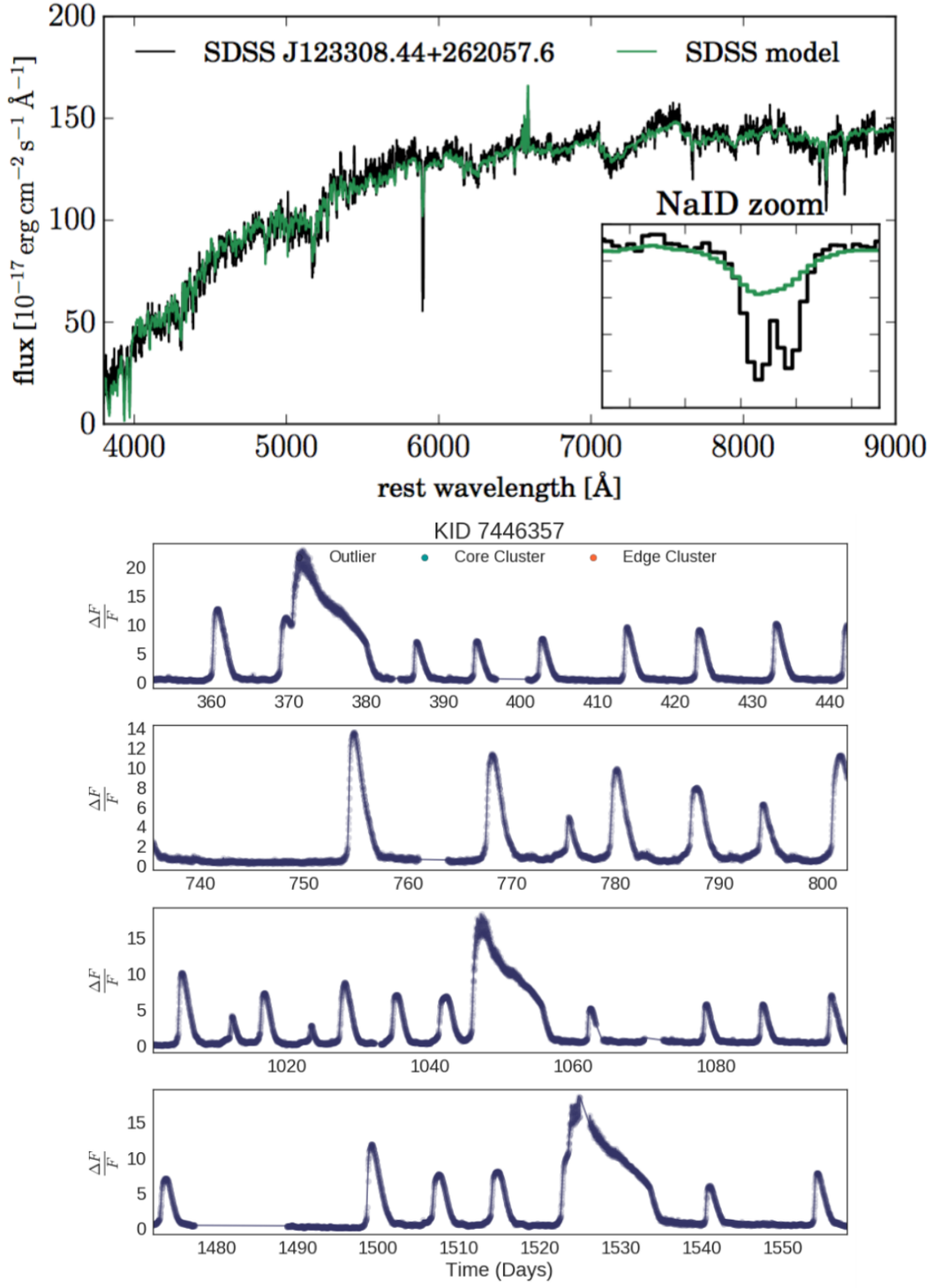
- This is a way to deal with the *unknown unknowns*
- Particularly pressing in astronomy, a science which is mostly exploratory
- Paradigm shifts in astronomy have come when discovery space is increased

.

Design large surveys. Open new discovery windows (e.g. multi-messenger). Data mining.

Making serendipitous discoveries systematic

Outlier (anomaly) detection algorithms offer a natural way to find the weirdest objects in large datasets for which only a fraction of the object has been assigned a class.

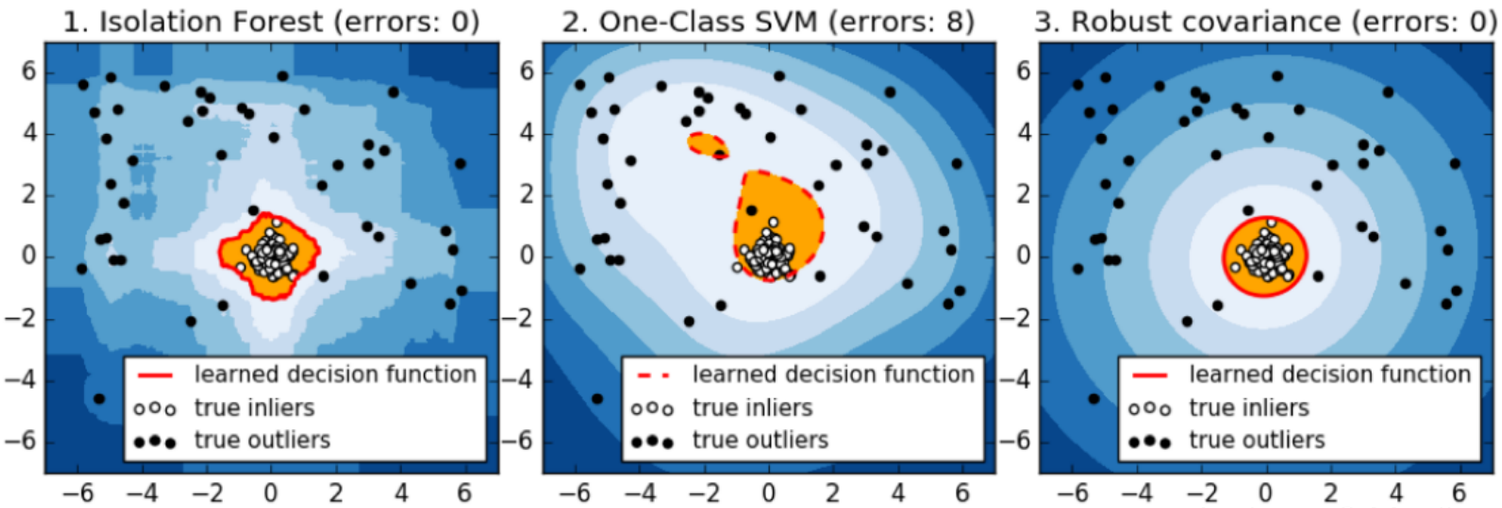


- Unsupervised Random Forest (URF) applied to SDSS spectra: complex velocity structures, extremely strong or rare absorption lines (Baron & Poznanski, 2017).

- Proximity clustering + dimensionality reduction applied to Kepler light curves: cataclysmic variables. (Giles & Walkowicz, 2019).

What is an outlier?

"An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism"
-- Hawkins (1980)



Andreas C Mueller

- Outliers/anomalies are hard to define/find.
- That is because “weirdness” is hard to define (and because datasets are huge)
- Weirdness of an object depends both on the features used to define it, and on the specific algorithm used.
- **Nevertheless, anomalies are scientifically interesting objects: they are often not explained by current models, and could lead to new discoveries.**

Finding “weird” objects

- We need a metric to measure similarity (or dissimilarity) between them. Not so easy in multidimensional parameter spaces.

- We also need a set of features to characterize each example in the dataset. These features can be defined by the scientist, or found through deep learning.

- We then rank all objects in our dataset according to their dissimilarity from other objects.

- Do science! What does “weirdness” mean in each case?

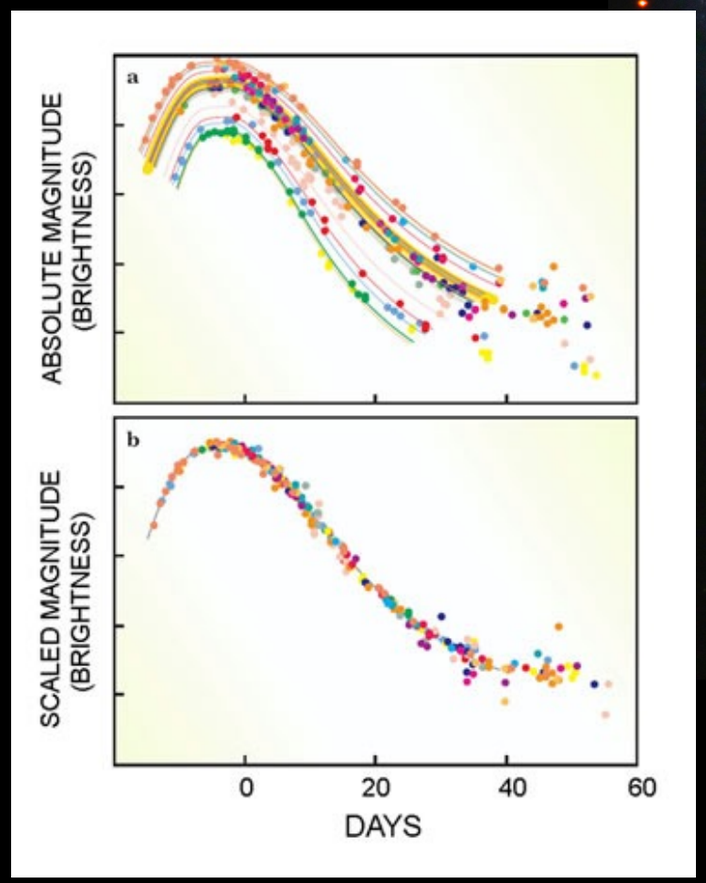
Time Domain Astronomy

image composite: BJ Fulton

Time Domain Astronomy

image composite: BJ Fulton

Time Domain Astronomy



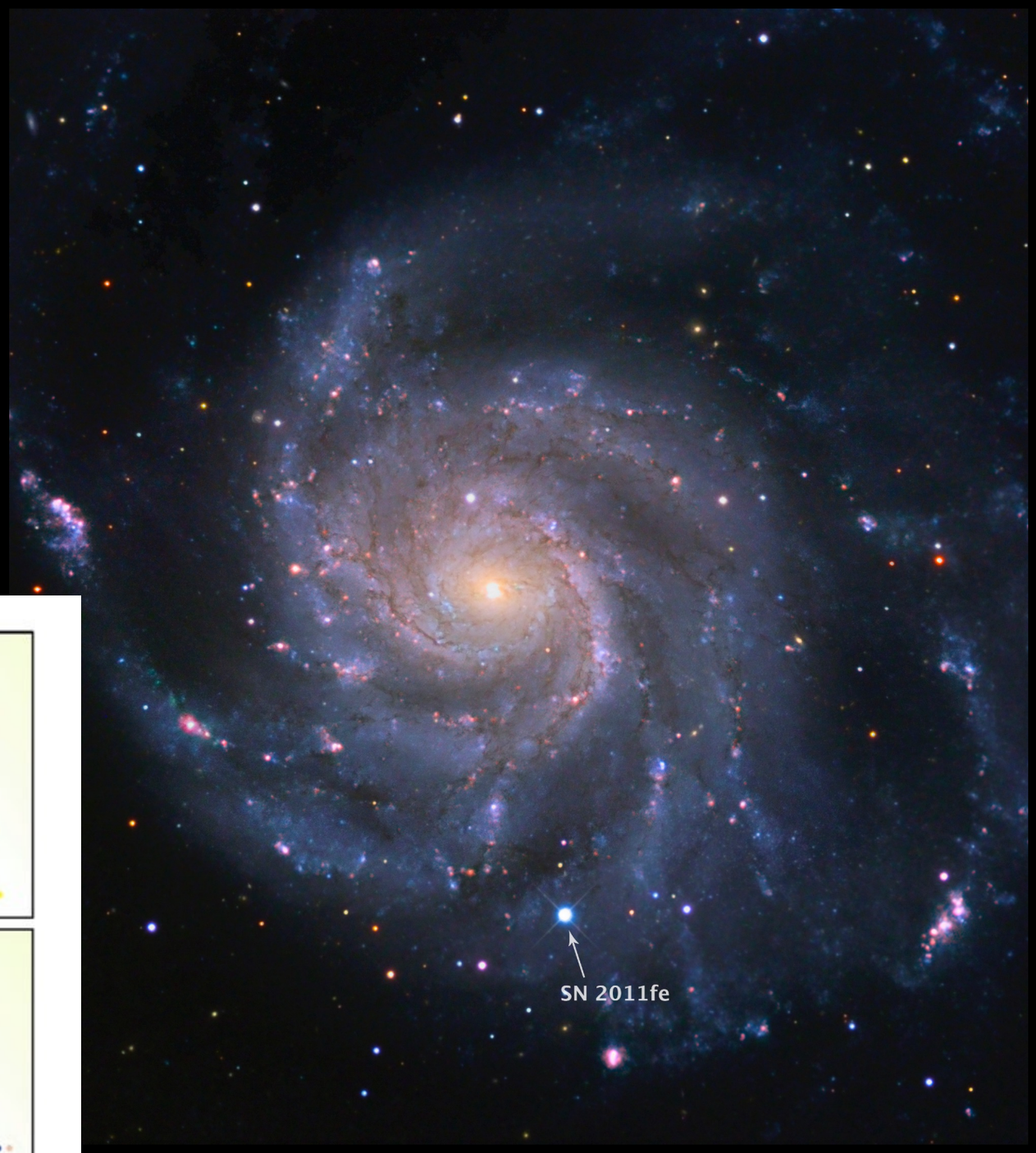
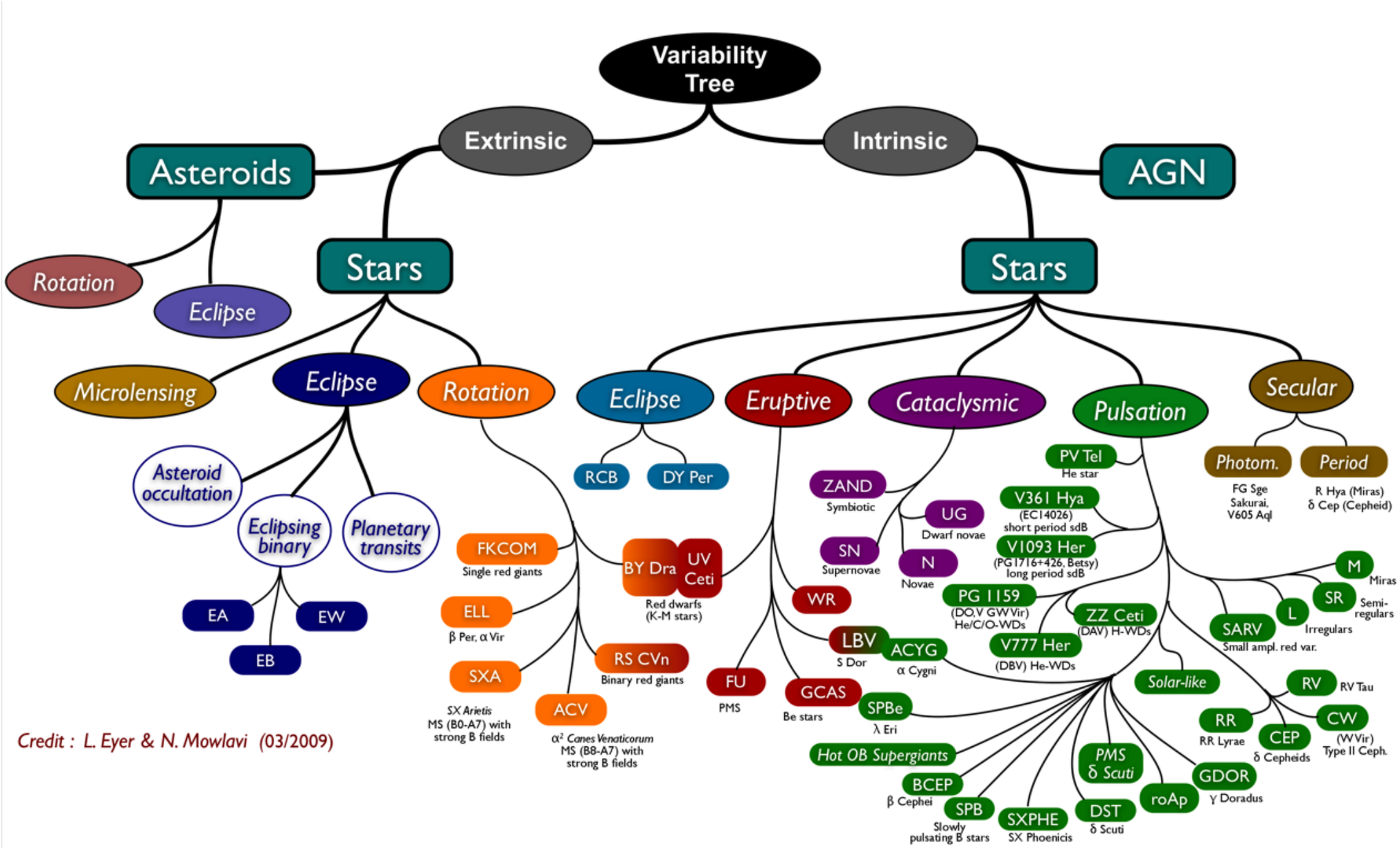


image composite: BJ Fulton

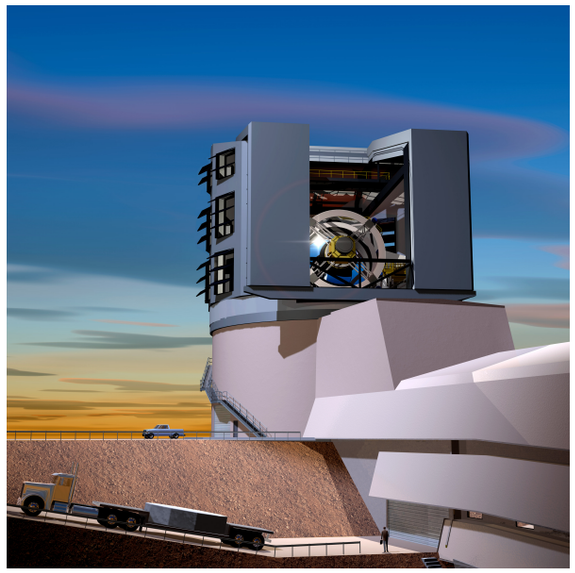
The variable universe

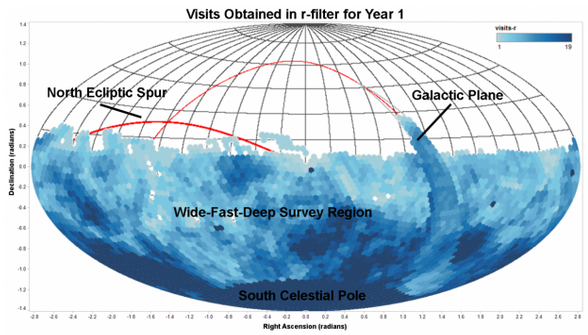


Variability is relevant to a wide range of fields of astrophysics, from understanding the internal composition of the stars and detection of extrasolar planets, to revealing the structure of galaxies and detecting invisible matter.

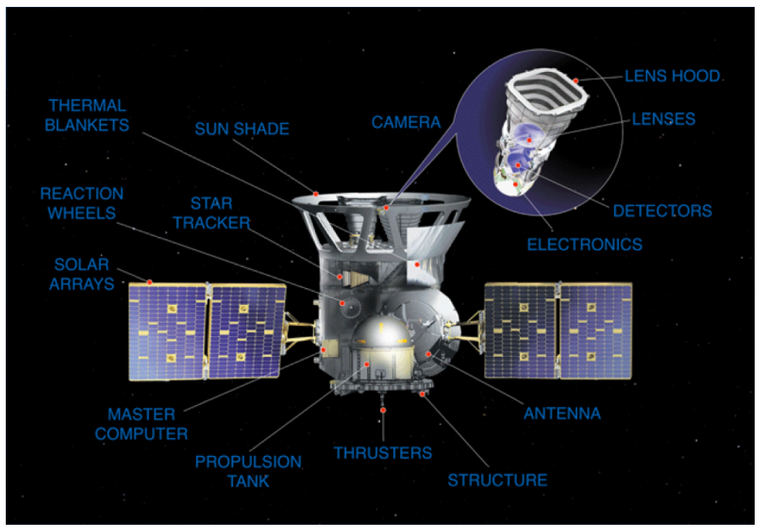
Datasets for discovery

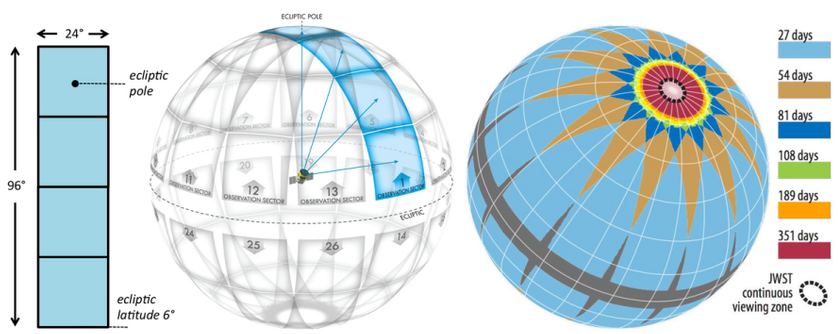
LSST



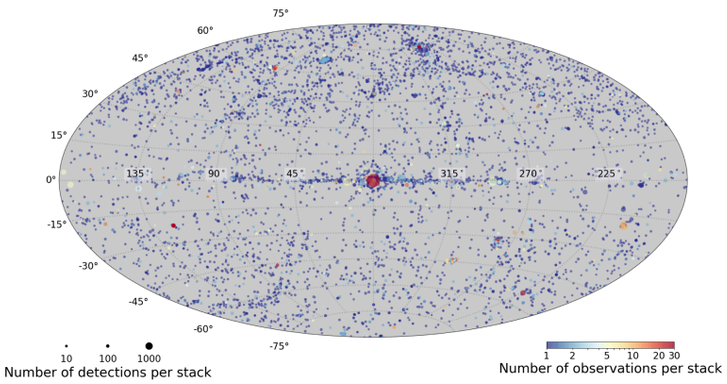


Kepler/TESS





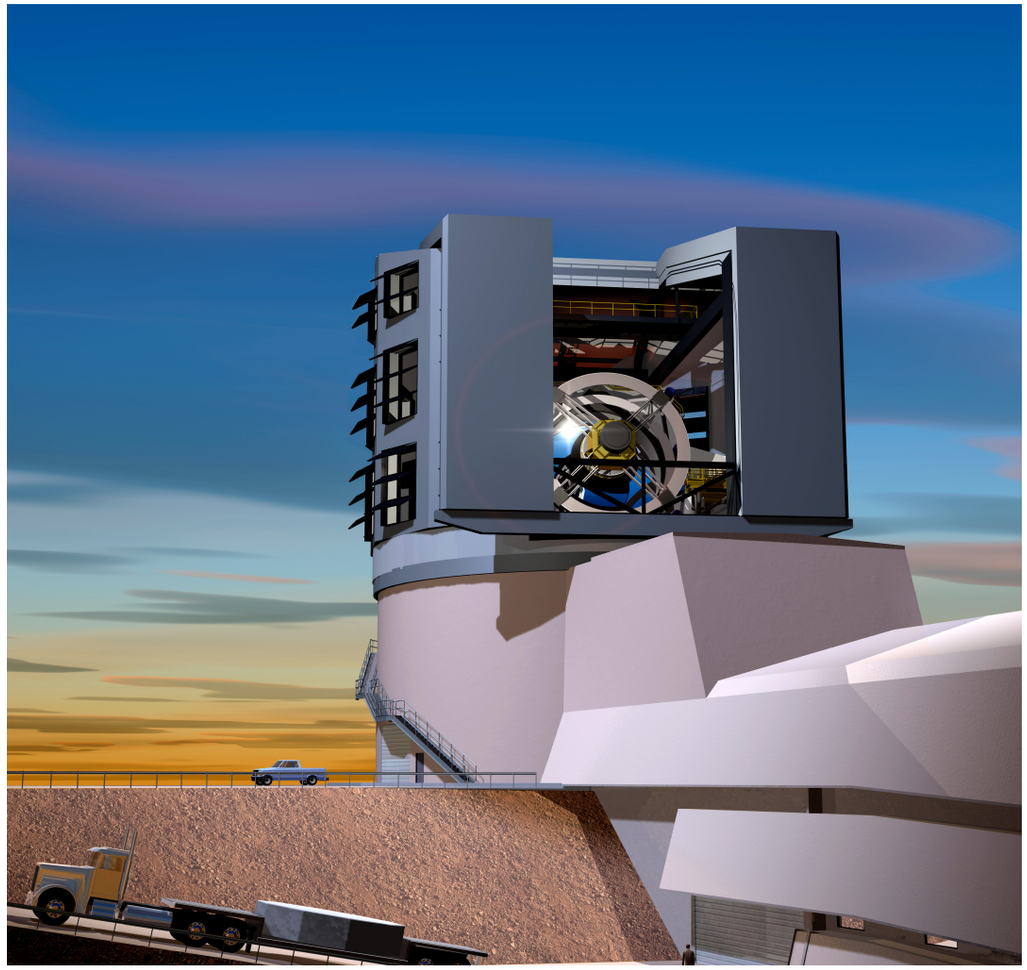
CSC2

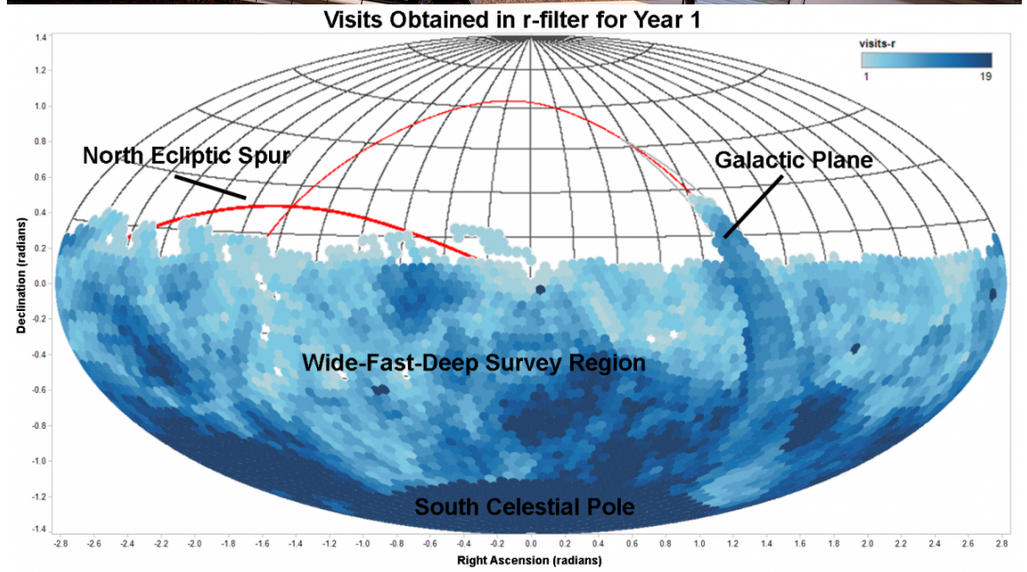


Unexplored discovery space. Majority of sources have not been classified, or in some cases even detected before.

LSST is coming

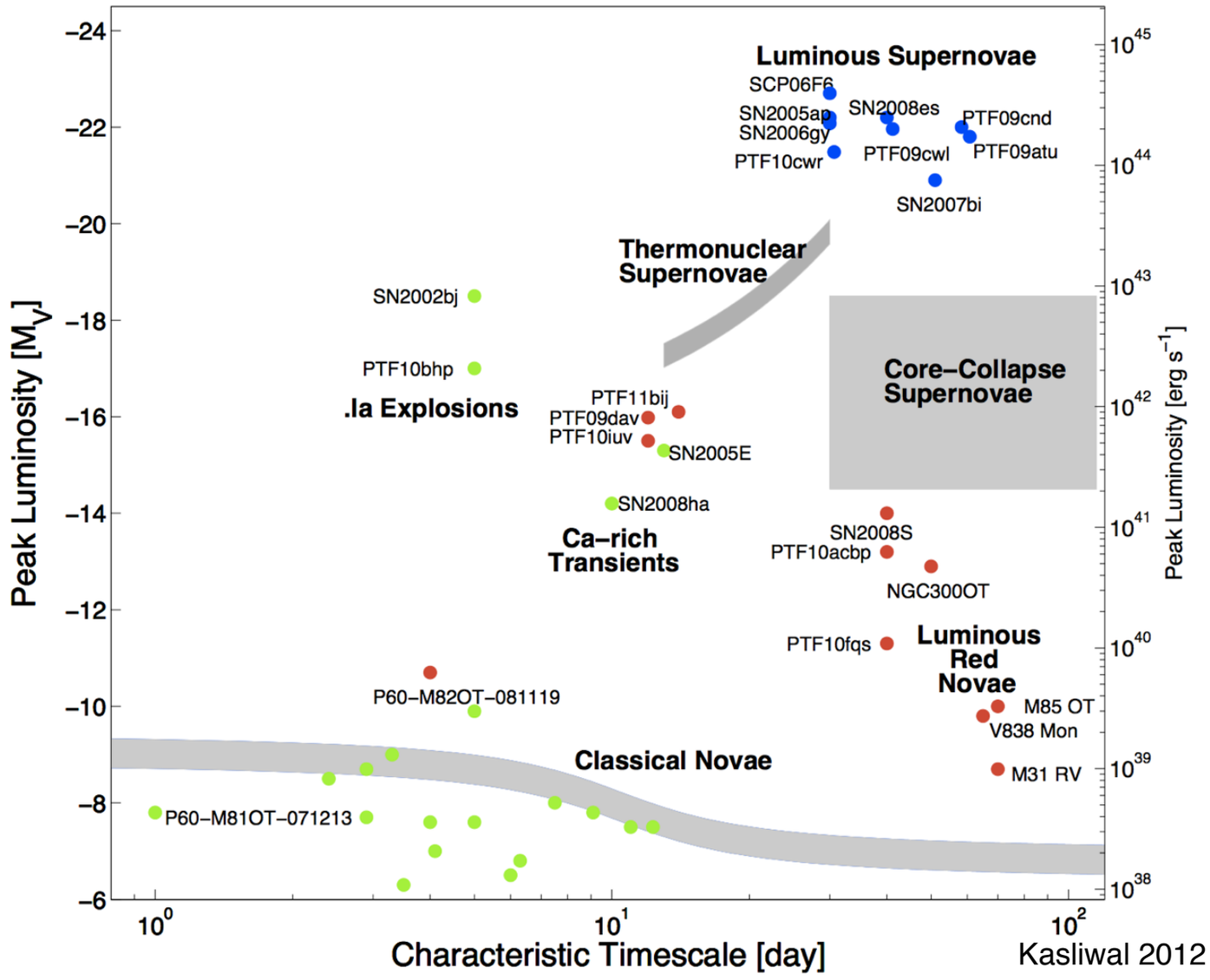
- The Large Synoptic Survey Telescope is a 8.4m reflector currently under construction in Chile (first light expected in 2021).
- Design concept: a survey that will take an image of every part of the entire visible sky every few nights, in six bands, for 10 years.
- Transients and variable stars: periodic and non-periodic variable sources will be studied in detail, and new types are expected at very short and very long timescales.





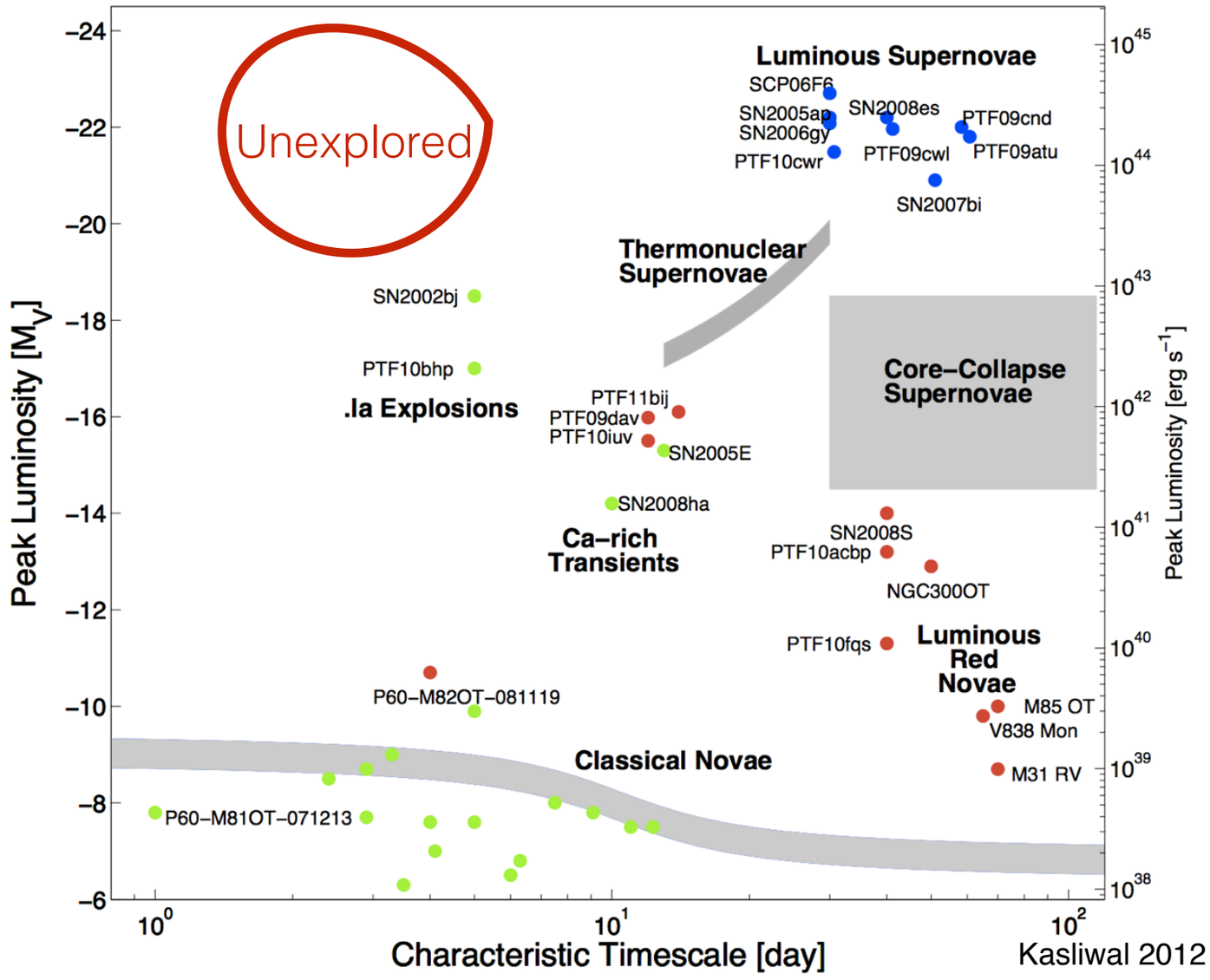
The (un)known transient universe

- New types of transients expected from last generation surveys.
- Fast and luminous transients of particular interest for gravitational wave astronomy.
- How do we identify those weirdos so that we can follow them up?

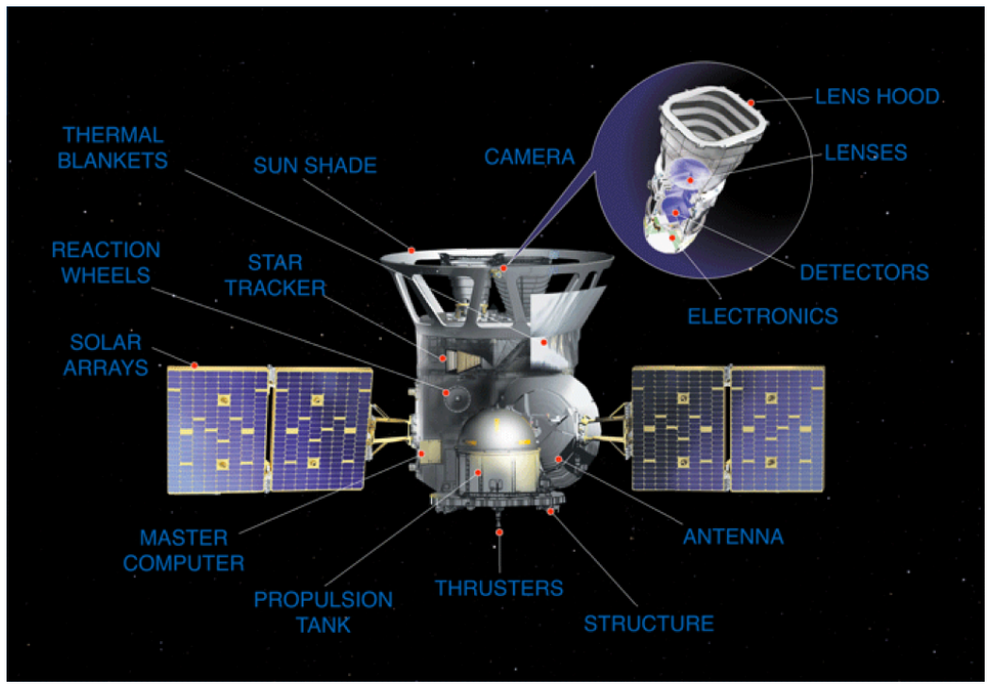


The (un)known transient universe

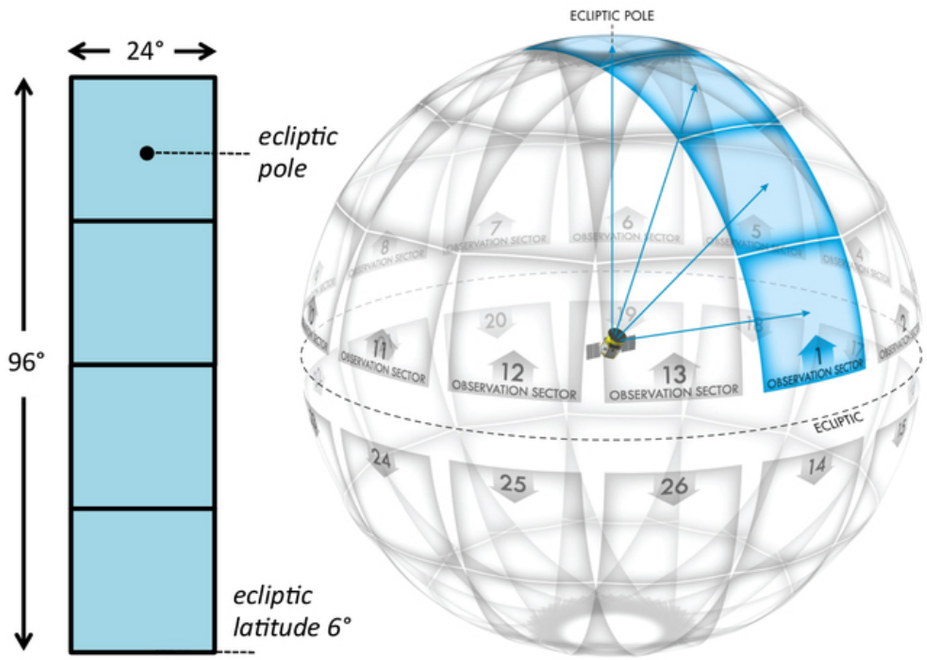
- New types of transients expected from last generation surveys.
- Fast and luminous transients of particular interest for gravitational wave astronomy.
- How do we identify those weirdos so that we can follow them up?

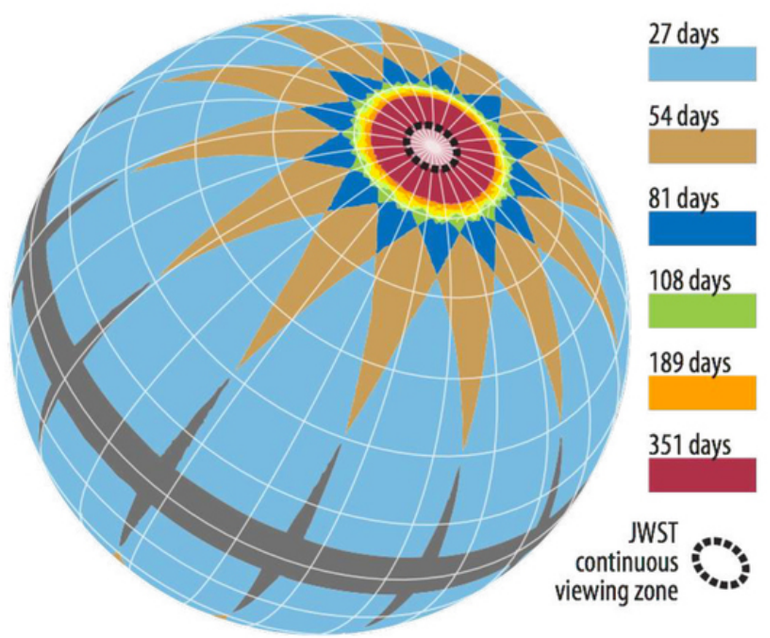


From Kepler to TESS

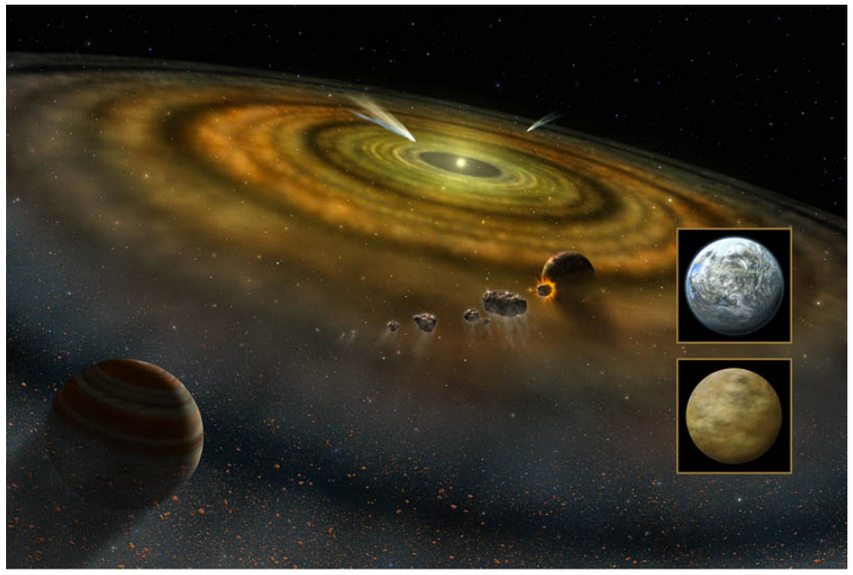


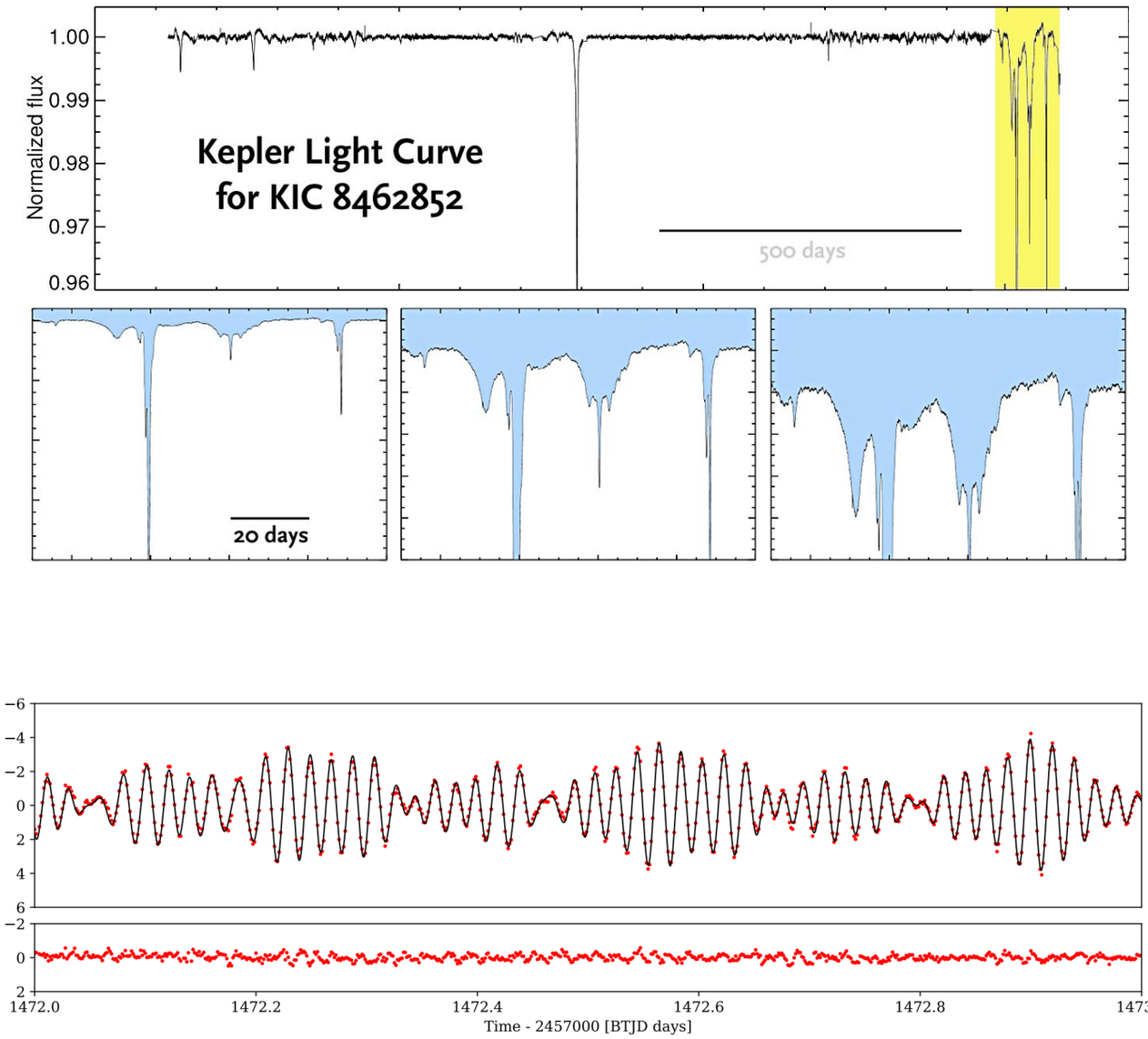
- Monitoring 200,000 stars
- Cadence: 2 minutes, 30 minutes
- Nearby stars of spectral type F5 to M5, at distances < ~60 pc.
- Search area 400 times larger than Kepler's
- Science: stellar astrophysics, exoplanets.

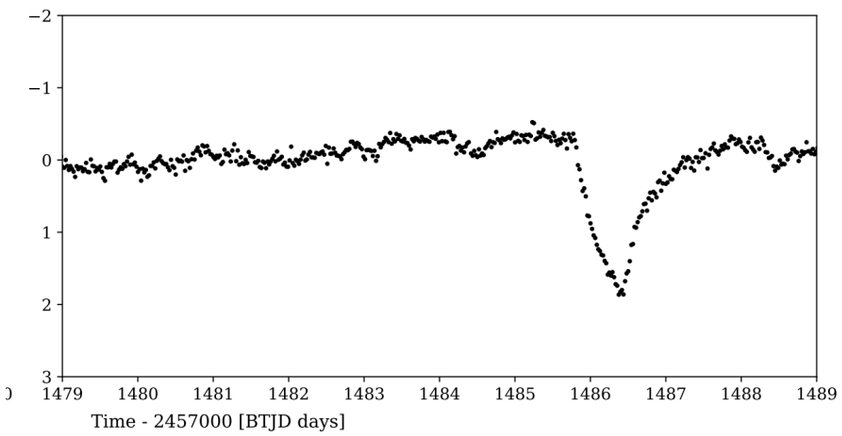




Exo-comets

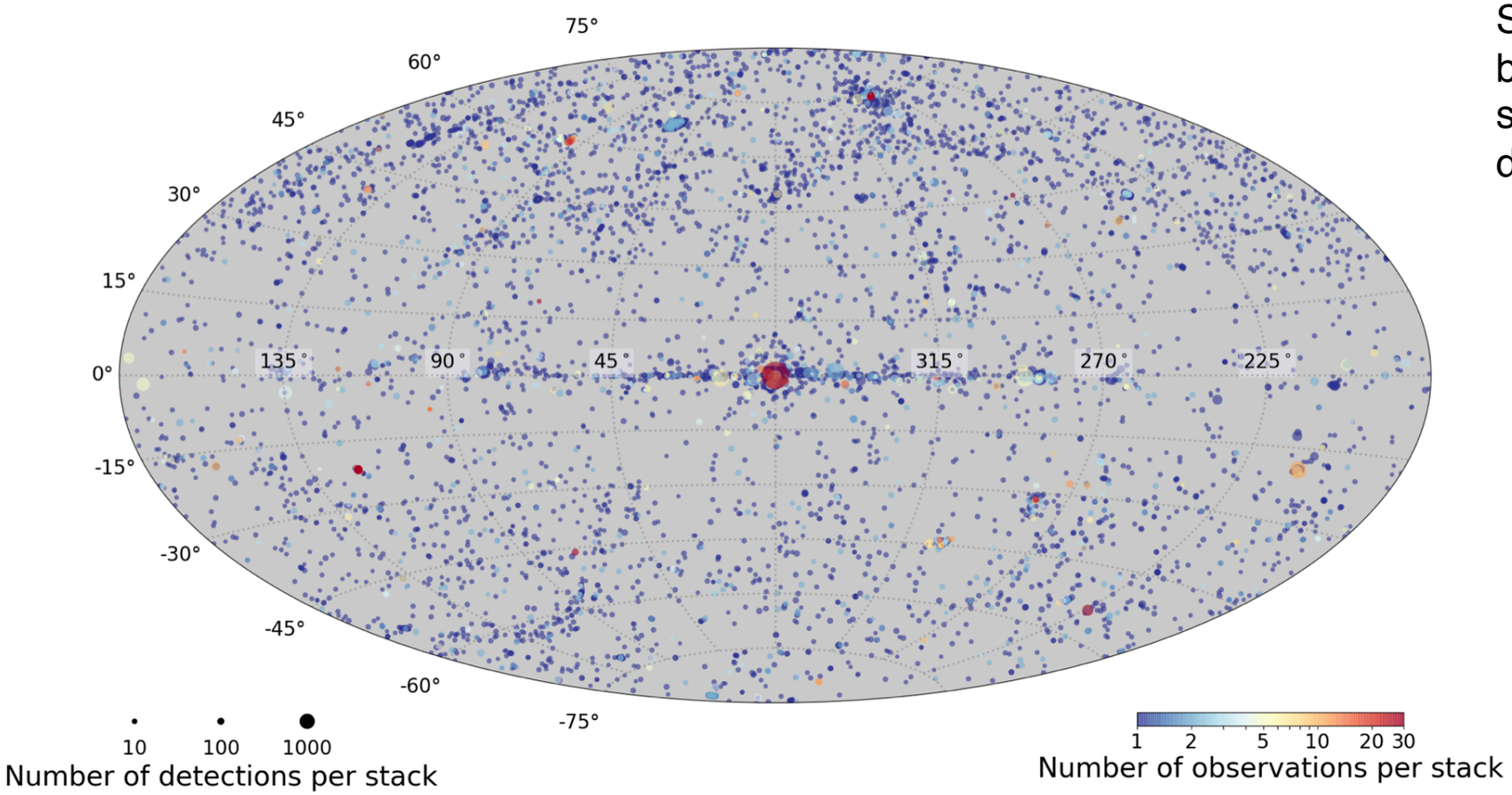






Exo-comets provide clues to the early stages of planet formation in extrasolar systems.

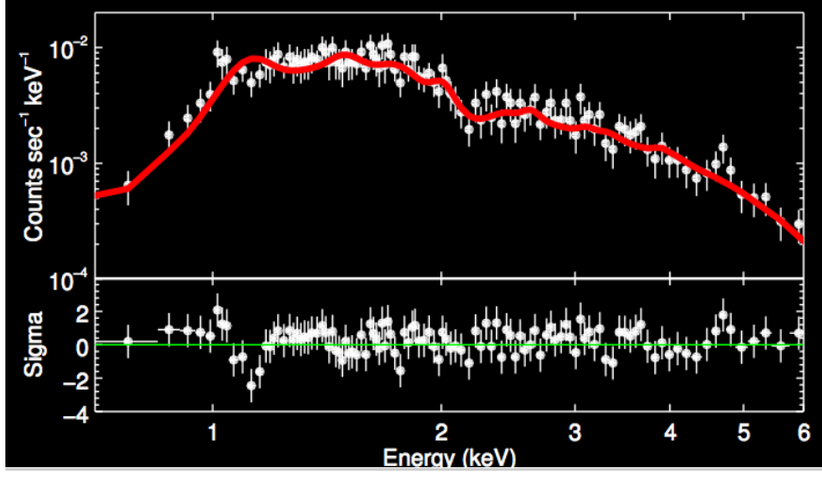
The Chandra Source Catalog 2.0



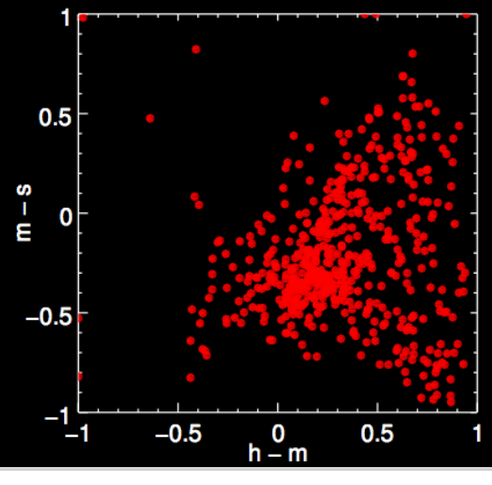
Source and detection positions, multi-band X-ray photometry, images, spectra, light-curves, and calibrated data products

315,875	number of sources
374,349	number of detections
10,382	individual observations
1999–2014	observation public release
~610	square deg sky coverage
245.8 Ms	total exposure
5.8 Ms	longest effective exposure
~5	limiting counts on-axis

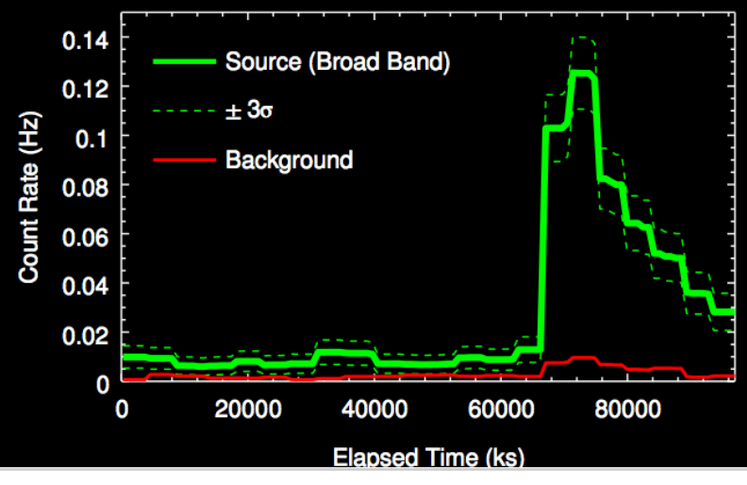
Spectral model fits and fluxes determined using multiple models



Hardness ratios

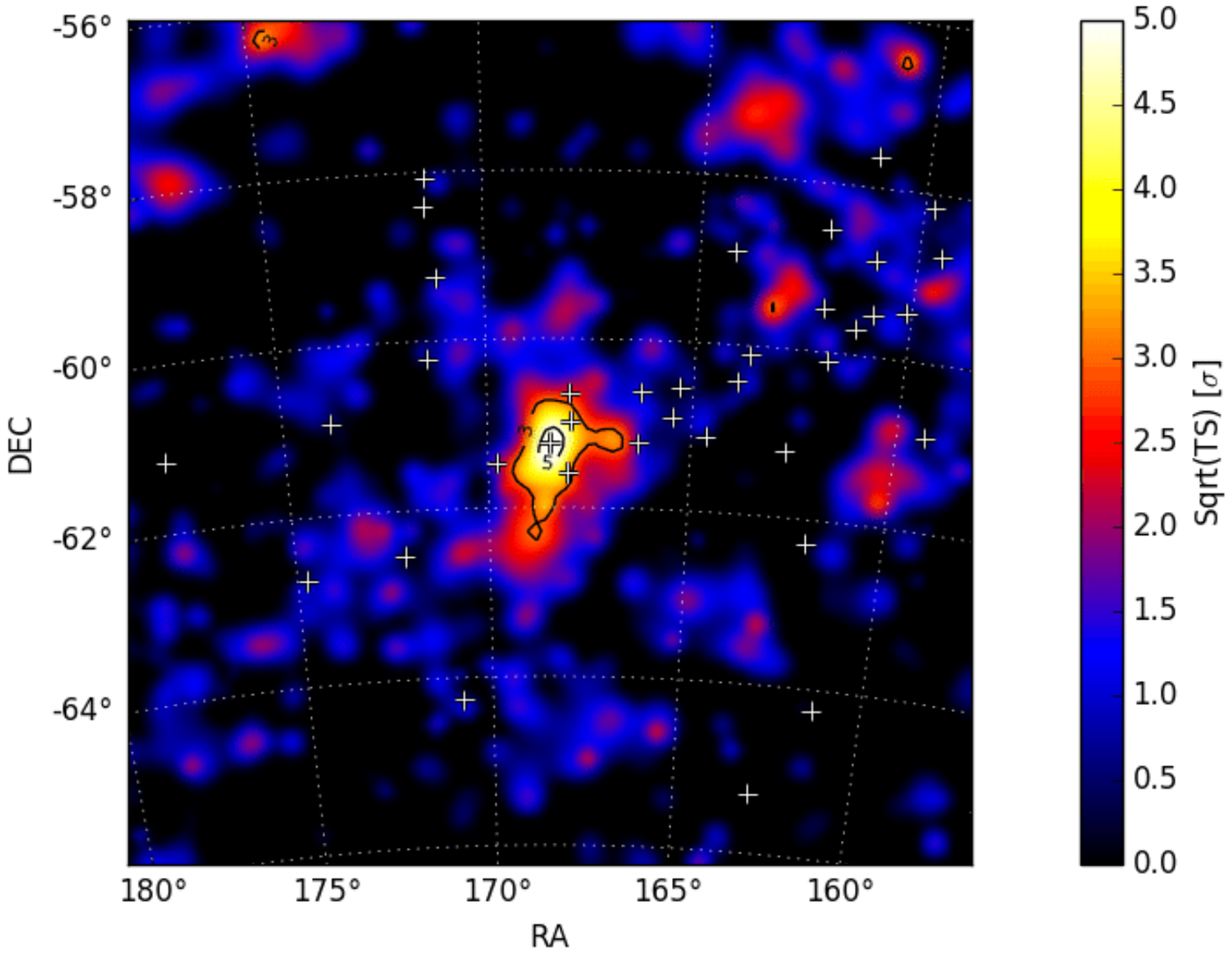


Intra- and inter-observation variability measures and light curves



High energy outliers

V679 Car

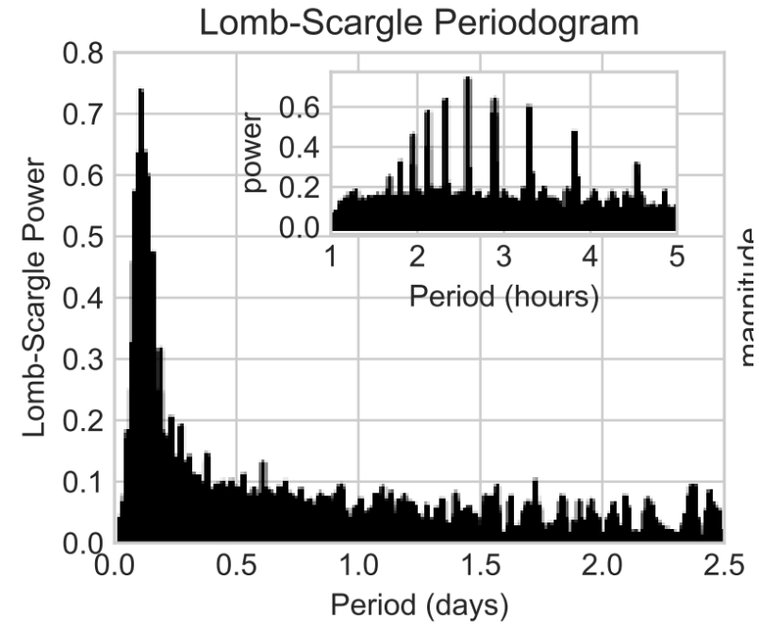


- Novae are runaway thermonuclear explosions at the surface of a white dwarf in a binary system (accretion).
- V679 Car is a rare case go γ-ray emitting novae. Reason for this is not well understood.
- Massive white dwarf: potential progenitor of type Ia supernovae.
- Type Ia supernovae are important for cosmology, as they are used as standard candles.

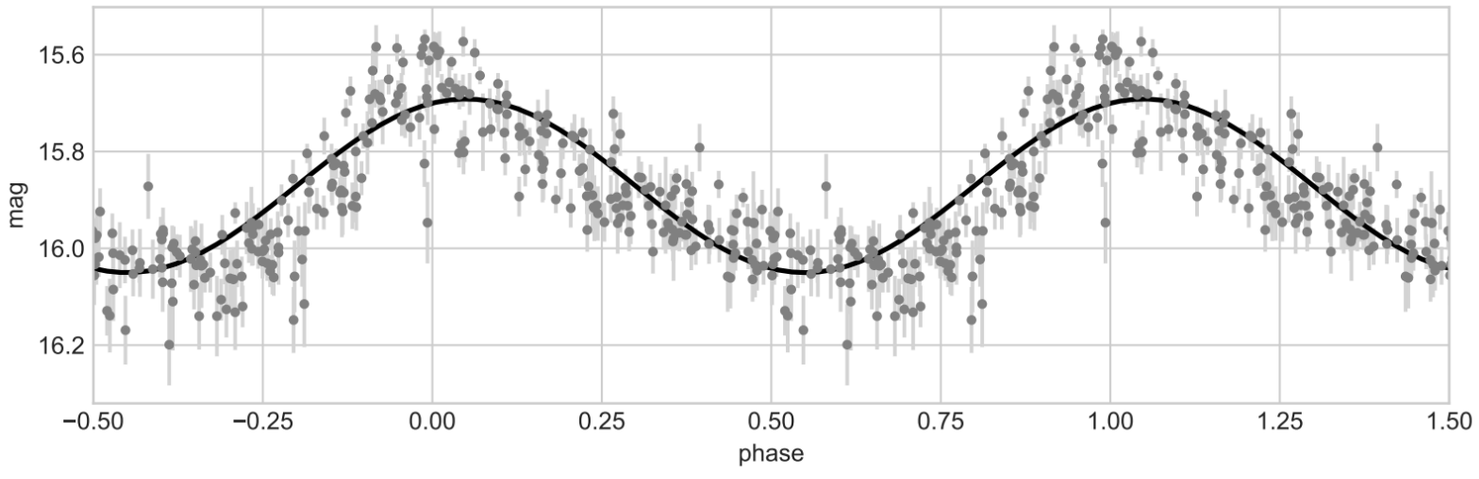
2. Feature engineering

Feature engineering for light curves

Combination of the LC magnitude measurements and the power values of the periodogram as the features to perform our analysis. The latter are the minimum X^2 values of fits to the light curve using sinusoidal curves of a given frequency.

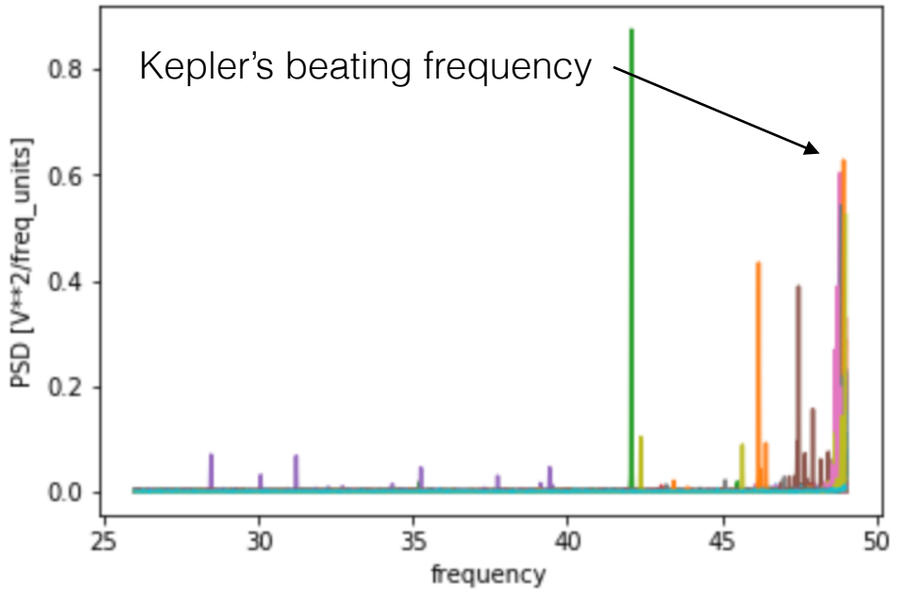


J. Vanderplas,. 2017

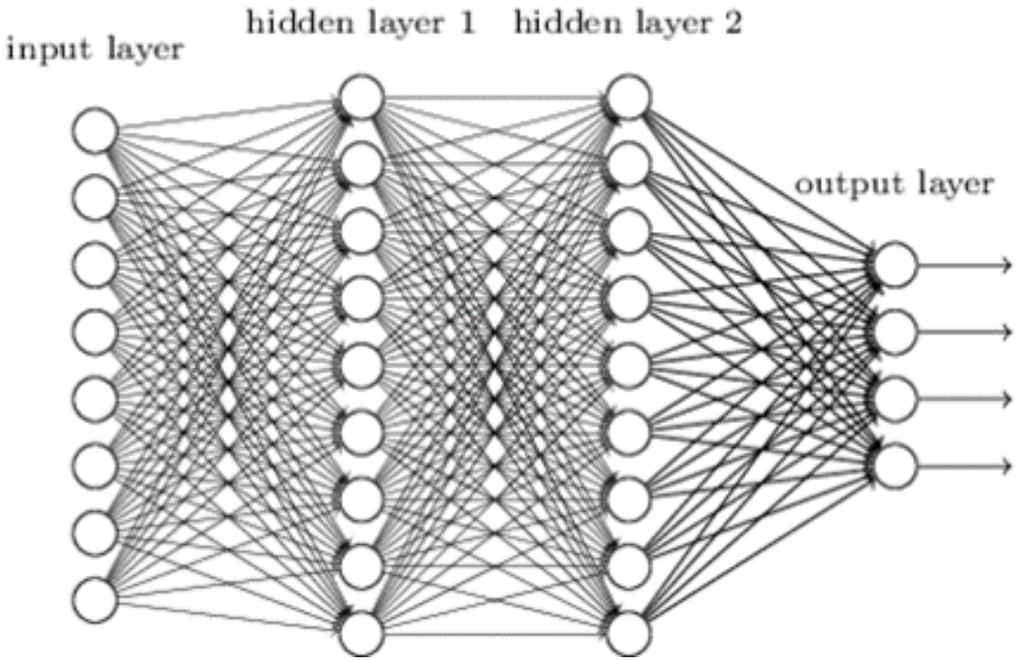


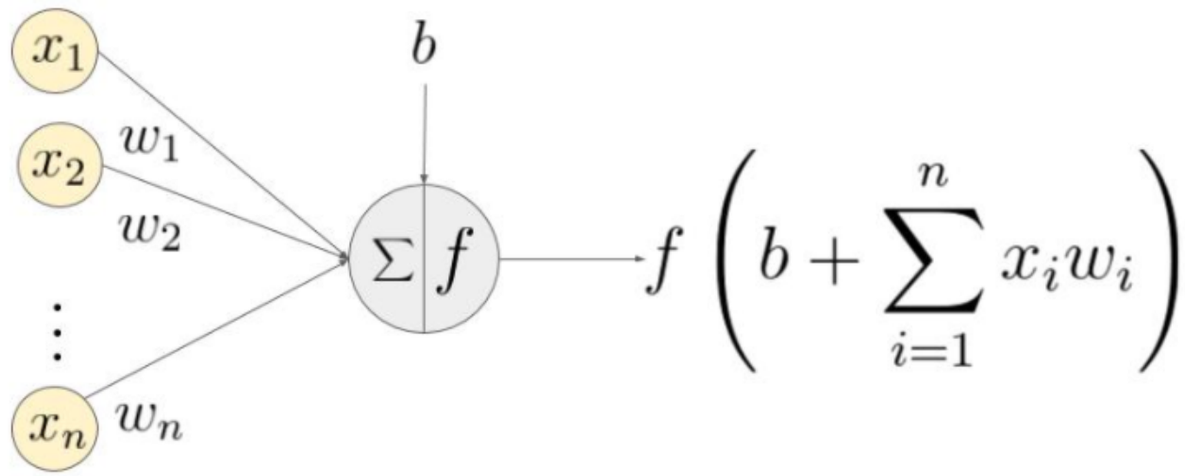
$$y(t;f) = A_f \sin(2\pi f(t-\phi_f)) \longrightarrow \chi^2(f) \equiv \sum_n \left(y_n - y(t_n;f)\right)^2$$

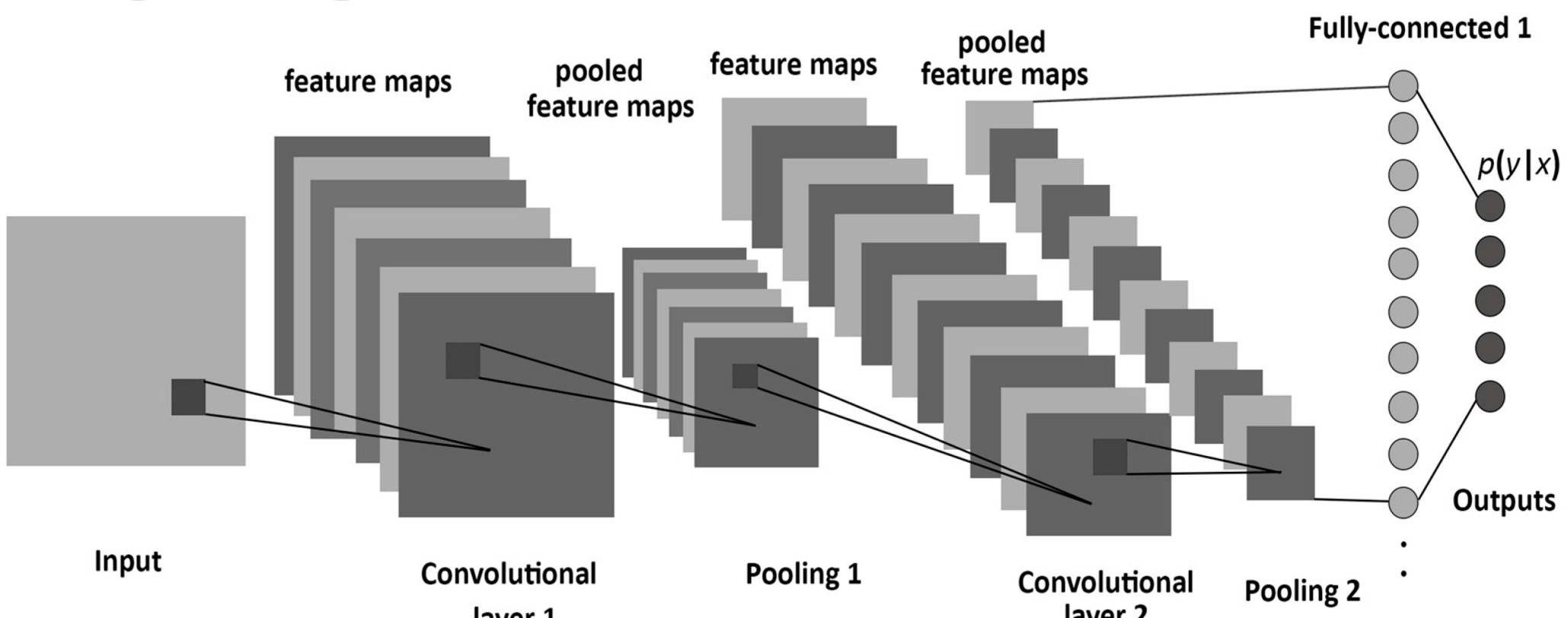
$$P(f) = \frac{1}{2}\left[\hat{\chi}_0^2 - \hat{\chi}^2(f)\right]$$



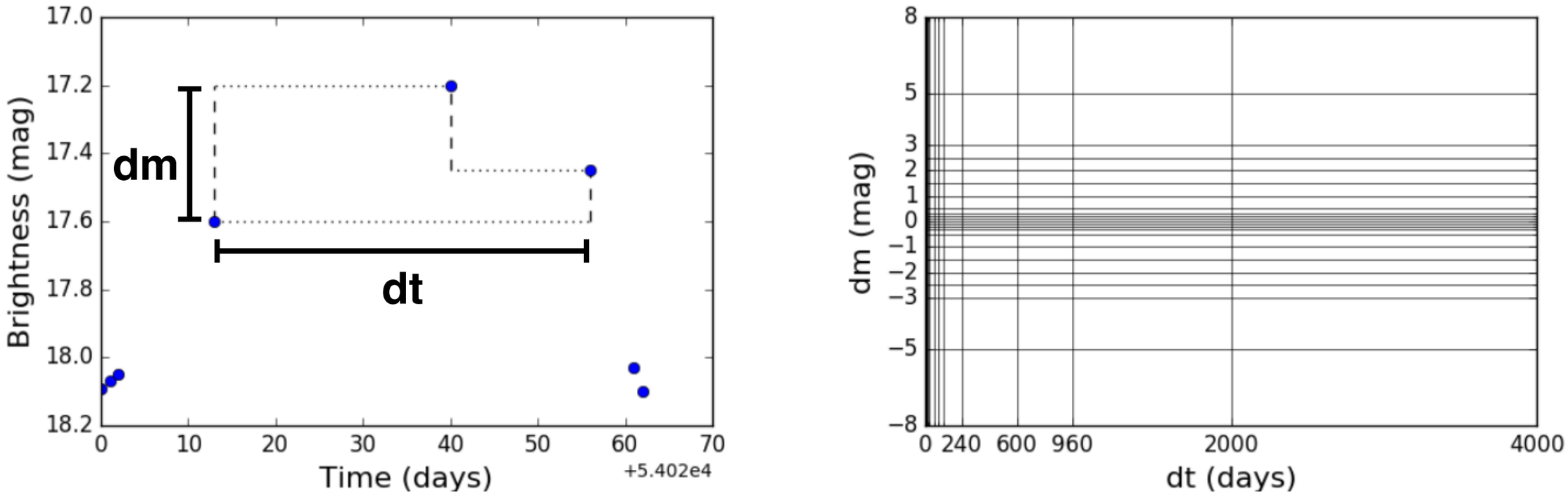
Deep Learning



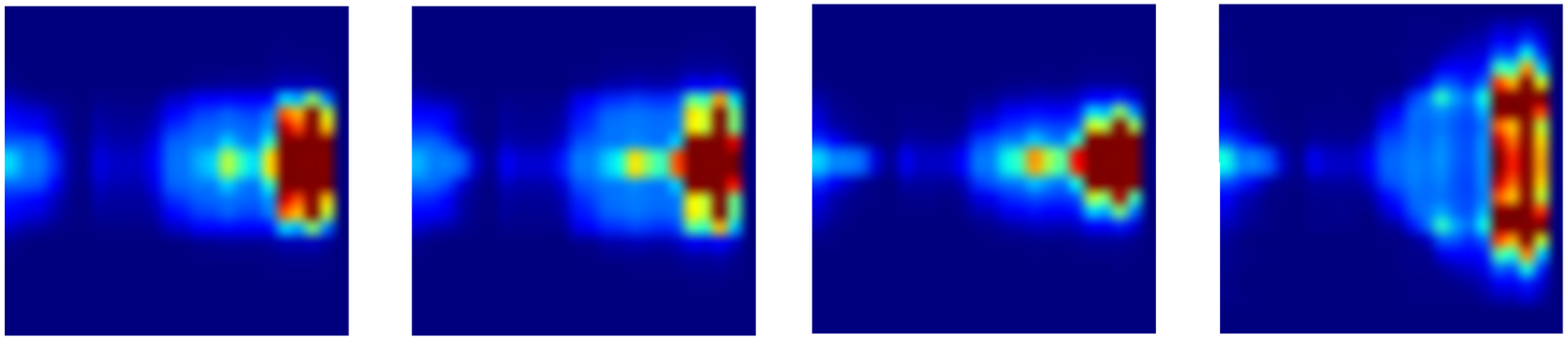




DMDT mappings + Deep Learning

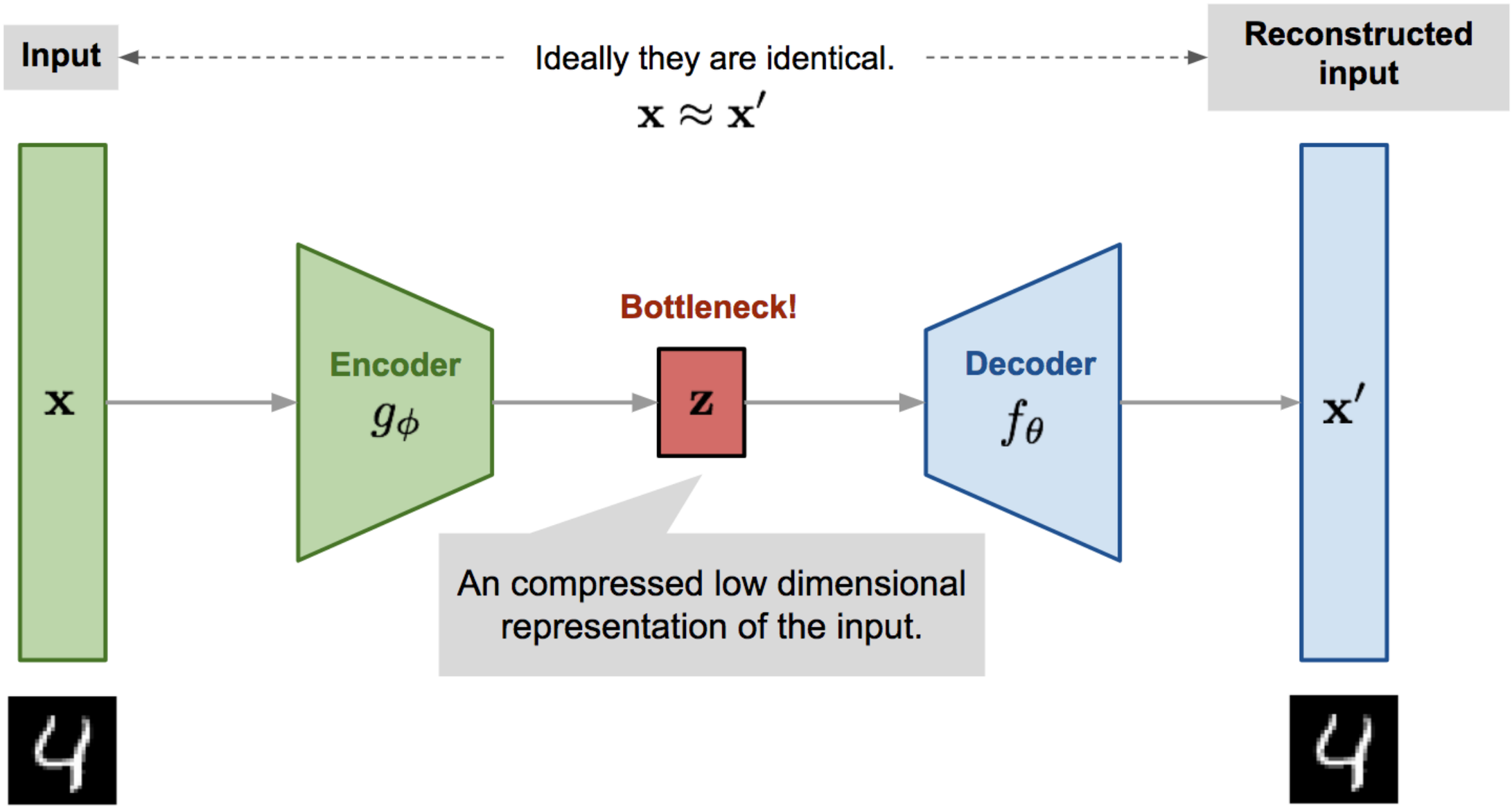


- For each pair of points, dm and dt measured.
- Resulting points binned in a semi-logarithmic bins to preserve frequency structure.
- Resulting unequal area bins transformed in equal area pixels.

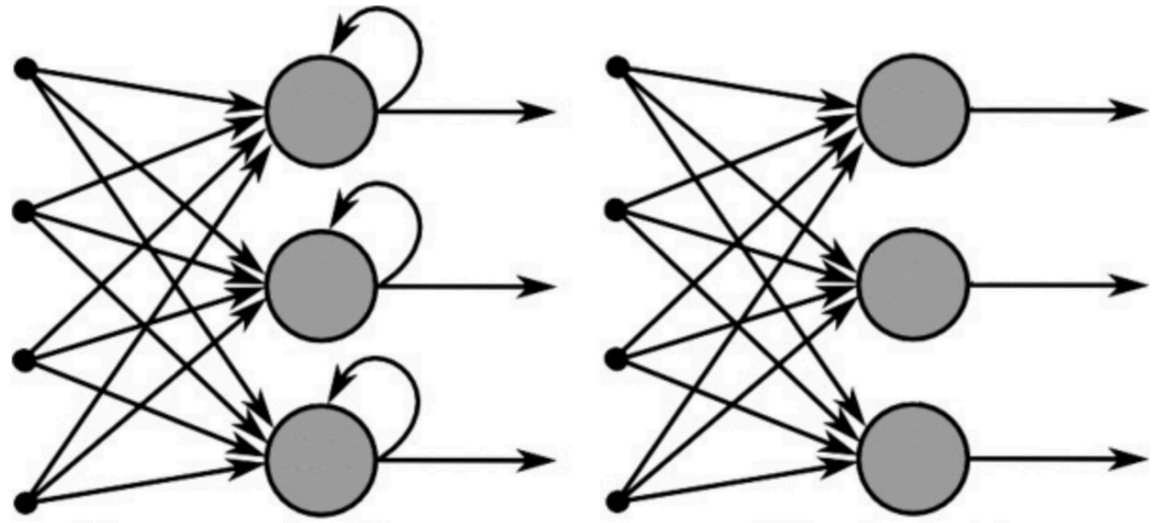


Mahabal et al. 2017

Autoencoders

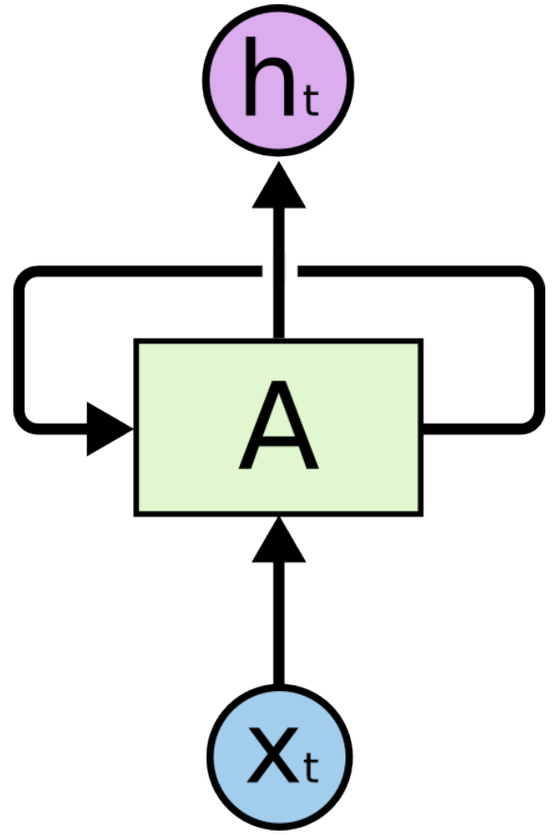


Recurrent Neural Networks



Recurrent Neural Network

Feed-Forward Neural Network

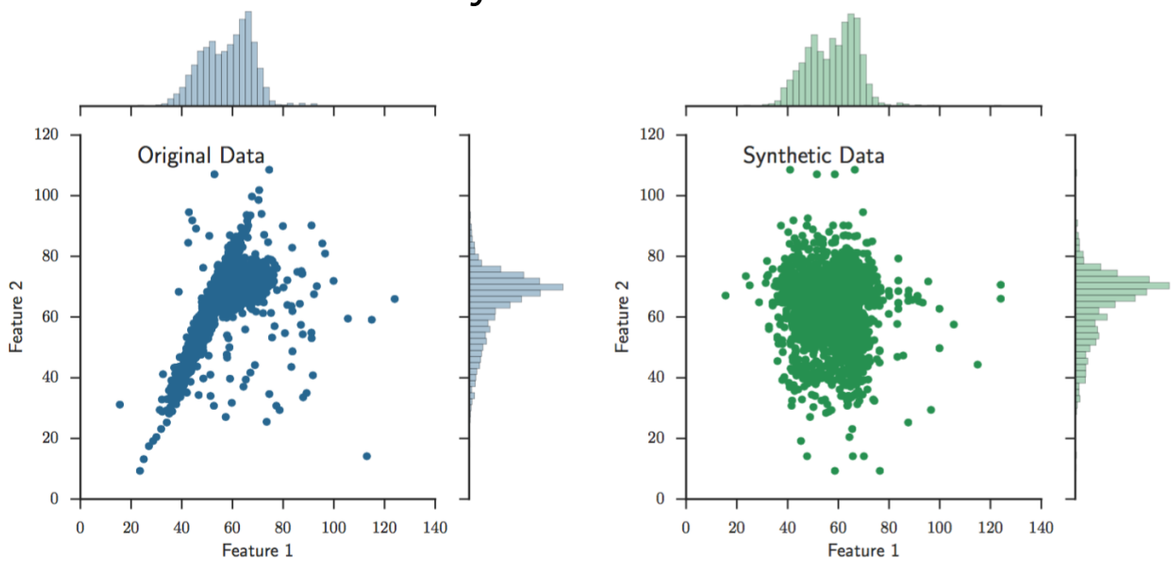


- RNNs back propagate information in time to update the network's weights.
- They have memory. They can remember information in a sequence.
- Two inputs for each node: Current own state and previous nodes.

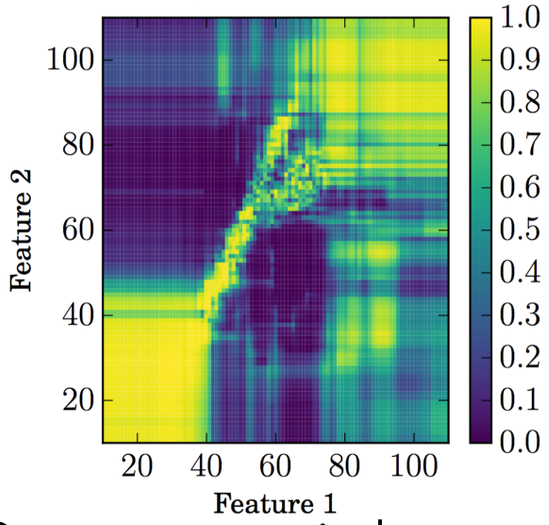
3. Methods

Unsupervised Random Forest

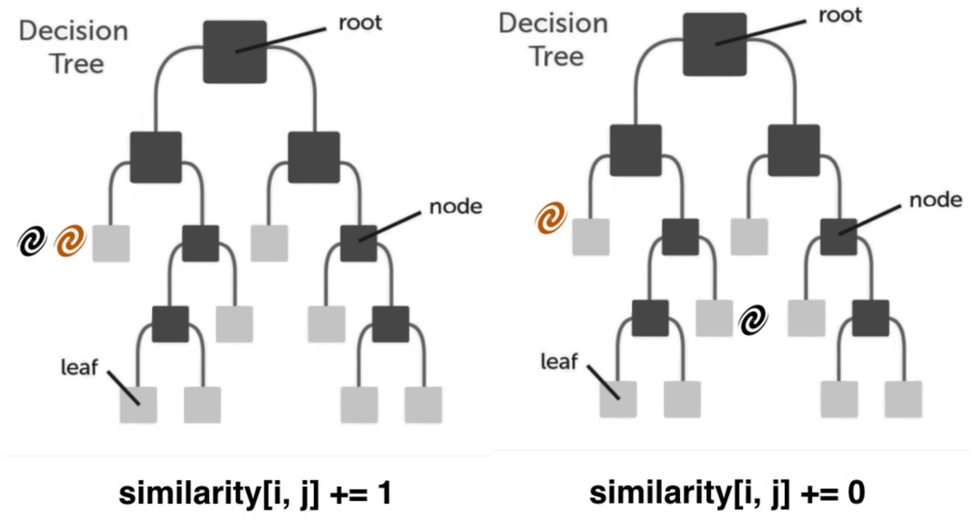
1. Generate synthetic dataset from



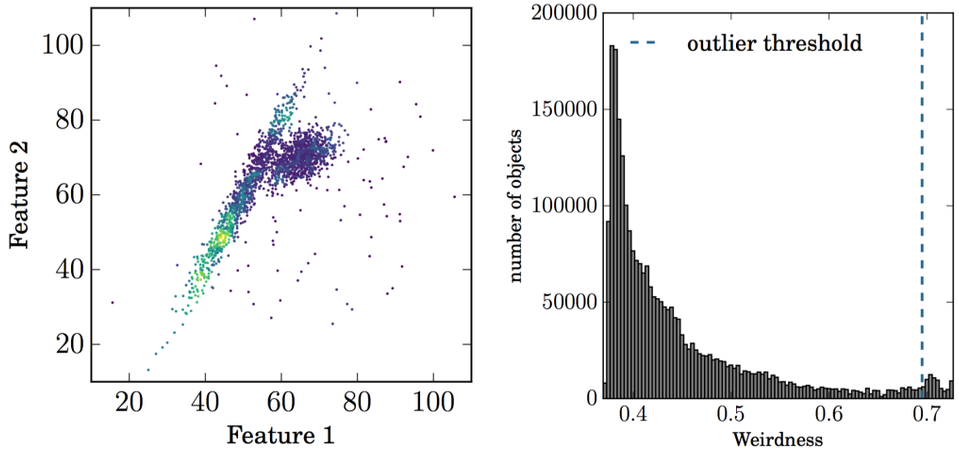
2. Train RF classifier with real+synthetic data



3. Propagate the real data



4. Compute weirdness



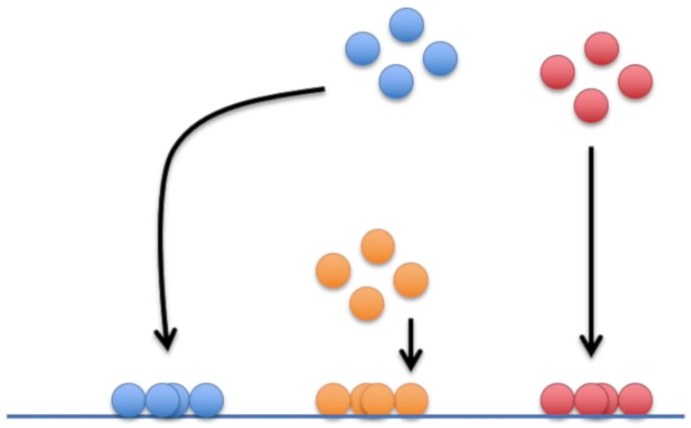
Baron & Poznanski , 2017

t-SNE

Outliers: euclidean distance
in lower dimensionality space

Courtesy StatQuest

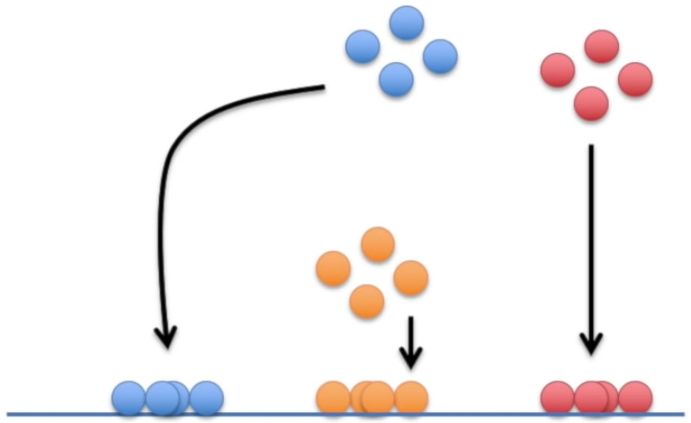
t-SNE

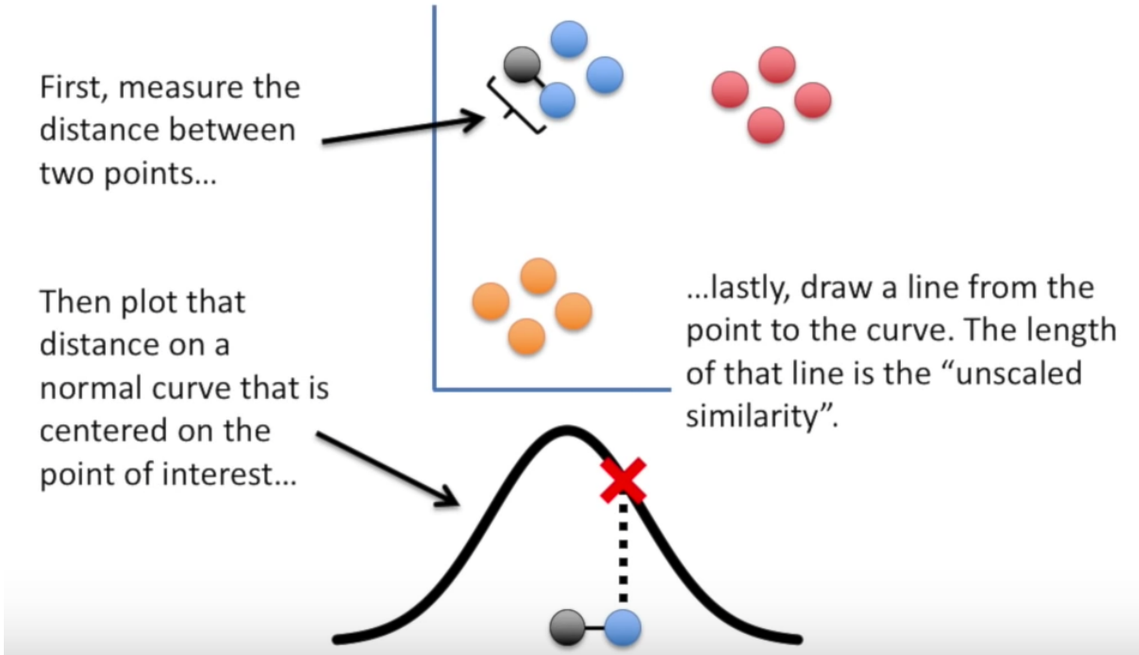


Outliers: euclidean distance
in lower dimensionality space

Courtesy StatQuest

t-SNE

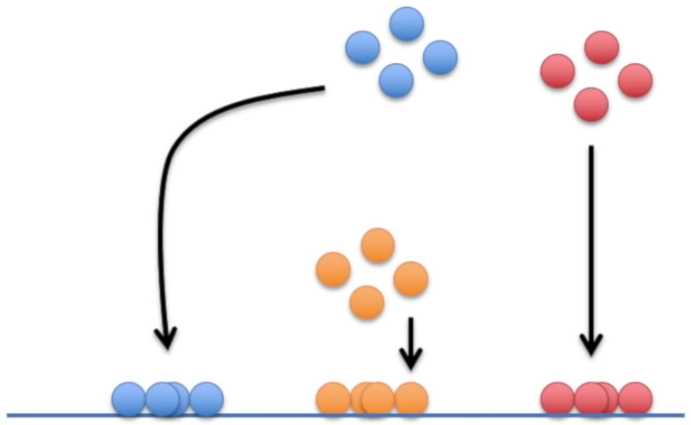


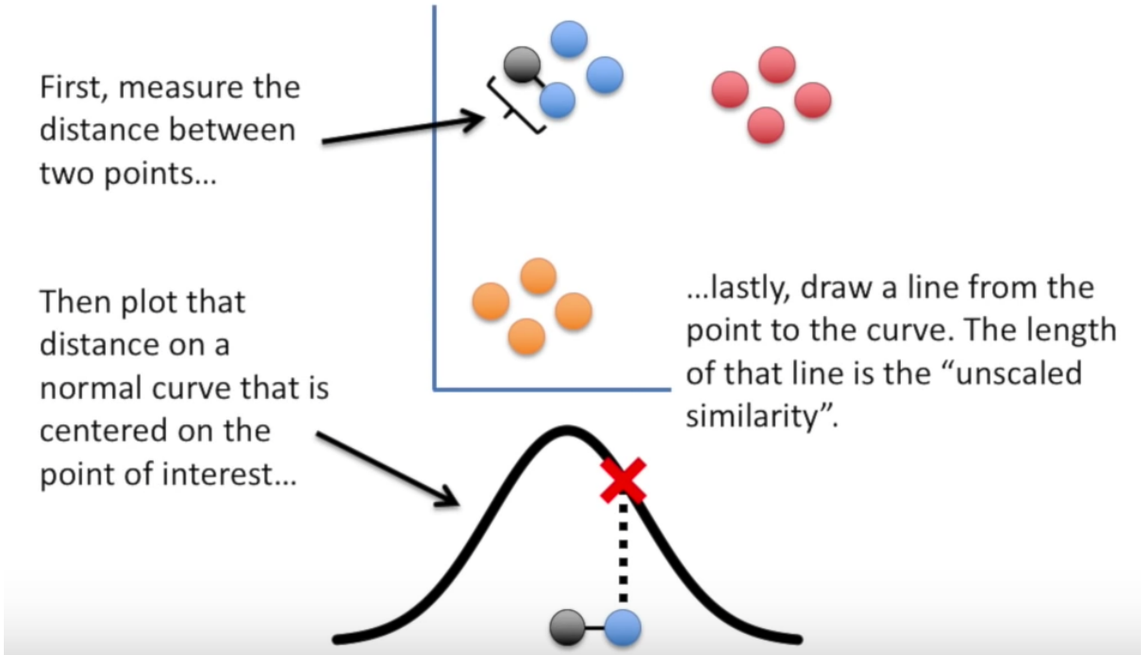


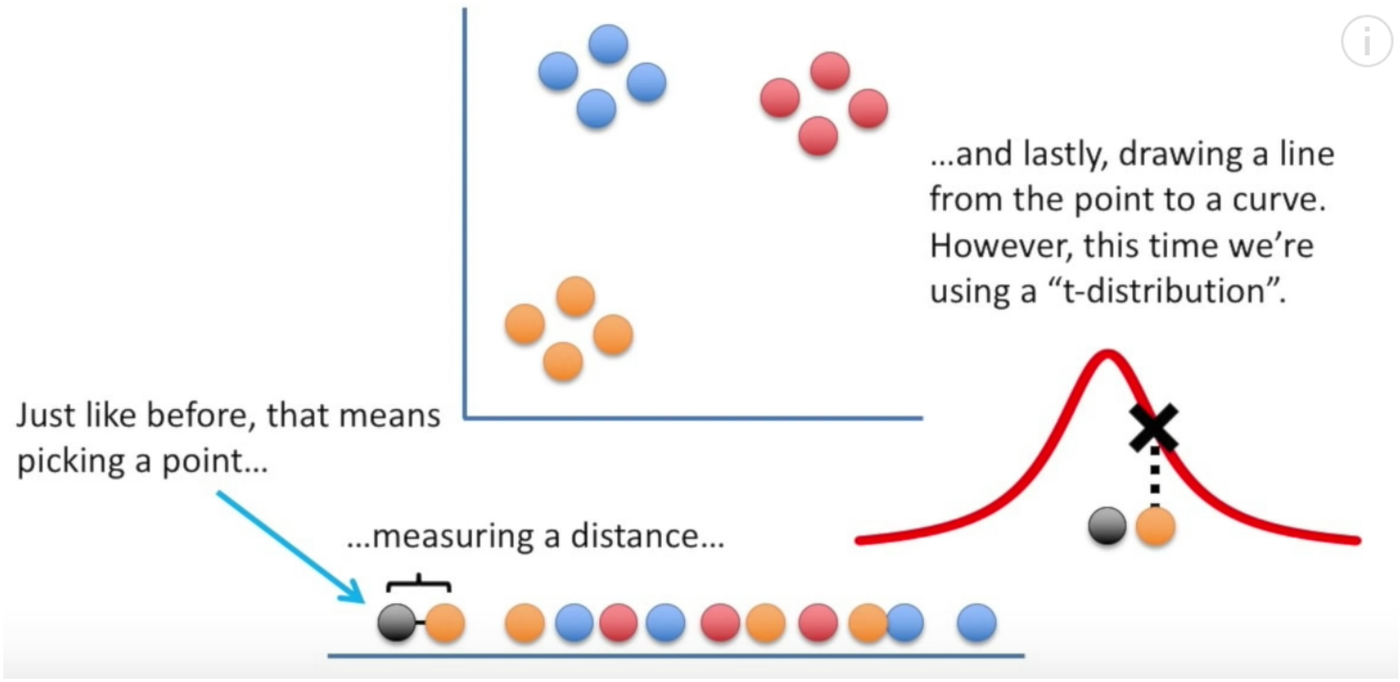
Outliers: euclidean distance
in lower dimensionality space

Courtesy StatQuest

t-SNE



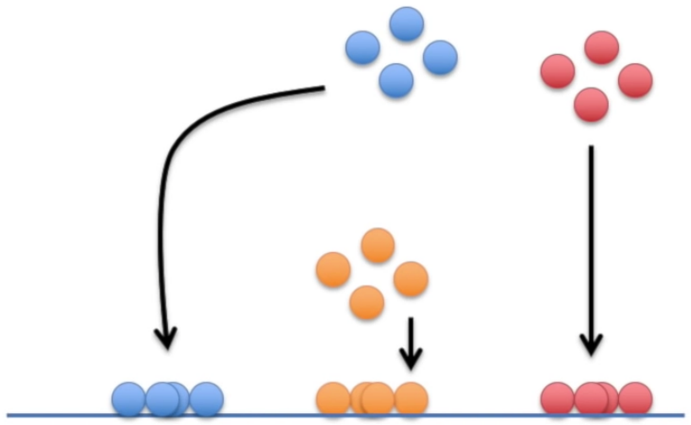


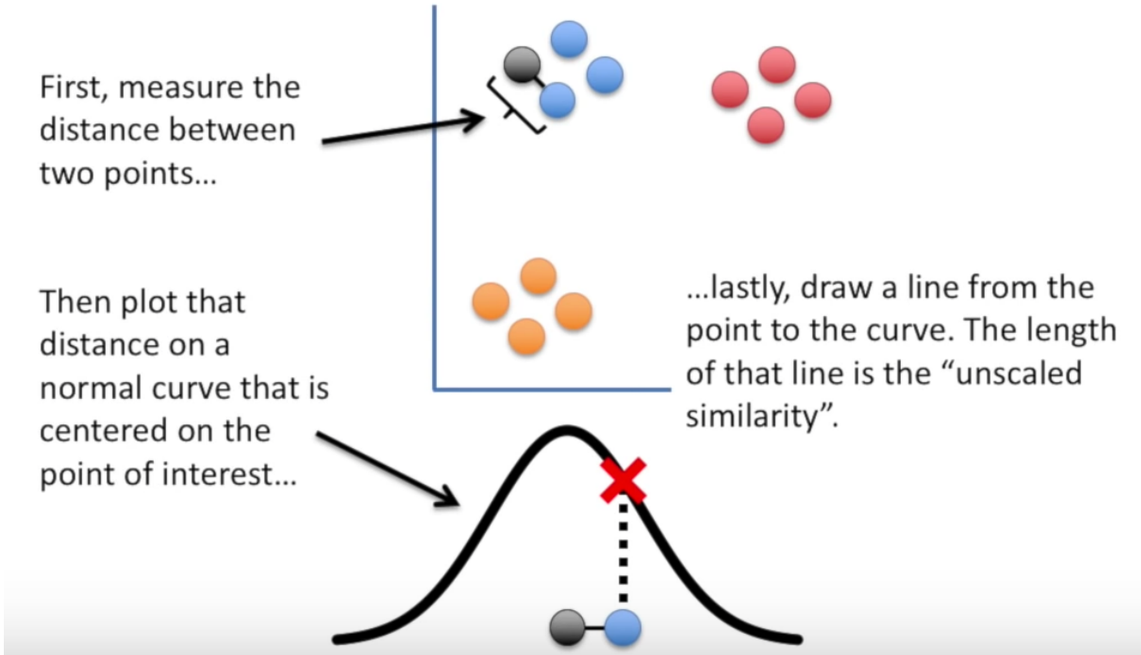


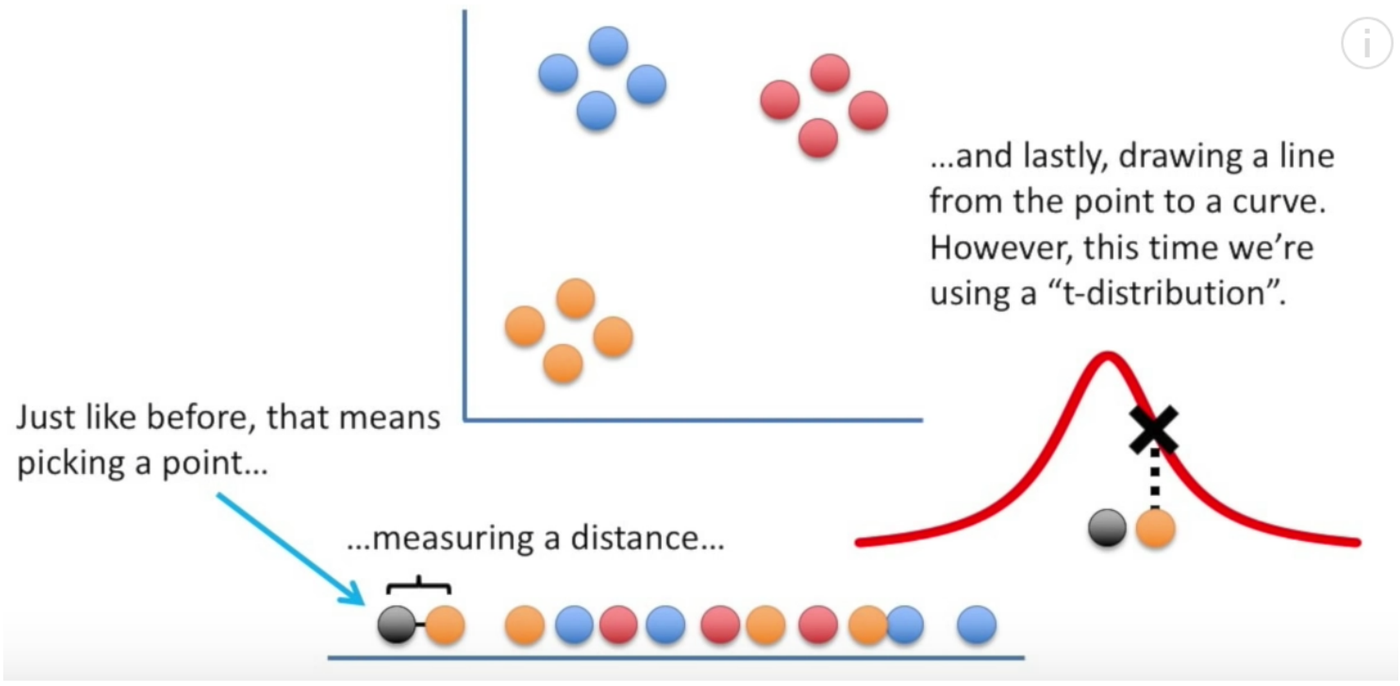
Outliers: euclidean distance in lower dimensionality space

Courtesy StatQuest

t-SNE





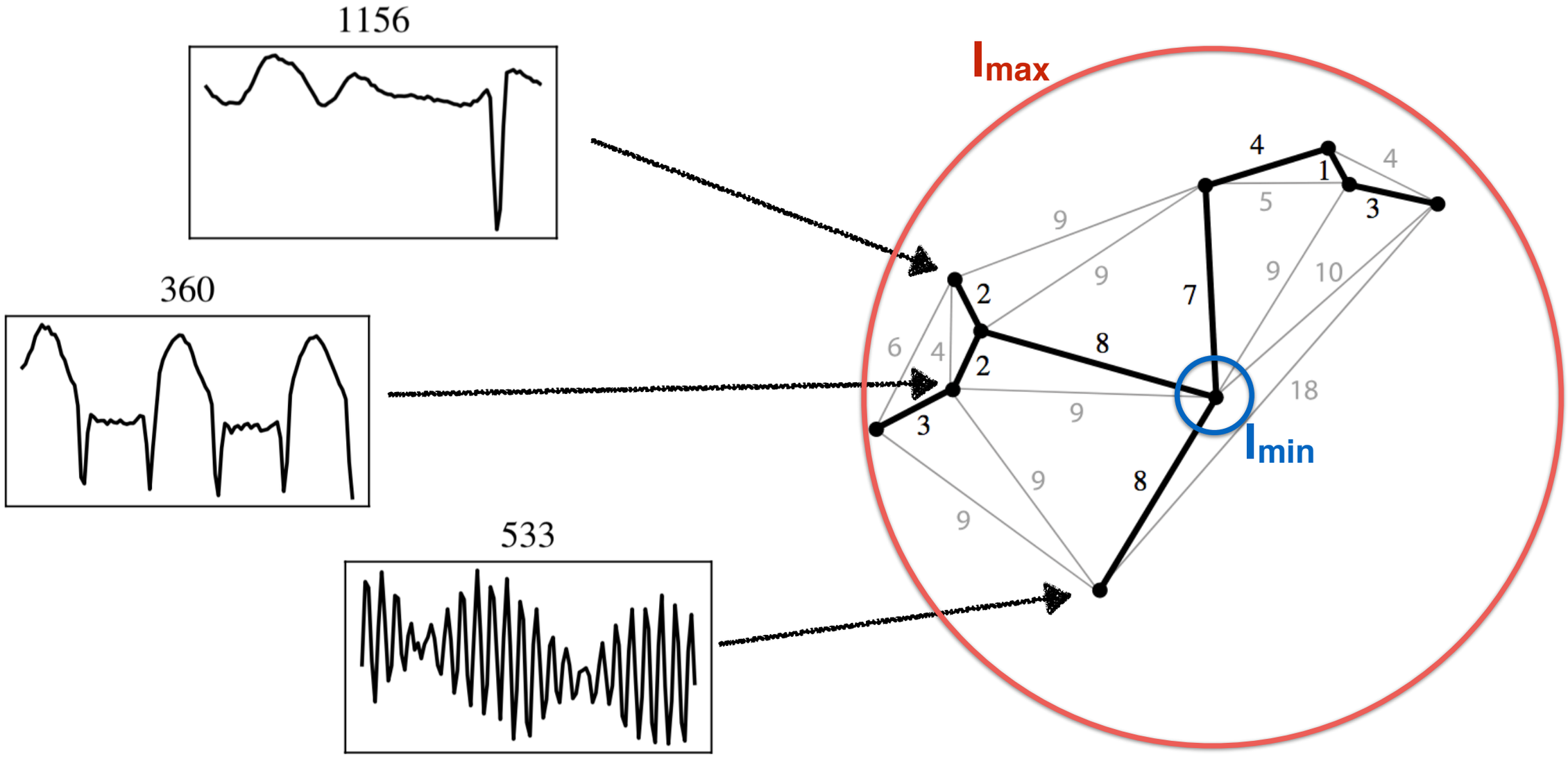


Minimize KL-divergence between two distributions (with gradient descent, for example)

Outliers: euclidean distance in lower dimensionality space

Courtesy StatQuest

Minimum Spanning Persistence

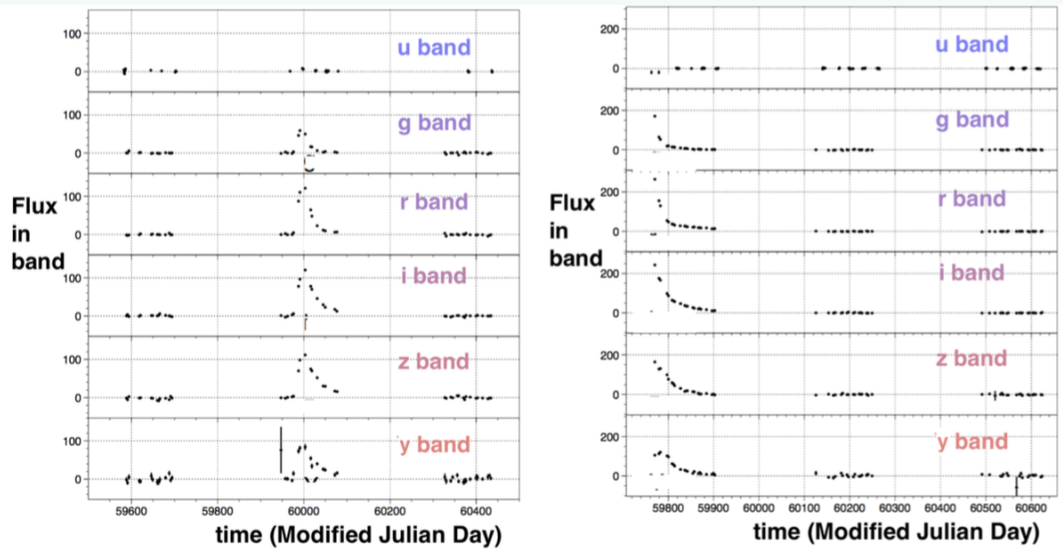


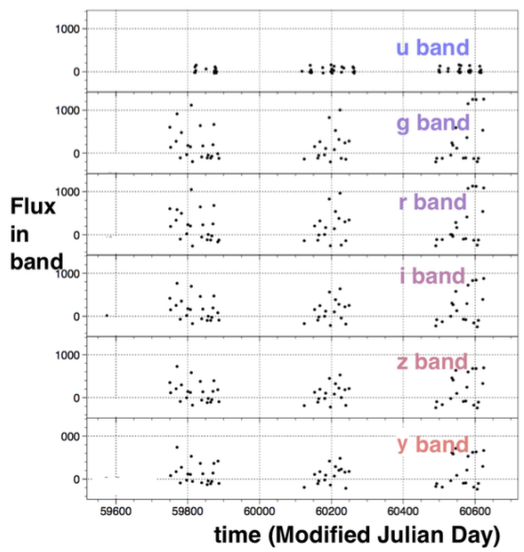
- l_{max}: length scale above which all nodes are connected
- l_{min}: length scale below which all nodes are separated
- In this case, persistence can be estimated as the fraction of l_{max}-l_{min} for which the node is isolated.

4. Results

Training set:

- ~8000 light curves in 6 bands, with LSST-like cadences.
- 14+ different classes, both galactic and extragalactic, both transients and variable sources.
- Feature extraction: multi-band periodogram





Hlozek et al. 2018

Summary of Models used in PLAsTiCC

(full article in preparation)

model class num[a]: name	model description	contributor(s)[b]	Nevent Gen[c]	Nevent train[d]	Nevent test[e]	redshift range[f]
90: SNIa	WD detonation, Type Ia SN	RK	16,353,270	2,313	1,659,831	< 1.6
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	< 0.9
52: SNIax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	< 1.3
42: SNII	Core Collapse, Type II SN	SG,LG:RK,JRP:VAV	59,198,660	1,193	1,000,150	< 2.0
62: SNIbc	Core Collapse, Type Ibc SN	VAV:RK,JRP	22,599,840	484	175,094	< 1.3
95: SLSN-I	Super-Lum. SN (magnetar)	VAV	90,640	175	35,782	< 3.4
15: TDE	Tidal Disruption Event	VAV	58,550	495	13,555	< 2.6
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	< 0.3
88: AGN	Active Galactic Nuclei	SD	175,500	370	101,424	< 3.4
92: RRL	RR lyrae	SD	200,200	239	197,155	0
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	0
16: EB	Eclipsing Binary stars	AP	220,200	924	96,572	0
53: Mira	Pulsating variable stars	RH	1,490	30	1,453	0
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2,820	151	1,303	0
991: μLens-Binary	μ-lens from binary lens	RD,AA	1,010	0	533	0
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1,702	< 0.4
993: CaRT	Calcium Rich Transient	VAV	2,834,500	0	9,680	< 0.9
994: PISN	Pair Instability SN	VAV	5,650	0	1,172	< 1.9
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	0
TOTAL	Sum of all models		117,128,700	7,846	3,492,888	—

[a] num>990 were all in unknown class 99 during the competition. An extra digit is added here to distinguish each model.
[b] Co-author initials. Colon separates independent methods.
[c] Number of generated events, corresponding to the true population without observational selection bias.
[d] Labeled subset from spectroscopic classification. 0 → predicted from theory, not convincingly observed, or very few observations.
[e] Unlabeled sample. `PLAsTiCC` goal is to label this sample.
[f] Redshift> 0 for extragalactic models; Redshift= 0 for Galactic models.

Model Contributors:

AA: Arturo Avelino (Harvard U.)
EB: Etienne Bachelet (LCO)
DC: David Chernoff (Cornell U.)
MD: Mi Dai (Rutgers U.)
SD: Scott Daniel (U.Washington)
RD: Rosanne Di Stefano (Harvard U.)
LG: Lluís Galbany (U.Pitt)
SG: Santiago González-Gaitán (U.Lisbon)
RH: Renée Hlozek (U.Toronto)
SJ: Saurabh Jha (Rutgers U.)
DK: Dan Kasen (U.C. Berkeley)
RK: Rick Kessler (U.Chicago)
GN: Gautham Narayan (STScI)
JRP: Justin Pierel (U. South Carolina)
AP: Andrej Prsa (Villanova U.)
VAV: Ashley Villar (Harvard U.)

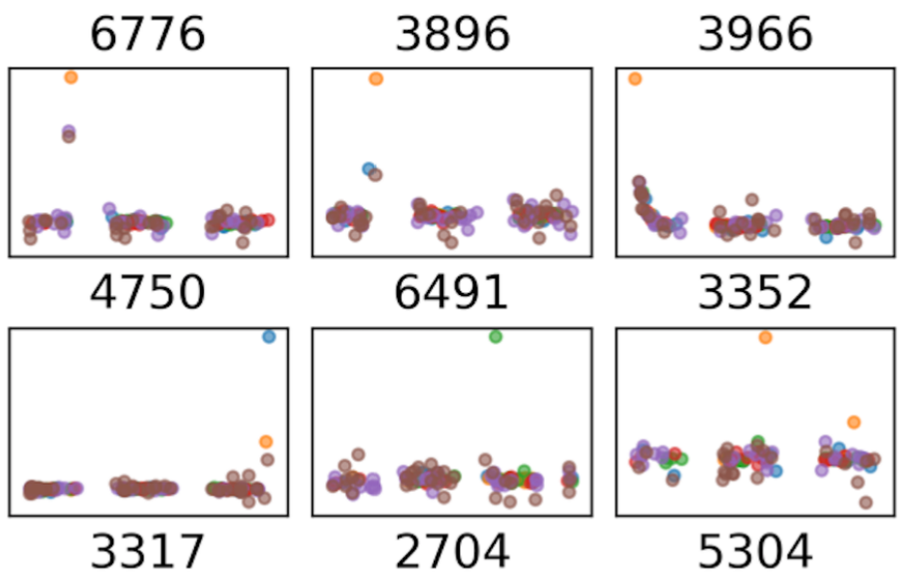
Unblinded Data Files: http://doi.org/10.5281/zenodo.2539456

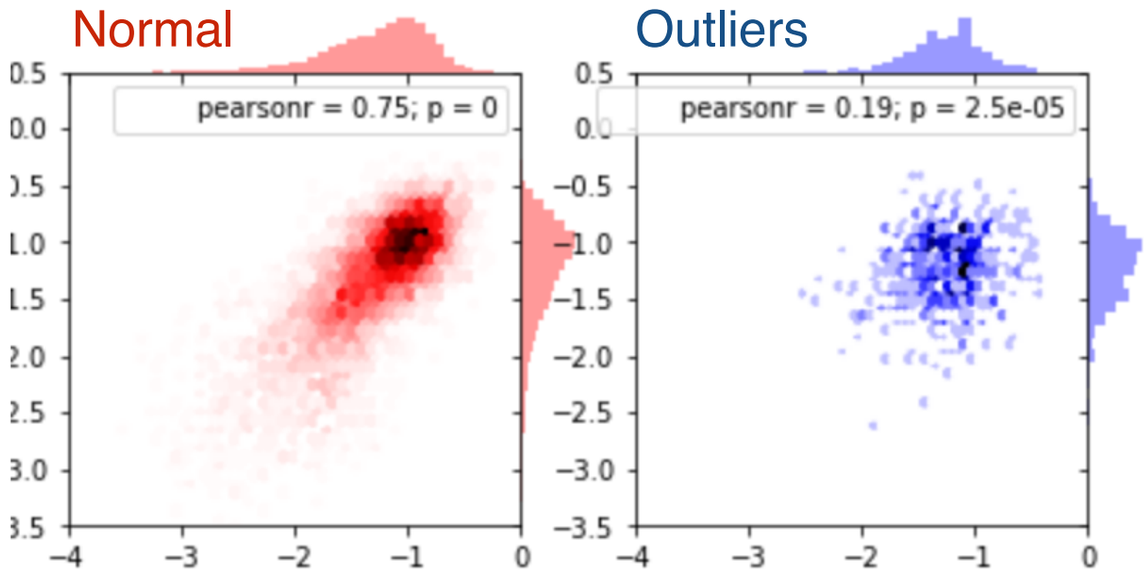
Simulation Source code: http://snana.uchicago.edu
(model libraries will be released with article)

URF: the meaning of weirdness

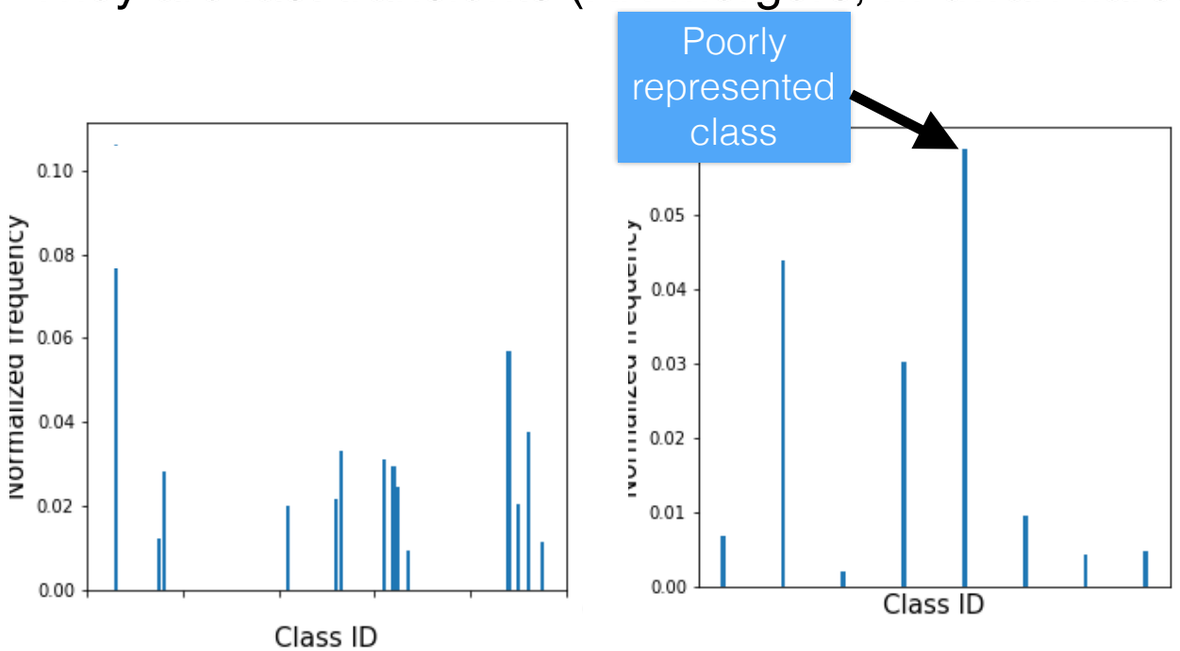
Outliers in LSST-like survey:

- Weirdest objects are fast transients.
- URF quite efficient at picking certain classes that are poorly represented in the training set.
- “Normal” objects, on the other hand, are fairly well distributed across classes.





Outliers have an uncorrelated power spectrum.
They are fast transients (KN mergers, M-dwarf flares).



KN mergers, flaring M-dwarf

Summary of Models used in PLAsTiCC

(full article in preparation)

model class num[a]: name	model description	contributor(s)[b]	Nevent Gen[c]	Nevent train[d]	Nevent test[e]	redshift range[f]
90: SNIa	WD detonation, Type Ia SN	RK	16,353,270	2,313	1,659,831	< 1.6
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	< 0.9
52: SNIax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	< 1.3
42: SNII	Core Collapse, Type II SN	SG,LG:RK,JRP:VAV	59,198,660	1,193	1,000,150	< 2.0
62: SNIbc	Core Collapse, Type Ibc SN	VAV:RK,JRP	22,599,840	484	175,094	< 1.3
95: SLSN-I	Super-Lum. SN (magnetar)	VAV	90,640	175	35,782	< 3.4
15: TDE	Tidal Disruption Event	VAV	58,550	495	13,555	< 2.6
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	< 0.3
88: AGN	Active Galactic Nuclei	SD	175,500	370	101,424	< 3.4
92: RRL	RR lyrae	SD	200,200	239	197,155	0
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	0
16: EB	Eclipsing Binary stars	AP	220,200	924	96,572	0
53: Mira	Pulsating variable stars	RH	1,490	30	1,453	0
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2,820	151	1,303	0
991: μLens-Binary	μ-lens from binary lens	RD,AA	1,010	0	533	0
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1,702	< 0.4
993: CaRT	Calcium Rich Transient	VAV	2,834,500	0	9,680	< 0.9
994: PISN	Pair Instability SN	VAV	5,650	0	1,172	< 1.9
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	0
TOTAL	Sum of all models		117,128,700	7,846	3,492,888	—

[a] num>990 were all in unknown class 99 during the competition. An extra digit is added here to distinguish each model.
[b] Co-author initials. Colon separates independent methods.
[c] Number of generated events, corresponding to the true population without observational selection bias.
[d] Labeled subset from spectroscopic classification. 0 → predicted from theory, not convincingly observed, or very few observations.
[e] Unlabeled sample. `PLAsTiCC` goal is to label this sample.
[f] Redshift> 0 for extragalactic models; Redshift= 0 for Galactic models.

Model Contributors:

AA: Arturo Avelino (Harvard U.)
EB: Etienne Bachelet (LCO)
DC: David Chernoff (Cornell U.)
MD: Mi Dai (Rutgers U.)
SD: Scott Daniel (U.Washington)
RD: Rosanne Di Stefano (Harvard U.)
LG: Lluís Galbany (U.Pitt)
SG: Santiago González-Gaitán (U.Lisbon)
RH: Renée Hlozek (U.Toronto)
SJ: Saurabh Jha (Rutgers U.)
DK: Dan Kasen (U.C. Berkeley)
RK: Rick Kessler (U.Chicago)
GN: Gautham Narayan (STScI)
JRP: Justin Pierel (U. South Carolina)
AP: Andrej Prsa (Villanova U.)
VAV: Ashley Villar (Harvard U.)

Unblinded Data Files: http://doi.org/10.5281/zenodo.2539456

Simulation Source code: http://snana.uchicago.edu
(model libraries will be released with article)

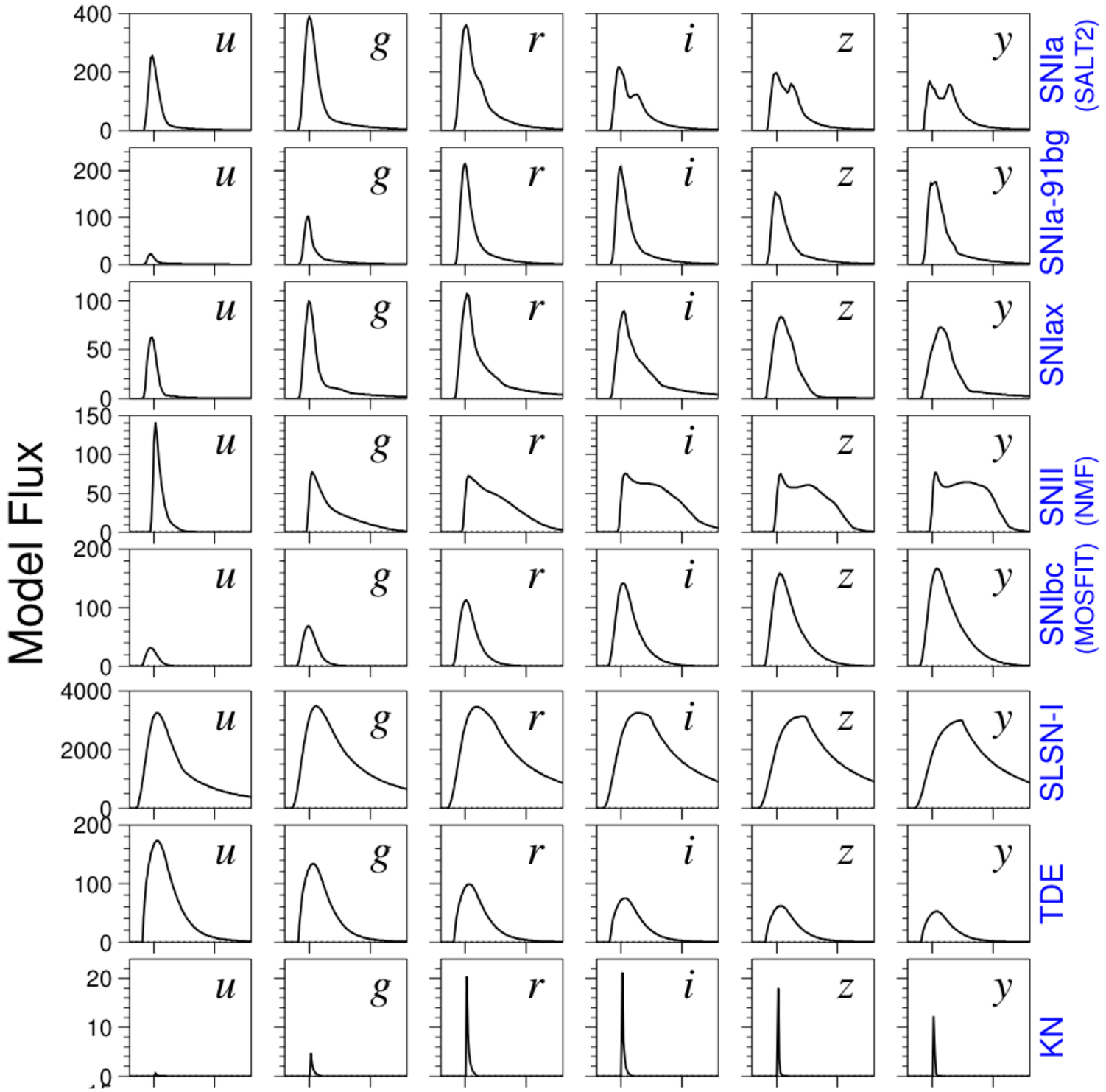
Summary of Models used in PLAsTiCC

model class num[a]: name	m de
90: SNIa	W
67: SNIa-91bg	Pe
52: SNIax	Pe
42: SNII	Cc
62: SNIbc	Cc
95: SLSN-I	Su
15: TDE	Ti
64: KN	Ki
88: AGN	Ac
92: RRL	RI
65: M-dwarf	M-
16: EB	Ec
53: Mira	Pu
6: μLens-Single	μ-
991: μLens-Binary	μ-
992: ILOT	In
993: CaRT	Ca
994: PISN	Pa
995: μLens-String	μ-
TOTAL	Su

[a]num>990 were all in unknow
[b]Co-author initials. Colon sep
[c]Number of generated events,
[d]Labeled subset from spectros
[e]Unlabeled sample. PLAsTiCC
[f]Redshift> 0 for extragalactic

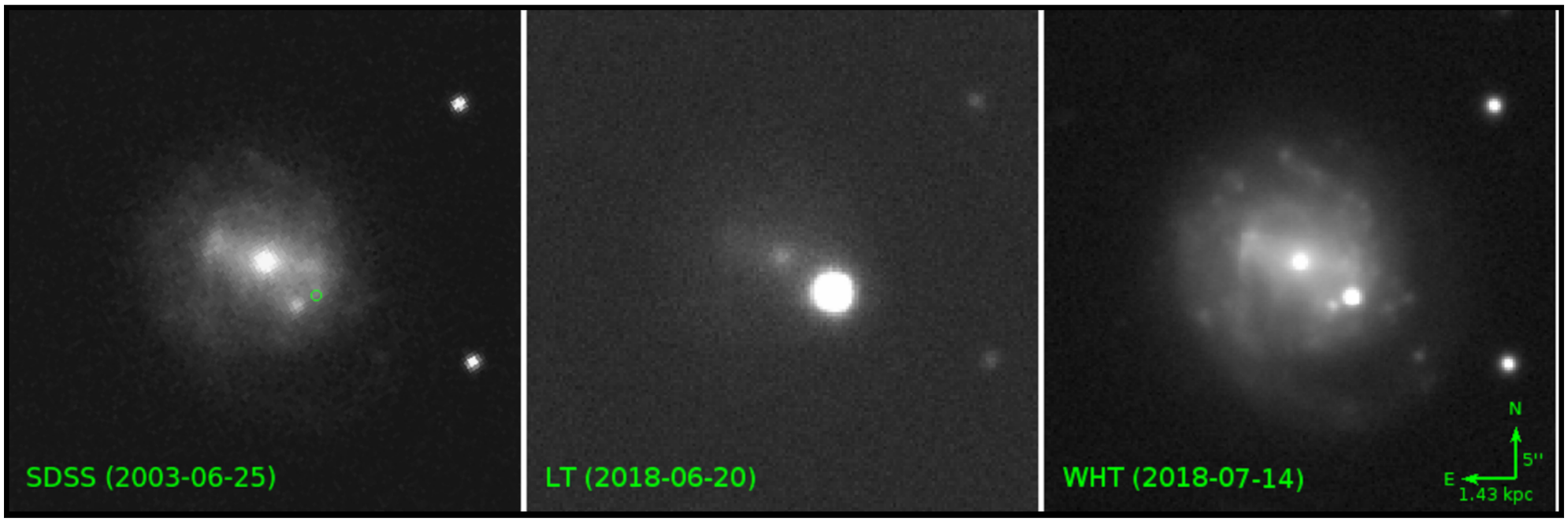
Unblinded D

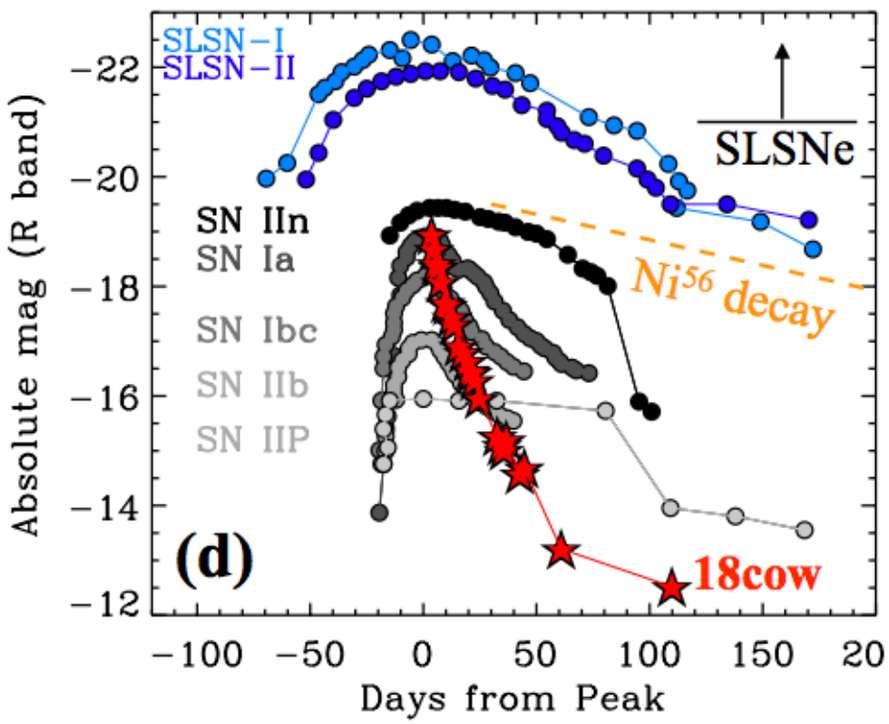
Simulat
(n



ntributors:
elino (Harvard U.)
chelet (LCO)
rnoff (Cornell U.)
utgers U.)
el (U.Washington)
Di Stefano (Harvard U.)
ny (U.Pitt)
González-Gaitán (U.Lisbon)
zek (U.Toronto)
a (Rutgers U.)
(U.C. Berkeley)
r (U.Chicago)
Narayan (STScI)
rel (U. South Carolina)
a (Villanova U.)
illar (Harvard U.)

AT2018cow

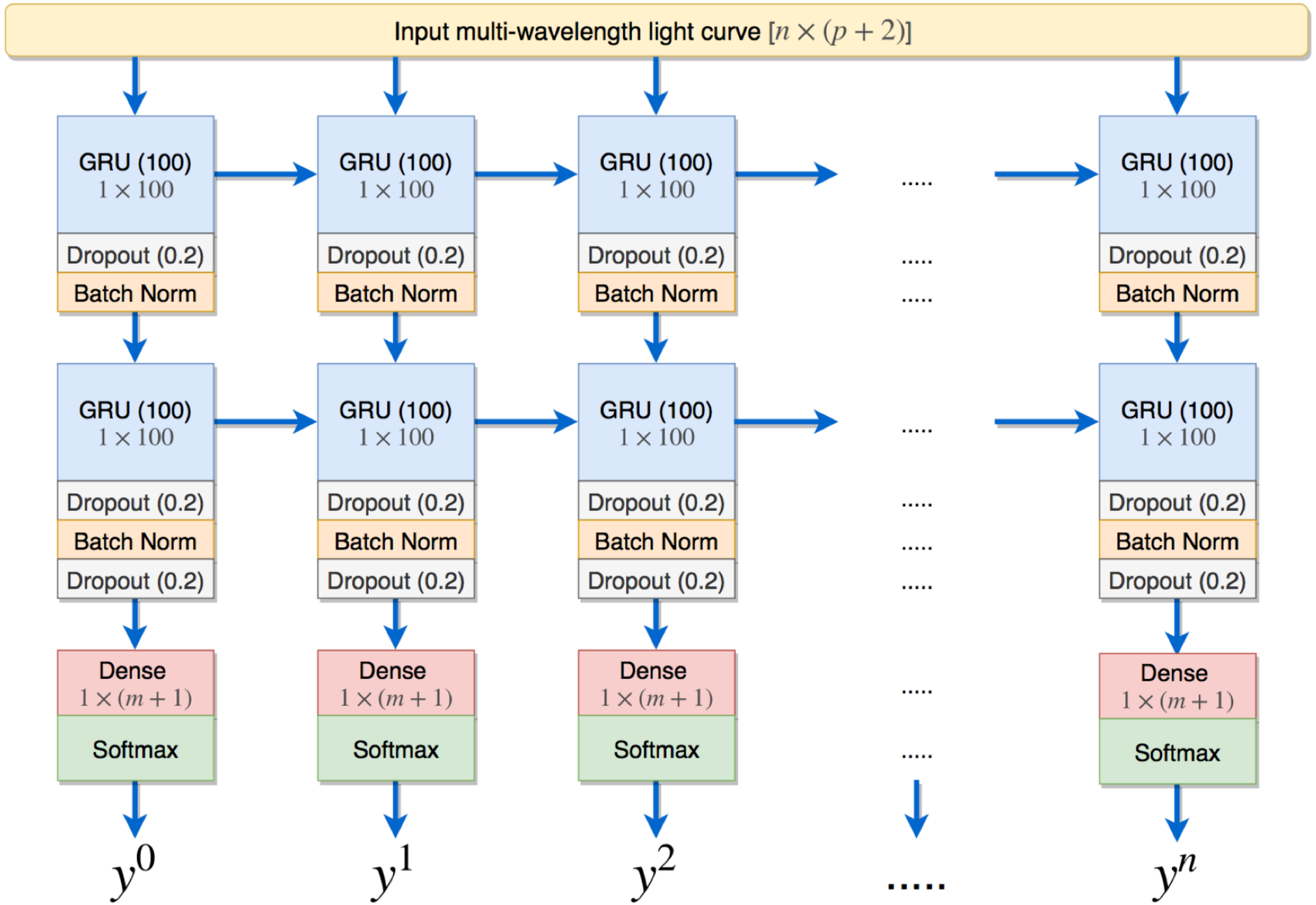




Marguitti et al. 2018

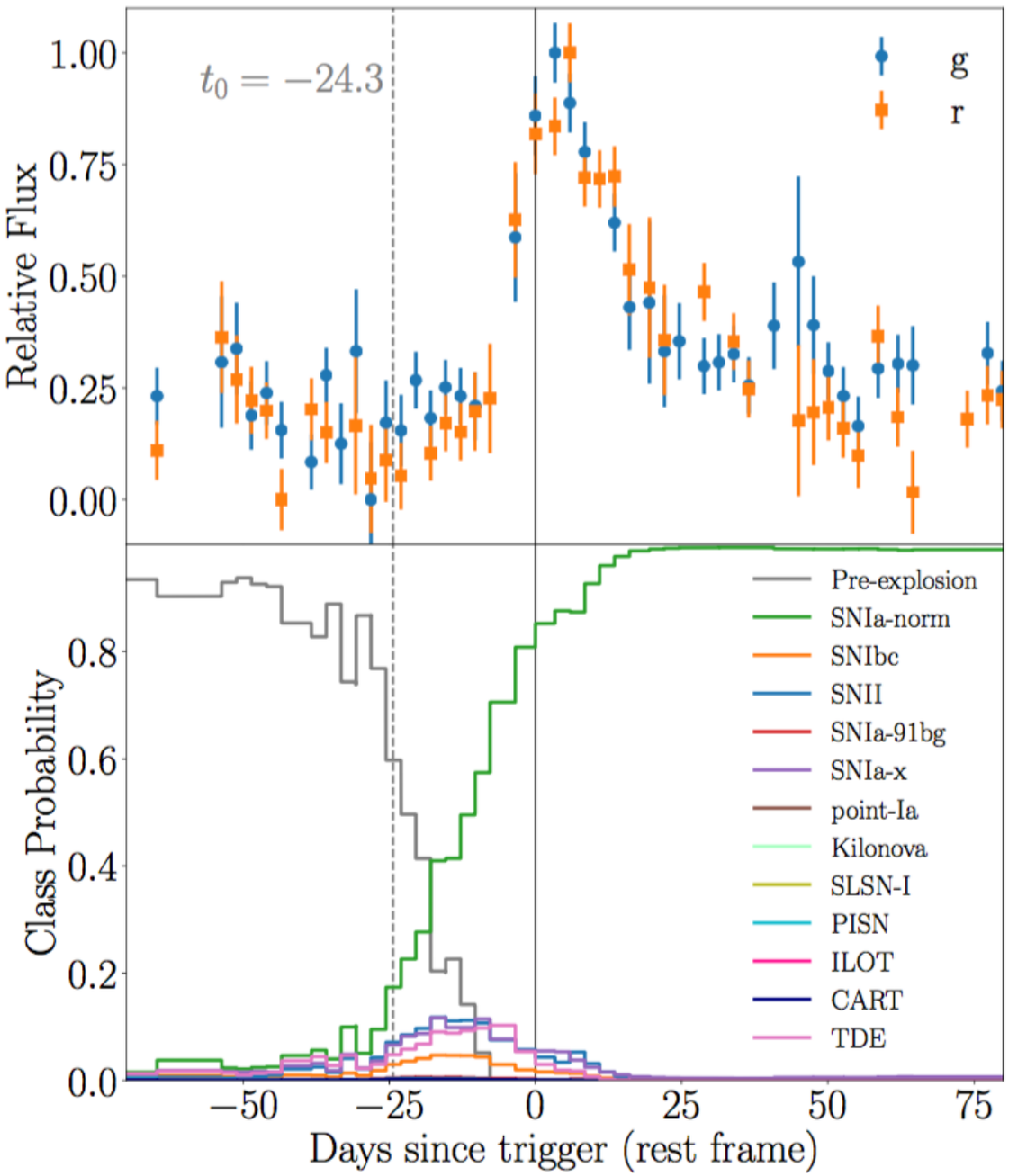
- Fast, luminous transient, first one observed live.
- Reached its peak brightness in days, not weeks.
- 10-100 times brighter than a normal supernova.
- Spectral variability does not match any know type of supernova (Perley et al. 2019).
- TDE by an intermediate-mass black hole?

RAPID



Muthukrishna et al. 2019

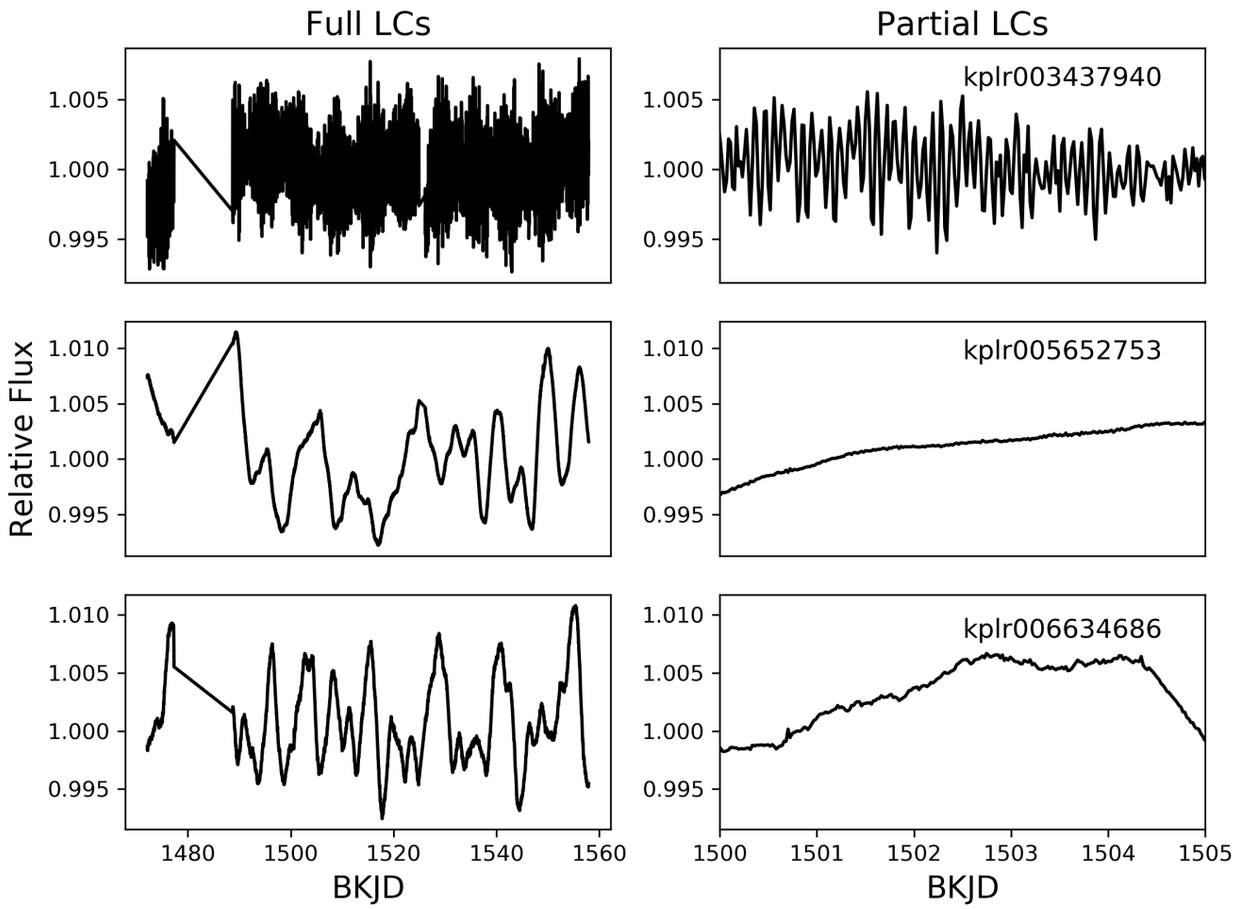
Real-time classification



Muthukrishna et al. 2019

- Classification of astronomical transients as a function of time.
- No user-defined features needed. LC points used directly.
- Early classification (within days) possible with good accuracy.
- Extremely important to prioritize spectroscopy follow ups.

Dataset: Kepler Q16 light curves



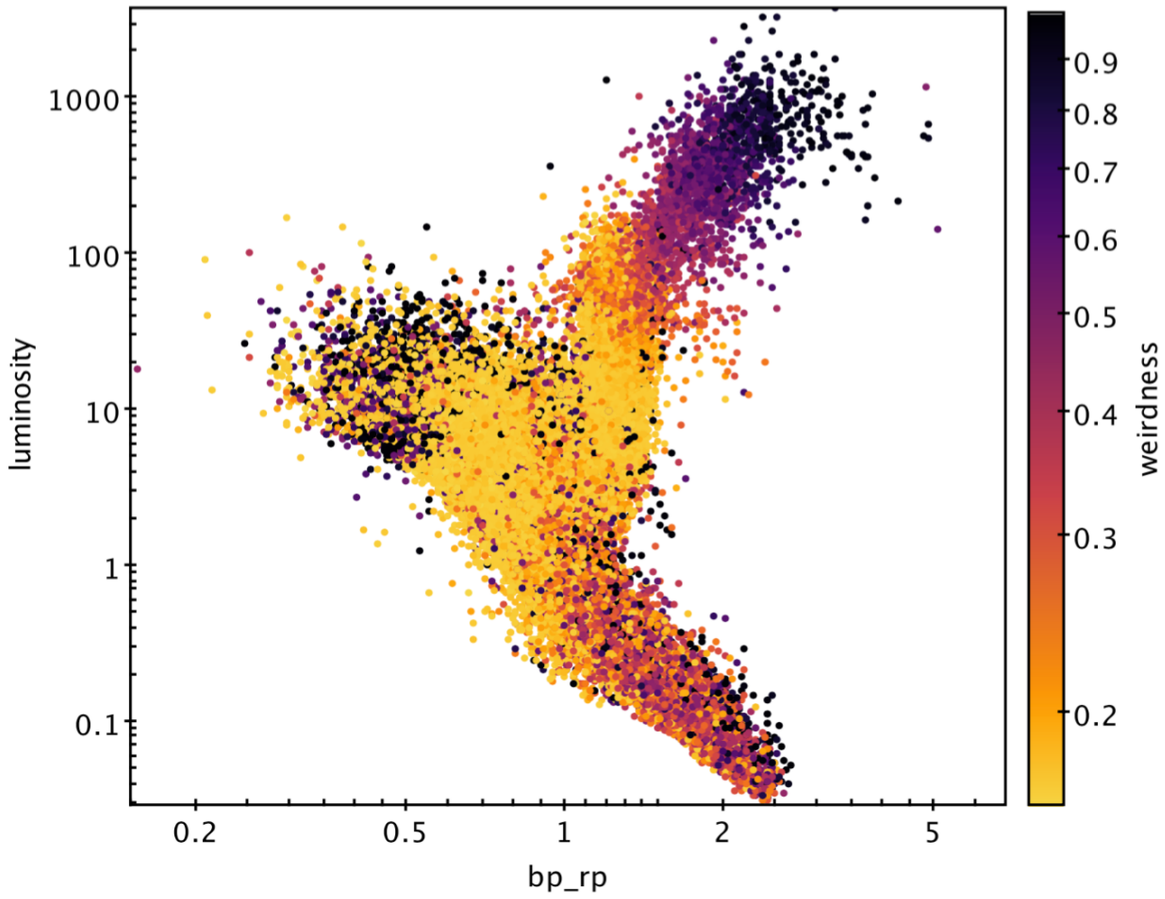
Exploratory analysis on Kepler light curves.

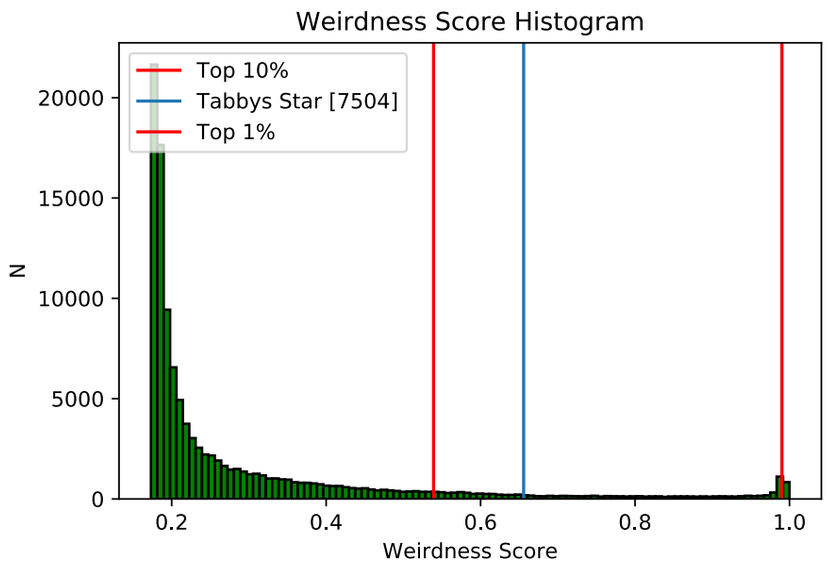
- 160,000 detrended light curves, as delivered by STScI's MAST archive.
- Light curves cover a time span of ~90 days, with a cadence of 30 minutes.
- Detrended using Kepler's pipeline.

This dataset is representative of Kepler light curves. It was randomly selected from Kepler Quarter 16 observations, taken between January 12 and April 8, 2013:

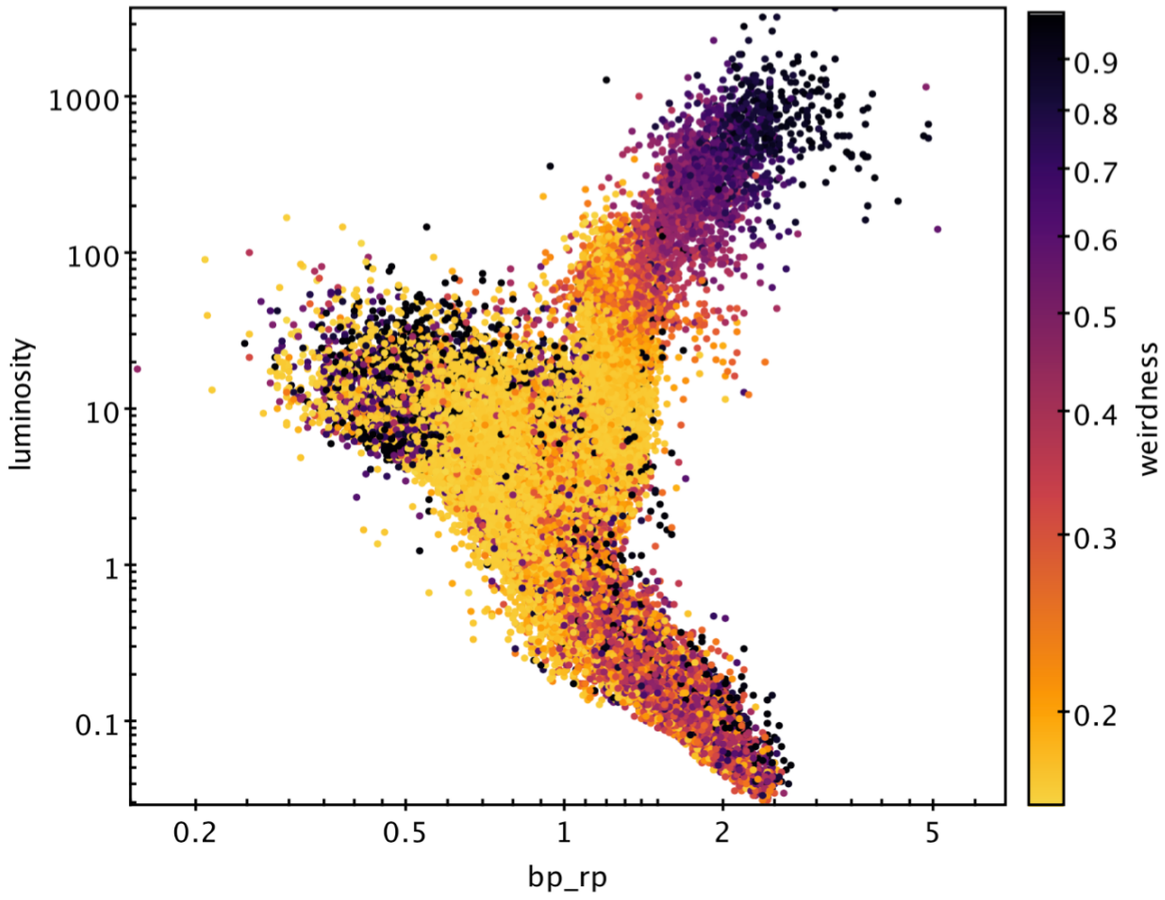
Roughly 60% are G-type main sequence stars, ~8% are giants, ~2% M-type dwarfs, <0.2% O and B-type stars, ~2% eclipsing binaries.

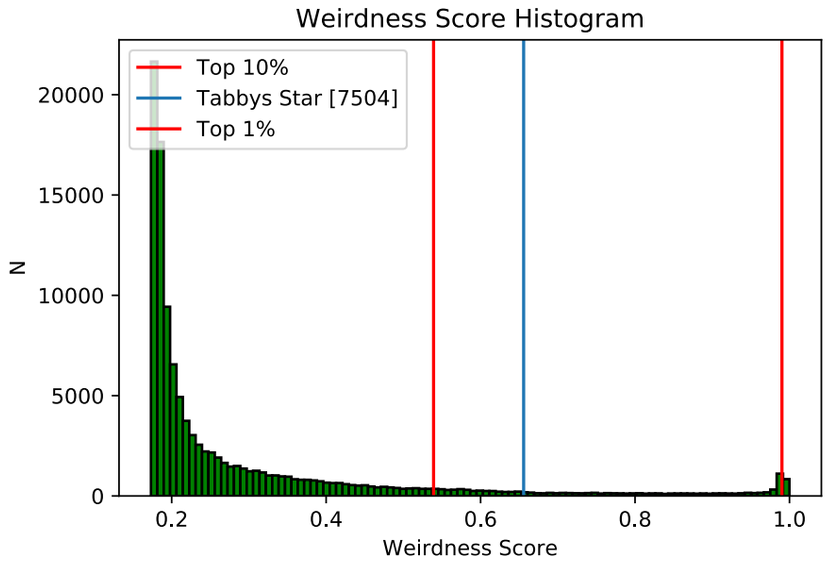
URF Outliers in the HR Diagram (Constructed from GAIA data)





URF Outliers in the HR Diagram (Constructed from GAIA data)

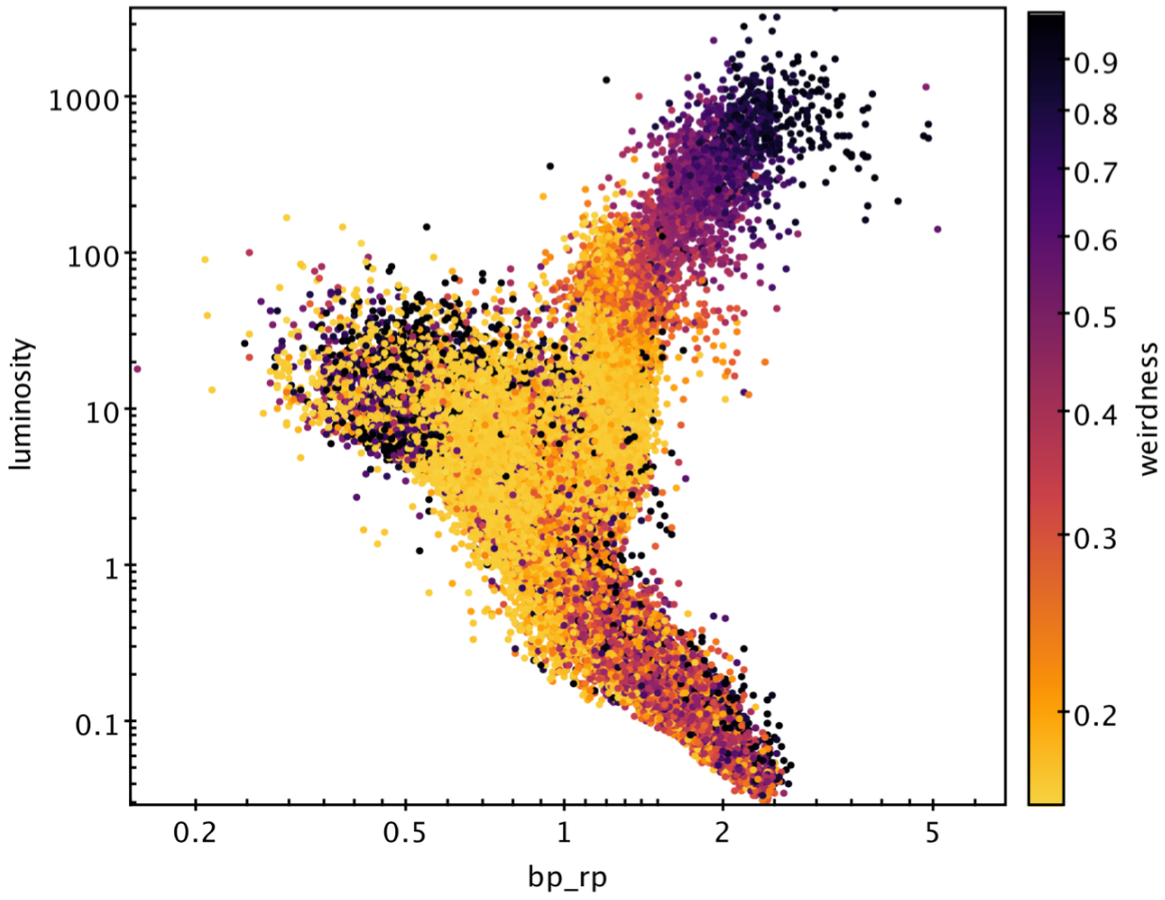


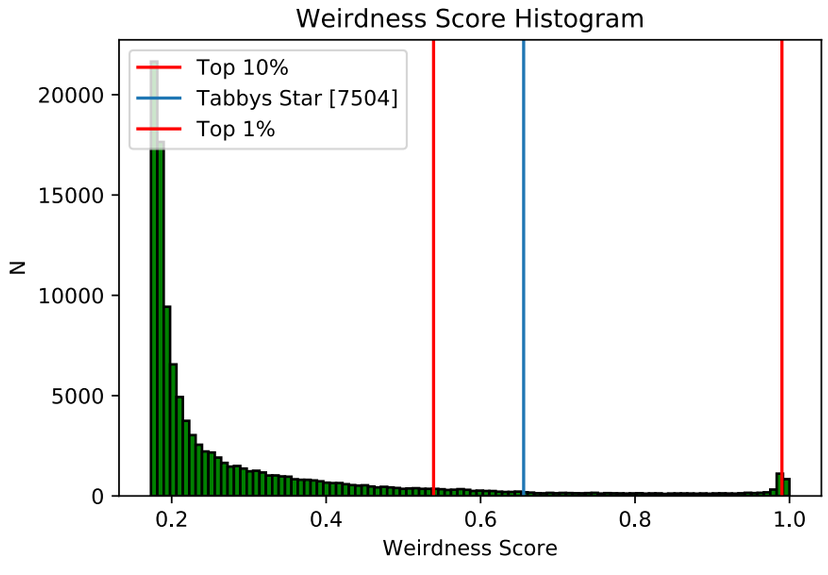


Top 1% outliers

- Contact binaries
- Rotational variables
- Cataclysmic variables
- Flaring brown dwarfs
- Dwarf novae
- X-ray binaries
- **Main sequence stars with deeps**

URF Outliers in the HR Diagram (Constructed from GAIA data)



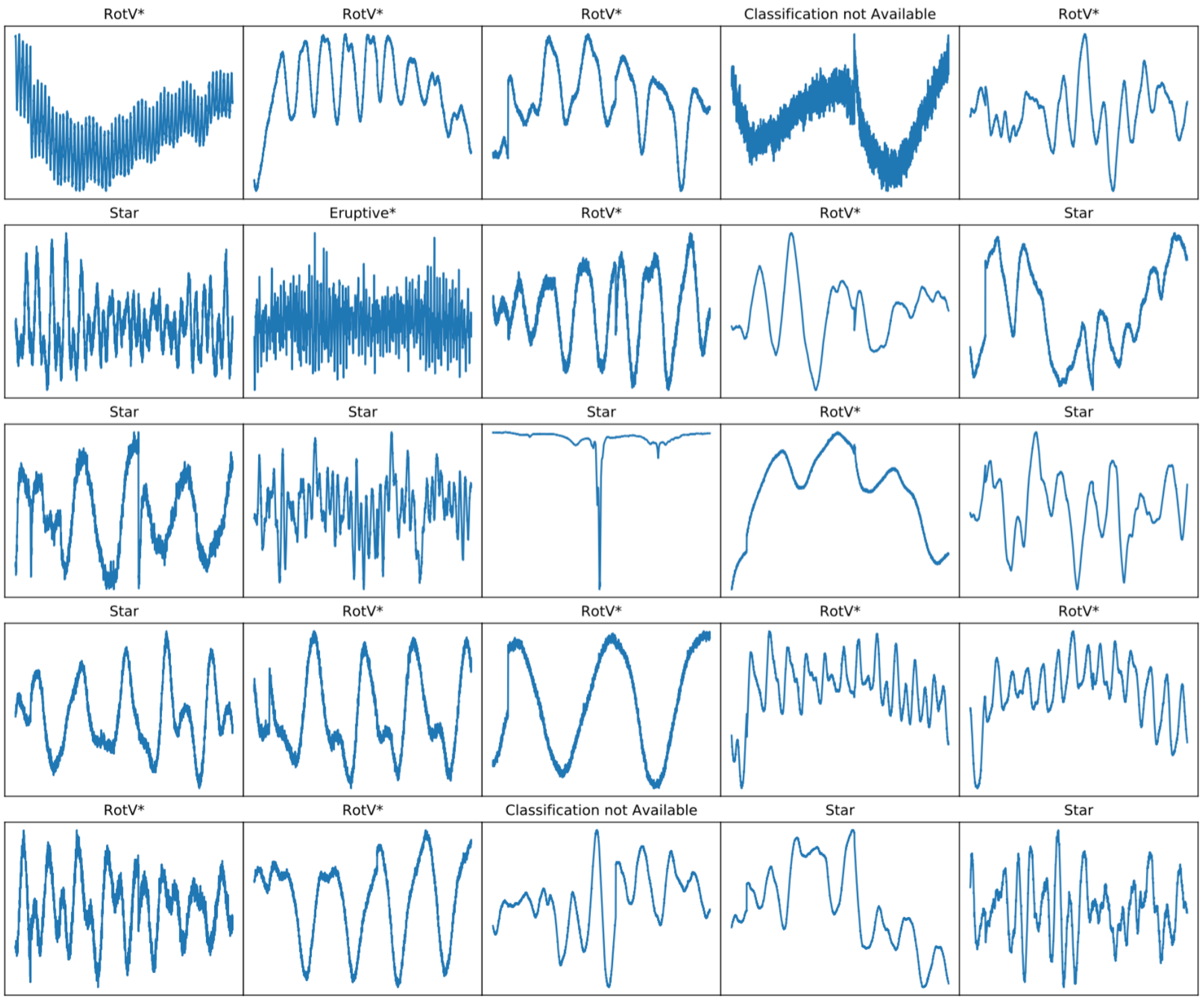


Top 1% outliers

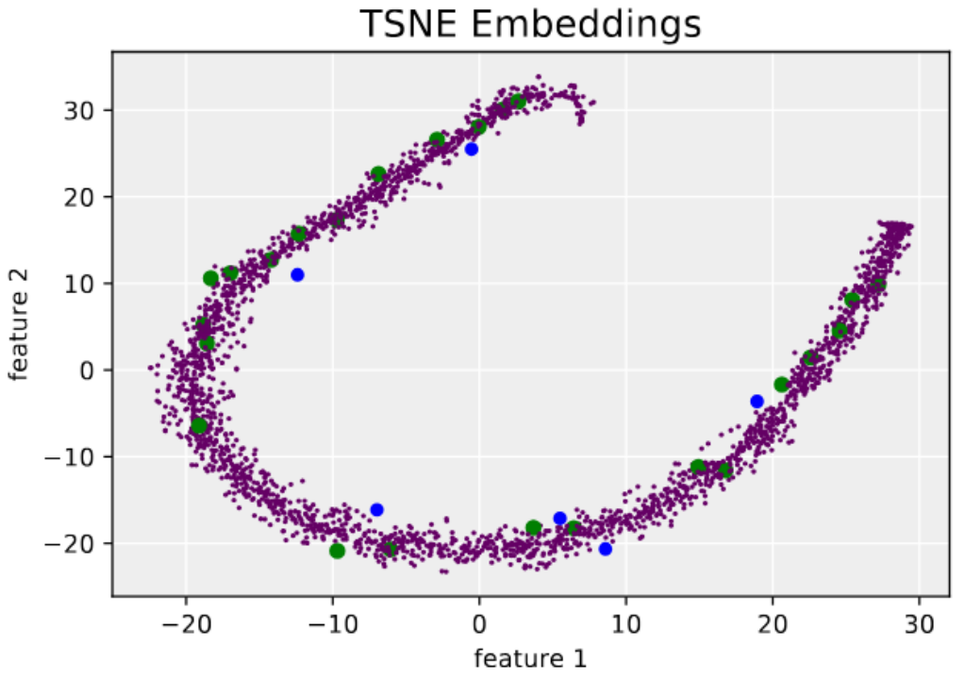
- Contact binaries
- Rotational variables
- Cataclysmic variables
- Flaring brown dwarfs
- Dwarf novae
- X-ray binaries
- **Main sequence stars with deeps**

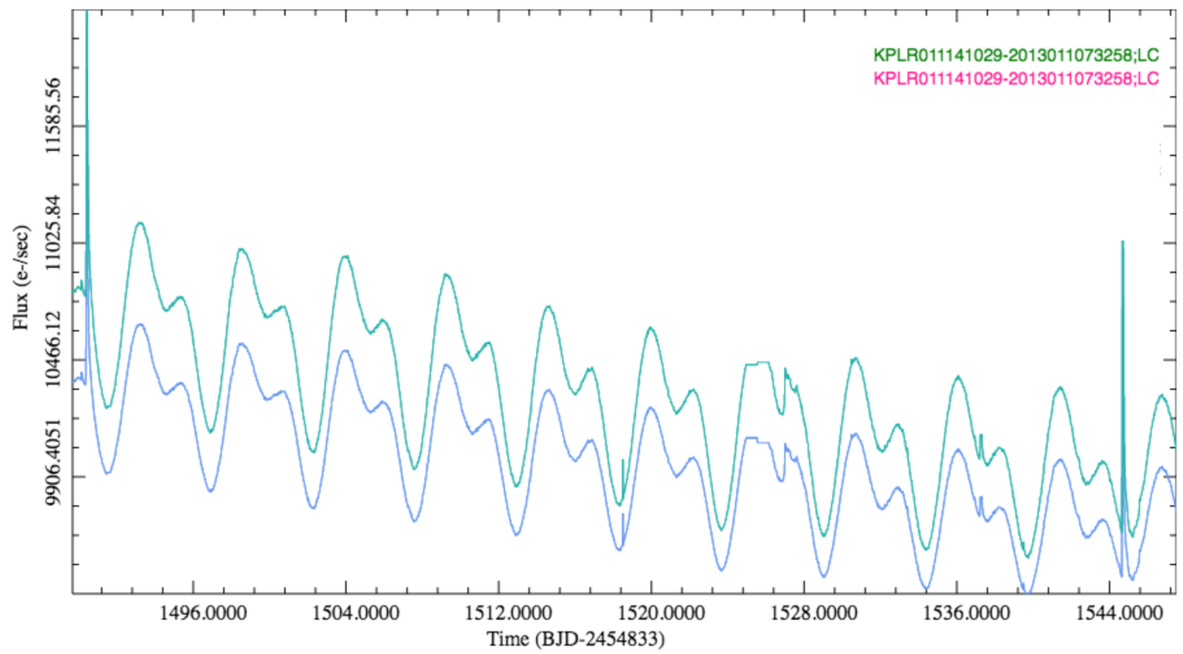
We are currently designing an observational program to follow up these interesting unclassified objects

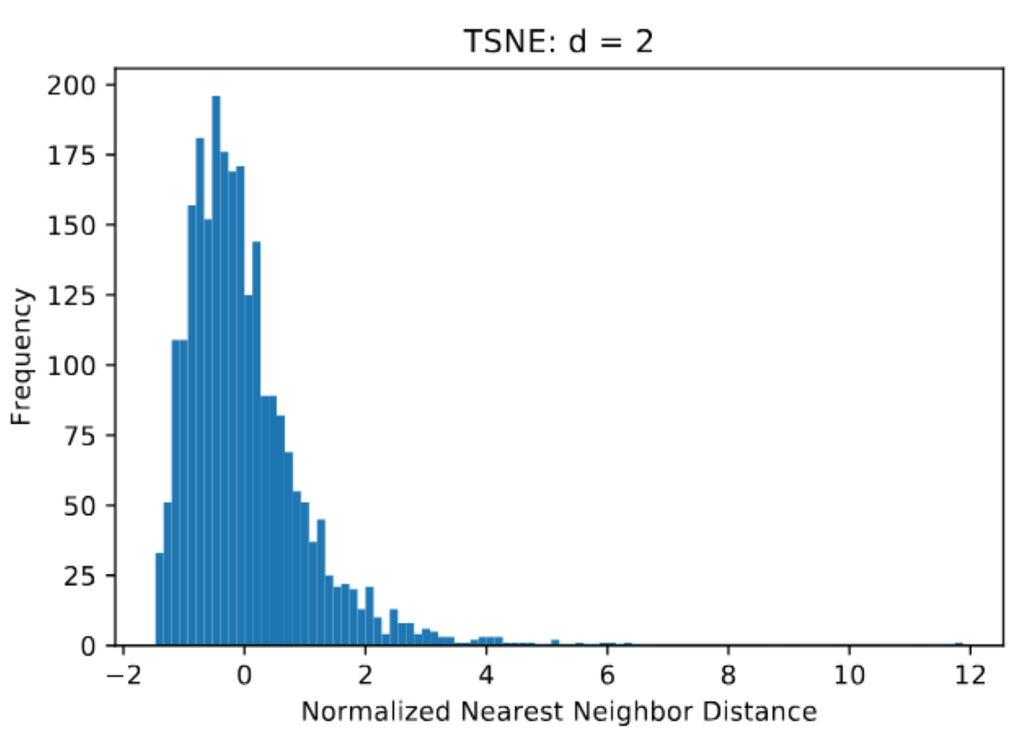
Light curves of outliers



Outliers with t-SNE



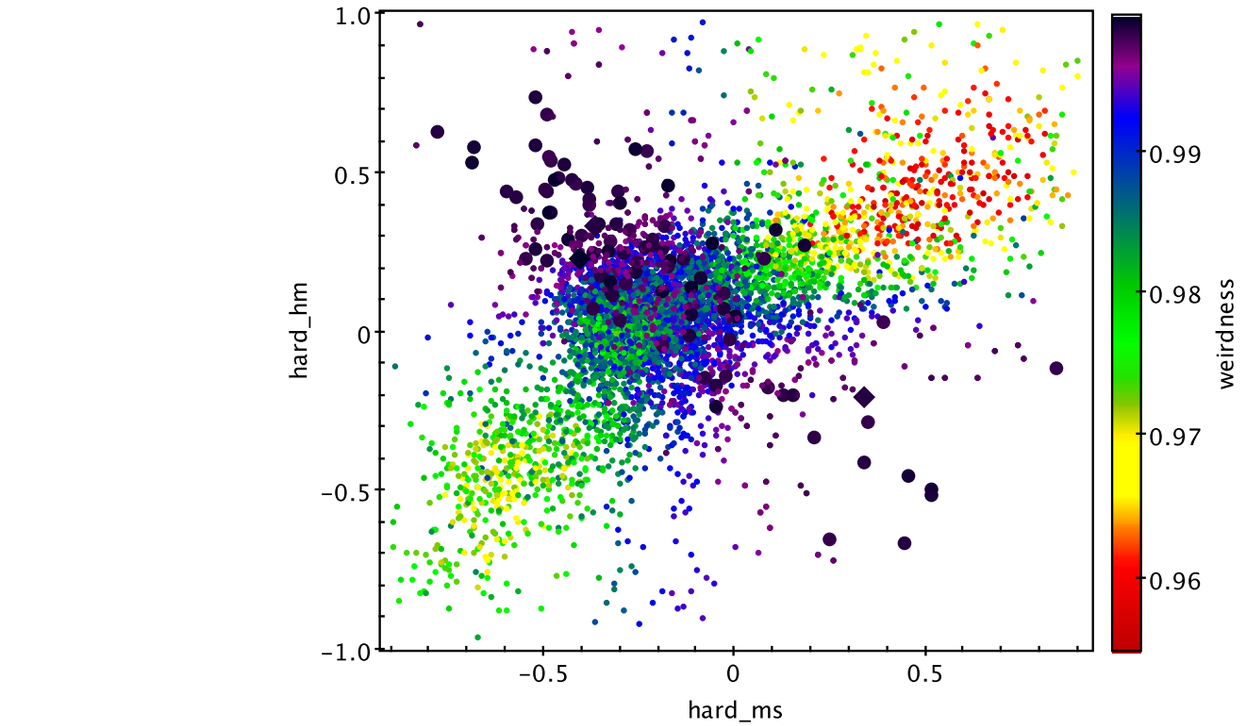


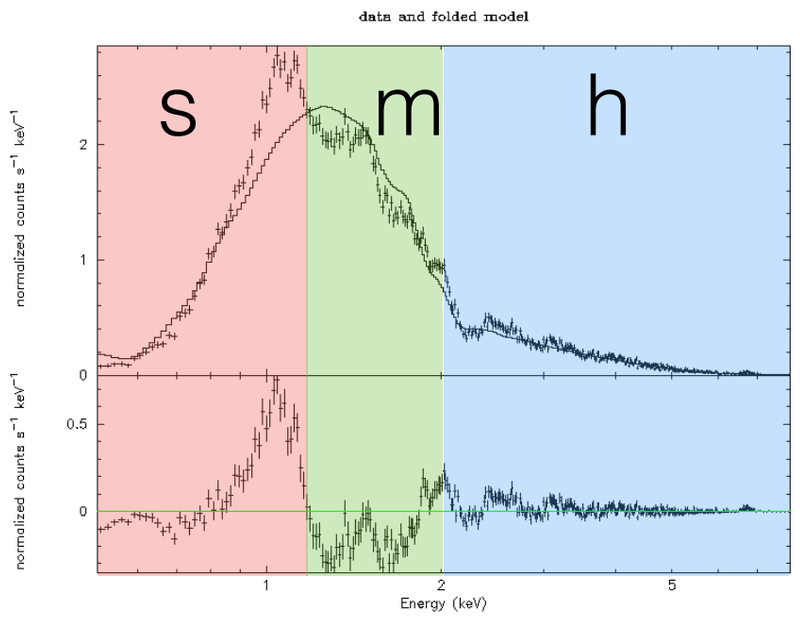


Features: flattened DMDT maps

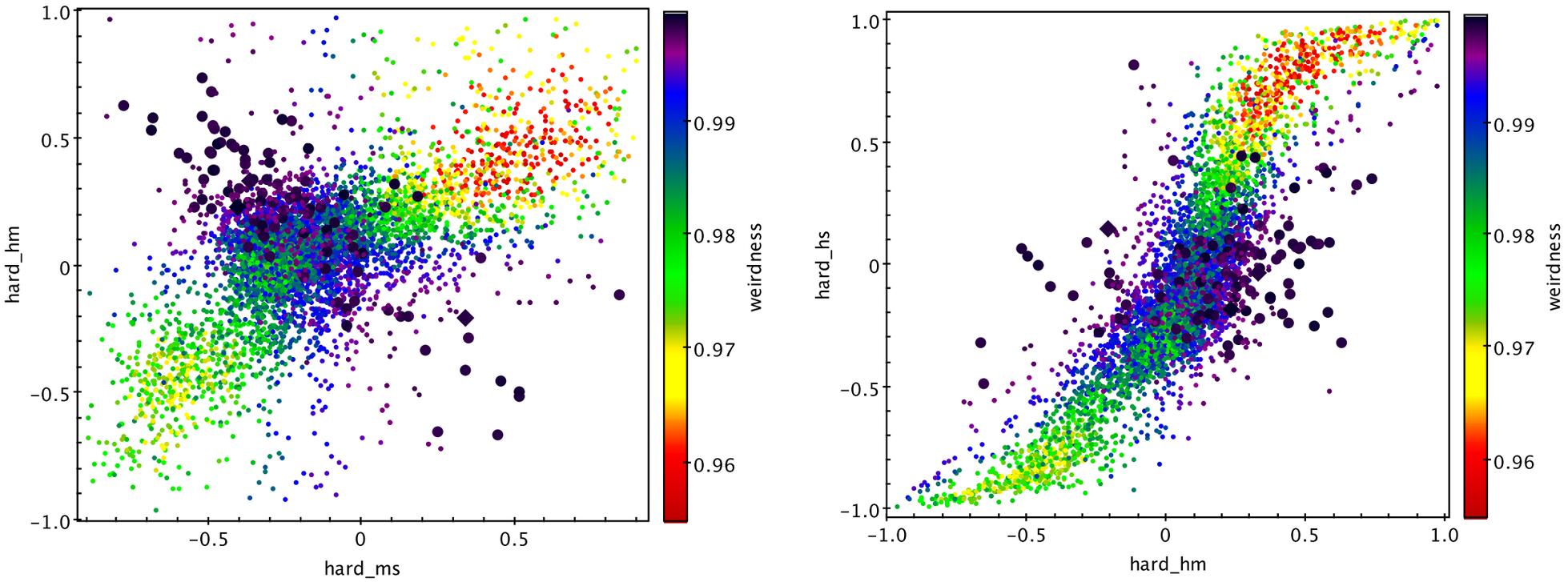
Weirdness dominated by smooth variability, with sudden deeps or flares

URF X-ray Outliers



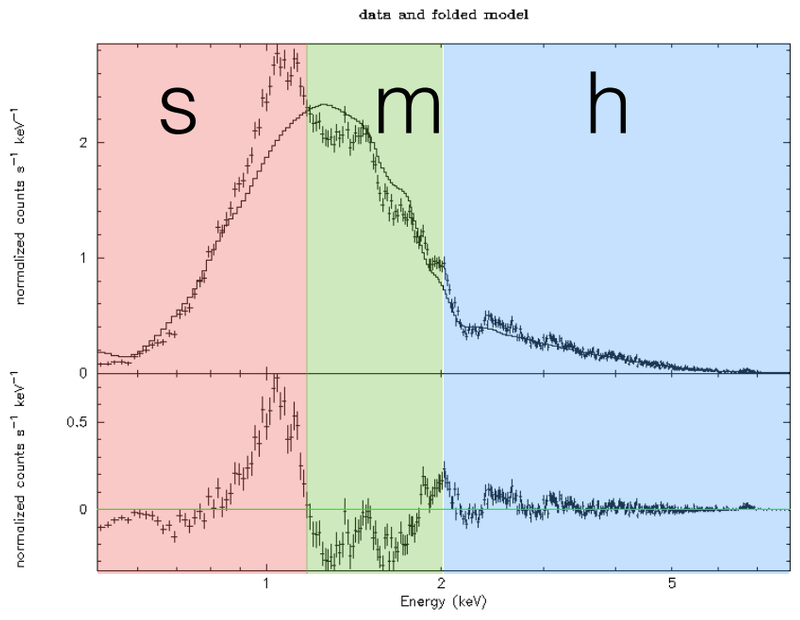


URF X-ray Outliers



Top 2% outliers

- Gamma-ray emitting novae (V679 Car)
- Super-Eddington ULX accretors (e.g. NGC 7793 P13).
- Black hole-WD binaries.
- Some T-tauri stars
- **Large majority still unclassified**



Take away messages

- Outlier detection is an excellent way to do “exploration approach” science.

- Astronomical outliers (anomalies) represent extreme stages in the evolutionary history of astrophysical objects - they used to be hard to find.

- Thanks to deep learning, feature engineering becoming obsolete.

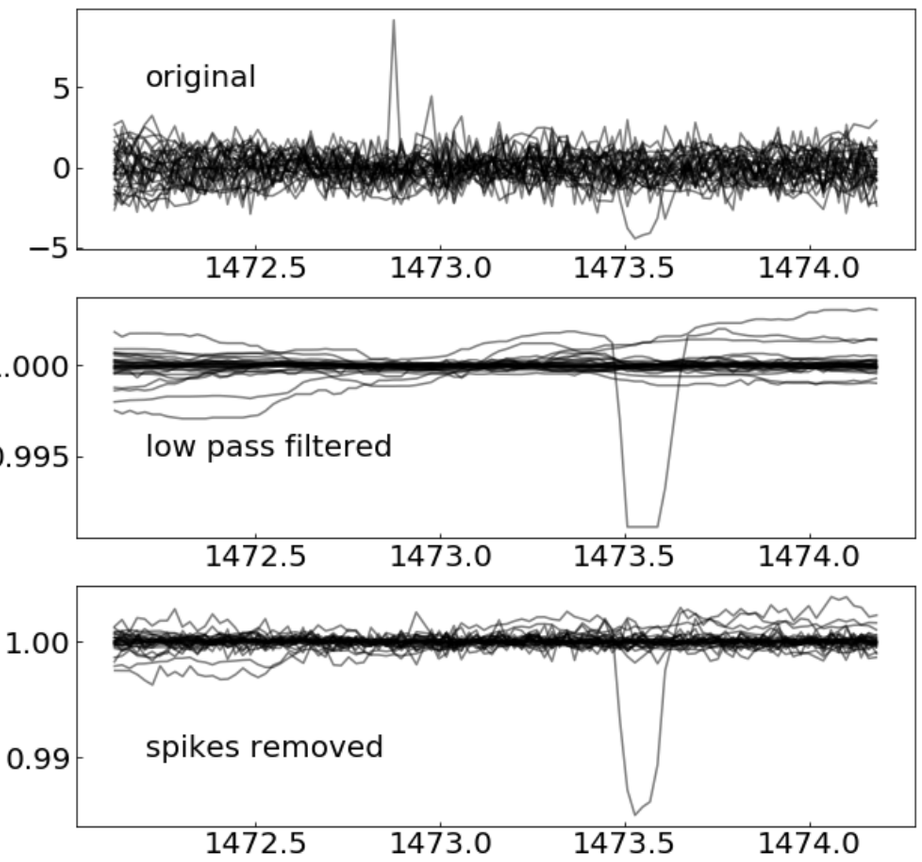
- Flexibility of algorithms allows to find anomalies of different nature.

- Early classification of light curves now possible.

- Excellent synergy of recent algorithms with upcoming surveys such as LSST, etc.

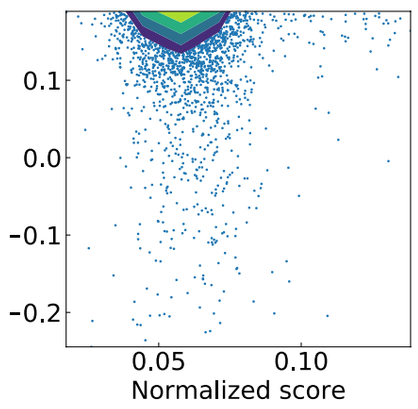
“Normality” score affected by artifacts

30 light curves, with three different pre-processing approaches

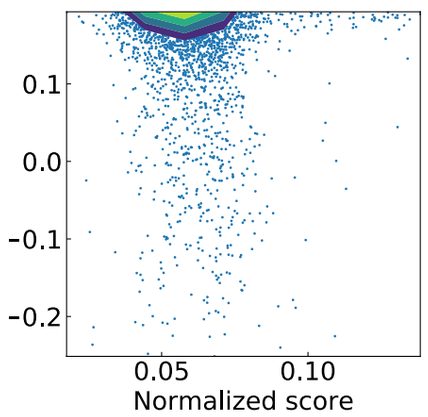


Martínez-Galarza et al. 2019 (in prep.)

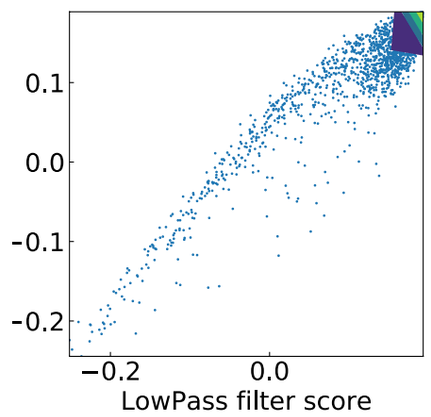
Isolation forest normality scores



Normalized vs. LPF



Normalized vs. clean



LPF vs clean

Preprocessing does have an impact on our ability to find outliers, and on what kind of outliers we find!

Need to make sure that “normality” or “weirdness” not dictated by systematic data artifacts!

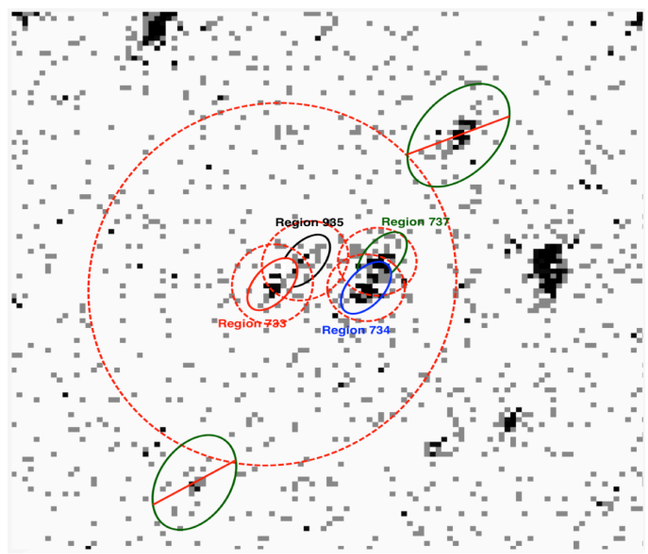
X-ray features in CSC2

1. Aperture photometry

Joint posterior for source fluxes and background:

$$P(s_1 \ldots s_n, b | C_1 \ldots C_n, B) = K \times P(b) P_{Pois}(B \,|\, \phi) \prod P(s_i) P_{Pois}(C_i \,|\, \theta_i)$$

$$\theta_i = E_i \times \left[\sum_{j=1}^{n} f_{ij} s_j \; + \; \Omega_i b \right] ; \; \phi = E_b \times \left[\sum_{i=1}^{n} g_i s_i \; + \; \Omega_b b \right]$$



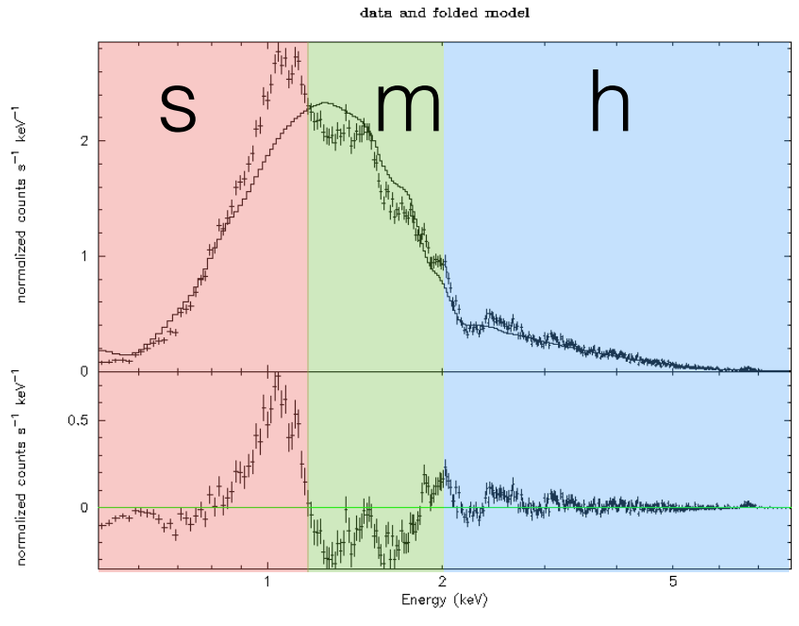
2. Hardness ratios

Define the H_{xy} for bands x, y as the ratio: $(F_x\text{-}F_y)/(F_x+F_y)$.
Then:

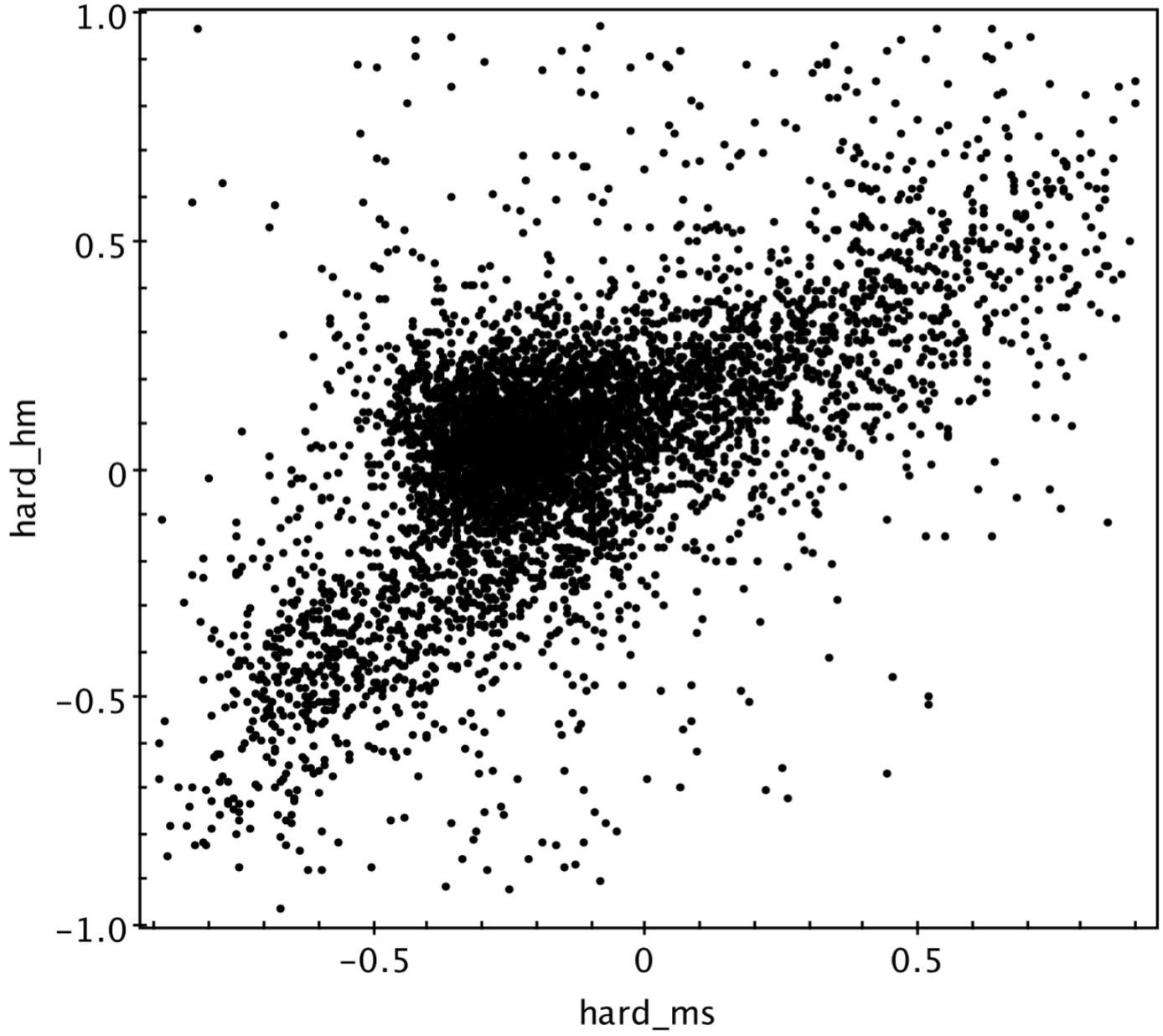
$$P_{H_{xy}}(H_{xy})\, dH_{xy} \quad = \quad \int_{F_{xy}=0}^{\infty} P_x \left(\frac{(1+H_{xy})F_{xy}}{2} \right) P_y \left(\frac{(1-H_{xy})F_{xy}}{2} \right) \frac{F_{xy}}{2} \, dH_{xy} dF_{xy}$$

Hardness ratios hard_ms, hard_hm, hard_hs are those values that maximize the PDF above.

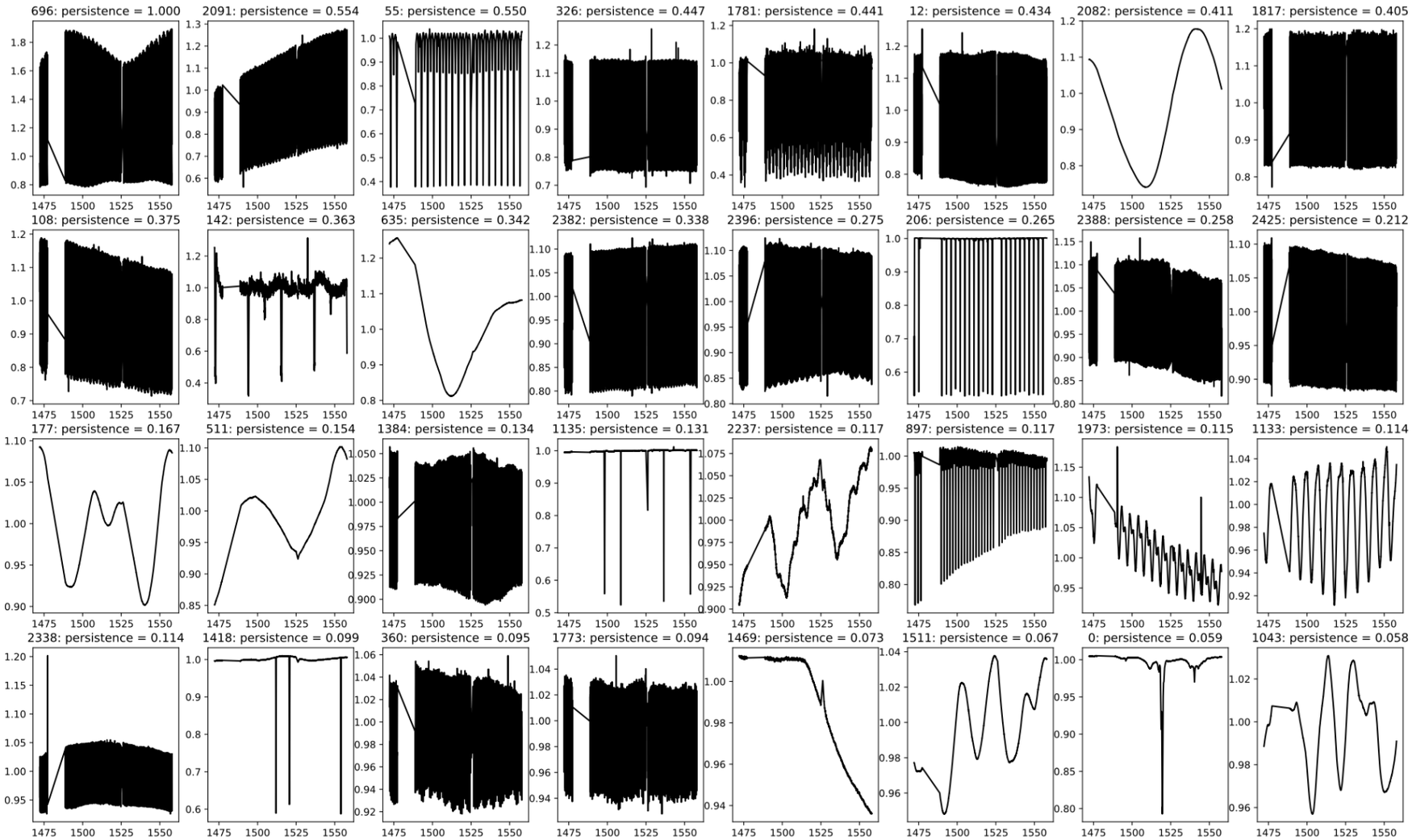
Used for spectral characterization when fitting is limited by few counts



Distribution of hardness ratios
Subset of CSC2 sources



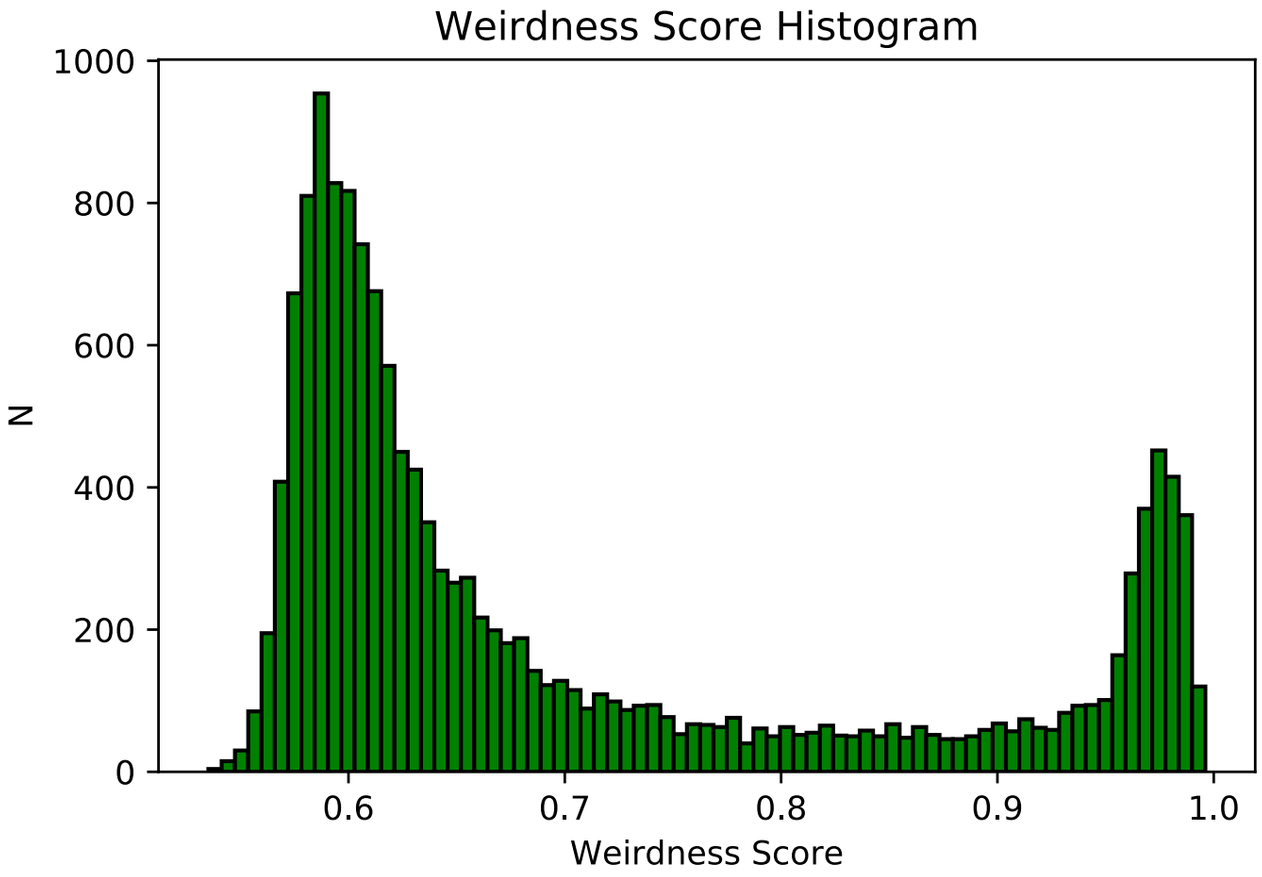
Outliers with MSP

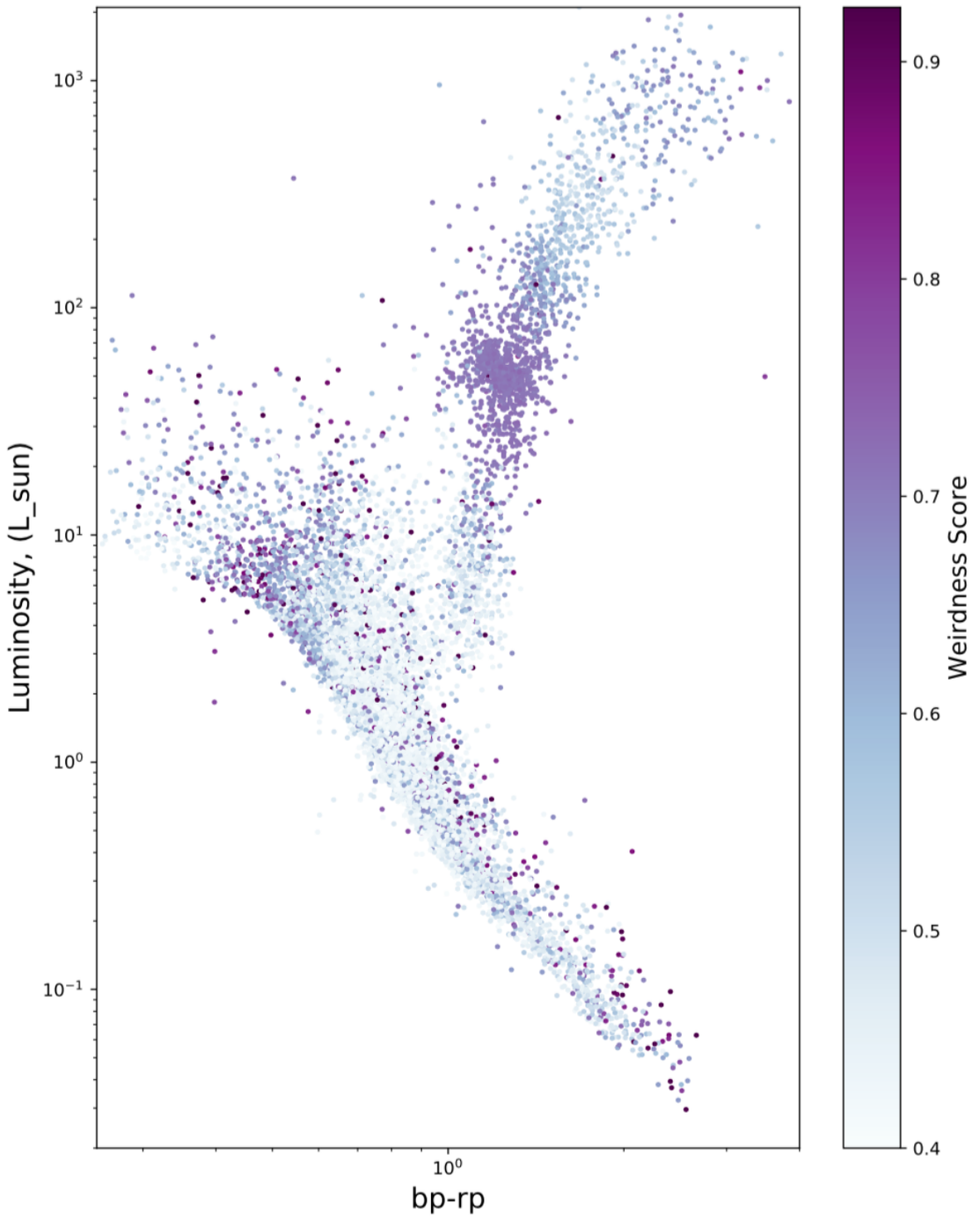


Tabby’s star, rotationally variable stars, seismic red giants

Mostly amplitude weirdos

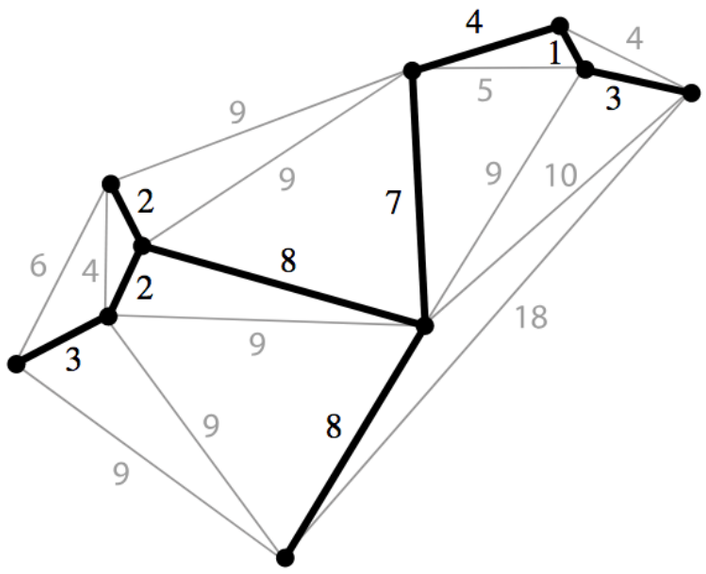
Preliminary results on public TESS data





More coming soon!

Minimum Spanning Persistence



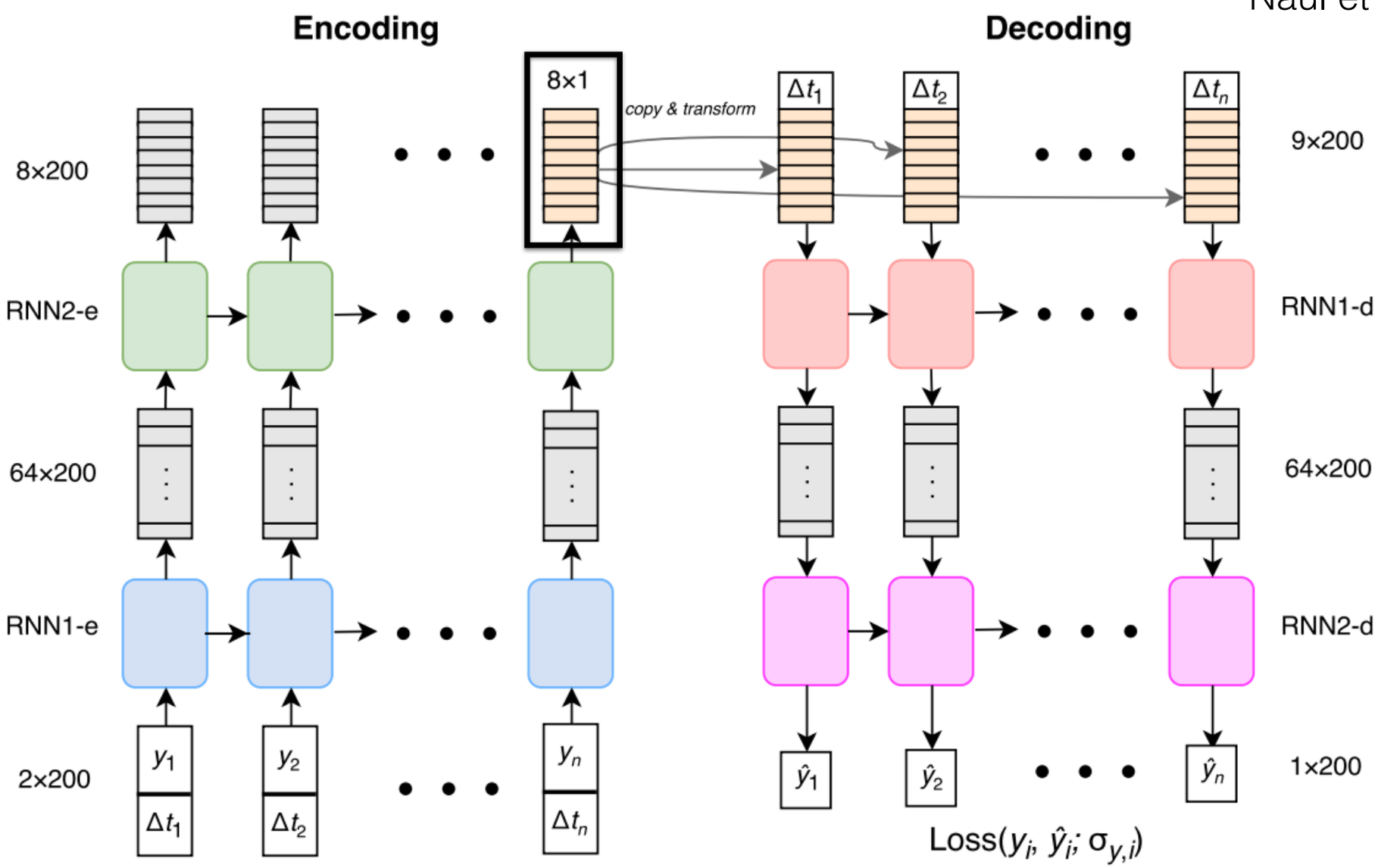
- Persistent homology aims at determining how stable (persistent) data structures are over a range of scales.
- Minimal spanning trees are unique constructs that encode all the information about connected components at all resolutions.

Outliers using MST persistence

- Significant outliers should be persistent components with minimal connectivity
- Lifetime of outlier is the range of filtration steps over which it remains unconnected.

RNNs, autoencoders

Naul et al. 2018



Latent space used as feature vector.