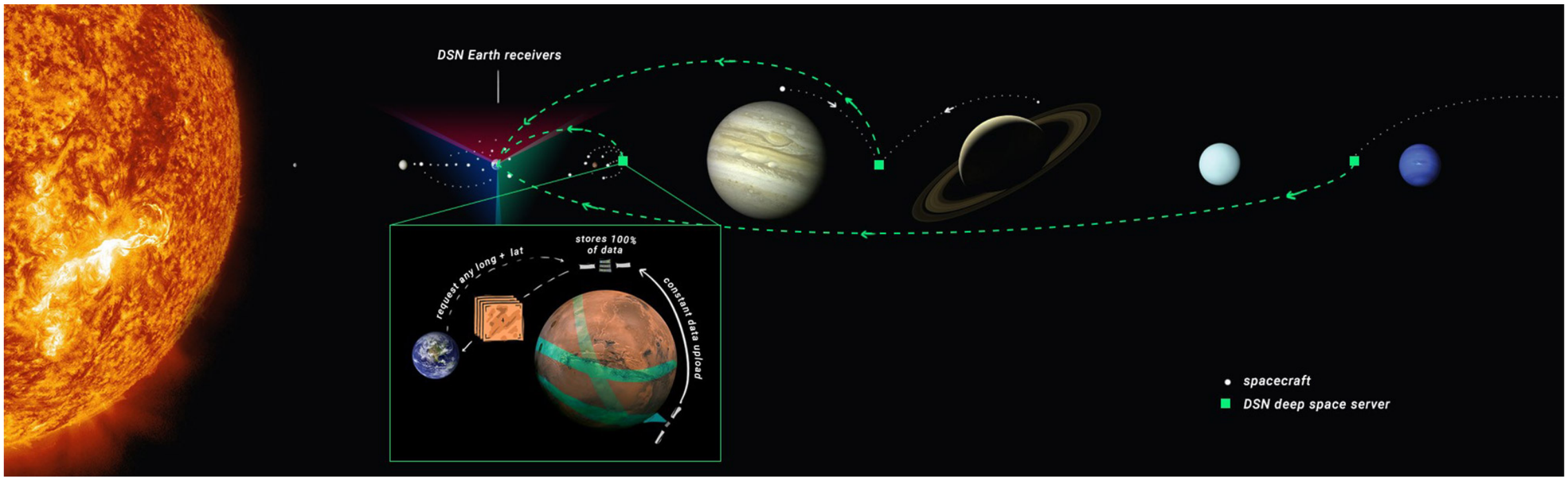
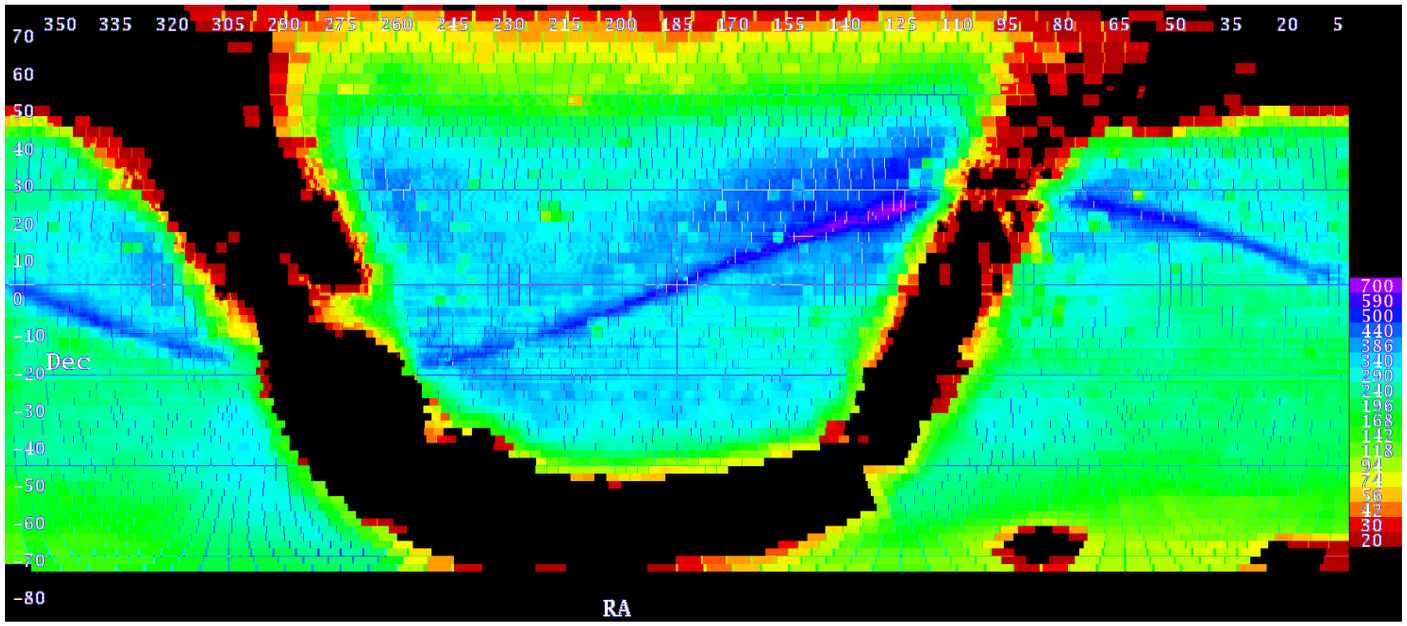
Machine Learning and Artificial Intelligence for Science Data



Ashish Mahabal
<ashish@caltech.edu>
Lead Computational and Data Scientist
Astronomy/Center for Data Driven Discovery, Caltech

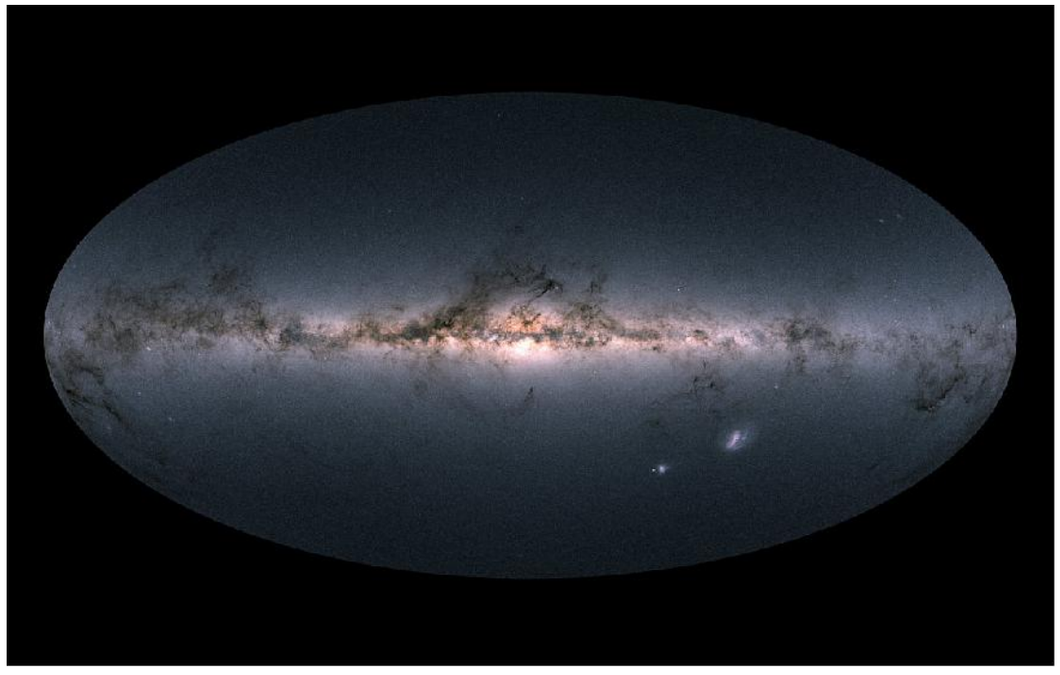
KISS - Space Nebulae, 2019-08-28

From snapshots to (slow) movies of the sky

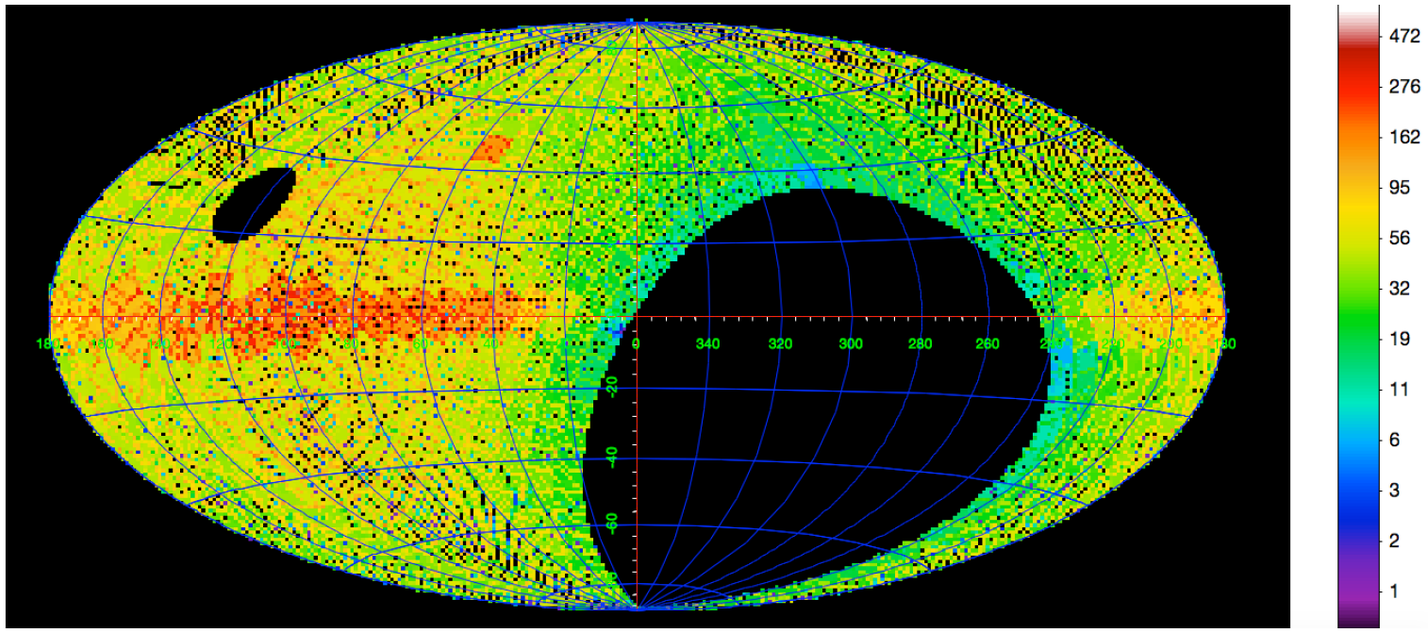


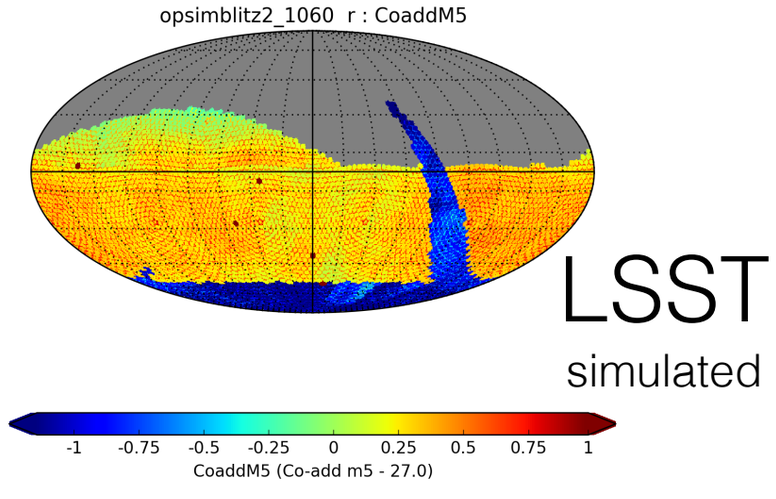
CRTS

Gaia



ZTF





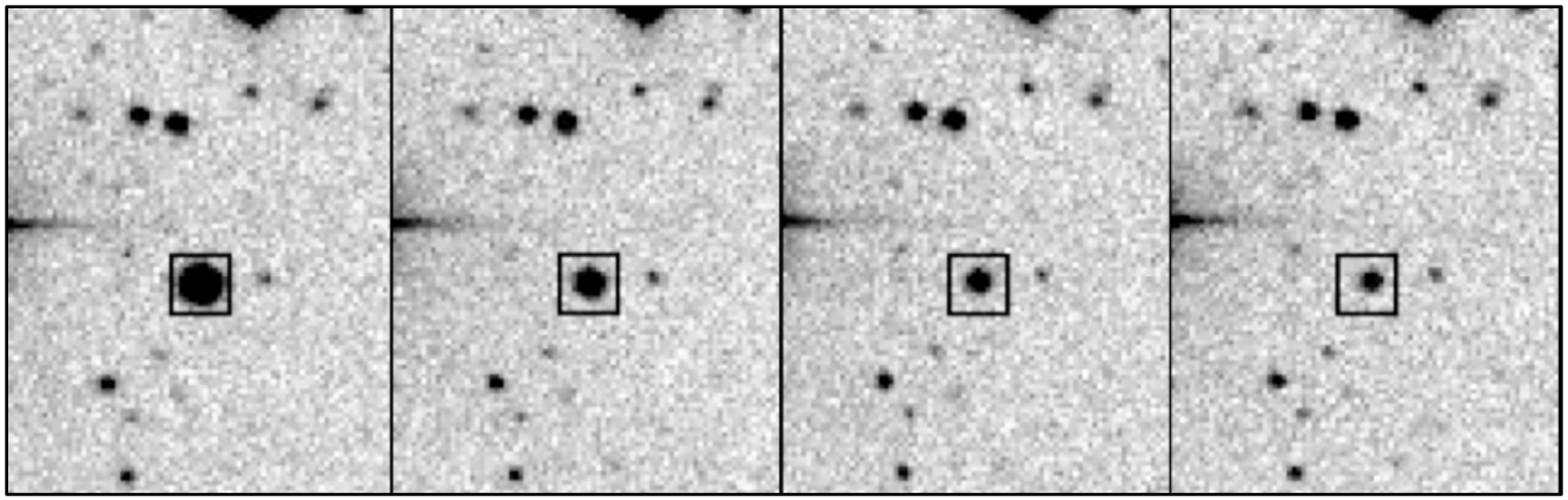
LSST

simulated

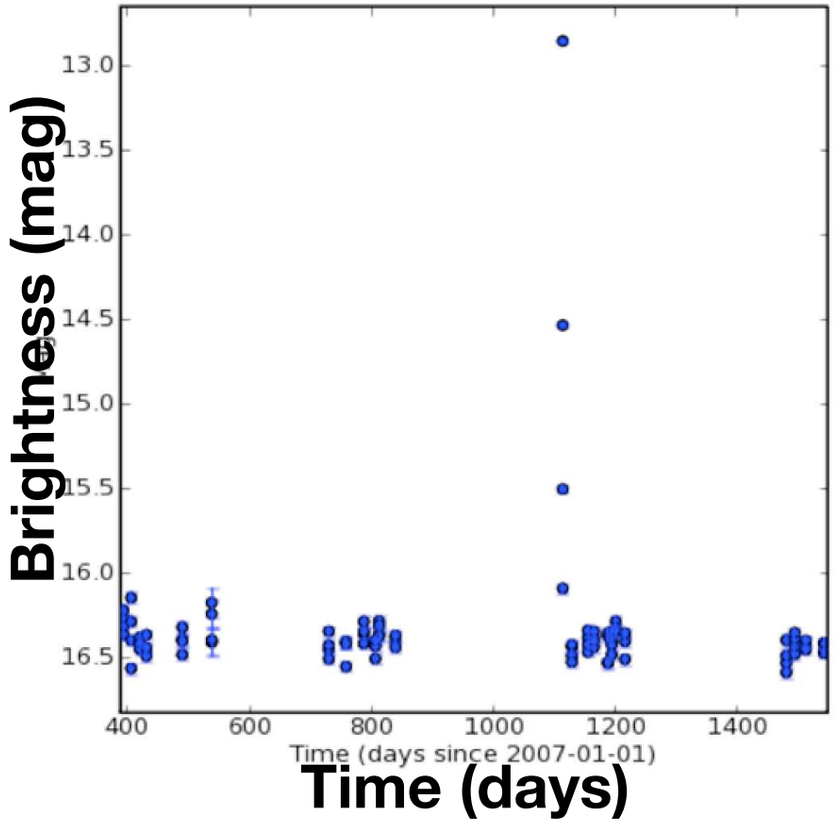
SDSS, Pan-STARRS, ASAS-SN, Skymapper, … (just in the optical)

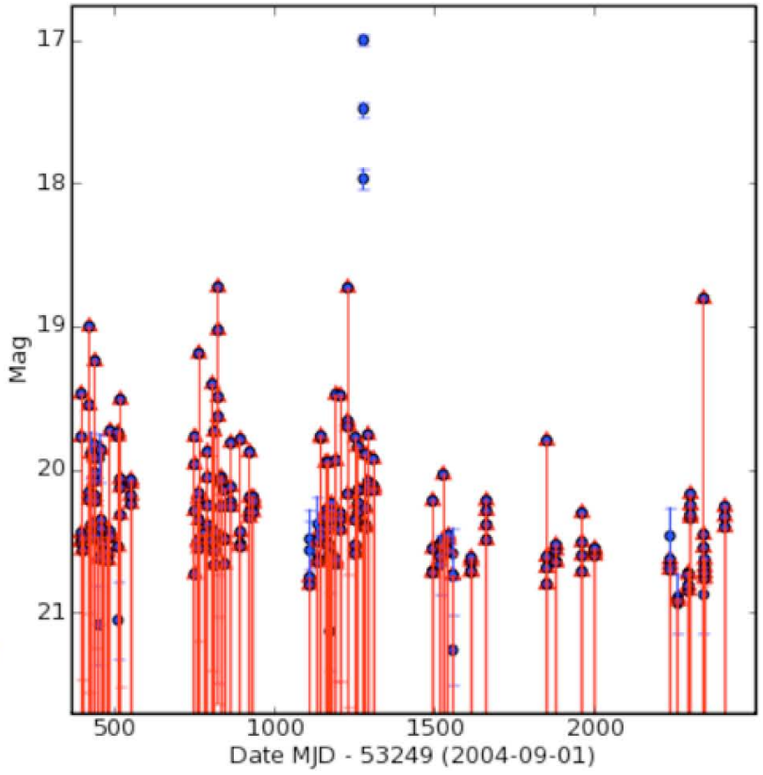
SKA

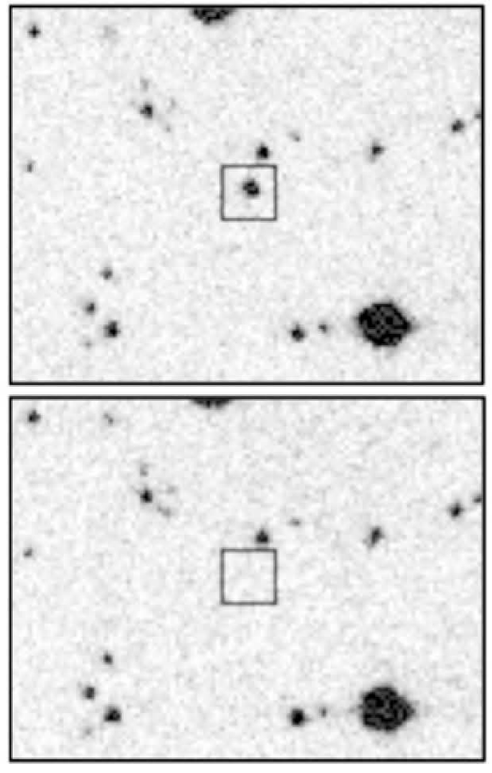
Catalina Real-time Transient Survey (CRTS)



4 individual exposures, separated by 10 **min**

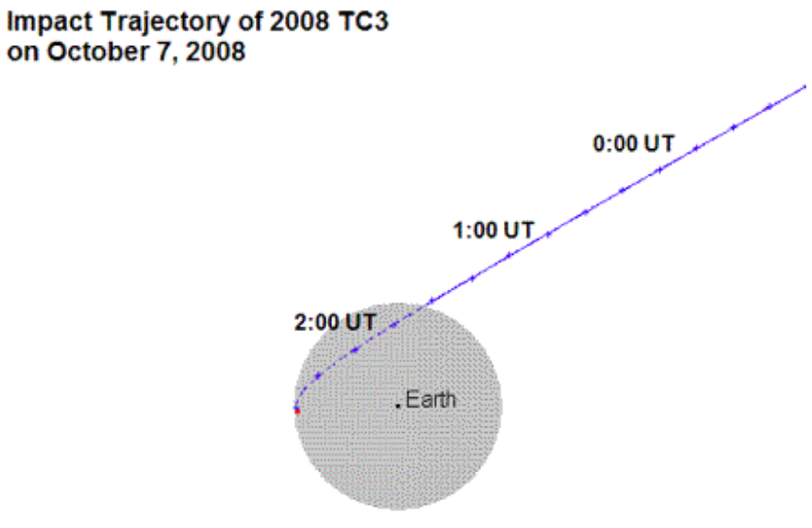


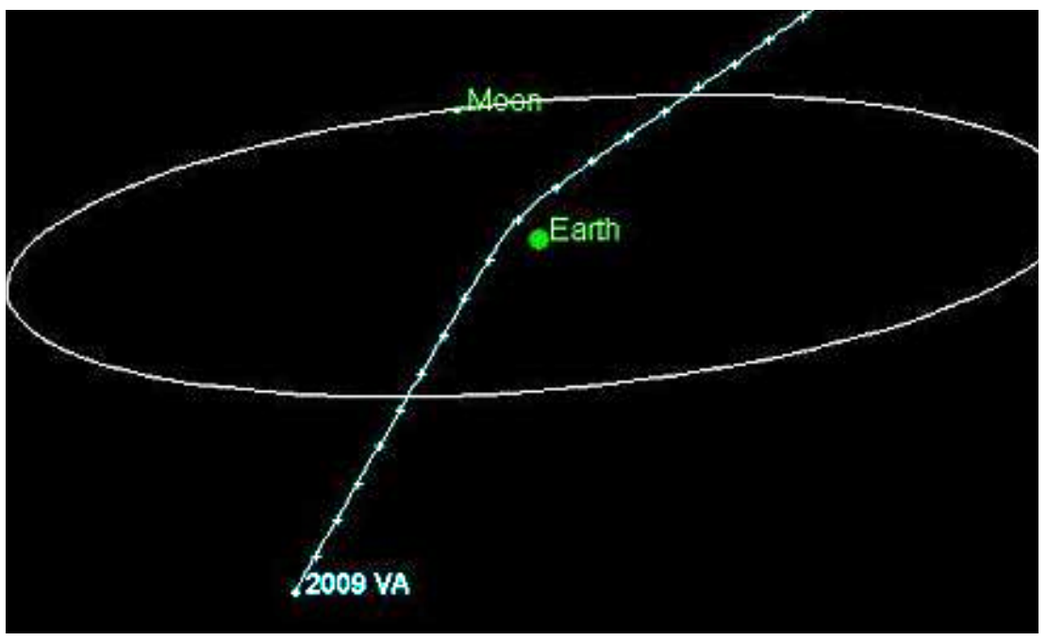


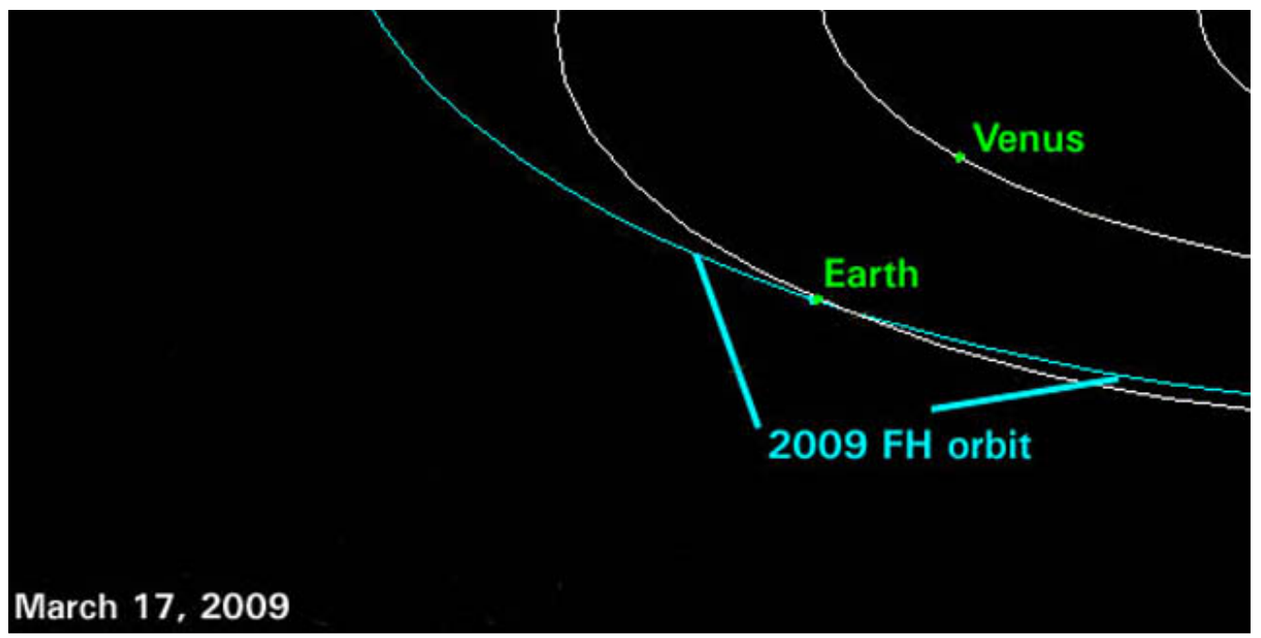


Most (but not all!) are flaring dwarf stars (UV Ceti)

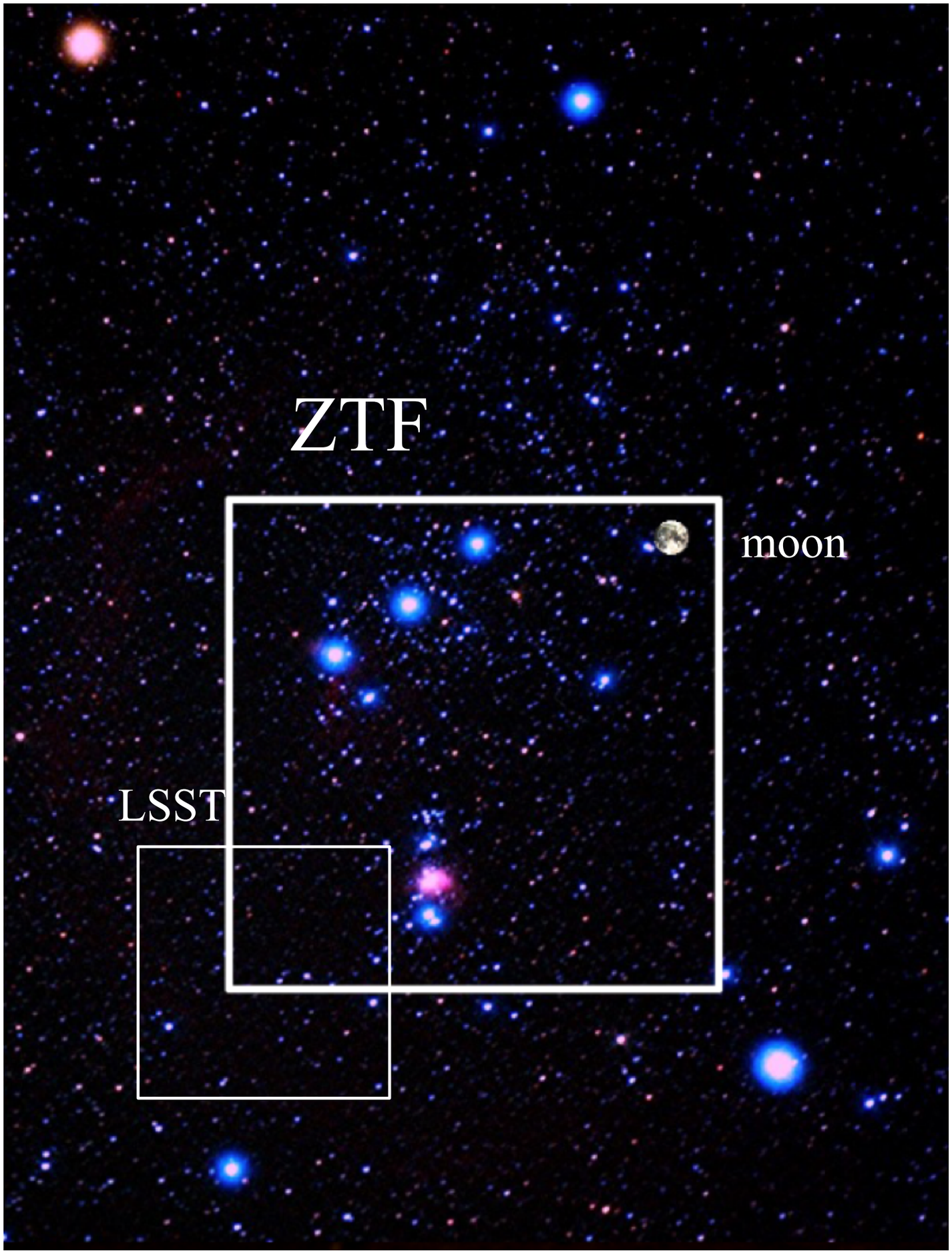
2008 TC3 discovered by CSS on 7 Oct 2008



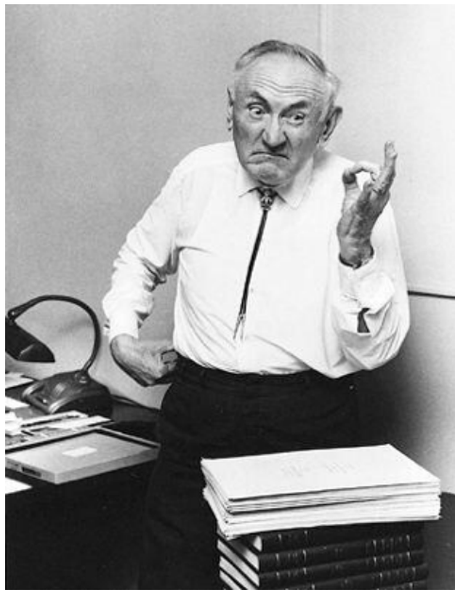




Low cost ‘sample return mission’



Zwicky Transient Facility



Area: ~ 47 deg²
(576M pixels)
Rate: 3760 deg² / hour
Depth (5σ): *r* ~ 20.5 mag.
Filters: 3 (*g, r, i*)
Public survey:
~ 15k deg²/3 nights

DR1 MSIP (public) data

ztf.caltech.edu

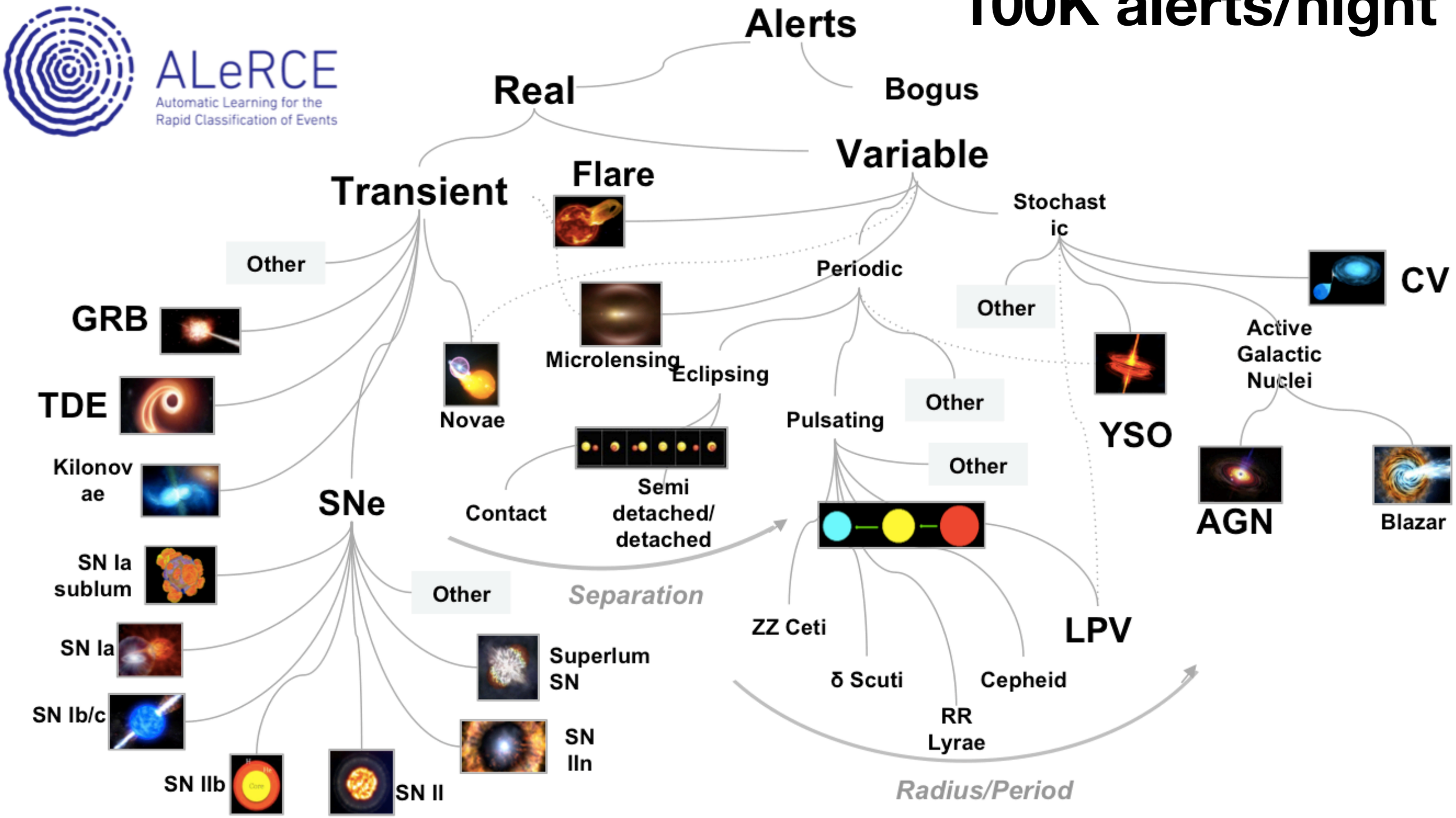
Filter(s)	#lightcurves with $N_{obs} \geq 2$	#lightcurves with $N_{obs} \geq 5$	#lightcurves with $N_{obs} \geq 10$	#lightcurves with $N_{obs} \geq 20$
g	704,000,504	589,547,084	508,917,850	391,041,883
r	1,334,687,993	1,142,671,302	1,013,283,728	852,773,692
g + *r*	2,038,688,497	1,732,218,386	1,522,201,578	1,243,815,575

Time series for over a billion sources

Variability tree: Many nodes have further subdivisions

100K alerts/night

ALeRCE
Automatic Learning for the Rapid Classification of Events



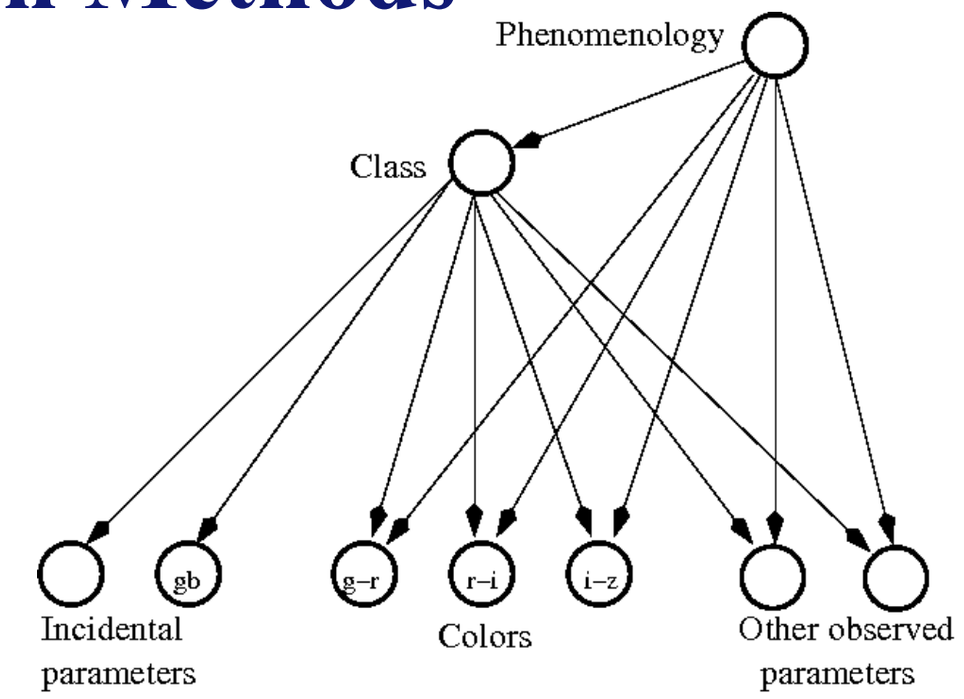
Scheduling follow-up and brokers

A Variety of Classification Methods

Bayesian Networks

Can incorporate heterogeneous and/or missing data

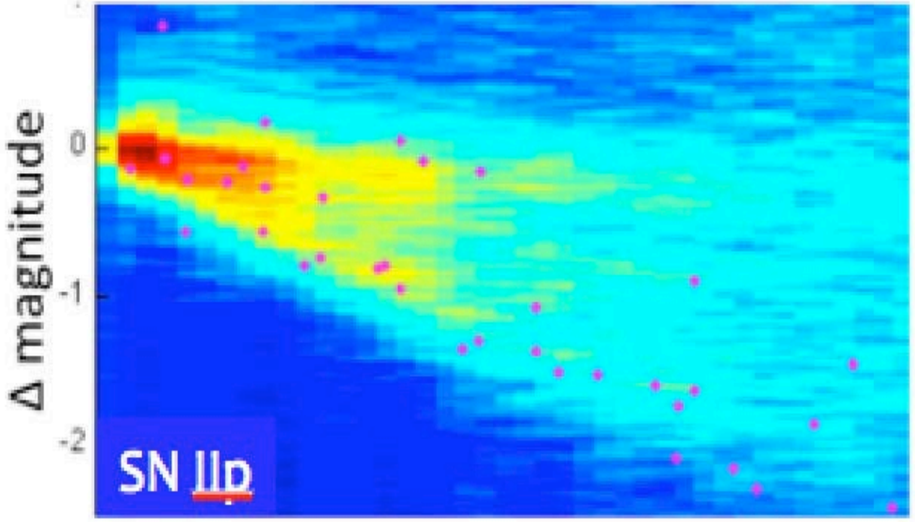
Can incorporate contextual data, e.g., distance to the nearest star or galaxy



Probabilistic Structure Functions

A new method, based on 2D [Δt_1, Δm] distributions

Now expanding to data point triplets: Δt_{12}, Δm_{12}, Δt_{23}, Δm_{23}, giving a 4D histogram



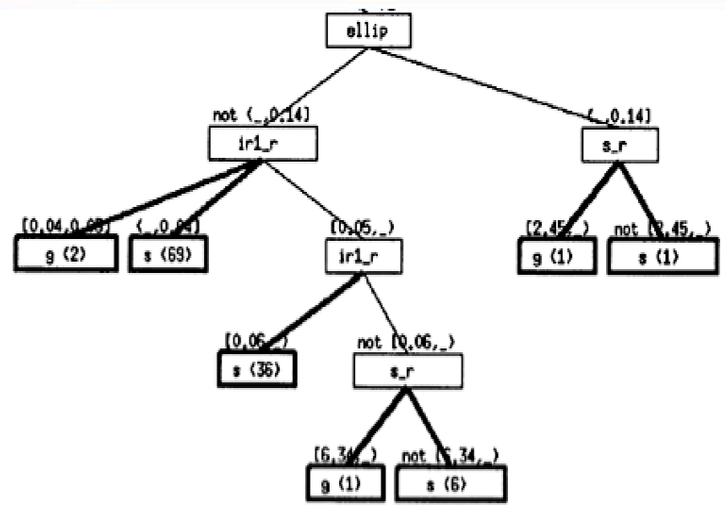
Random Forests

Ensembles of Decision Trees

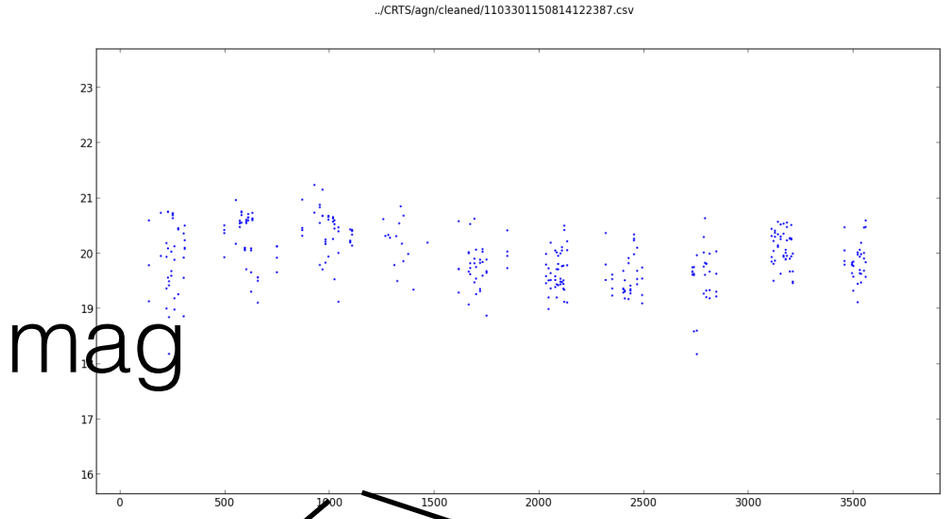
Feature Selection Strategies

Optimizing classifiers

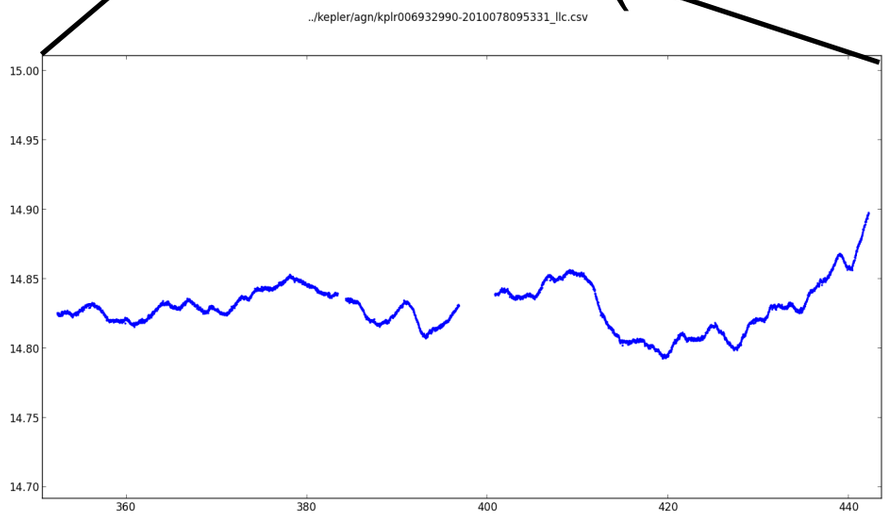
Machine-Assisted Discovery



Sparse Data

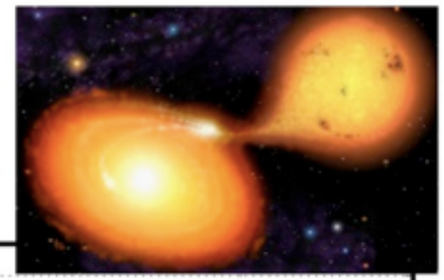


CRTS



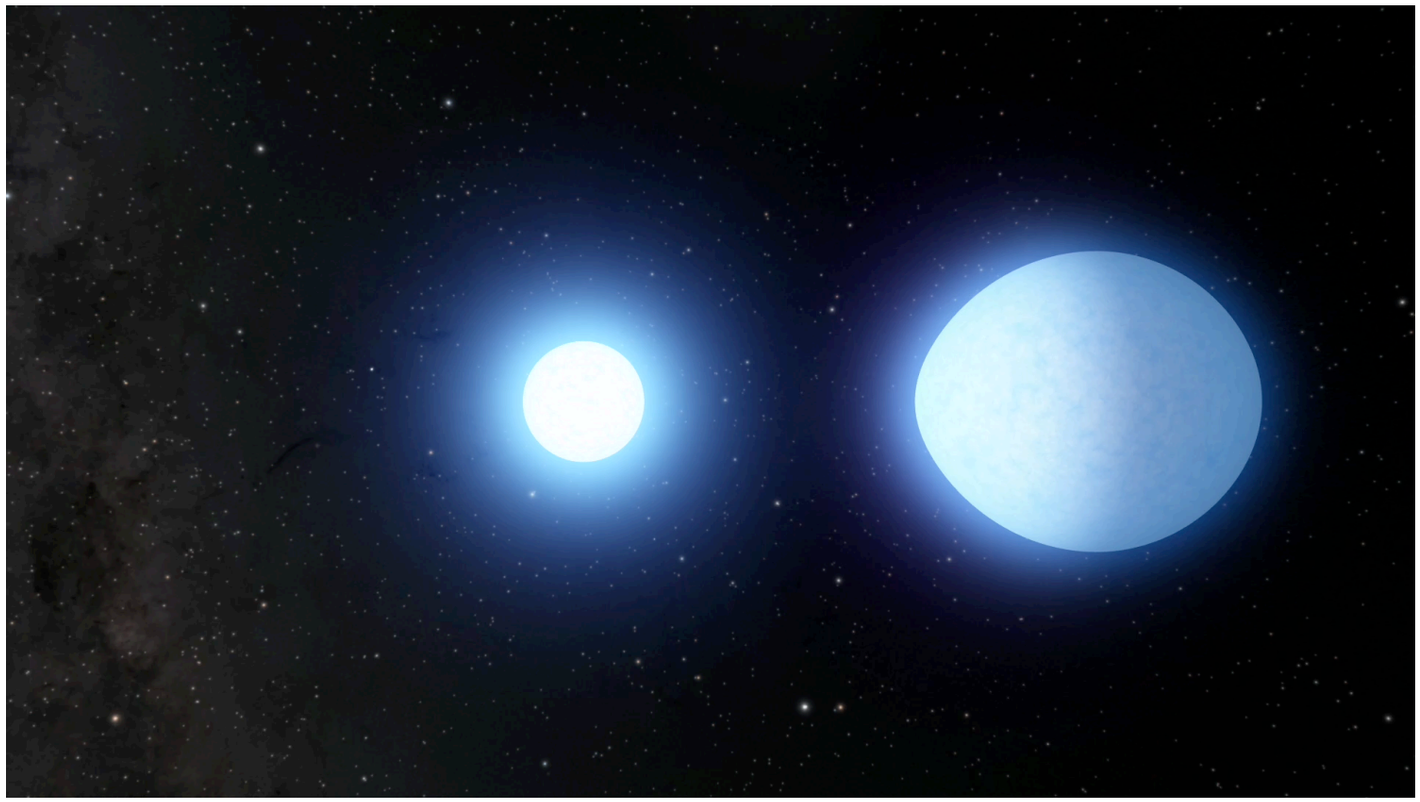
~100 days

Kepler - small area non-sparse

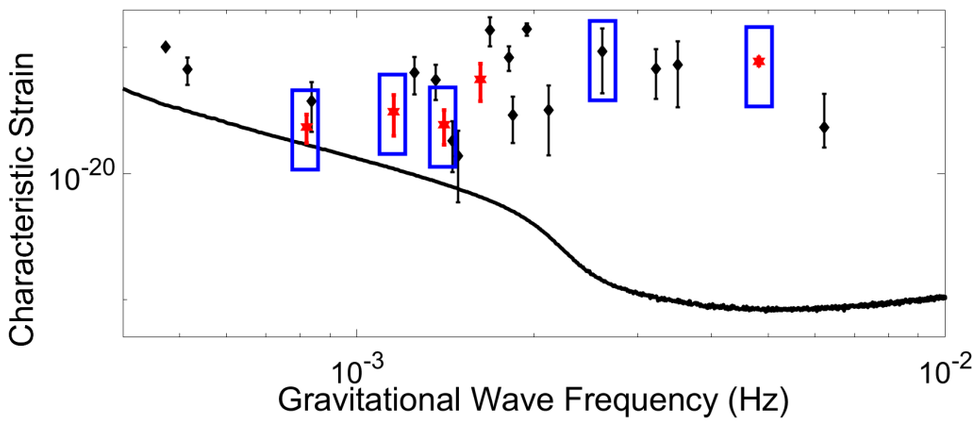


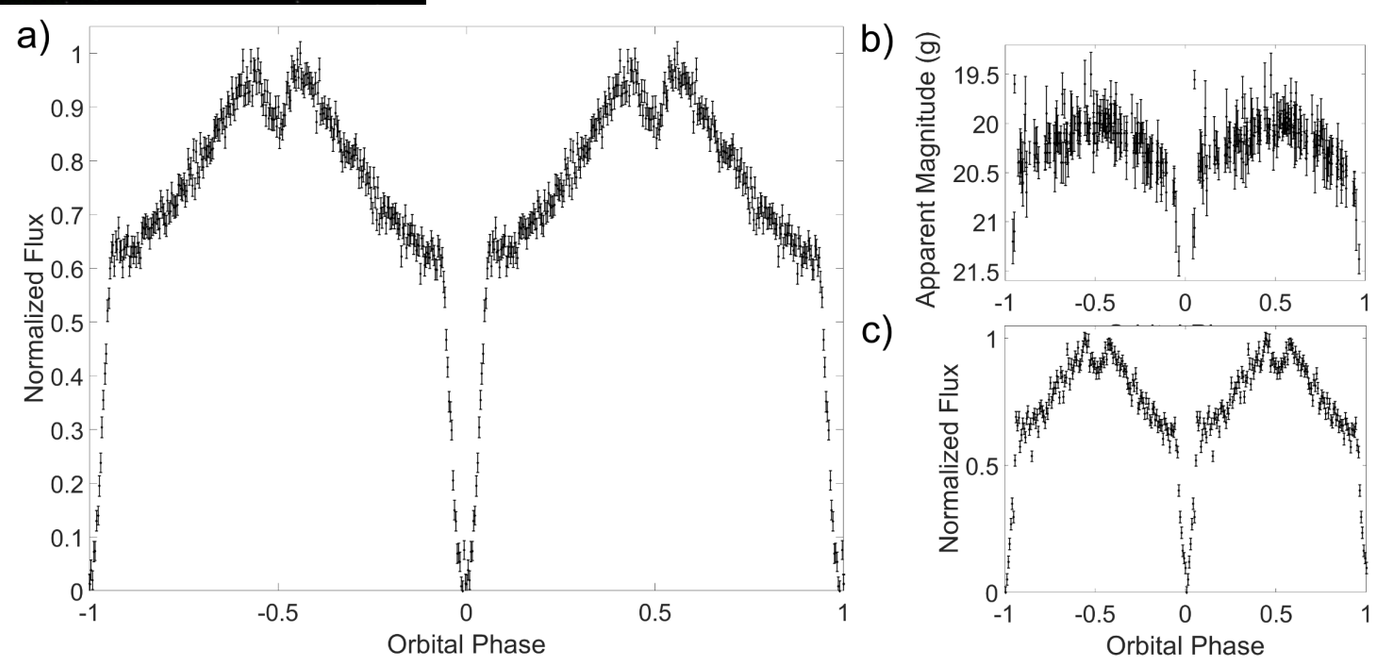
ZTF J153932.16+502738.8

Burdge et al, Nat 24 Jul 2019



hot primary ≈ 0.6 M WD (likely C-O), cool secondary ≈0.2 M WD (likely He).



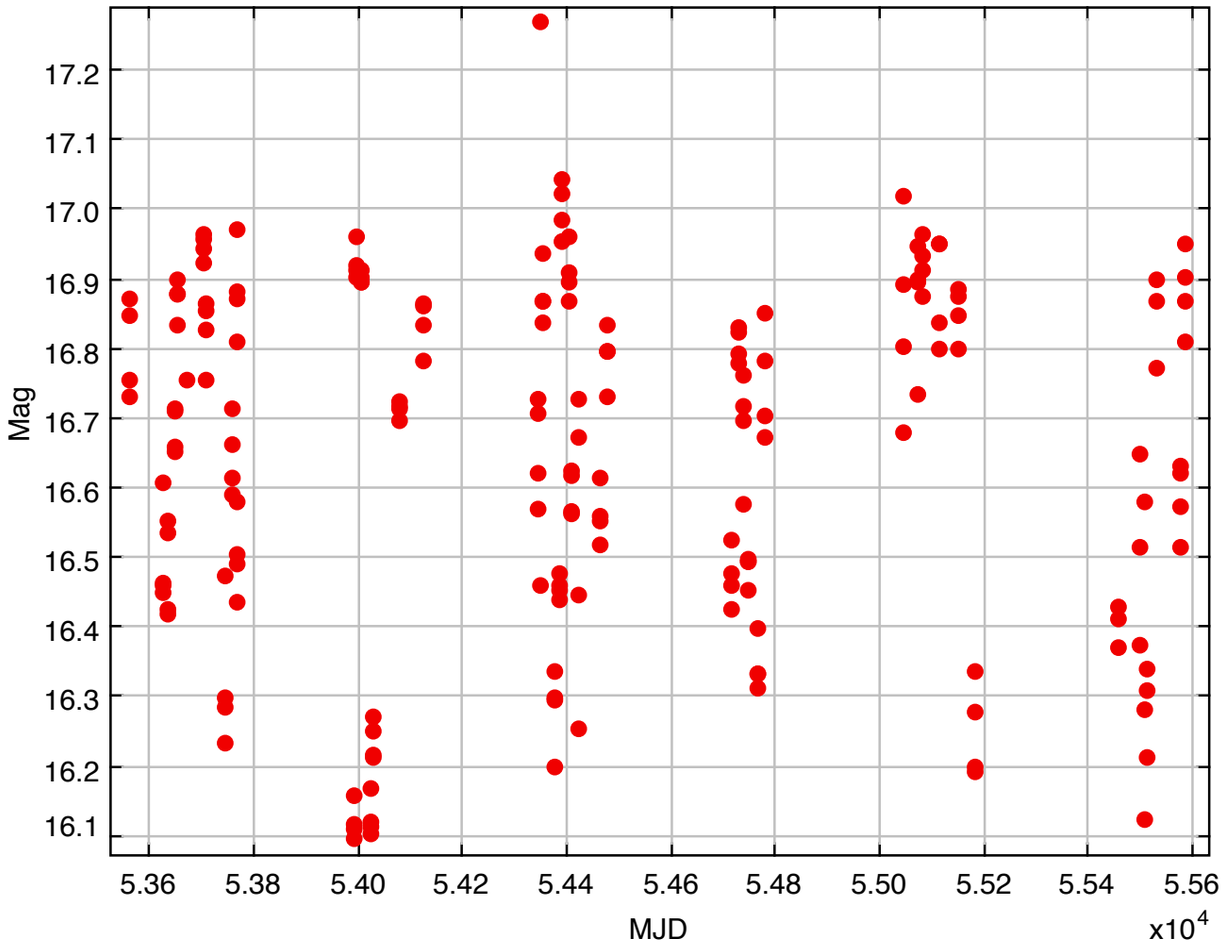


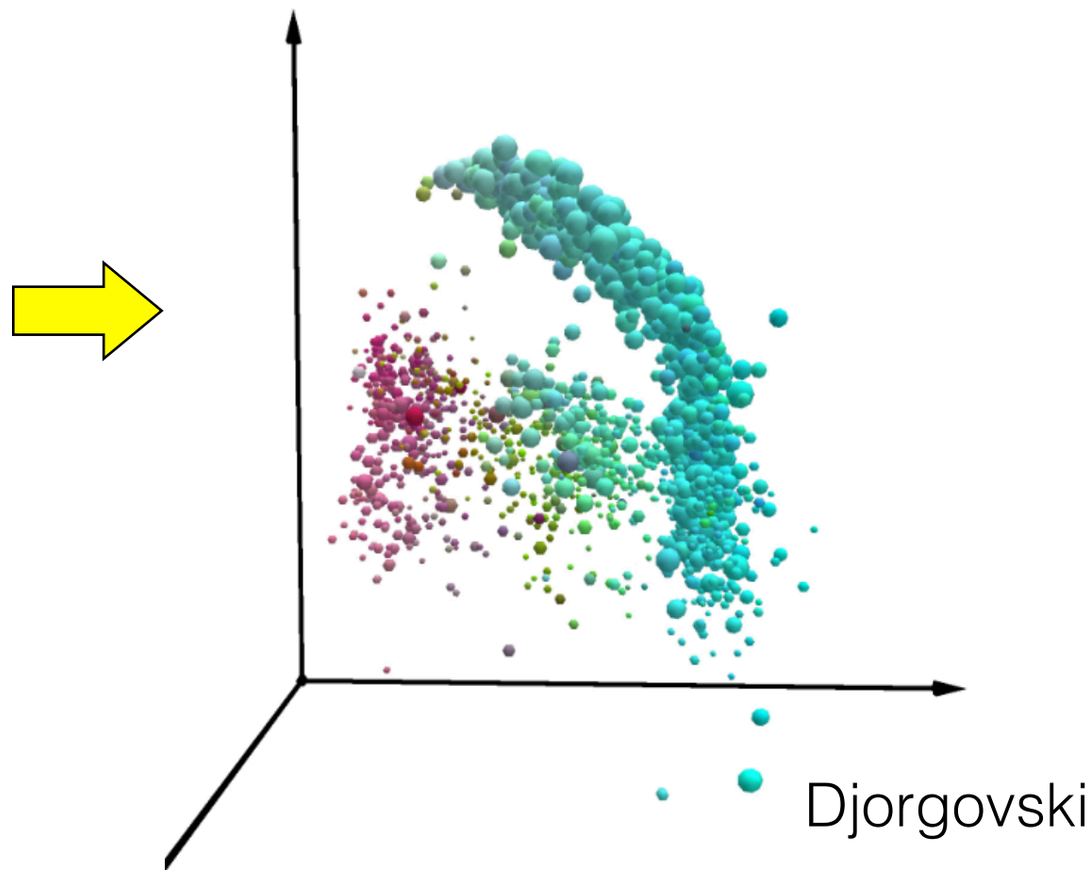
light curves from CHIMERA, ZTF, KPED (Kitt Peak)

Statistical features

Streaming updates!

Compute features (statistical measures) for each light curve:
amplitudes, moments, periodicity, etc.
Converts heterogeneous light curves into homogeneous
feature vectors in the parameter space
Apply a variety of automated classification methods





Many features
- not all are independent

Adam Miller

flux_%_mid20
flux_%_mid35
flux_%_mid50
flux_%_mid65
flux_%_mid80

scatter_res_raw

fold_2p_slope_10%
fold_2p_slope_90%

medperc90_p2_p

QSO
non_QSO

p2p_scatter_pfold_over_mad

p2p_scatter_2praw

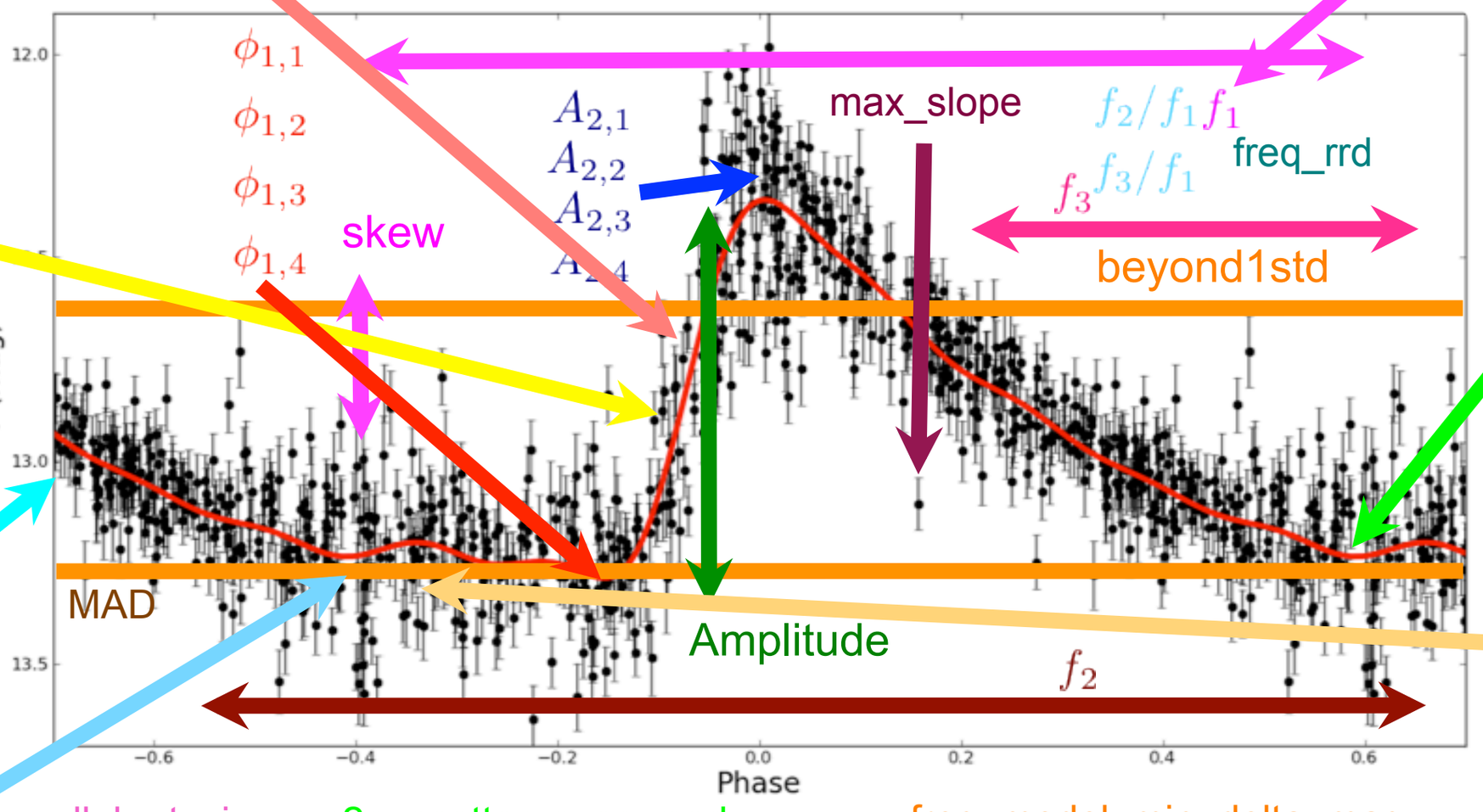
pair_slope_trend

percent_difference_flux_percentile

freq_signif

freq_n_alias

freq_varrat



$A_{1,1}$
$A_{1,2}$
$A_{1,3}$
$A_{1,4}$
$A_{2,1}/A_{1,1}$
$A_{3,1}/A_{1,1}$

freq_y_offset

stetson_j
stetson_k

$\phi_{3,1}$
$\phi_{3,2}$
$\phi_{3,3}$
$\phi_{3,4}$

small_kurtosis

p2p_scatter_over_mad

freq_model_min_delta_mag
freq_model_max_delta_mag
freq_model_phi1_phi2
p2p_ssqr_diff_over_var

percent_amplitude

linear_trend

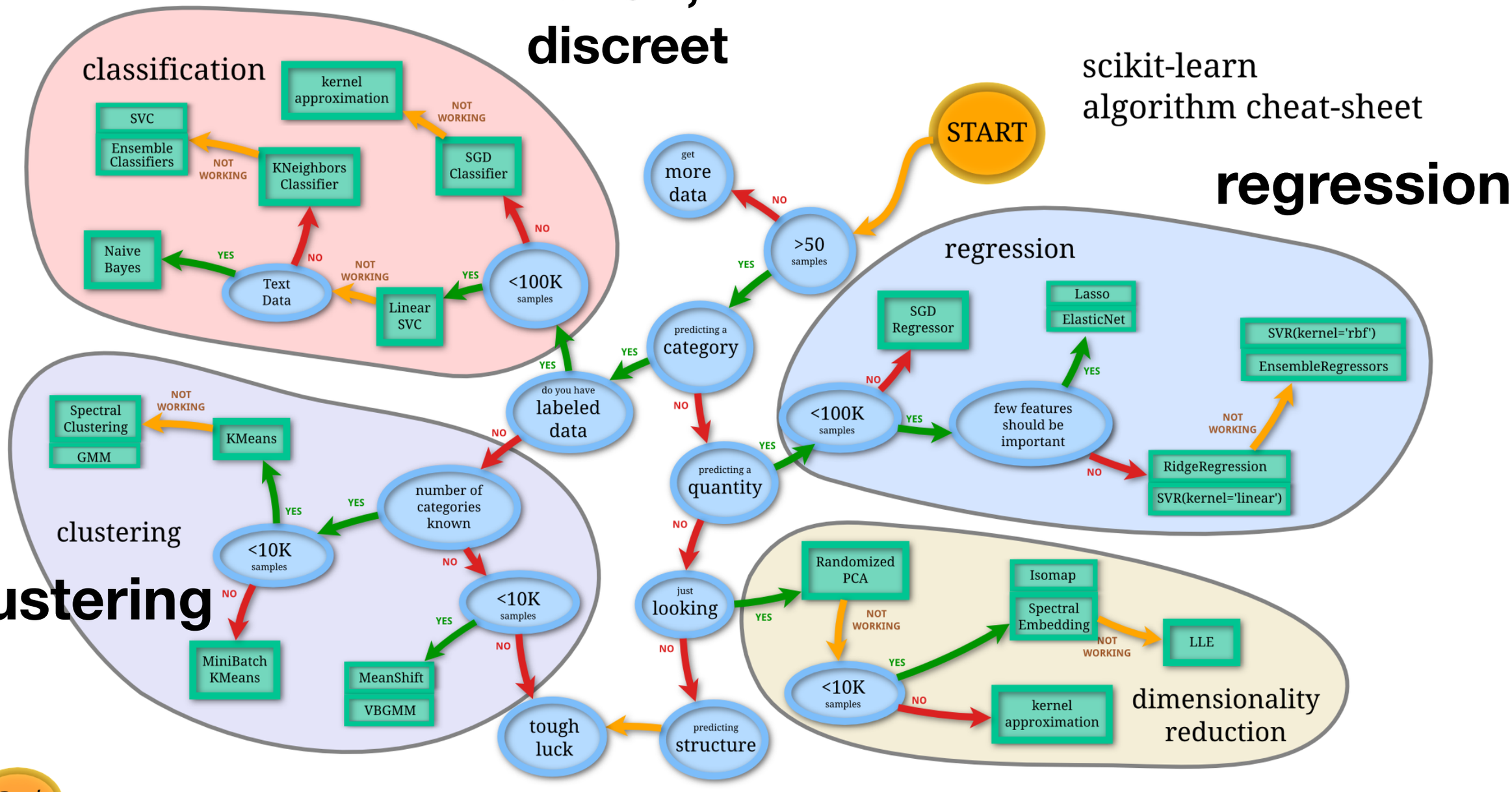
median_buffer_range_percentage

classification

labels, >10K, discreet

scikit-learn
algorithm cheat-sheet

regression



clustering

dimensionality reduction

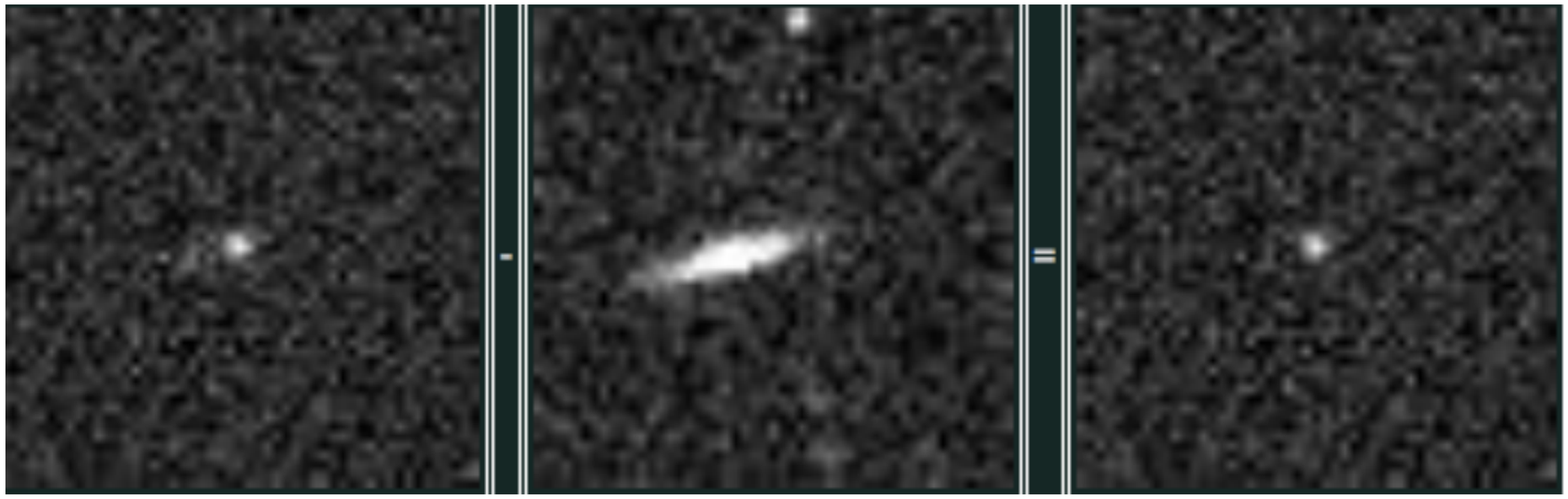
Back

scikit learn

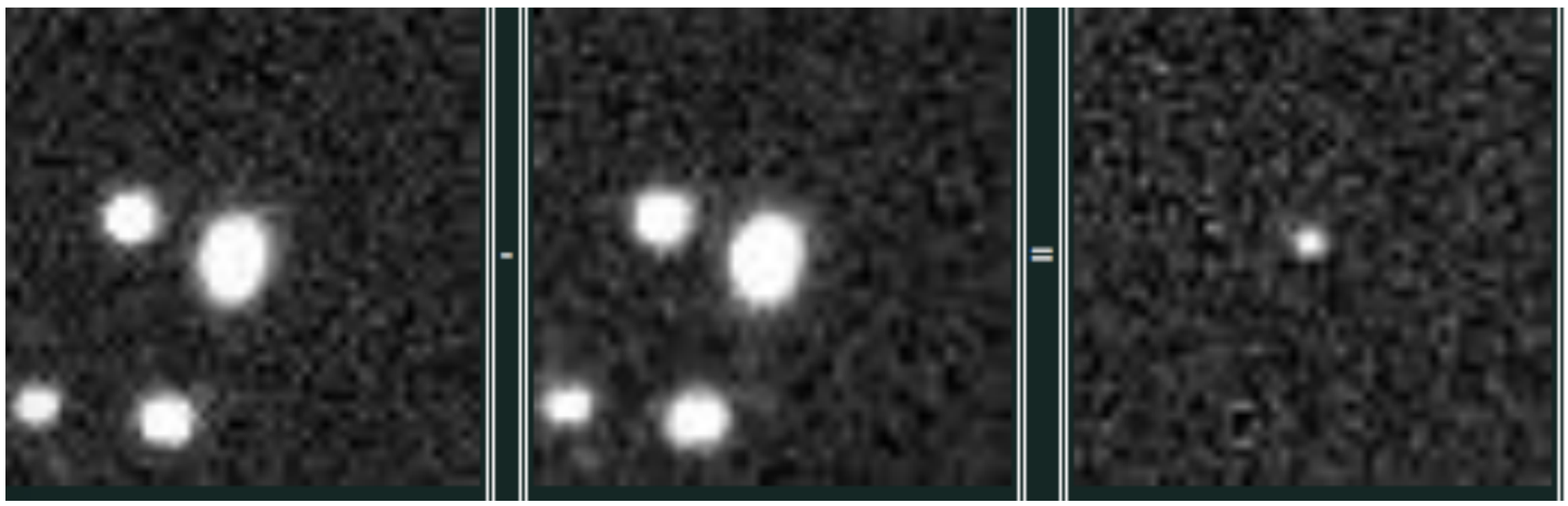
Labeled data, versus continuous variables

We will concentrate on supervised classification

Change/event detection (ZTF)

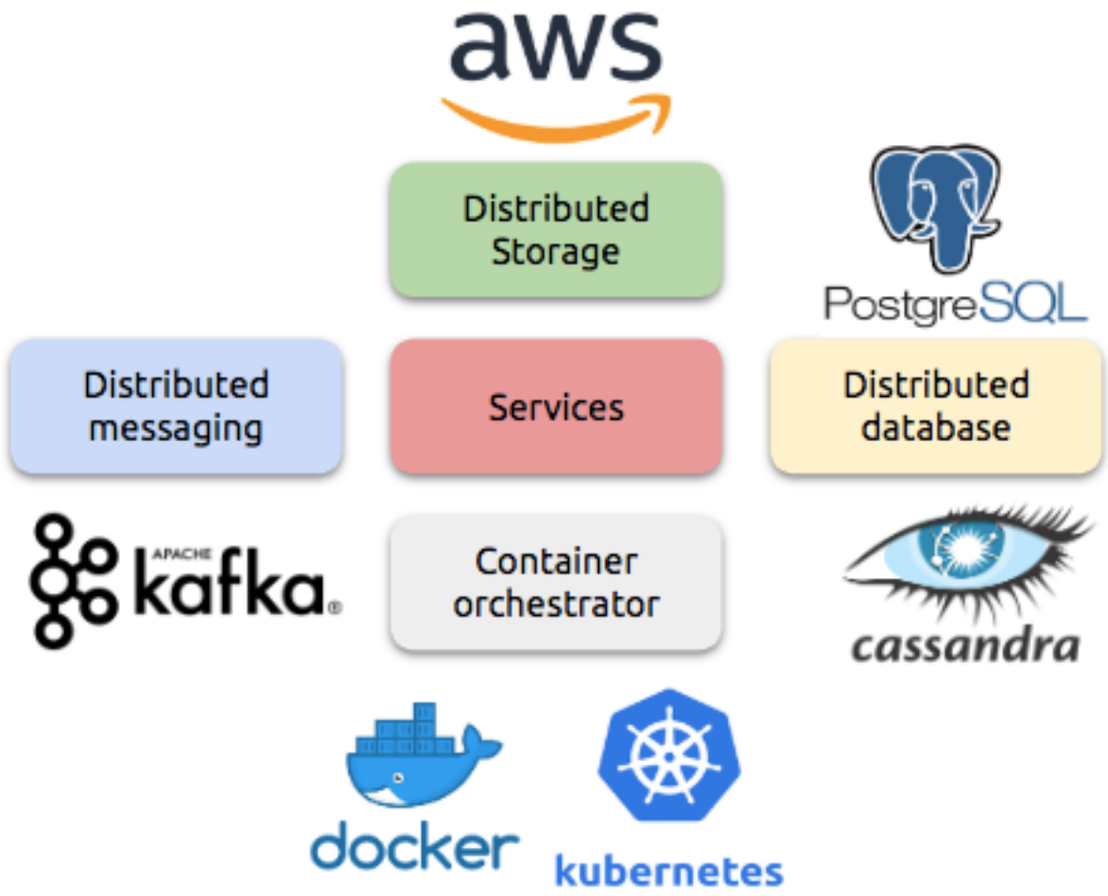


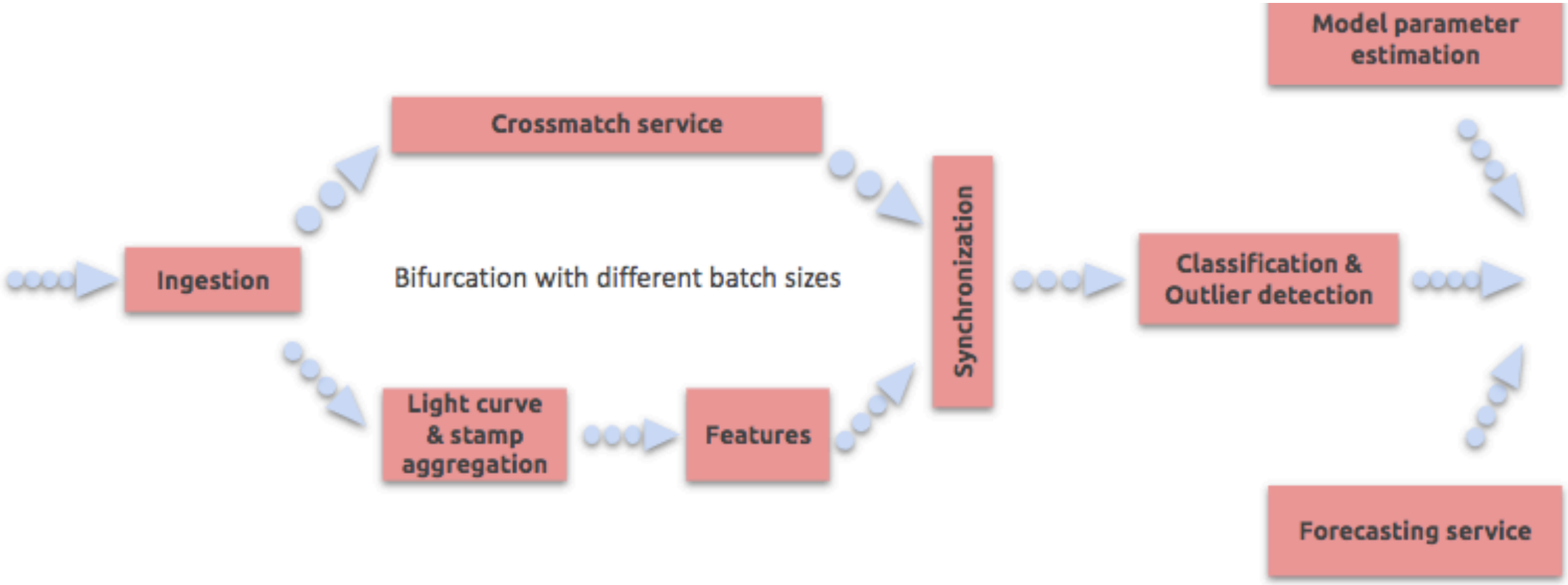
Sci Ref Diff



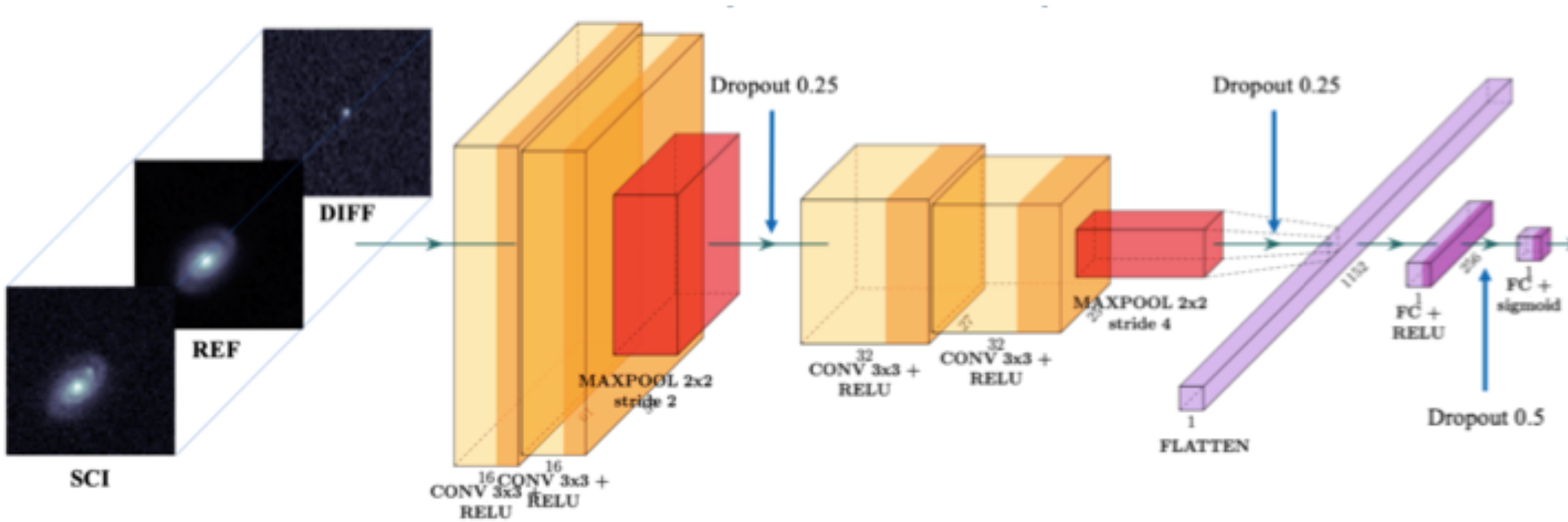
Brokers

ALeRCE





‘braai’ the real-bogus separator

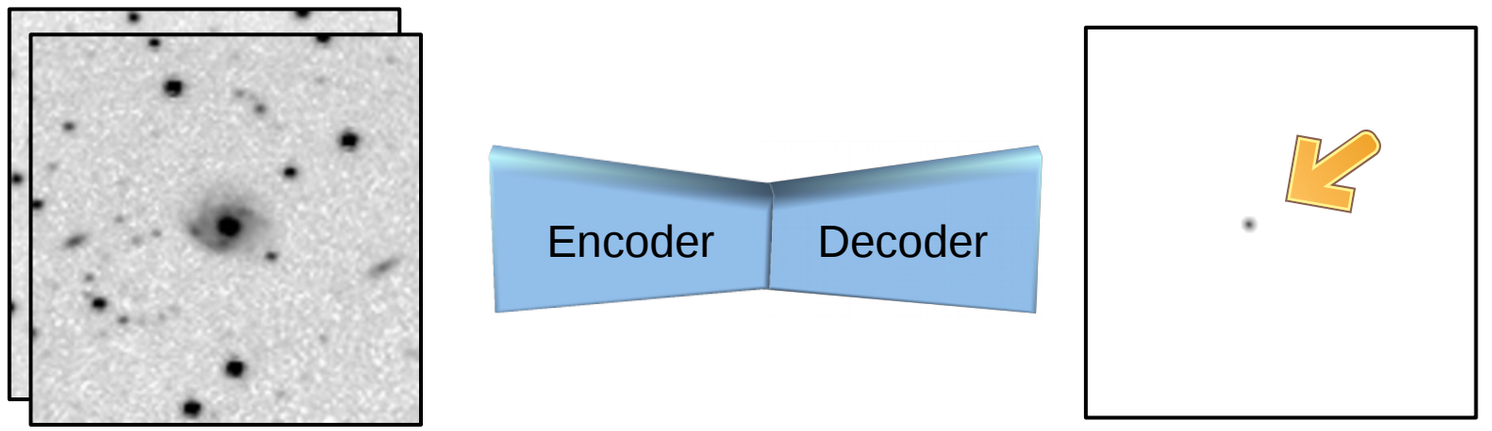


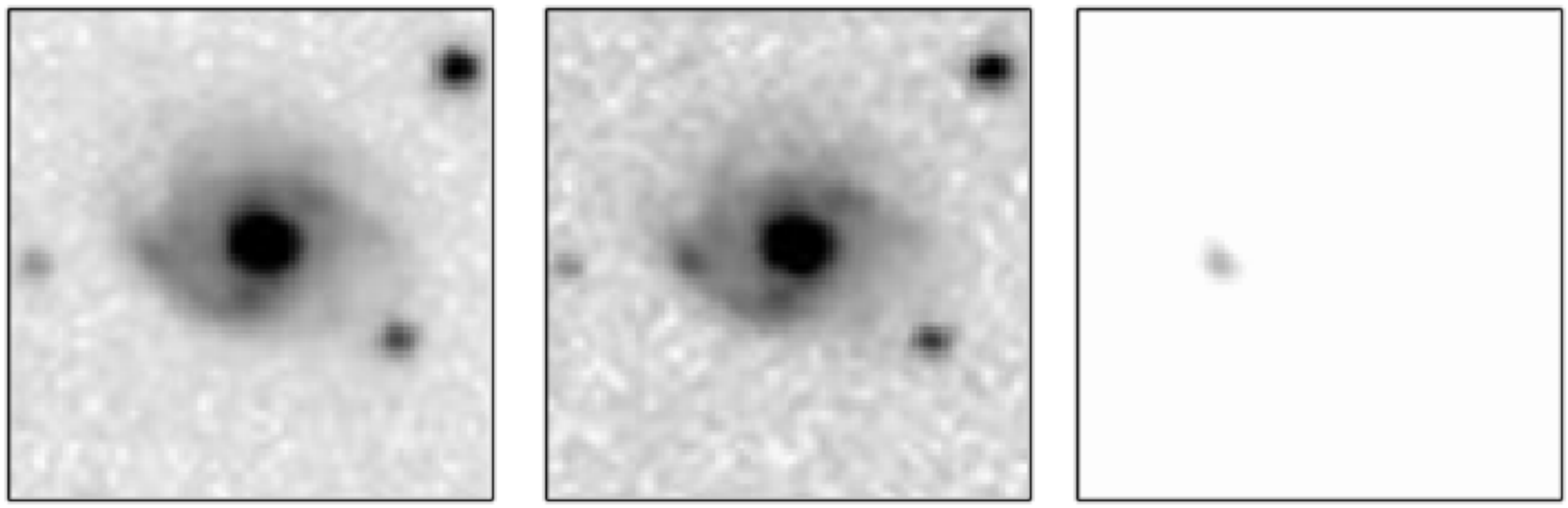
Also demonstrated with TPUs

Duev et al.

Image subtraction for hunting transients without subtraction

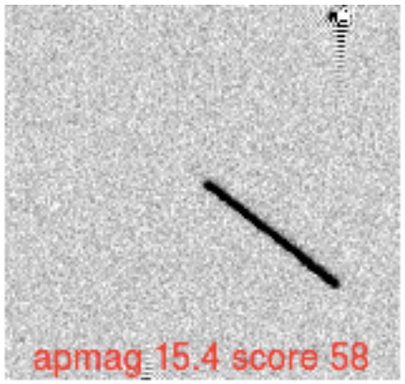
Encoder/decoder

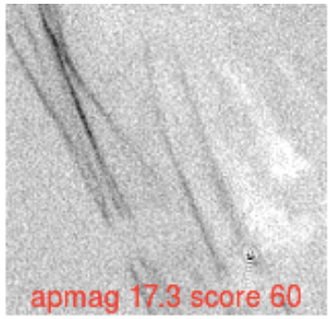




arXiv:1710.01422

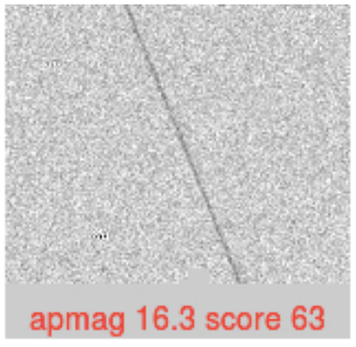
Deep Learning with AStreaks





These are ghosts and dementors

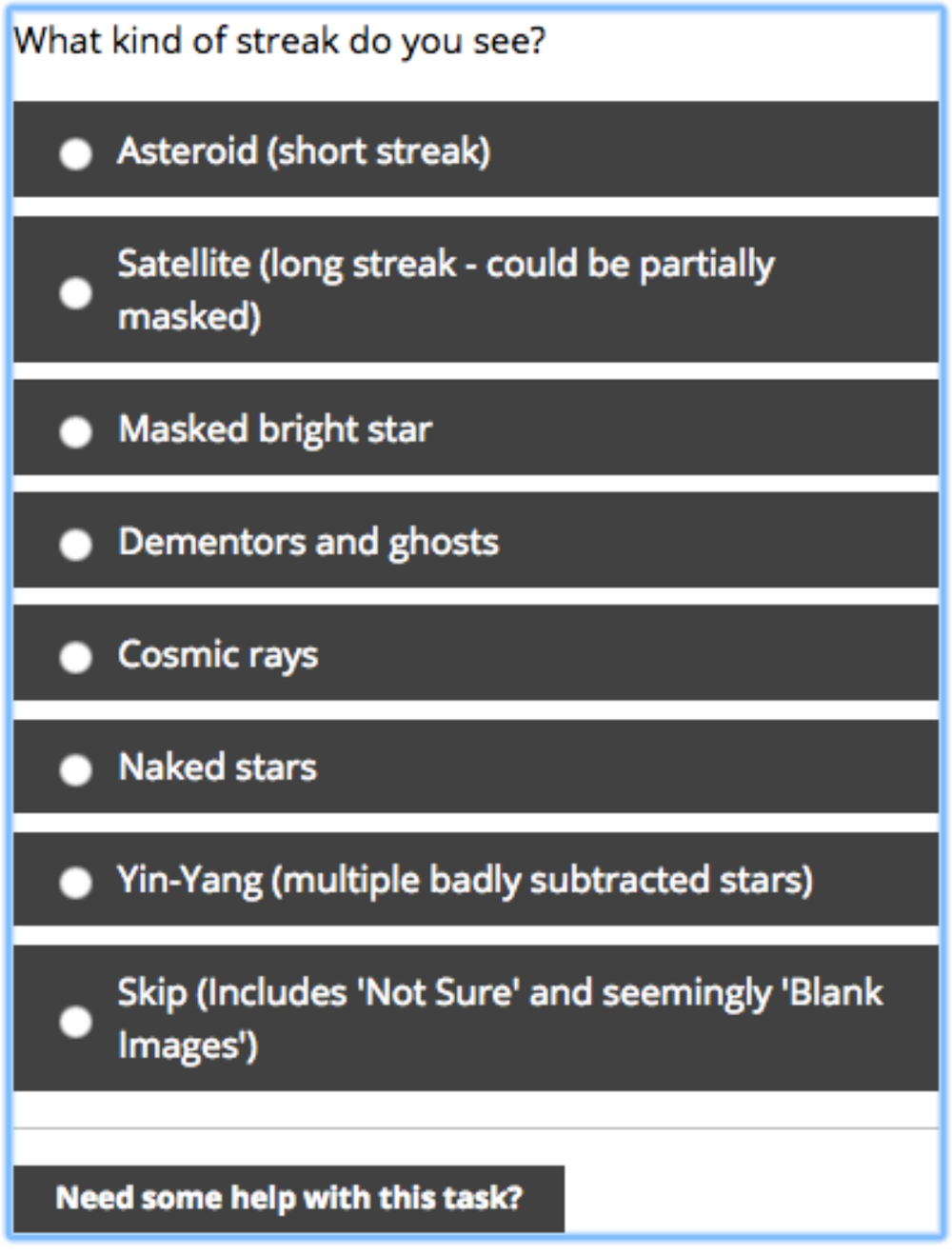
This is how a real asteroid would look. Short streak.



Another satellite trail

A satellite trail. Note that part of it is masked out, and the unmasked trail is longer.

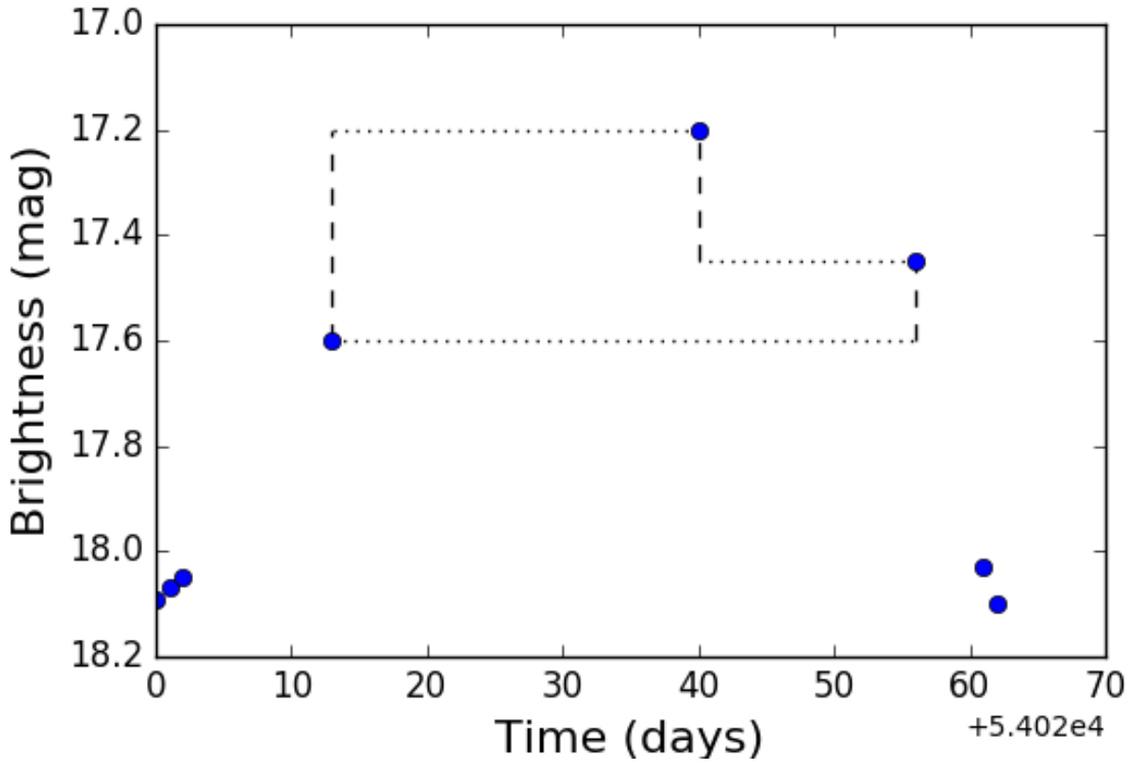
A masked bright star



Duev et al.

(dmdt) Image representation

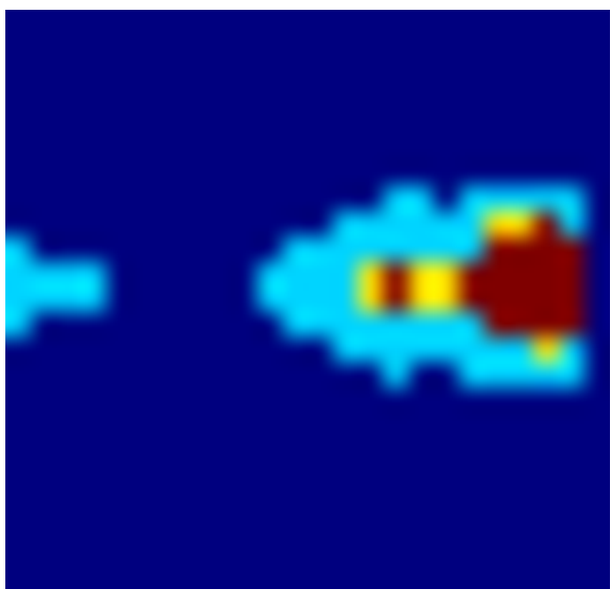
Mahabal, Sheth et al., 2017
1709.06257



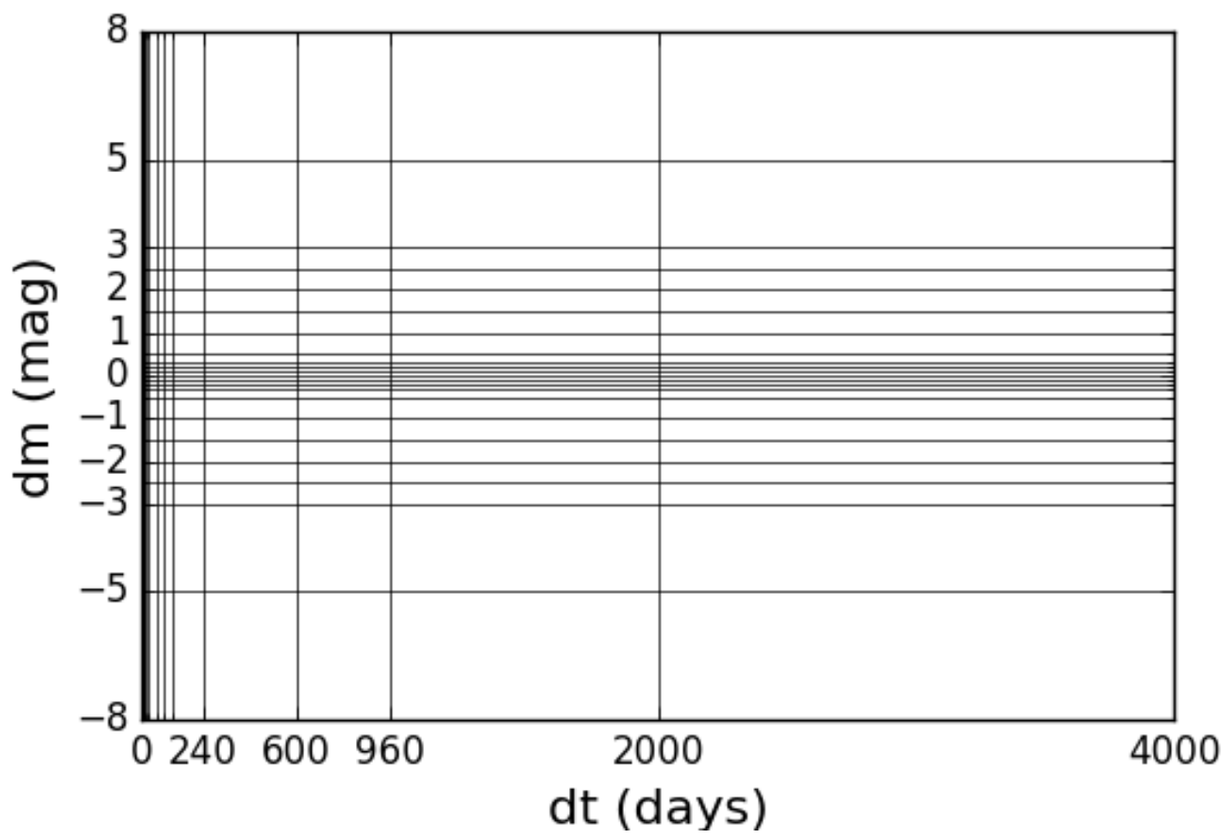
light curve with n points

23 x 24 output grid

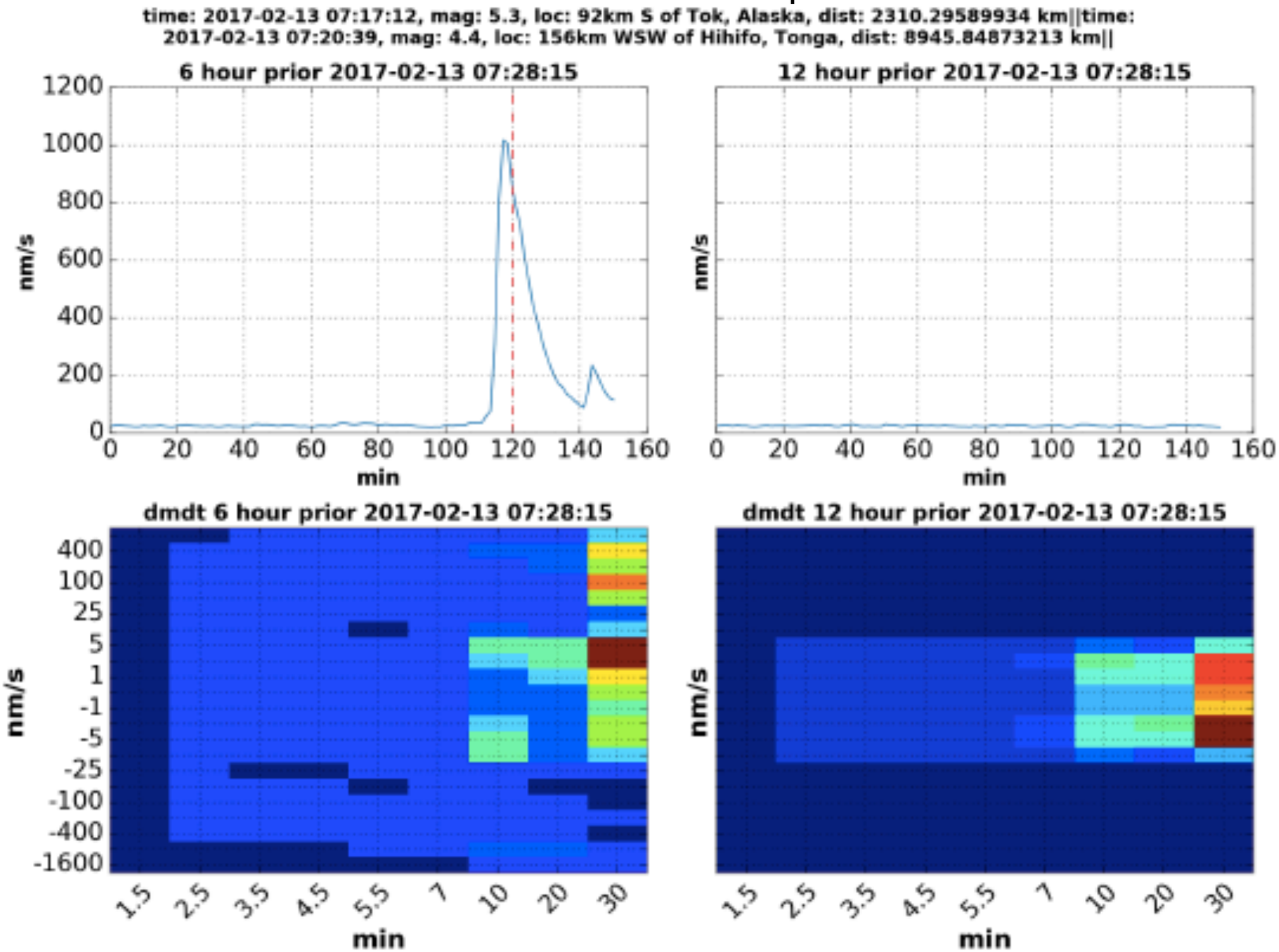
n * (n-1)/2 points



Area equalized pixels

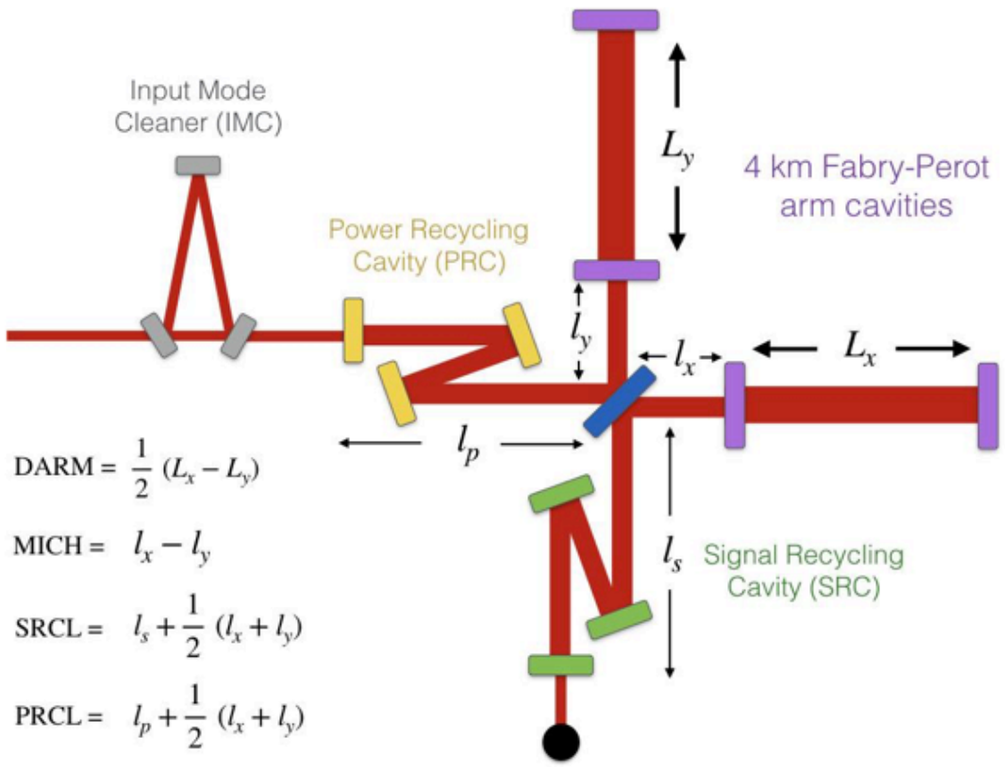


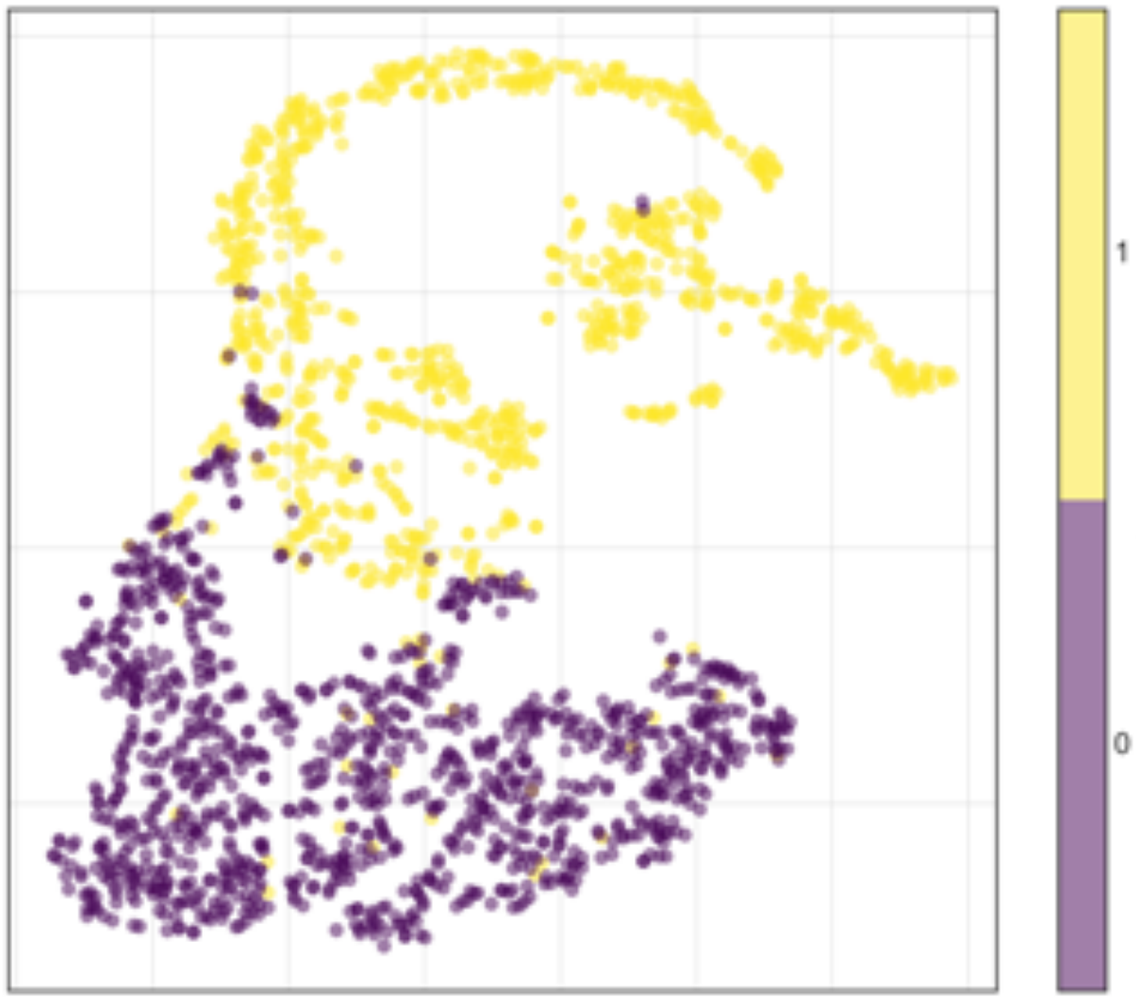
Effect of earthquakes



LIGO **Putting the instrument in safe mode in case of an adverse event**

Nearly clean separation of lock-loss events in GW detectors using cavity channels





POP+SRCL+MITCH

Unsupervised classification

- t-SNE
- UMAP

We may need more unsupervised learning (combined with not-so-easy-validation)

When we're learning to see, nobody's telling us what the right answers are - we just look. Every so often, your mother says "that's a dog", but that's very little information, You'd be lucky if you got a few bits of information - even one bit per second - that way. The brain's visual system has 10^14 neural connections, And you only live for 10^9 seconds. So it's no use learning one bit per second. You need more like 10^5 bits per second. And there's only one place you can get that much information: from the input itself, Geoffrey Hinton, 1996

**Human *on* the loop,
transfer learning
and all that**

Reuters

ZTF ~0.1 LSST

	ZTF	LSST 2022
No. of sources	1 billion	37 billion
No. of detections	1 trillion	37 trillion
Annual visits per source	1000 (2+1 filters)	100 (6 filters)
No. of pixels	600 million (1320 cm^2 CCDs)	3.2 billion (3200 cm^2 CCDs)
Field of view	47 deg^2	9 deg^2
Hourly survey rate	3750 deg^2	1000 deg^2
Nightly alert rate	1 million	10 million
Nightly data rate	1.4 TB	15 TB

42!

Astronomy's continuing battle with bigdata

Volume: TB -> PB -> EB -> ZB

Velocity: Real-time analysis/publishing/follow-up (**partly 'Volatility' too**). Variability on ms to s to days

Variety: 400 classes - multiband images; time series; spectra; polarization; …

….

ZTF: 1.4TB/day
SKA: EB/day

Veracity: error-bars; fuzzy classifications

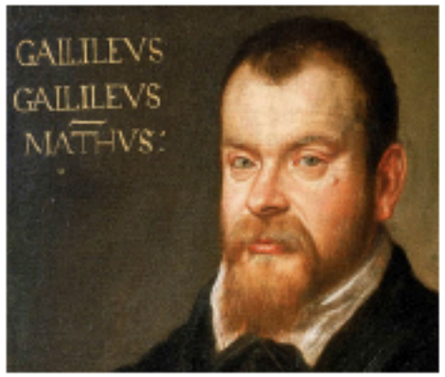
The Evolving Paths to Knowledge

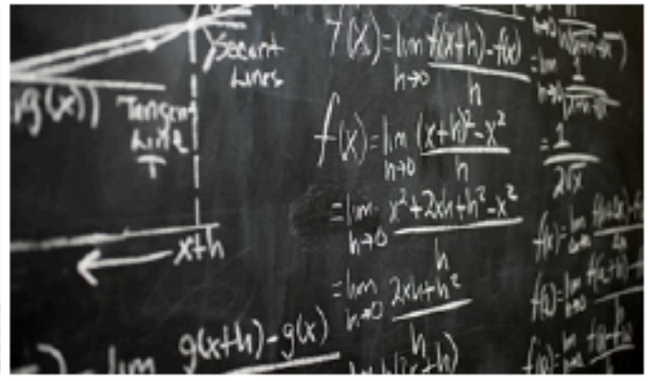
- The First Paradigm:
  Experiment/
  Measurement

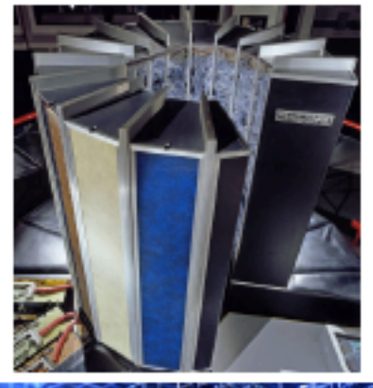
- The Second Paradigm:
  Analytical Theory

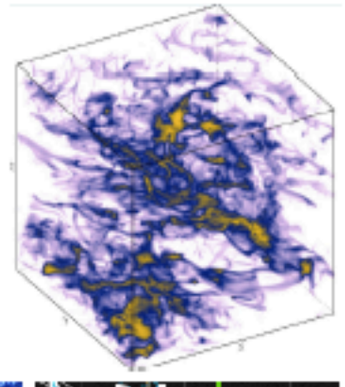
- The Third Paradigm:
  Numerical Simulations

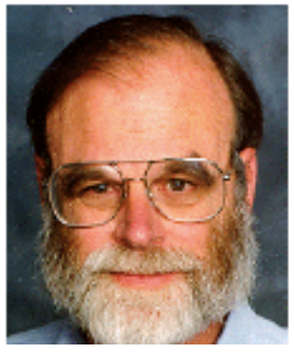
- The Fourth Paradigm:
  Data-Driven Science

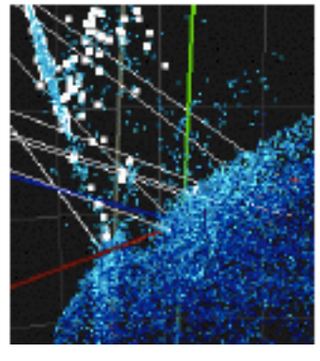




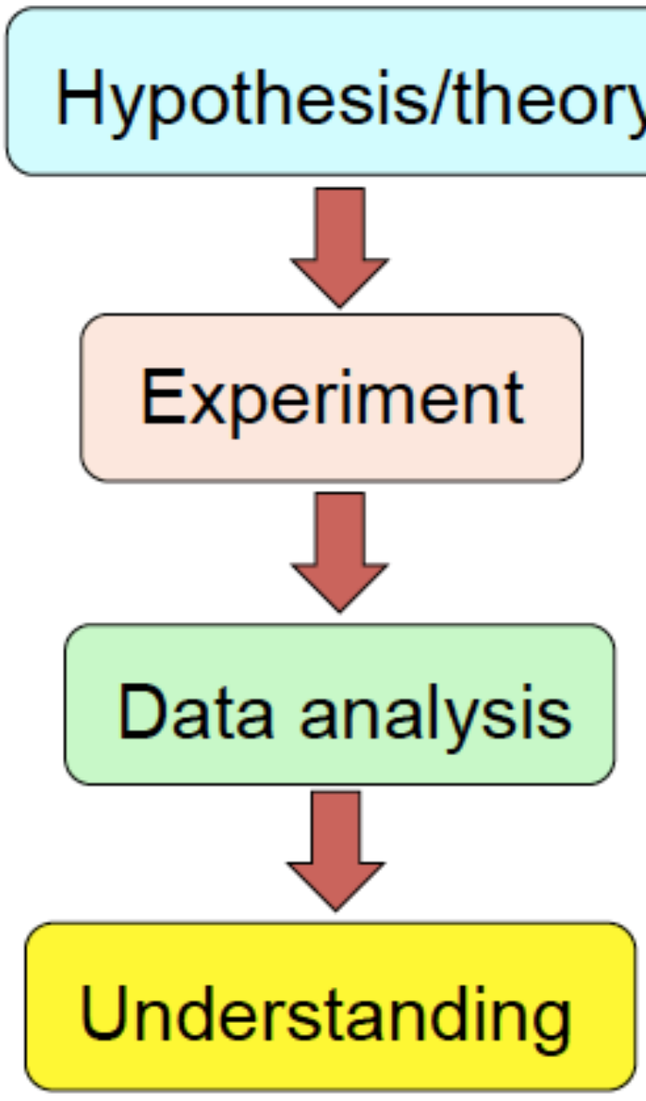




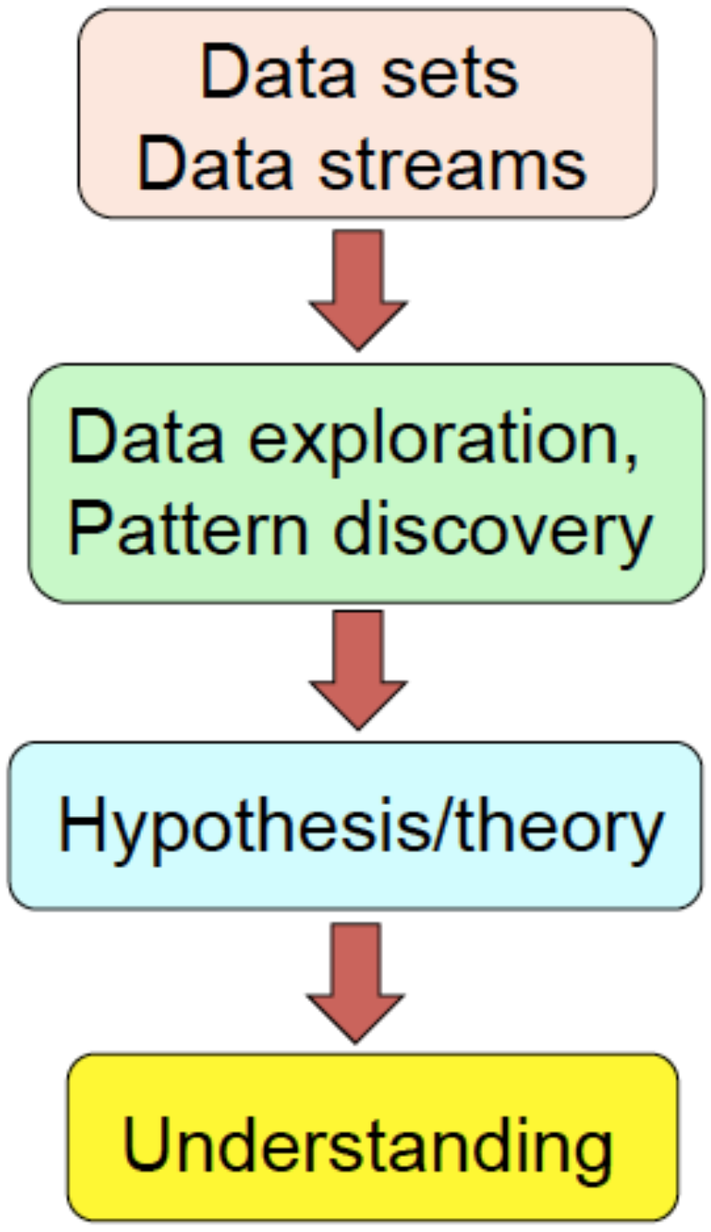




Hypothesis-driven science



Data-driven science



The two approaches are complementary

A Modern Scientific Discovery Process

Data Gathering (finstruments, sensor networks, their pipelines…)

↳ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability
} Databases, Data grids

↳ **Data Mining**

Key Technical Challenges

Pattern or correlation search
Clustering analysis, classification
Outlier / anomaly searches
Hyperdimensional visualization

↳ **Data Understanding**

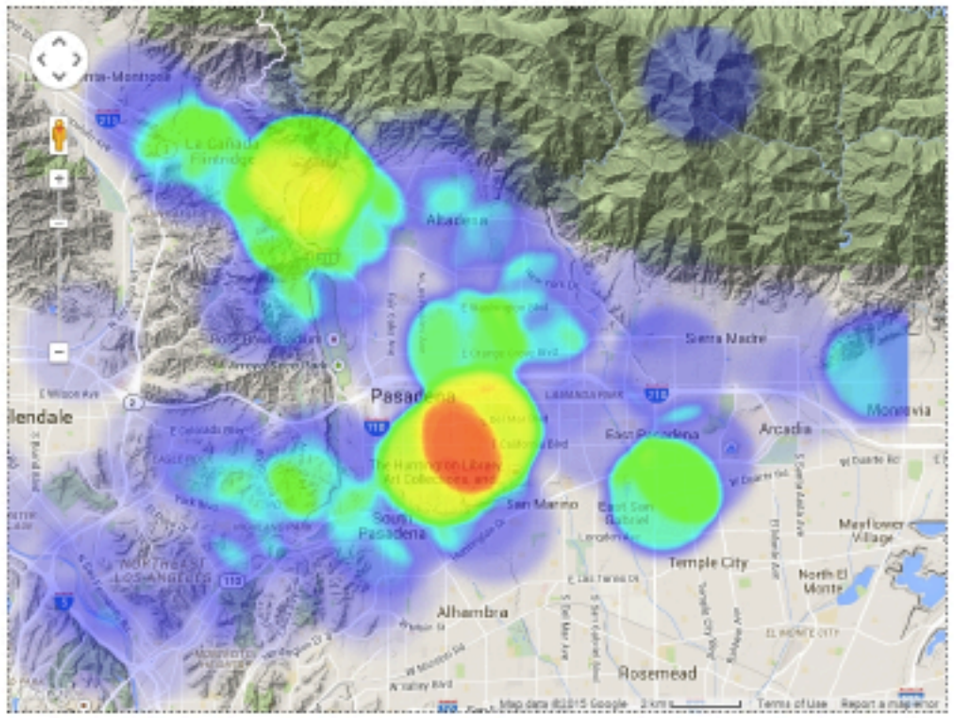
+feedback

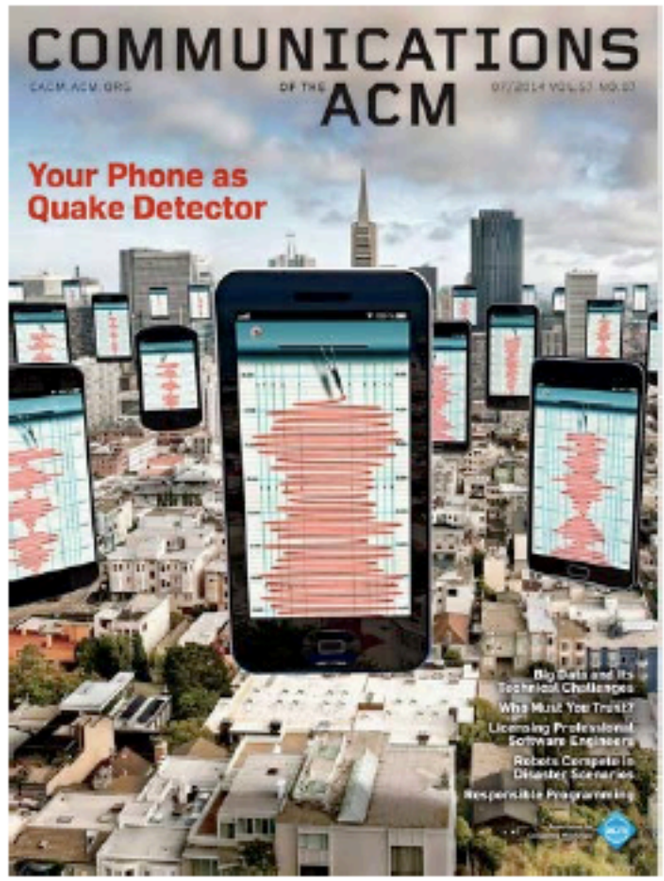
↳ **New Knowledge**

Real Time Classification and Response

Seismology:
Cell phones as a sensor network

Lake Castaic M4.2 Jan 4 2015 Heatmap



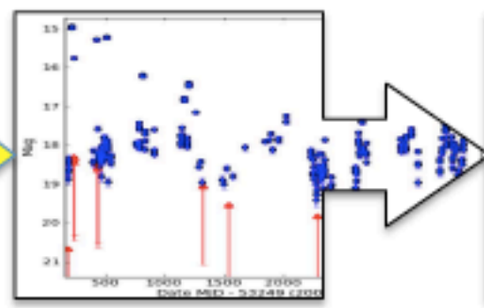


Time domain astronomy

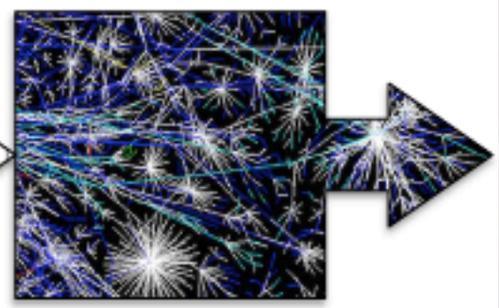
Event

Detection

Classification

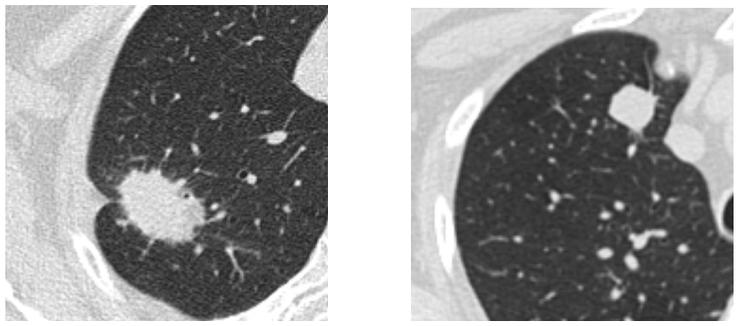


Decision making

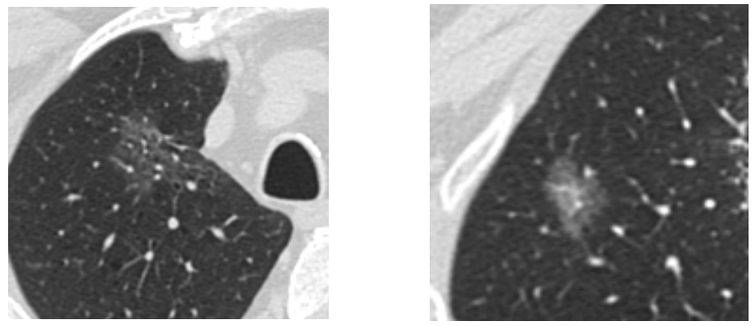


Follow-up

Consistency - solid

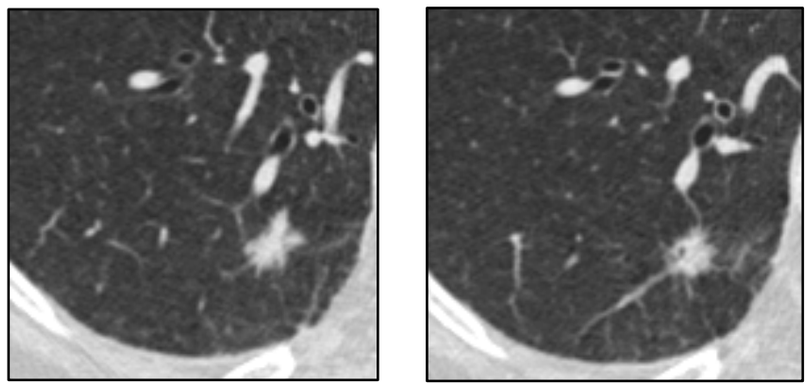


Consistency - pure GGN



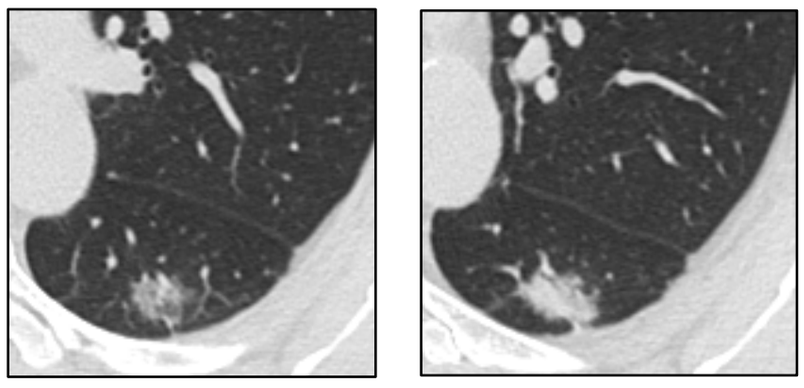
Solid or PSN

(Part Solid Nodule)

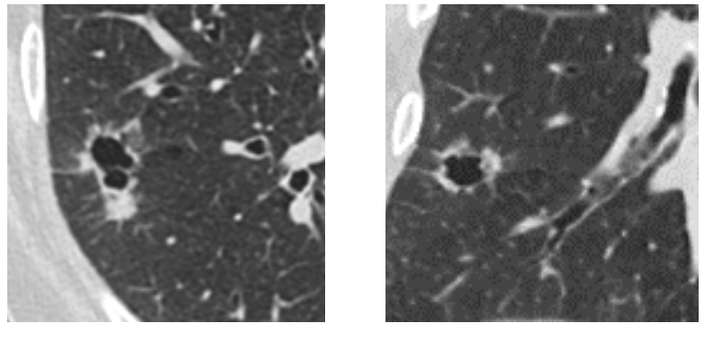


Diff. axial levels - PSN

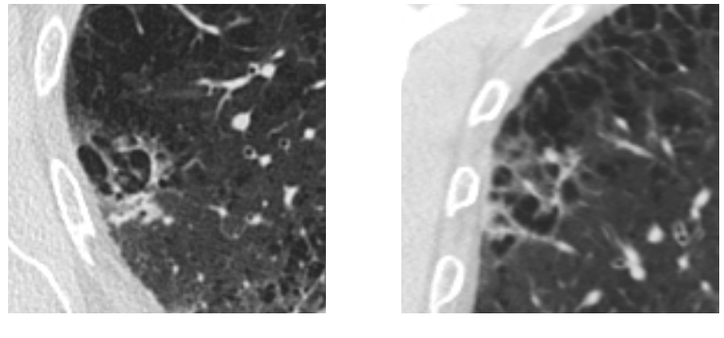
(by consensus)



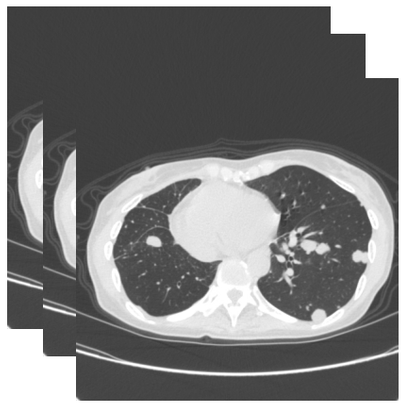
Per-cystic or cystic



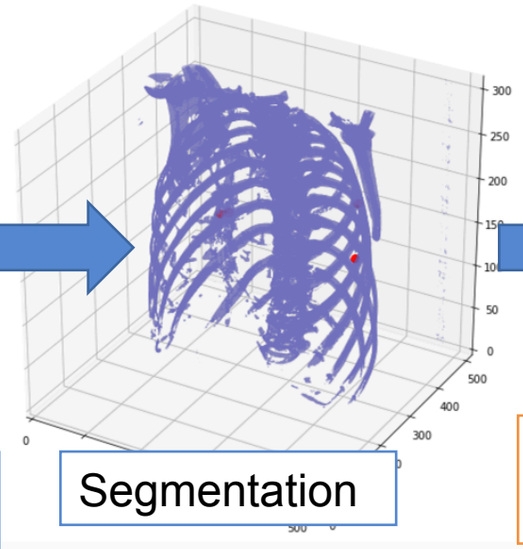
Axial **Coronal**



Axial **Coronal**

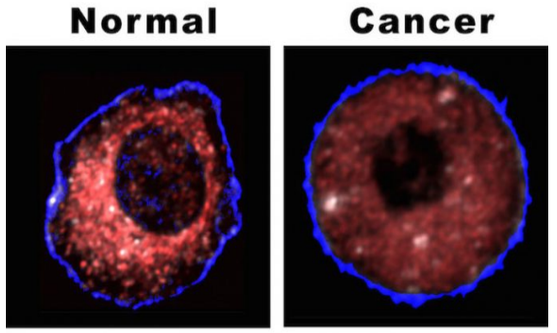


NLST Dicom Slices:
~1000 Patients



Segmentation

Extract bounding boxes of nodule



LUNA16/Kaggle1 trained model predicts cancer

GRT123

Fangzhou, L. (2017)

Domain adaptation and transfer learning

Accuracy 87% on GRT1
Repeat on NLST data
Retrain final layer with NLST data to improve

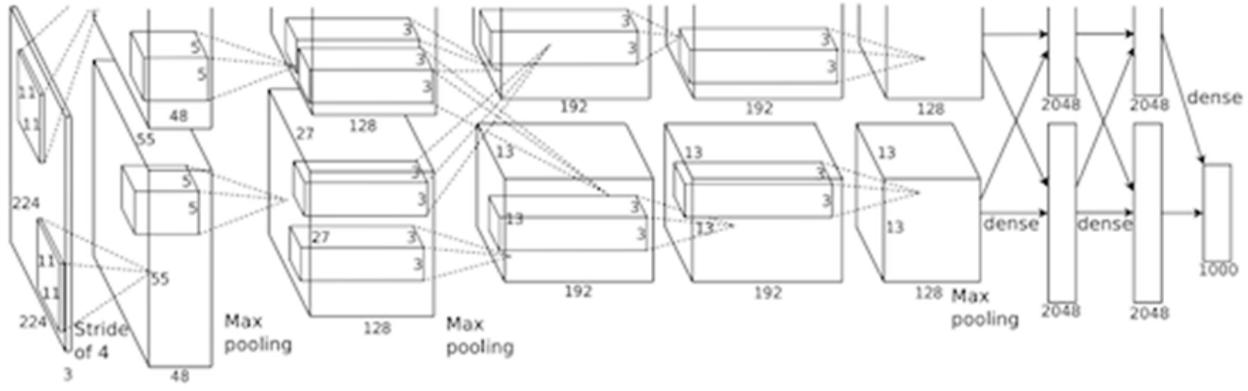
Also protein folding Bioinformatics,

…

Explainability/Interpretability!

IMAGENET

Large Scale Visual Recognition Challenge

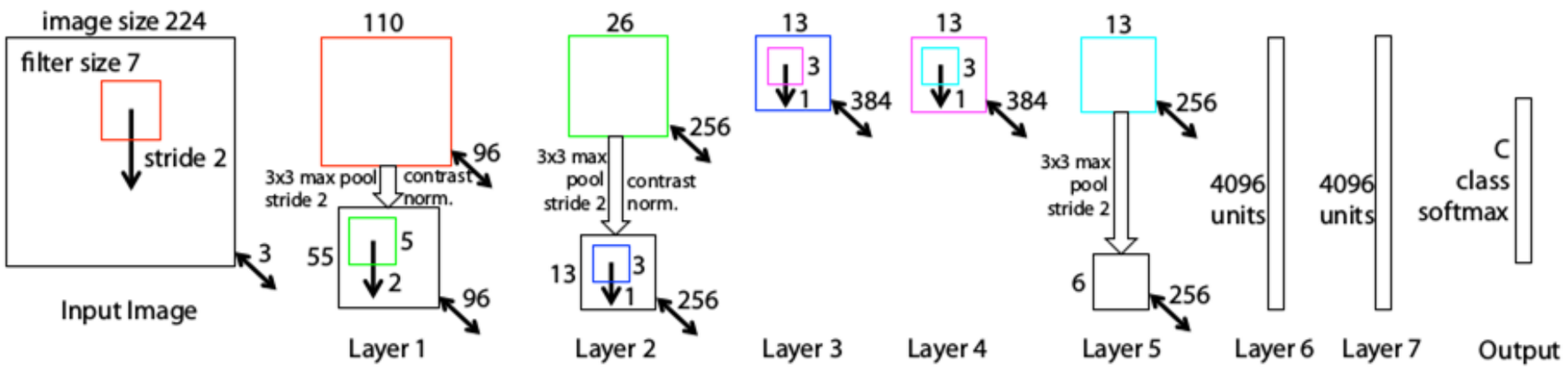


AlexNet architecture (May look weird because there are two different "streams". This is because the training process was so computationally expensive that they had to split the training onto 2 GPUs)

- 2012: Alexnet (error rate 15.4%)
- 2013: ZFnet (error rate 11.12%)

ILSVRC

- DeConvNets (Caffe)



ZF Net Architecture

Adit Deshpande

https://adeshpande3.github.io/adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html

2016 ILSVRC leaderboard

Team name	Entry description	**Number of object categories won**	mean AP
CUImage	Ensemble of 6 models using provided data	109	0.662751
Hikvision	Ensemble A of 3 RPN and 6 FRCN models, mAP is 67 on val2	30	0.652704
Hikvision	Ensemble B of 3 RPN and 5 FRCN models, mean AP is 66.9, median AP is 69.3 on val2	18	0.652003
NUIST	submission_1	15	0.608752
NUIST	submission_2	9	0.607124
Trimps-Soushen	Ensemble 2	8	0.61816
360+MCG-ICT-CAS_DET	9 models ensemble with validation and 2 iterations	4	0.615561
360+MCG-ICT-CAS_DET	Baseline: Faster R-CNN with Res200	4	0.590596
Hikvision	Best single model, mAP is 65.1 on val2	2	0.634003
CIL	Ensemble of 2 Models	1	0.553542
360+MCG-ICT-CAS_DET	9 models ensemble	0	0.613045

Classification error:
0.02991

Natural Adversarial Examples

Hendryks et al.
arXiv:1907.07174v2

Figure 1: Natural adversarial examples from IMAGENET-A. The red text is a ResNet-50 prediction with its confidence, and the black text is the actual class. Many natural adversarial examples are incorrectly classified with high confidence, despite having no adversarial modifications as they are examples which naturally occur in the physical world.

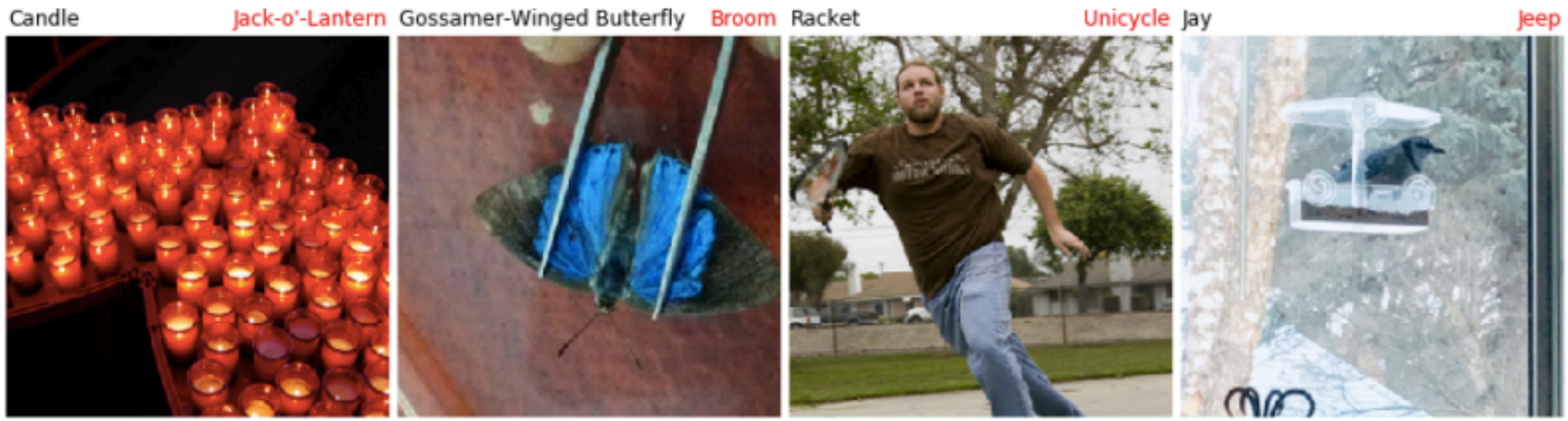
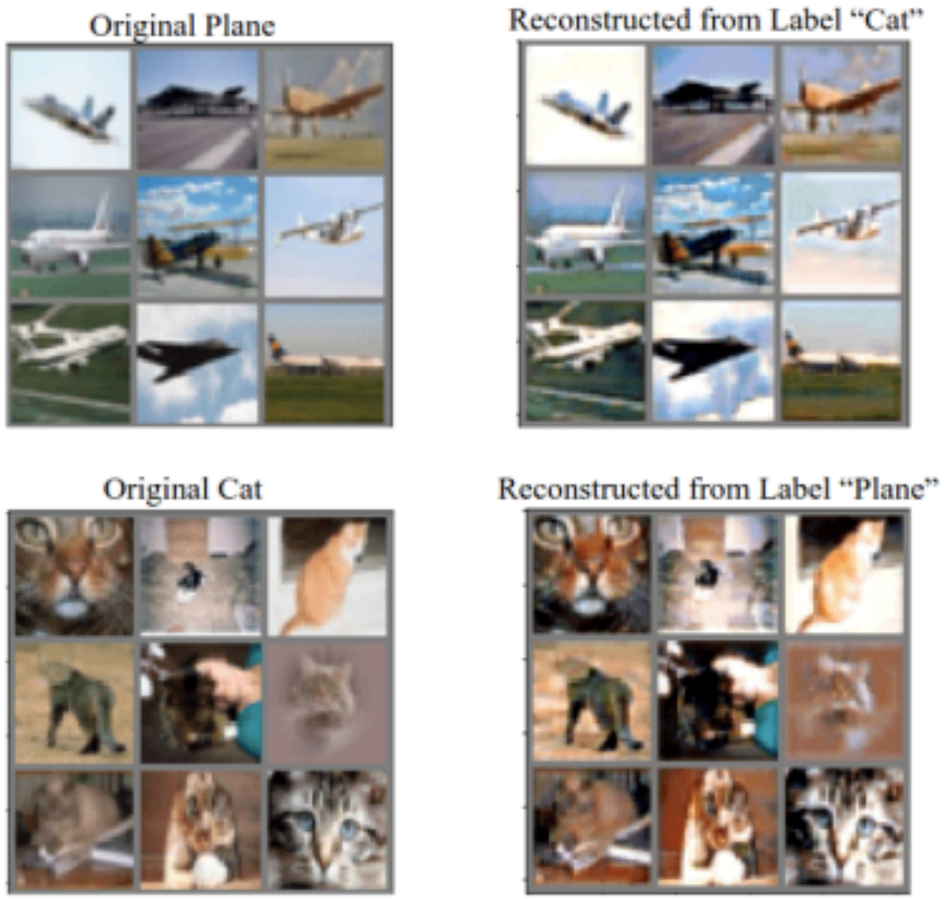


Figure 2: IMAGENET-A examples demonstrating that classifiers may predict a class even without a plausible shape in the image to support its prediction. The red text is a ResNet-50 prediction, and the black text is the actual class.

Out-of-Distribution Detection Using Neural Rendering Generative Models

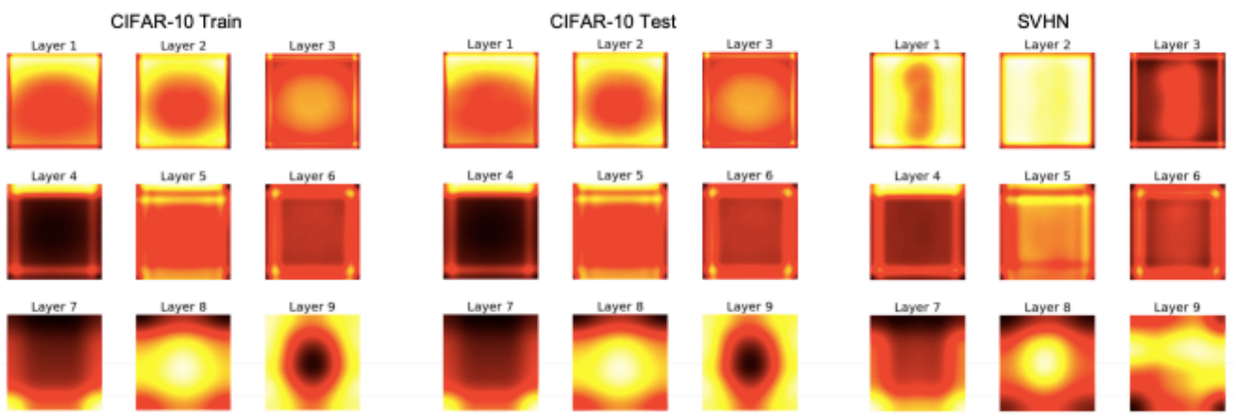


Huang et al.
arxiv: 1907.04572

Reconstruction in presence of wrong labels shows how this is different from deep-dreaming

Figure D.3: Top row: Original airplane image (left) vs airplane image reconstructed from label "Cat" using optimal latents for reconstructing original plane (right). Bottom row: Original cat image (left) vs cat image reconstructed from label "Airplane" using optimal latents for reconstructing original cat

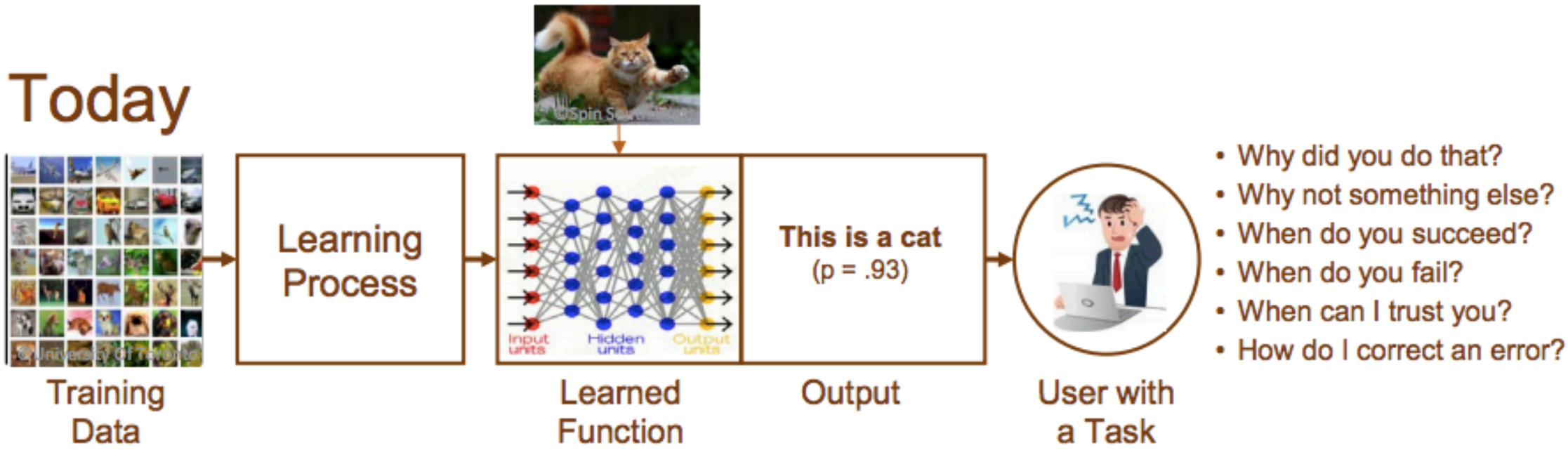
Latent variable visualization

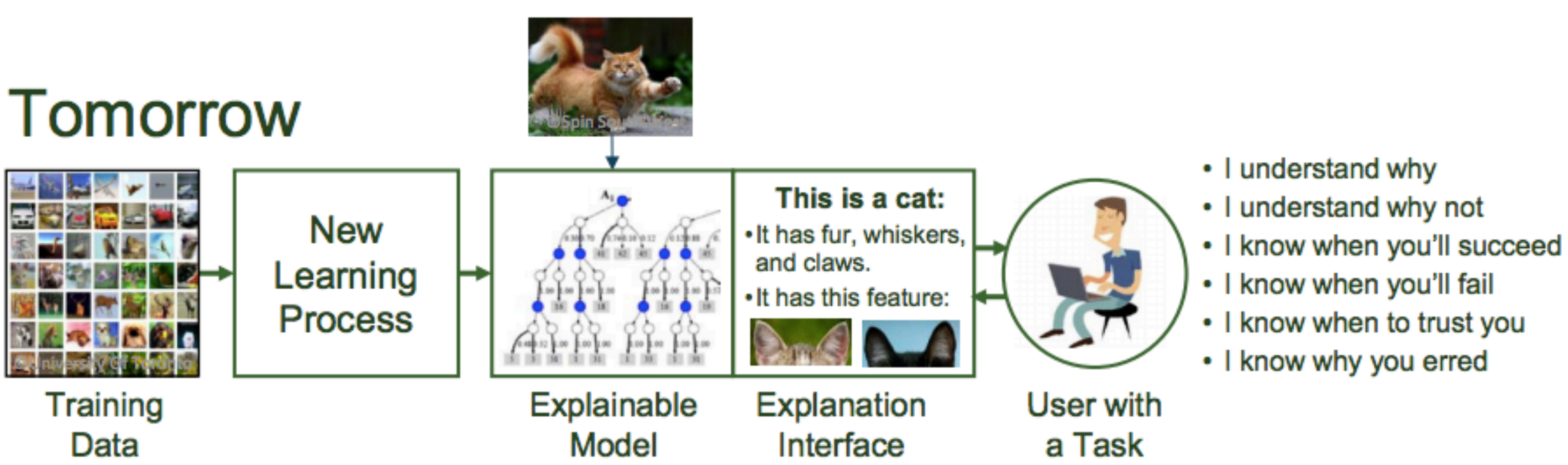


Fewer labels could be sufficient

Figure E.4: Mean of rendering latent variable $s(\ell)$ at each layer. We see that for CIFAR-10 train and test sets, the mean latents are almost the same. The mean latents for SVHN show conspicuous difference from those of CIFAR-10.

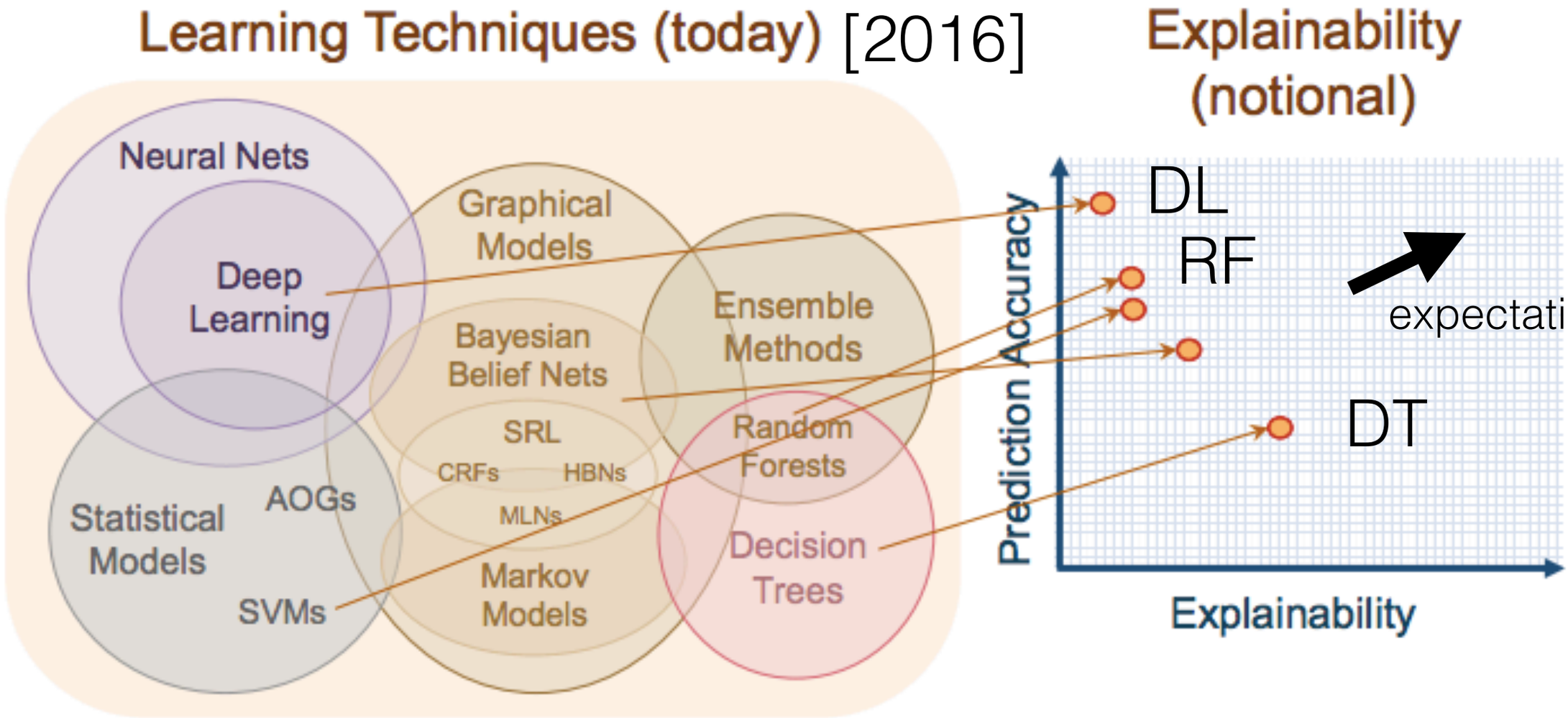
Interpretability





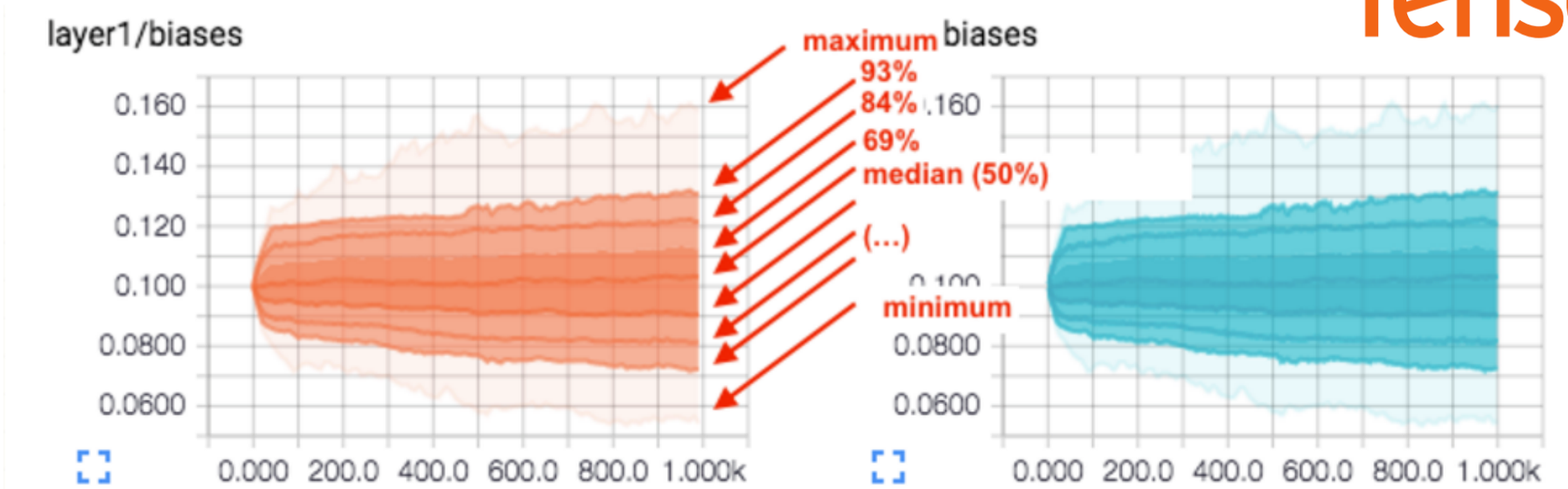
David Gunning (DARPA/I2O)

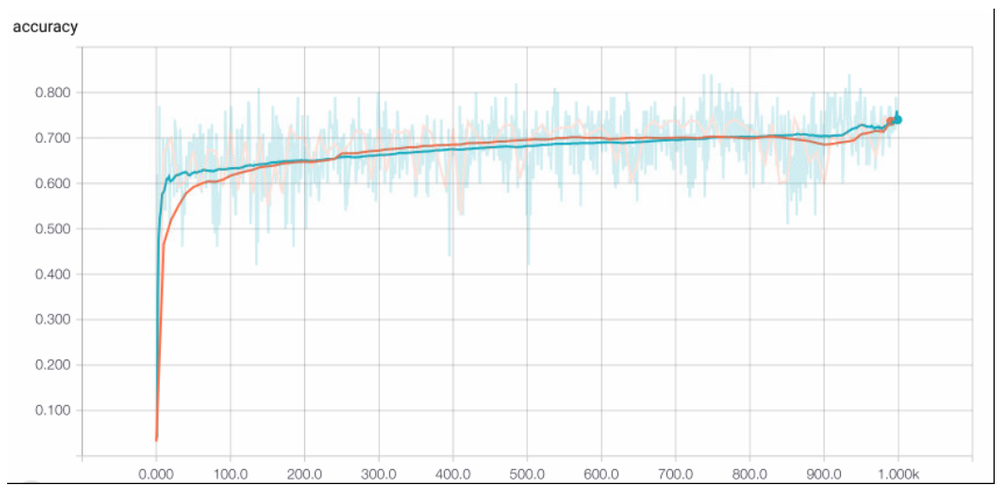
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf



David Gunning (DARPA/I2O)

Distribution Summaries





Percentile distributions over the data:
max, 93, 84, 69, 50, 31, 16, 7, min

Interactivity

Visualization for interpretability

A. Activation Maximization

https://raghakot.github.io/keras-vis/

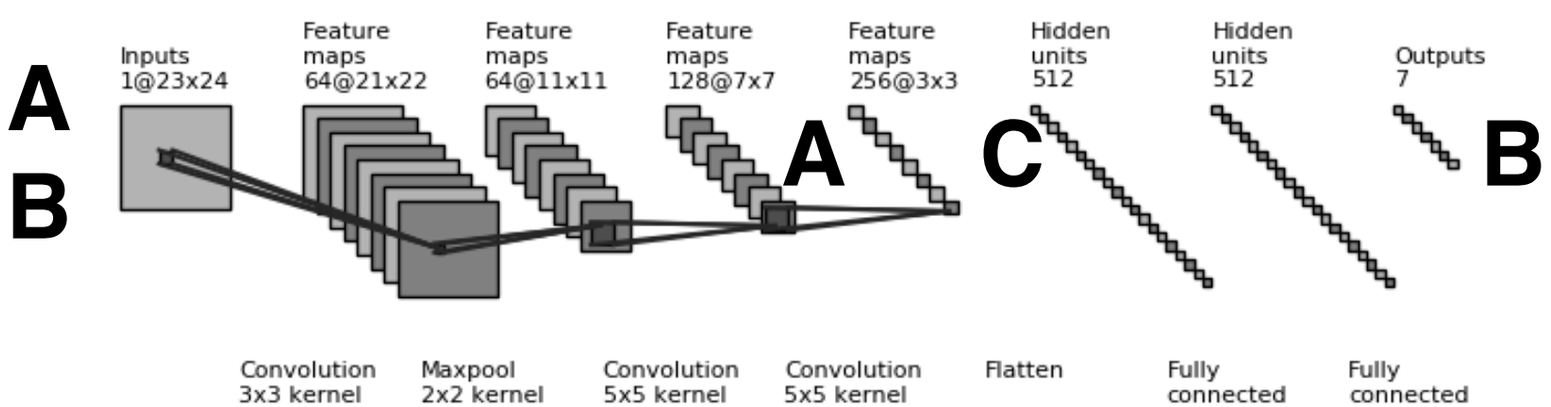
- Initial layer filters easy to visualize
- Generate input image that activates later filters

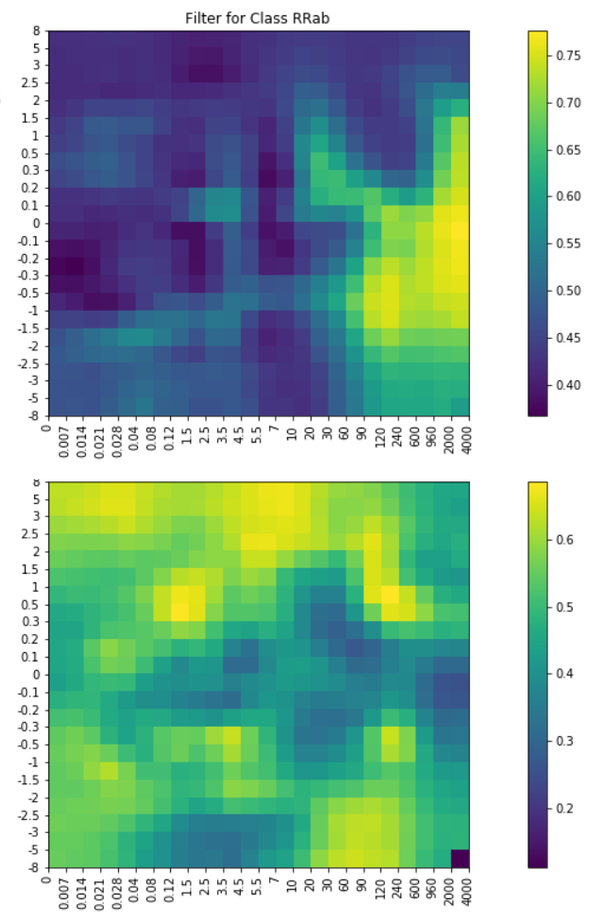
B. Saliency Maps

- Gradient of o/p category wrt input image
- Understanding attention of the classifier

C. Class Activation Maps

- Gradients based on first dense layer
- Spatial information still intact





Astronomy time-series

Towards the Glass Bead Game

This is just a
low-dimensional projection

Classification

CRTS

ZTF

Gaia

Pan-STARRS

LSST

SKA

surveys

TNS

babamul

Antares

braai

ALeRCE

TransiNet

Lasair

Data Science/ML

Science

Holy grail

Sky surveys opening exciting windows
A lot more possible than what is presently done

JPL has established a program focused on building and implementing an institution-wide strategy for data science

- Expanding from archives to enable data analytics as a first class activity
- Methodology transfer across disciplines
- Research partnerships with academia, government, and industry

JPL Data Science
Rich Doyle/Dan Crichton

George Djorgovski

CD3 Caltech

Data Science pilots

Pilot Title	Domain	Lead, Organization
Data-driven Model Adaptation: From Theory to Application	Robotics and	Ali Agha, 347
Active Learning and Importance Sampling Applied to Monte	Science	Wayne Chi, 397
Unpacking the black box of Machine Learning for	Mission Operations	S. Davidoff, 397
Self-improving hybrid retrieval schemes that learn from long-	Science	Anthony Davis, 329
Automatic Per-Pixel Classification of UAVSAR Imagery	Mission Operations	Michael Denbina, 334
Teaching Machines the Way of the CMB Toward efficient	Science	Olivier Dore, 326
The Big Climate Data Pipeline (BCDP): a data processing	Mission Operations	Alex Goodman, 398
AutoML for Microwave Instrument Science	Engineering	Tanvir Islam, 386
Accelerating The Efficiency Of Galaxy Formation	Science	Jeff Jewell, 398
GFO Data Analytics	Mission Operations	Lukas Mandrake, 398
Automatic Image Captioning and Annotation Capability for the	Science	Chris Mattmann, 170
Automatic AI-based Software Vulnerability & Risk Extractions	Engineering	Michael Pajevski, 394
Diagnosing Failures in Scheduling using Visualizations for	Mission Operations	Emine Basak Alper
Infusion of Astronomical Source Vetting and Variable Star	Science	U. Rebbapragada,
Speeding up InSAR Unwrapper Using Convolutional Neural	Science	Gian Franco Sacco, 398
Enhancing NASA Data Applications in High Societal Impact	Science	Hui Su, 329
Mission-Ready Prototype of an Advanced Bayesian Level-2 for	Science	David Thompson, 382
Framework for Multi-Mission Rock Detection Pipeline	Mission Operations	Marshall Trautman, 397
Machine Learning to Understand Cloud Processes across	Science	Qing Yue, 329

2 FY19 DS projects

Mission Operations and Engineering

Operational Recommendations for Capturing History and Infusing Data Science (ORCHIDS) – Jack Lightholder, 39

Science

Develop automated multi-scale CH_4/CO_2 event/anomaly detection and classification – Riley Duren, 8X

Upcoming events

- NASA AI and Data Science Workshop – **March 24-26, 2020** at Caltech
- 2nd Planetary Informatics and Data Analytics Conference – **June 2020** at ESA Astronomy Center, Madrid, Spain

Summary

Nature of Astronomy (and other sciences) changing

Data complexity and not just volume is a challenge

Data driven science has already emerged

Extreme caution required when using canned solutions

(but outlook is positive)

Pertinent questions for this meeting:

- Taking advantage of what exists
- Driving towards use cases
- Identifying datasets that exist
- Transfer learning and data fusion