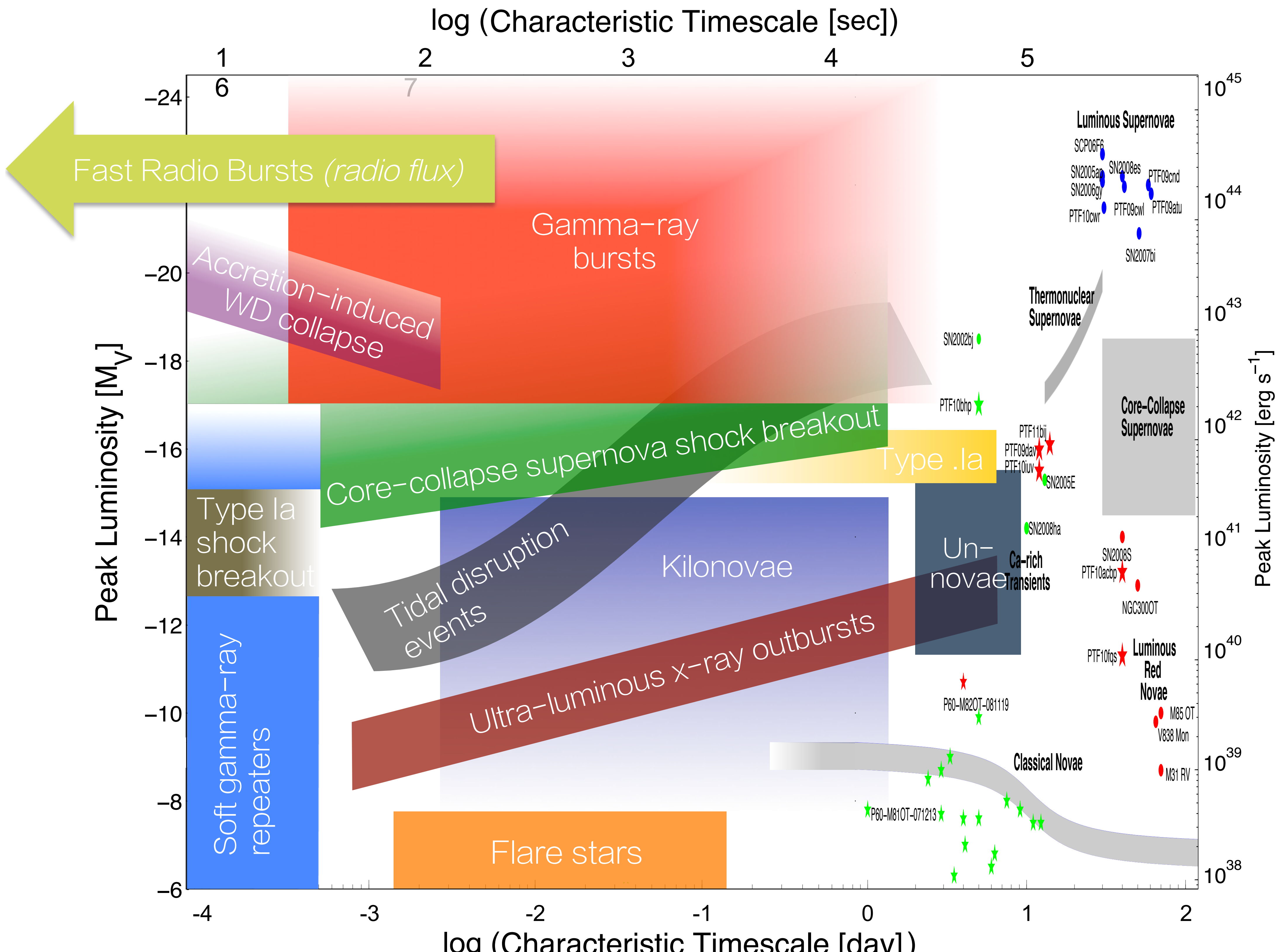
Deep Learning for classification in Astronomy and Biomedicine

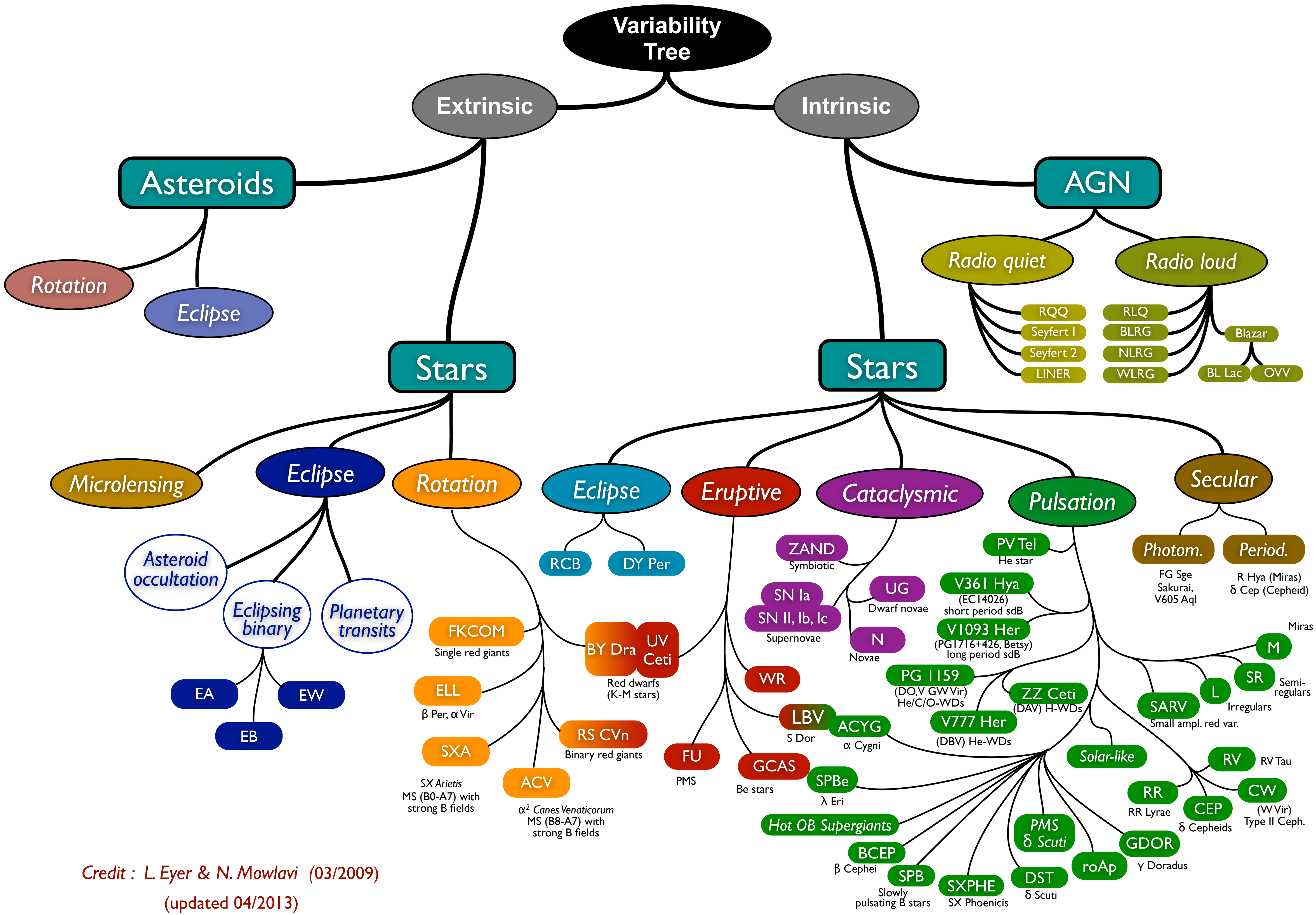
Ashish Mahabal, aam at astro.caltech.edu
Center for Data-Driven Discovery, Caltech
Astroinformatics, Caltech Pasadena 2019-06-25



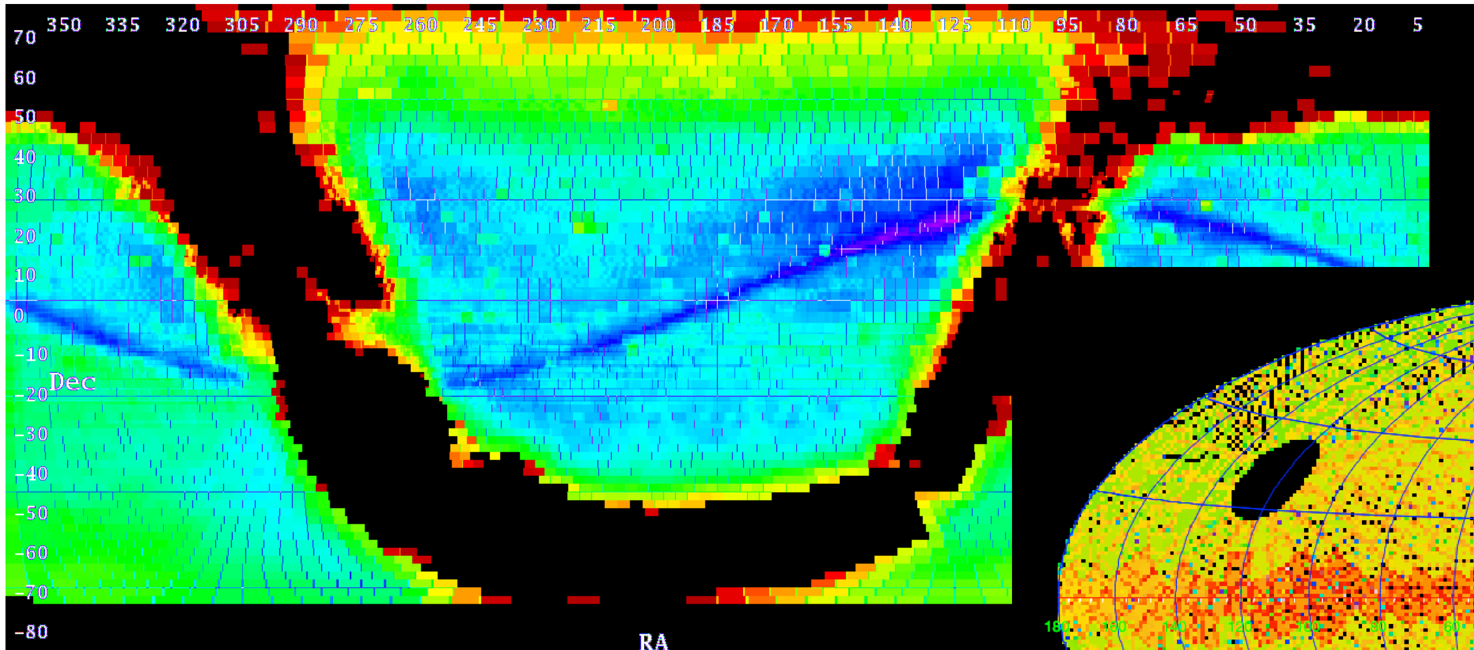
log (Characteristic Timescale [sec])
1
2
3
4
5
6
7
Fast Radio Bursts *(radio flux)*
Gamma-ray bursts
Accretion-induced WD collapse
Core-collapse supernova shock breakout
Type .Ia
Type Ia shock breakout
Tidal disruption events
Kilonovae
Un-novae
Ultra-luminous x-ray outbursts
Soft gamma-ray repeaters
Flare stars
Luminous Supernovae
SCP06F6
SN2005ap
SN2006gy
SN2008es
PTF09cnd
PTF09atu
PTF10cwr
PTF09cwl
SN2007bi
Thermonuclear Supernovae
SN2002bj
PTF10bhp
PTF11bij
PTF09dav
PTF10iuv
SN2005E
Core-Collapse Supernovae
SN2008ha
Ca-rich Transients
SN2008S
PTF10acbp
NGC300OT
PTF10fqs
Luminous Red Novae
P60-M82OT-081119
M85 OT
V838 Mon
Classical Novae
M31 RV
P60-M81OT-071213
Peak Luminosity [M_V]
−24
−20
−18
−16
−14
−12
−10
−8
−6
Peak Luminosity [erg s^{-1}]
10^{45}
10^{44}
10^{43}
10^{42}
10^{41}
10^{40}
10^{39}
10^{38}
-4
-3
-2
-1
0
1
2
log (Characteristic Timescale [day])

J Cooke

Variability tree: Many nodes have further subdivisions

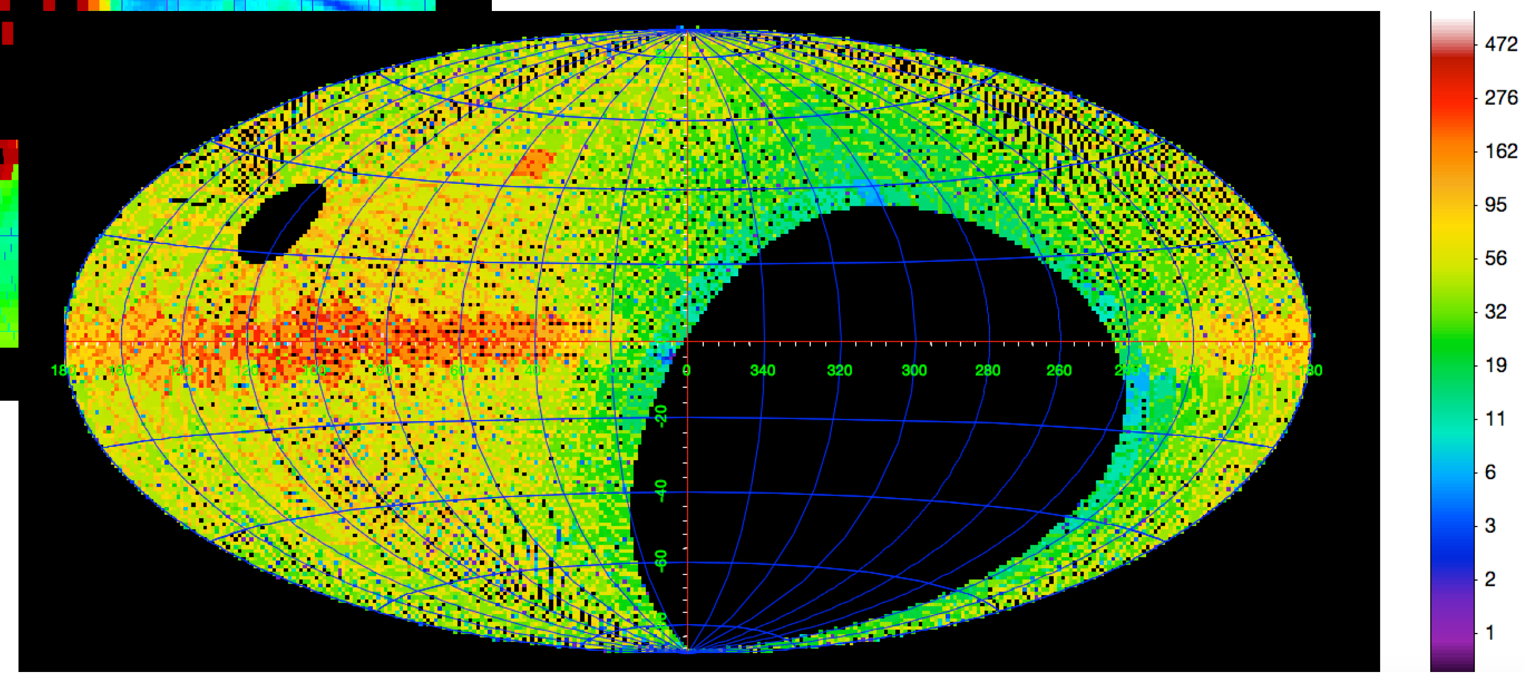


From snapshots to (slow) movies of the sky

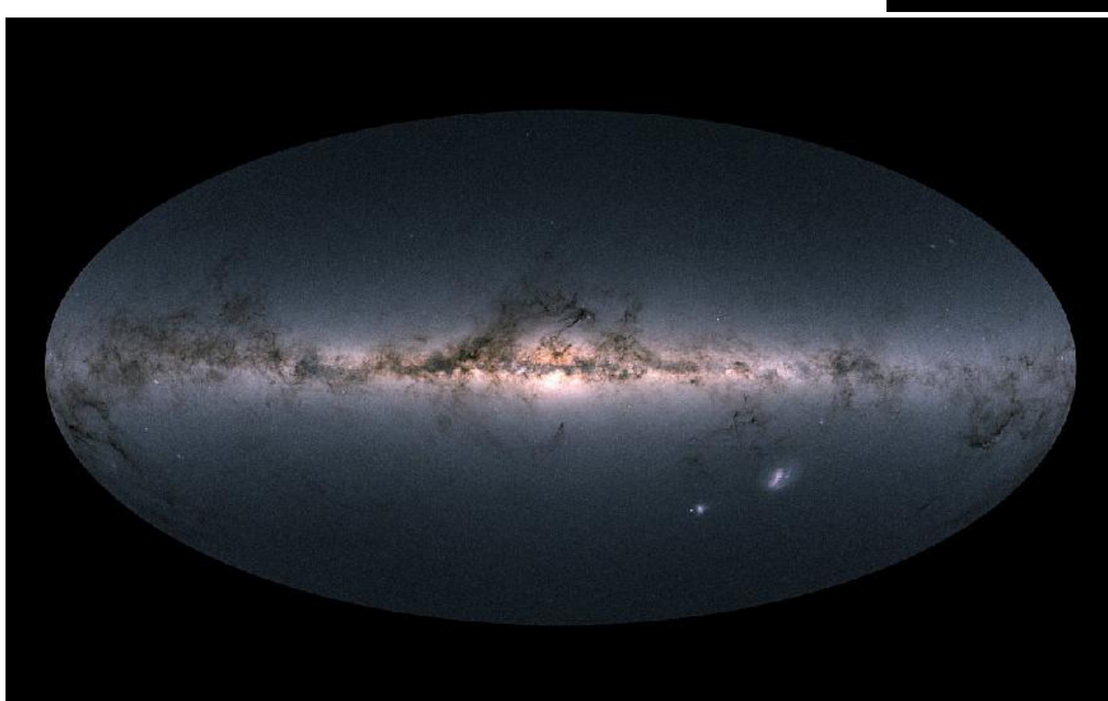


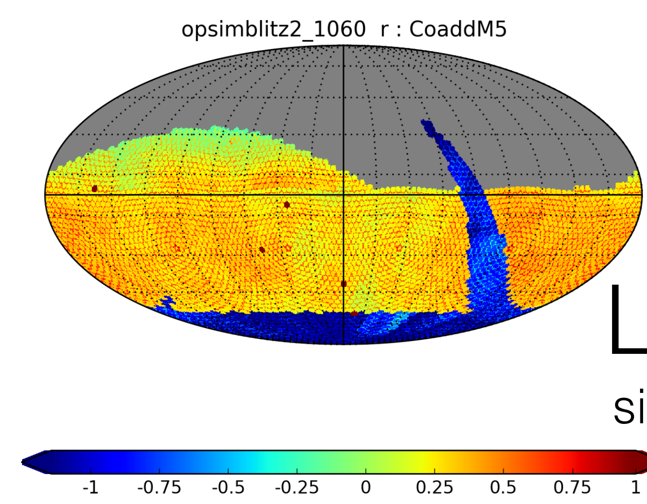
CRTS

ZTF



Gaia





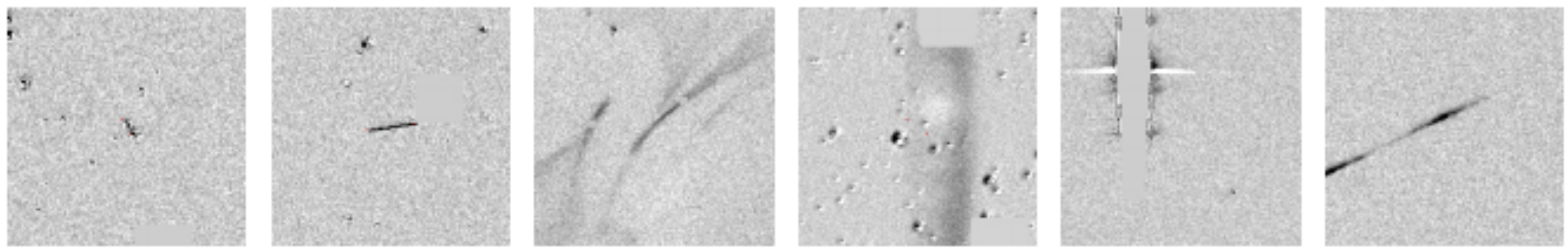
LSST

simulated

Identifying streaking asteroids

DeepStreaks: identifying FMOs in ZTF data 5

(See Dima's talk for details)

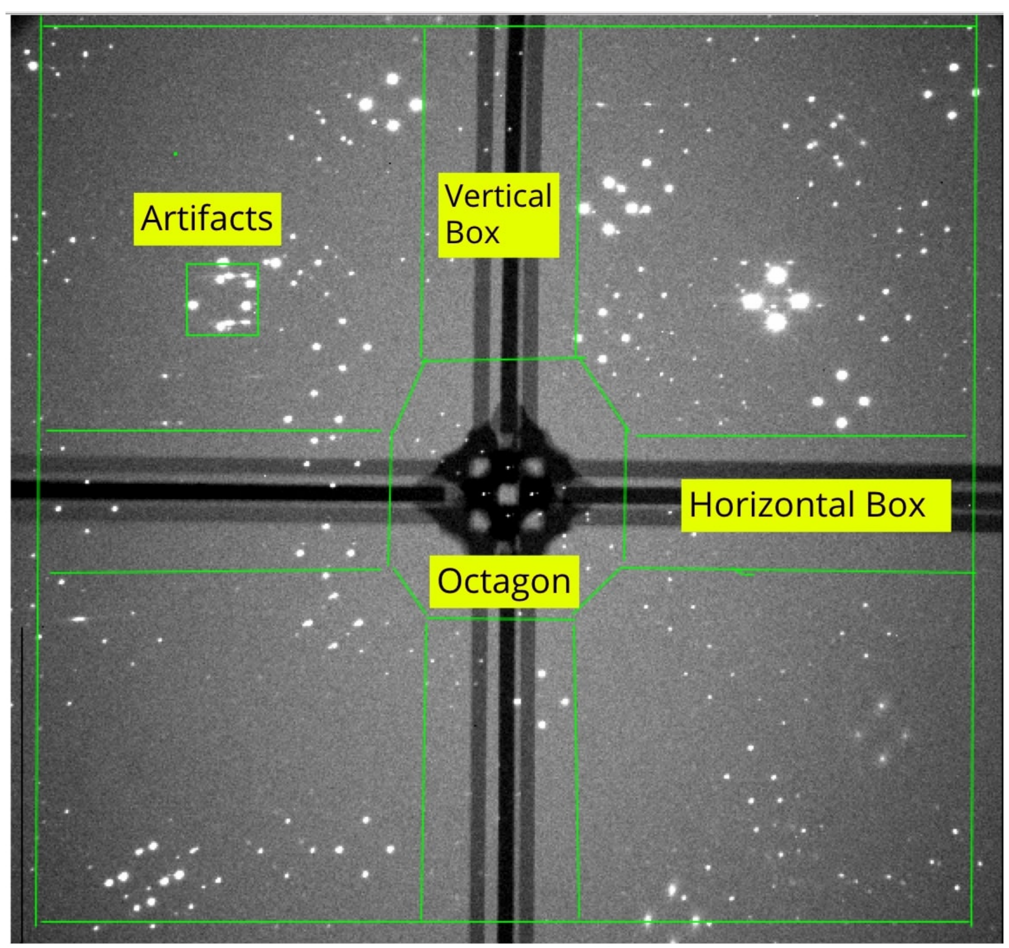


(a) Bad subtraction (b) Cosmic ray (c) "Dementor" (d) "Ghost" (e) Masked star (f) Satellite trail

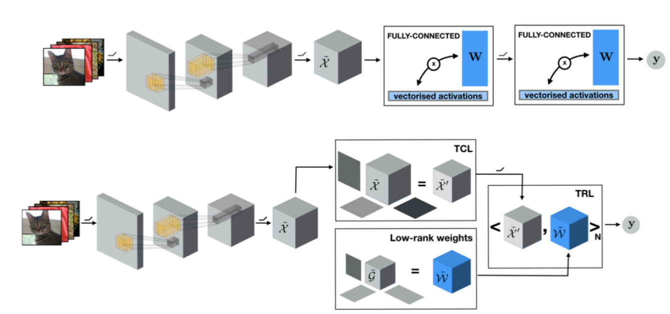
Figure 4. Examples of different classes of bogus streak detections.

Duev, Mahabal, … 2019
arxiv:1904.05920

Robopol



Tensor Regression Networks - a new concept!



Source: Tensor Regression Networks (2017) - J. Kossaifi et al

Dhruv Paranjpye
Gina Panopoulou
Robopol team

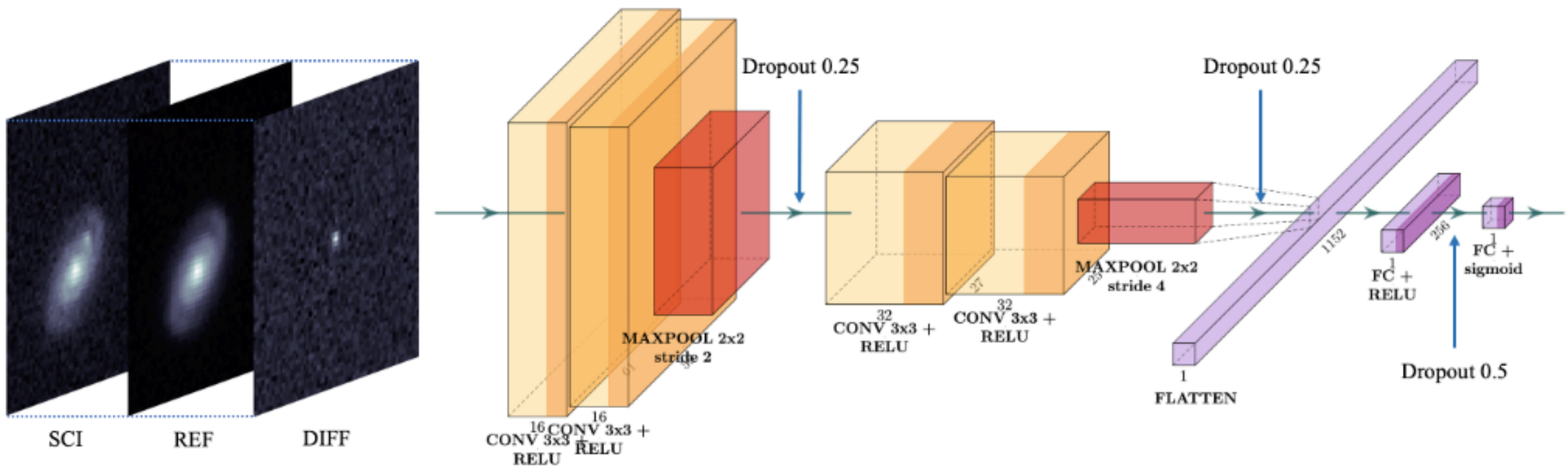
(See Dhruv's poster outside for details)

Extendable to Gattini data (large pixels, bright upper limits)

braai

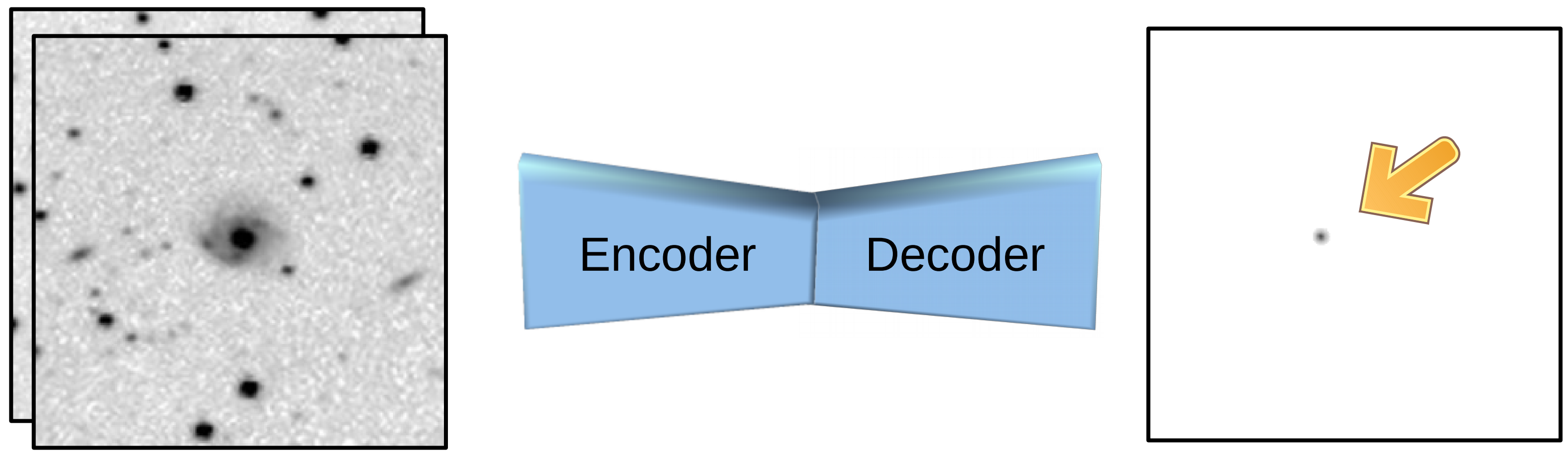
braai architecture

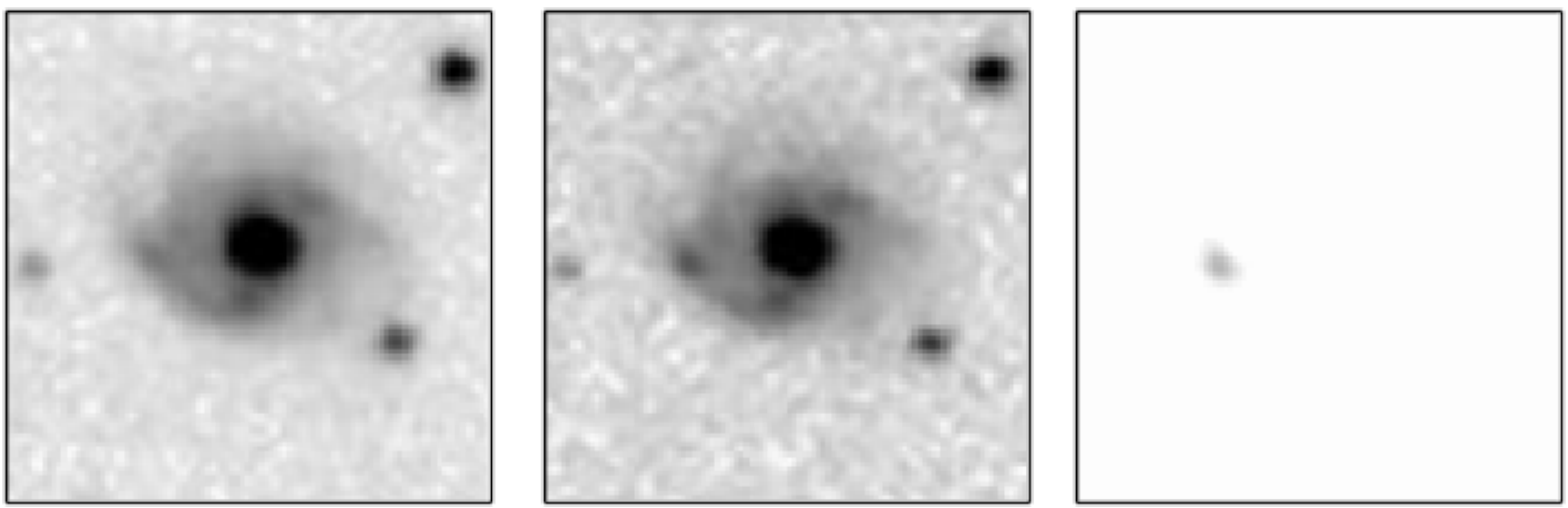
We will use a simple custom VGG-like sequential model ($VGG6$; this architecture was first proposed by the Visual Geometry Group of the Department of Engineering Science, University of Oxford, UK). The model has six layers with trainable parameters: four convolutional and two fully-connected. The first two convolutional layers use 16 3x3 pixel filters each while in the second pair, 32 3x3 pixel filters are used. To prevent over-fitting, a dropout rate of 0.25 is applied after each max-pooling layer and a dropout rate of 0.5 is applied after the second fully-connected layer. ReLU activation functions (Rectified Linear Unit – a function defined as the positive part of its argument) are used for all five hidden trainable layers; a sigmoid activation function is used for the output layer.



(See Dima's talk for details)

Image subtraction for hunting transients without subtraction





Sedaghat and Mahabal, 2017
arXiv:1710.01422

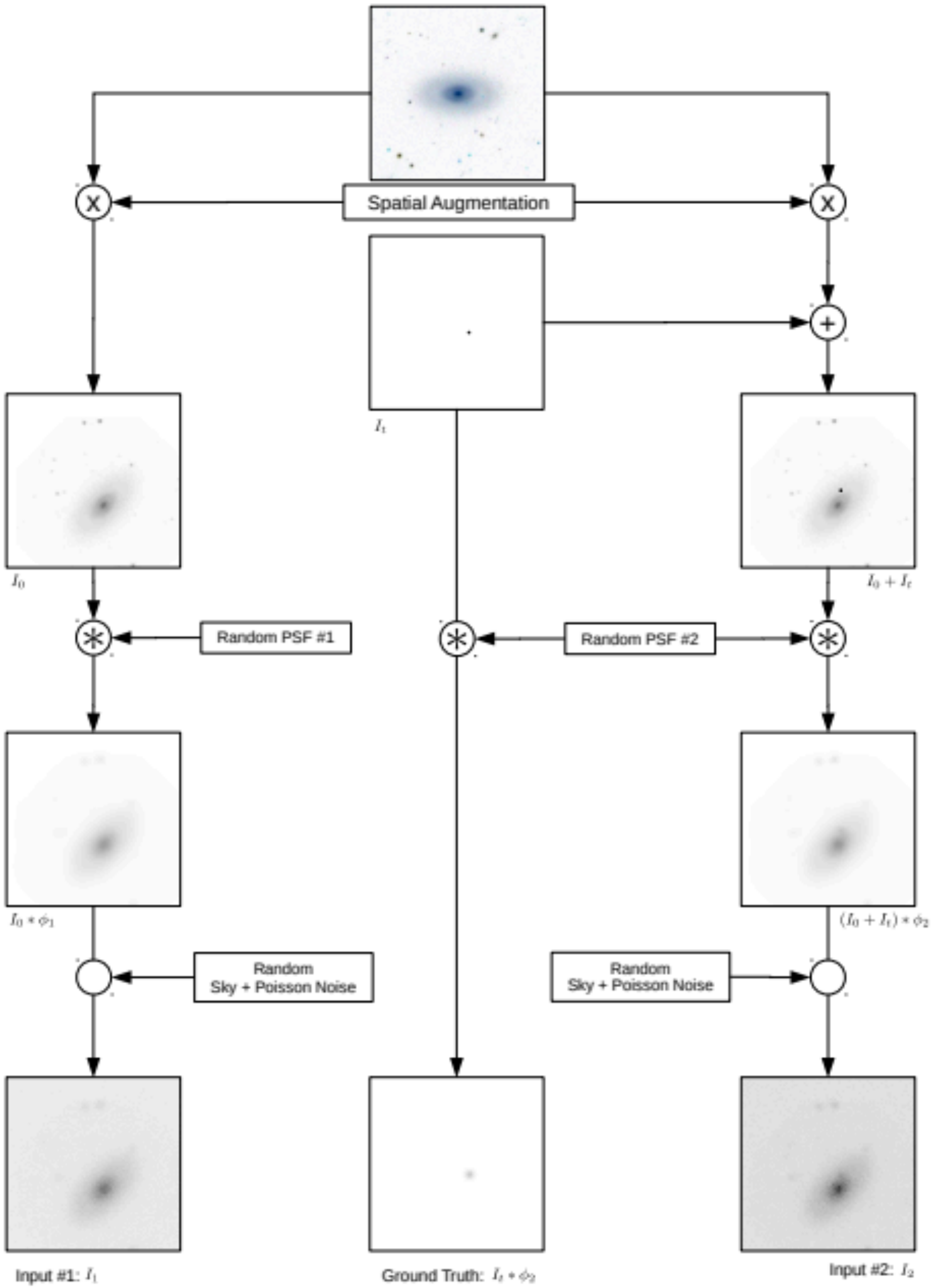
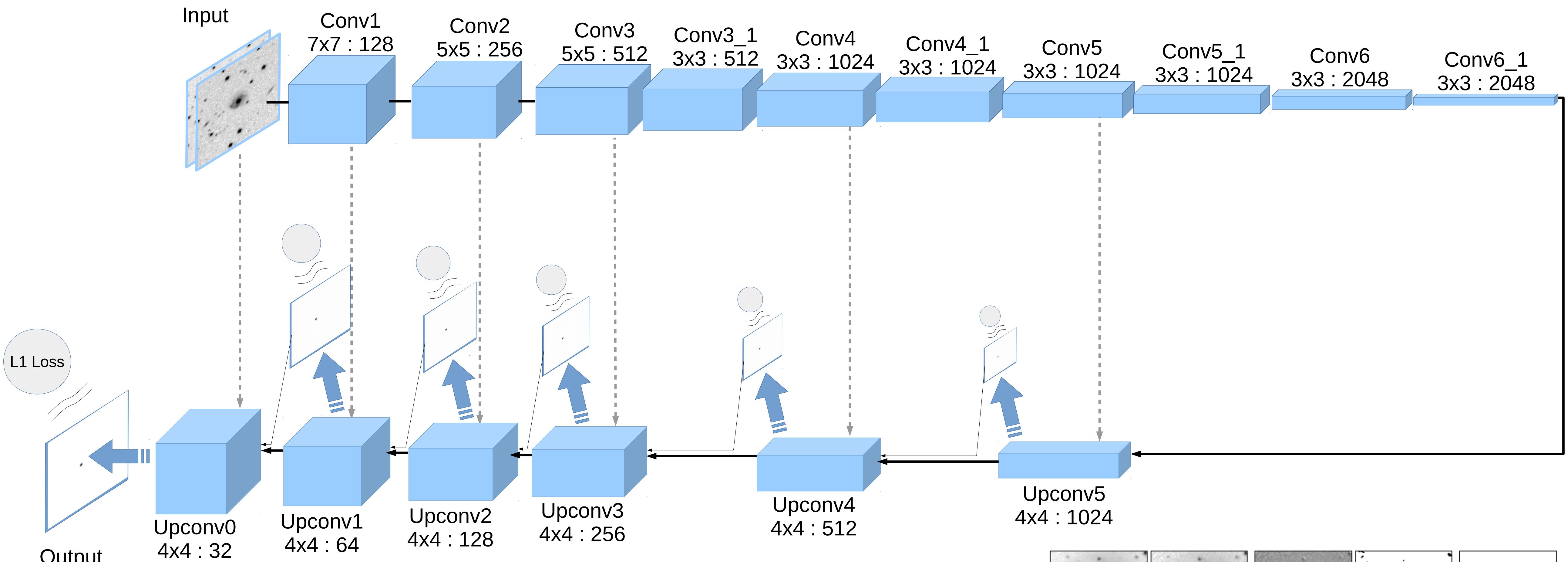
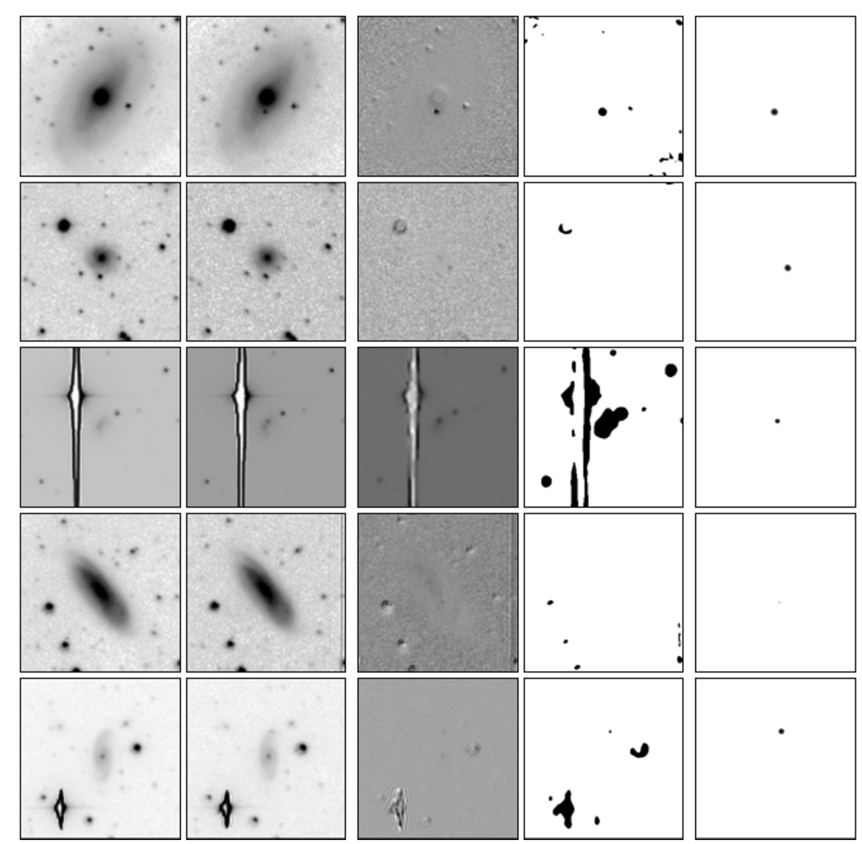


Figure 5. The synthetic sample generation procedure. The notations used here are described in Equations (1) and (2).

Training cycles involving different PSFs

Encoder-decoder network (fully convolutional)





Sedaghat and Mahabal, 2017

arXiv:1710.01422

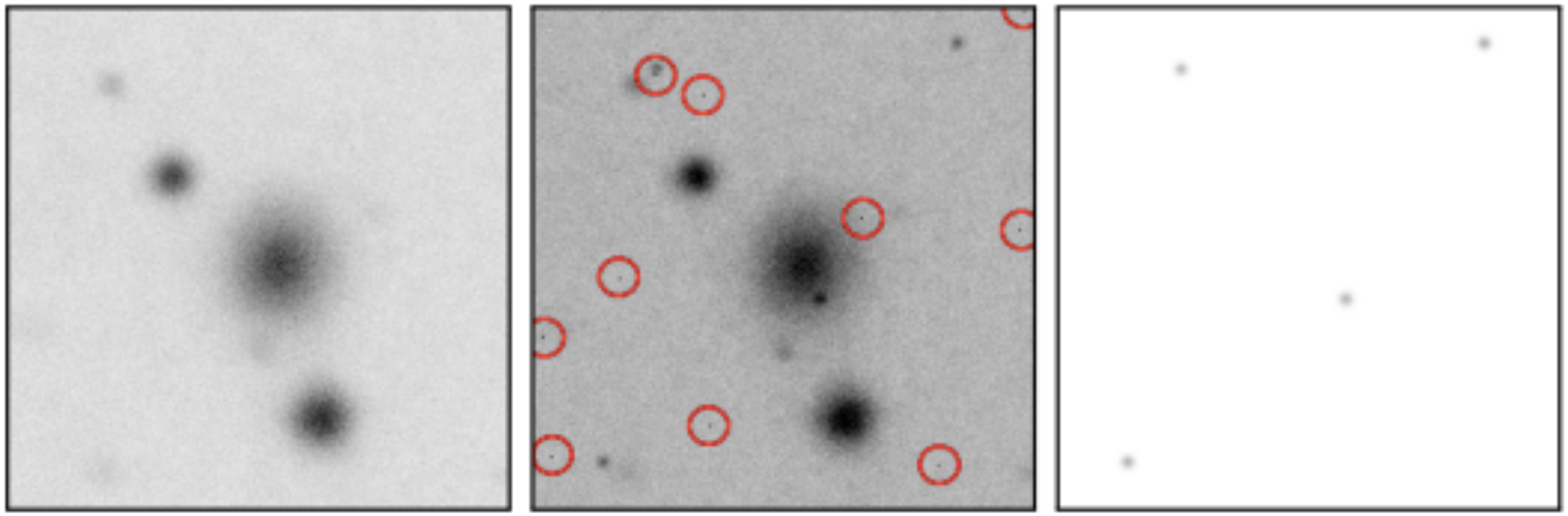
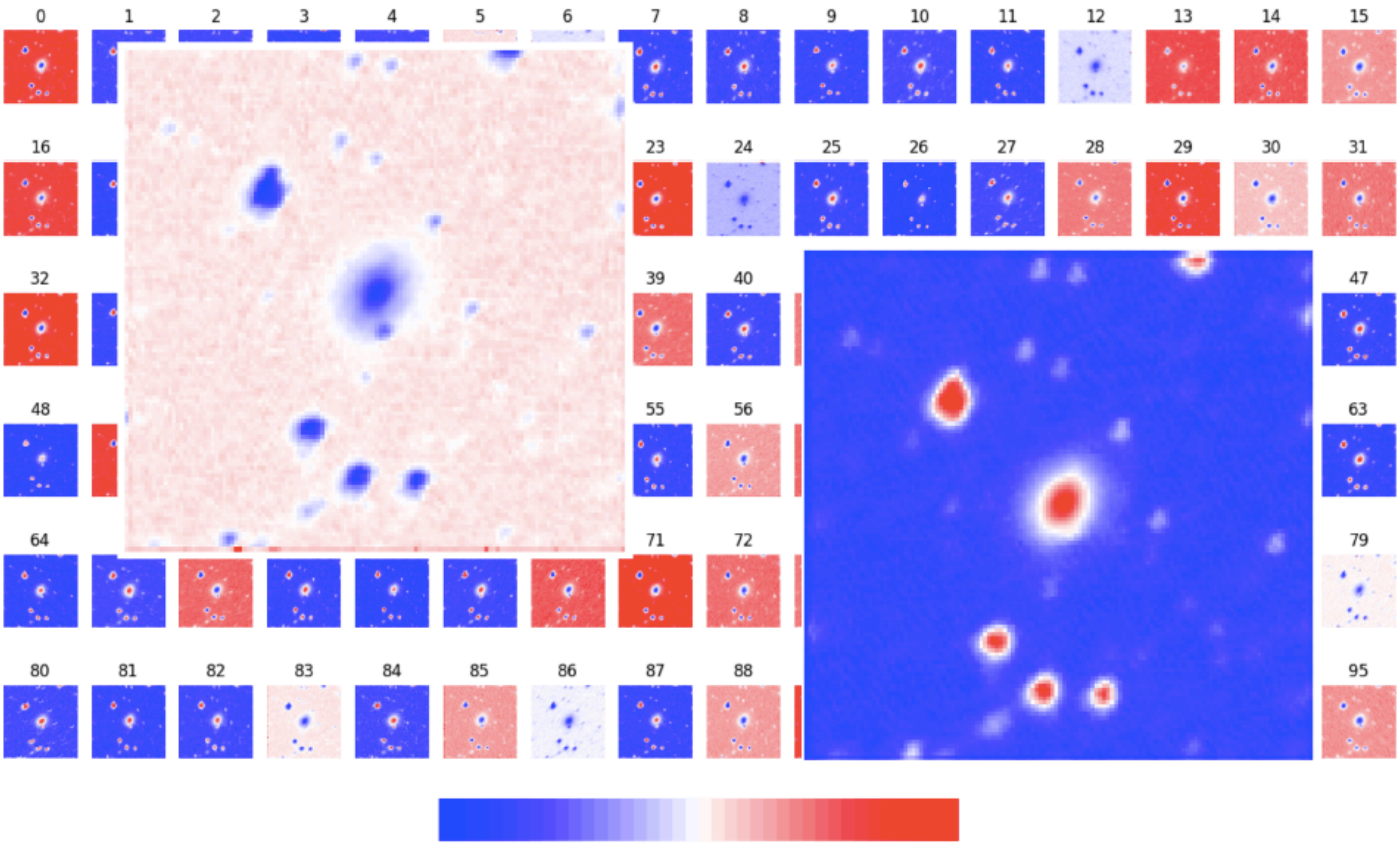


Figure 7. An exemplar multi-transient case from the zoo dataset. The reference image (left), science image (middle) with 10 single-pixel Cosmic Ray events, indicated by red circles, and four transients, and the network prediction (right) with all transients detected cleanly, and all CRs rejected.

reflections, rotations etc. standard techniques made ample use of

Under the Hood – going deeper



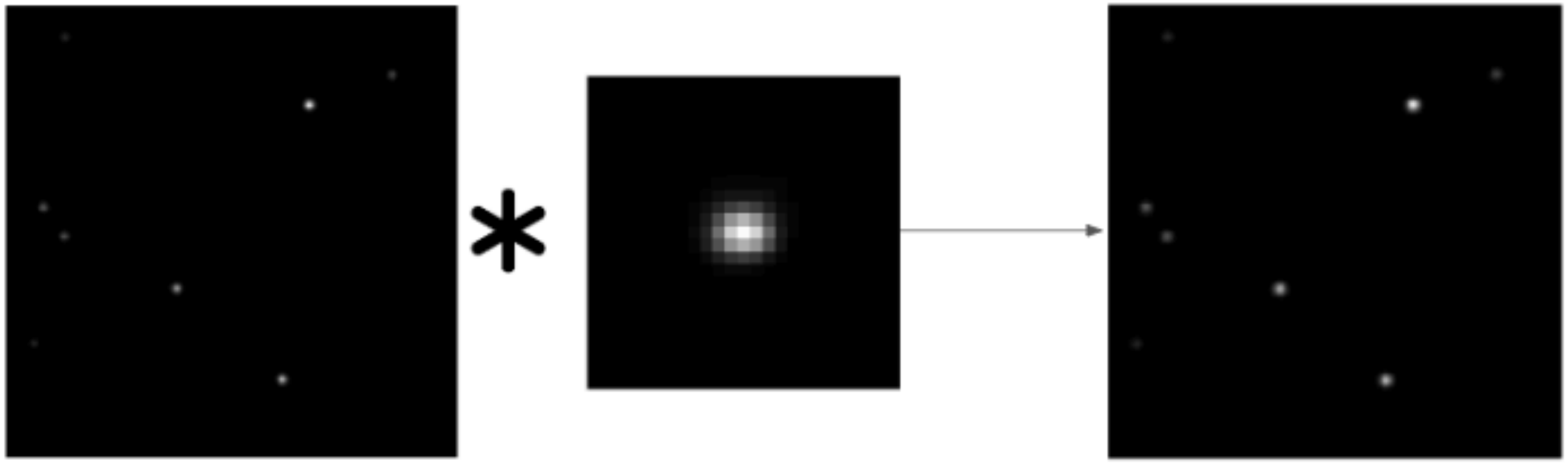
Moving towards ZTFs 3k x 3k images, 0-64K with NaNs, in real-time

Nima Sedaghat, Chaoran Zhang, …

Deconvolution using a encoder-decoder

Data generation from simulated data

Shubhranshu Singh



Ground truth image | PSF | Convolved image

Results on simulated data

Ground Truth | Convolved Image (PSNR = 43.39 dB) | Output of the network (PSNR = 52.97 dB)

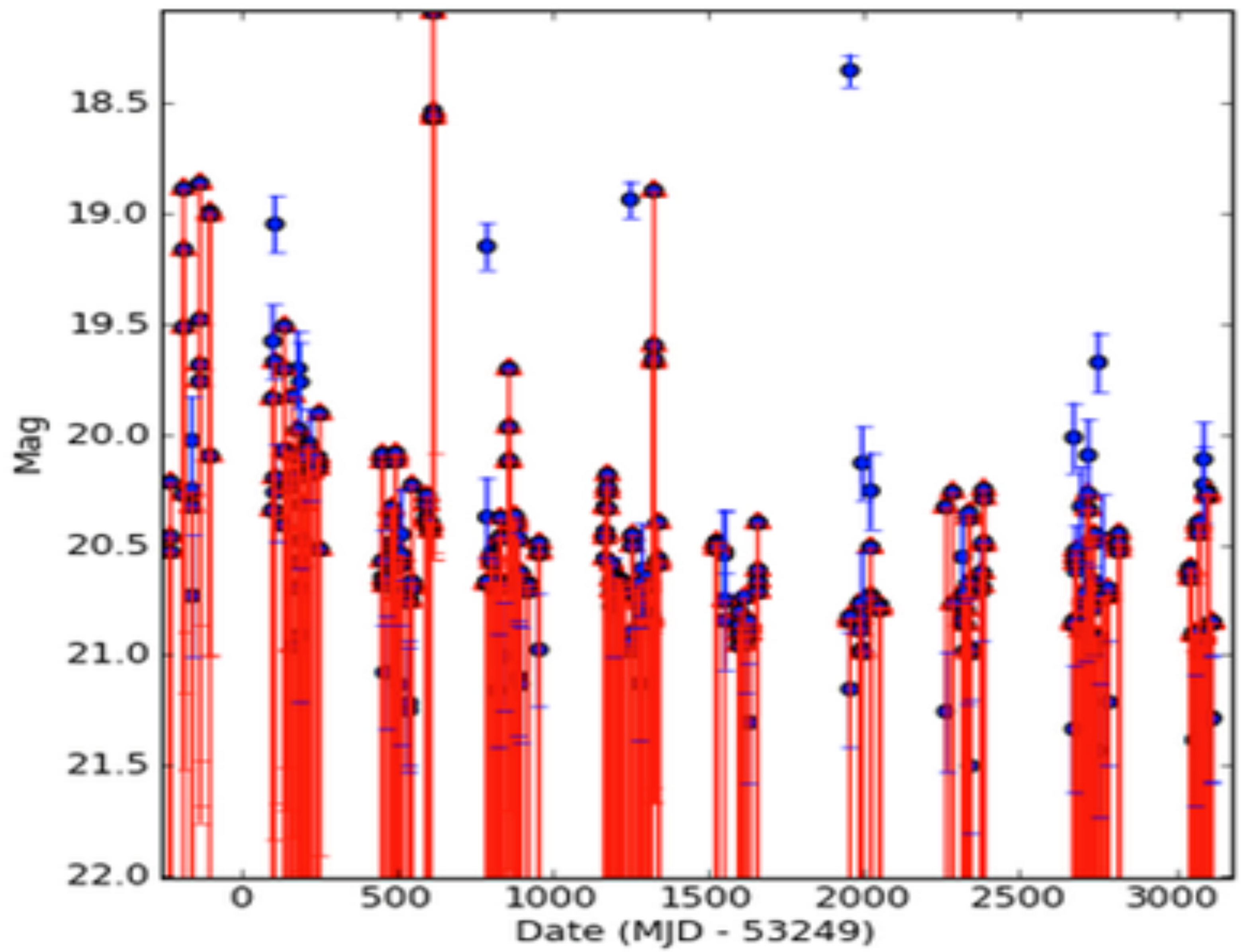
Overwhelming (amounts of) data

CRTS: 500+M light curves over 15 years
ZTF: 1B light curves (just over a year)

LSST, SKA

Properties of light-curves

- Gappy
- Irregular
- Heteroskedastic



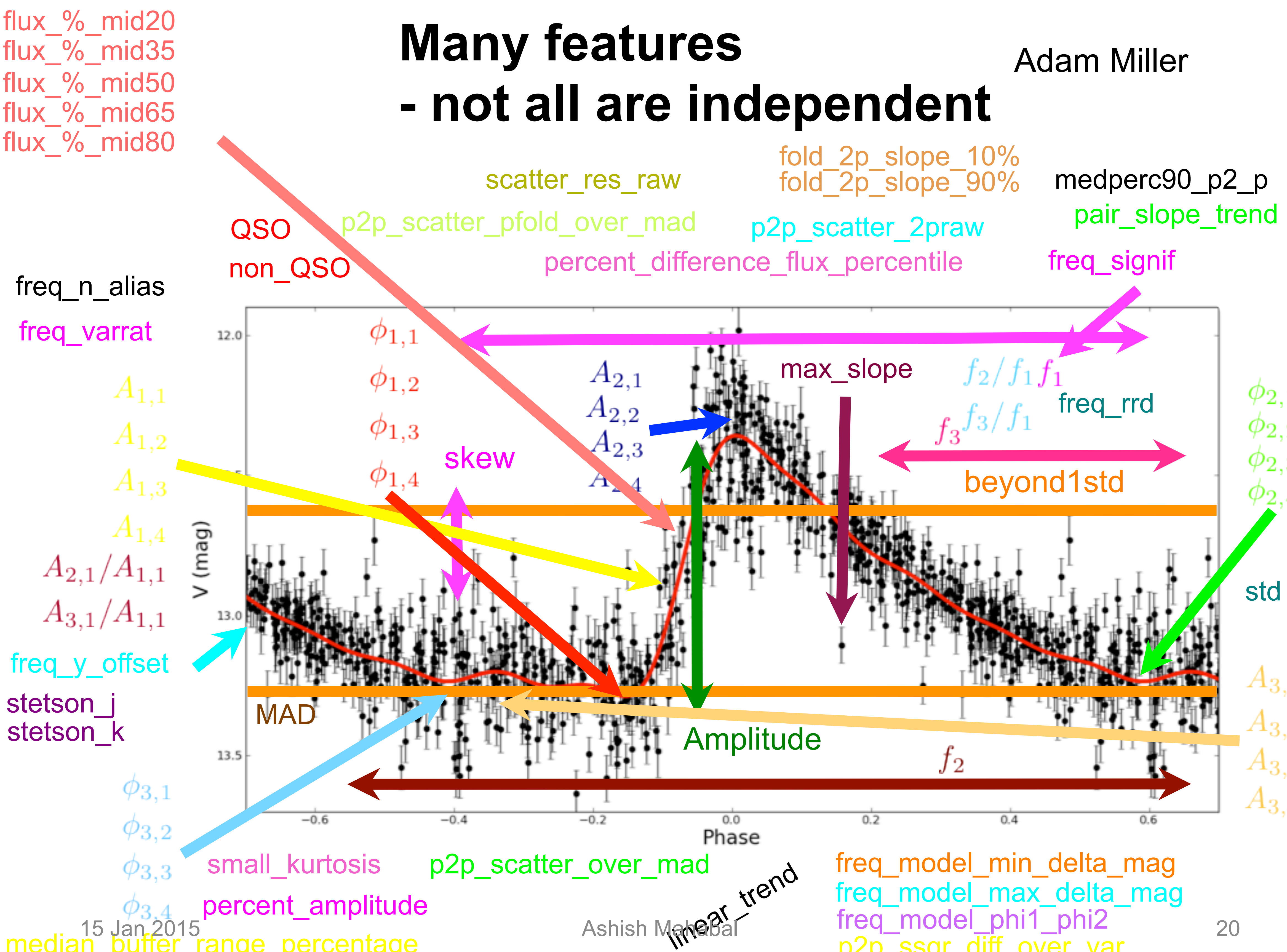
Reasons:

- **expense, rotation/revolution of Earth, moon**
- **science objectives, weather, moon**
- **weather, moon, airmass**

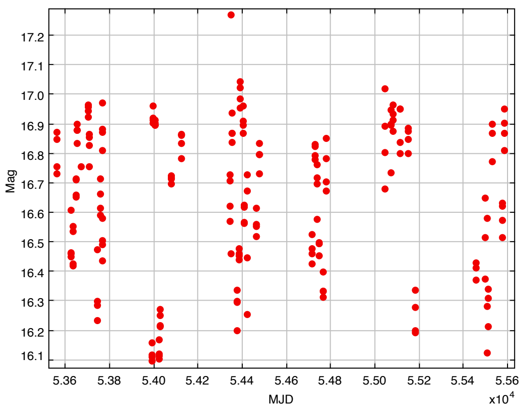
errors ignored by many methods

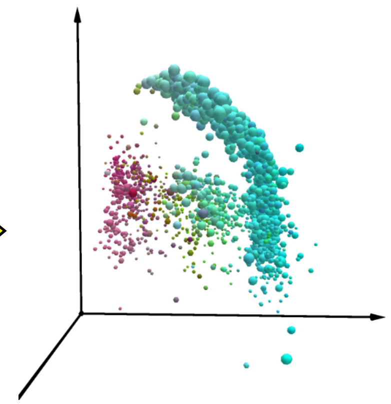
Many features
- not all are independent

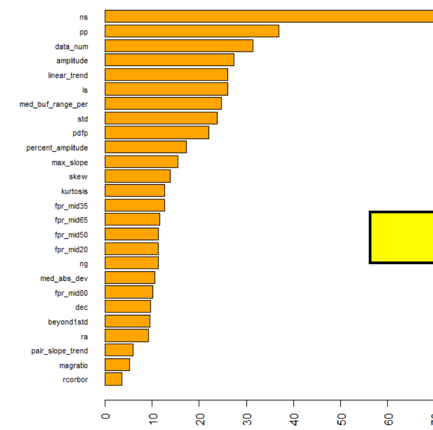
Adam Miller

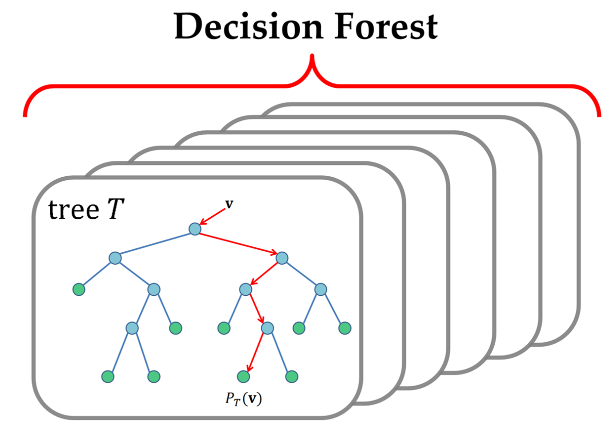


Classification Workflow









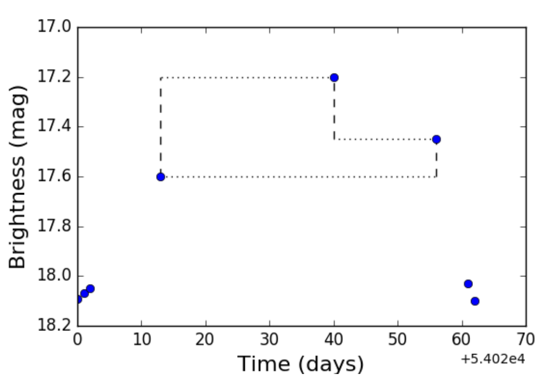
Light curves

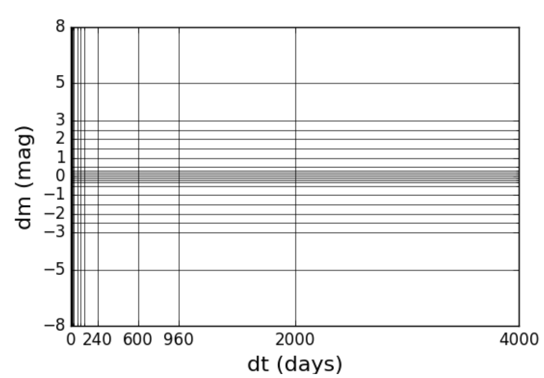
Feature vectors

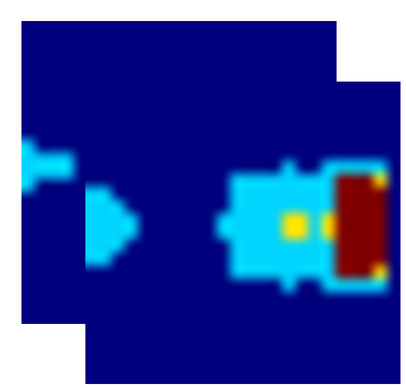
Dimensionality Reduction

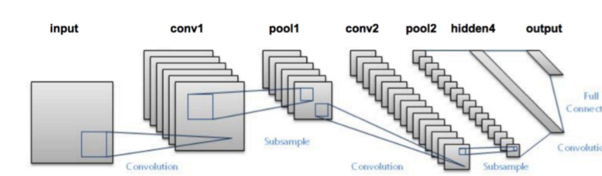
Classification

Domain knowledge/subjectivity







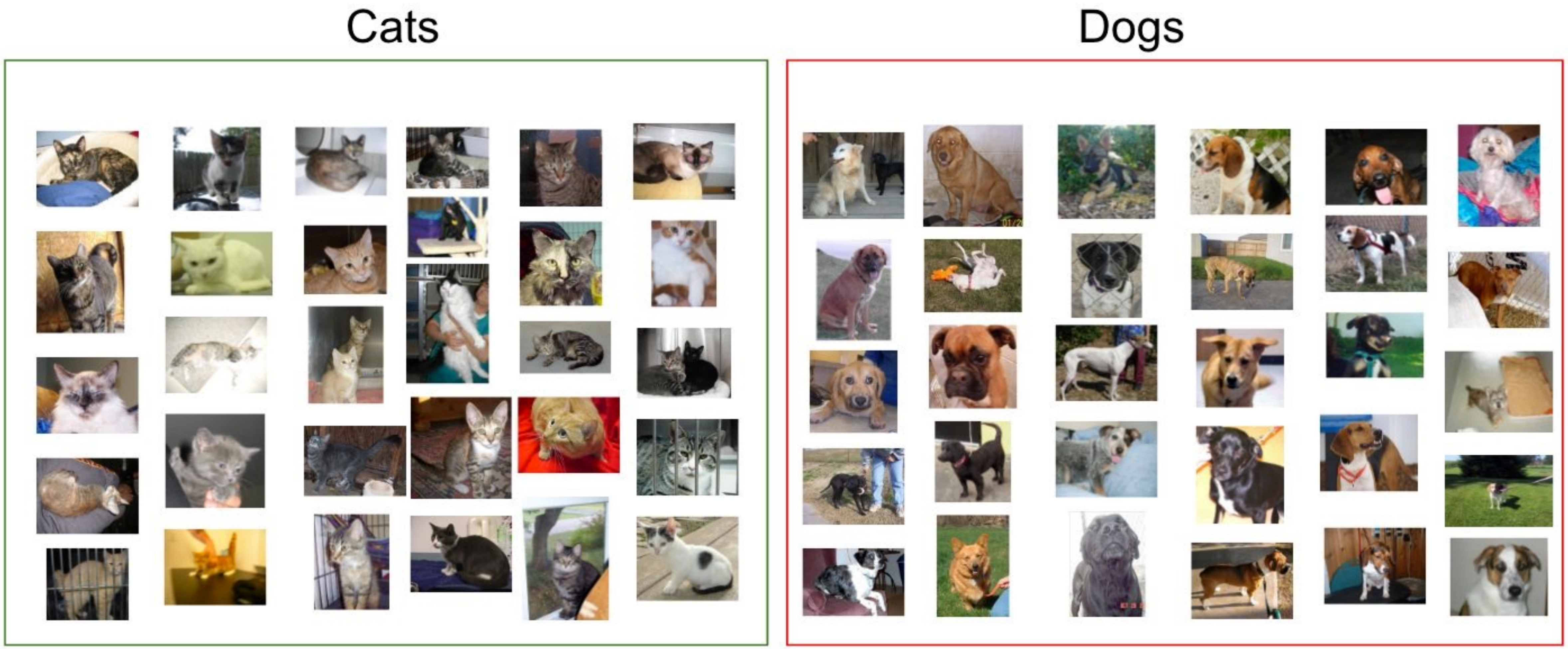


Light curves

Density representation

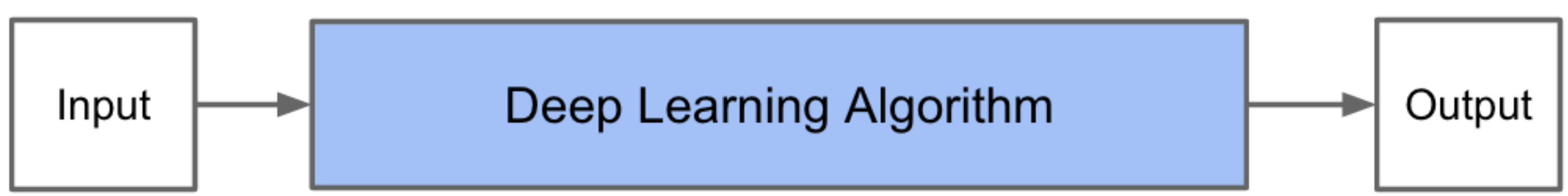
Equi-area images

Convolutional Neural Network



Sample of cats & dogs images from Kaggle Dataset

Traditional Machine Learning Flow



Deep Learning Flow

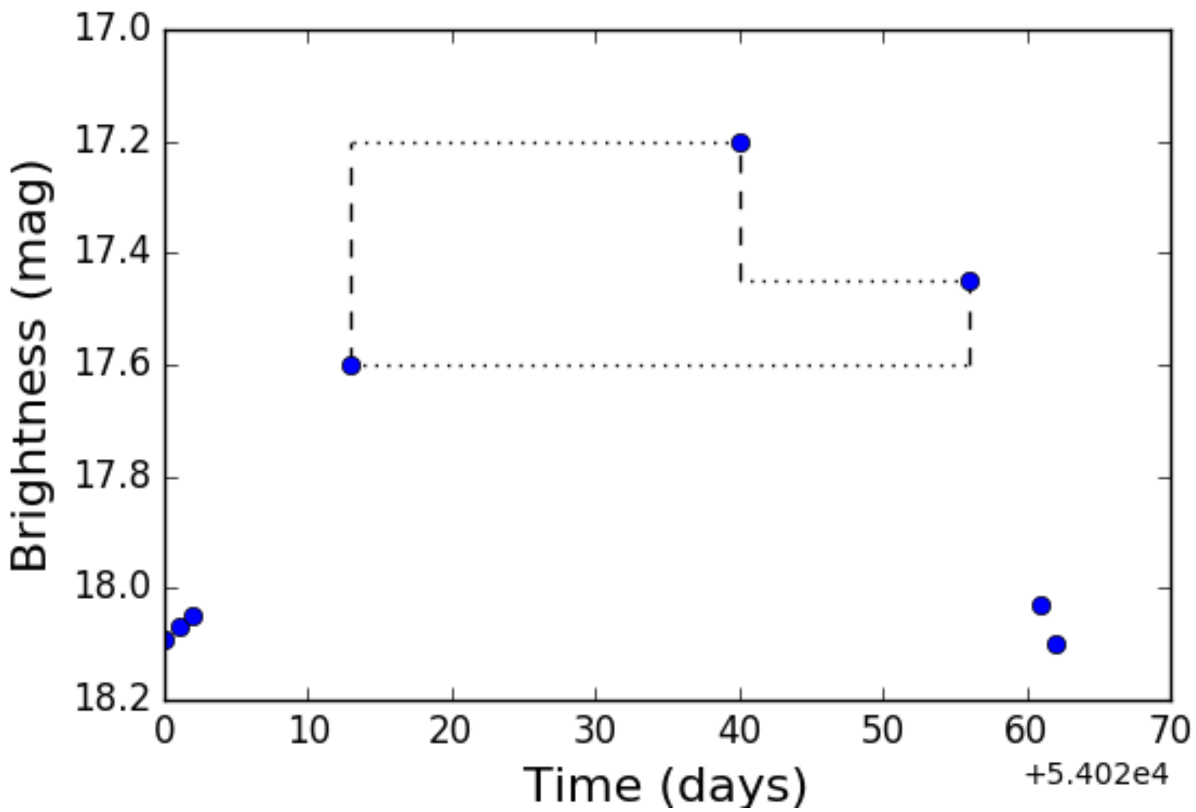
Promise: Works better

Pitfall: Blacker box

Adil Moujahid

(dmdt) Image representation

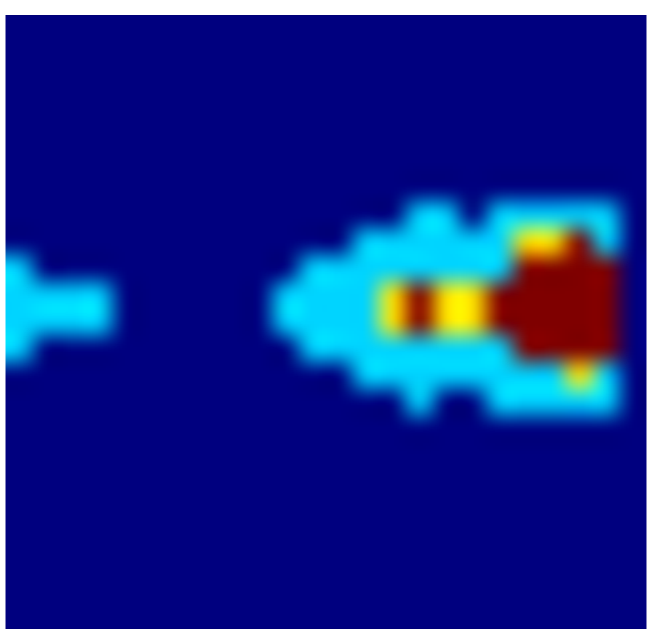
Mahabal, Sheth et al., 2017
1709.06257

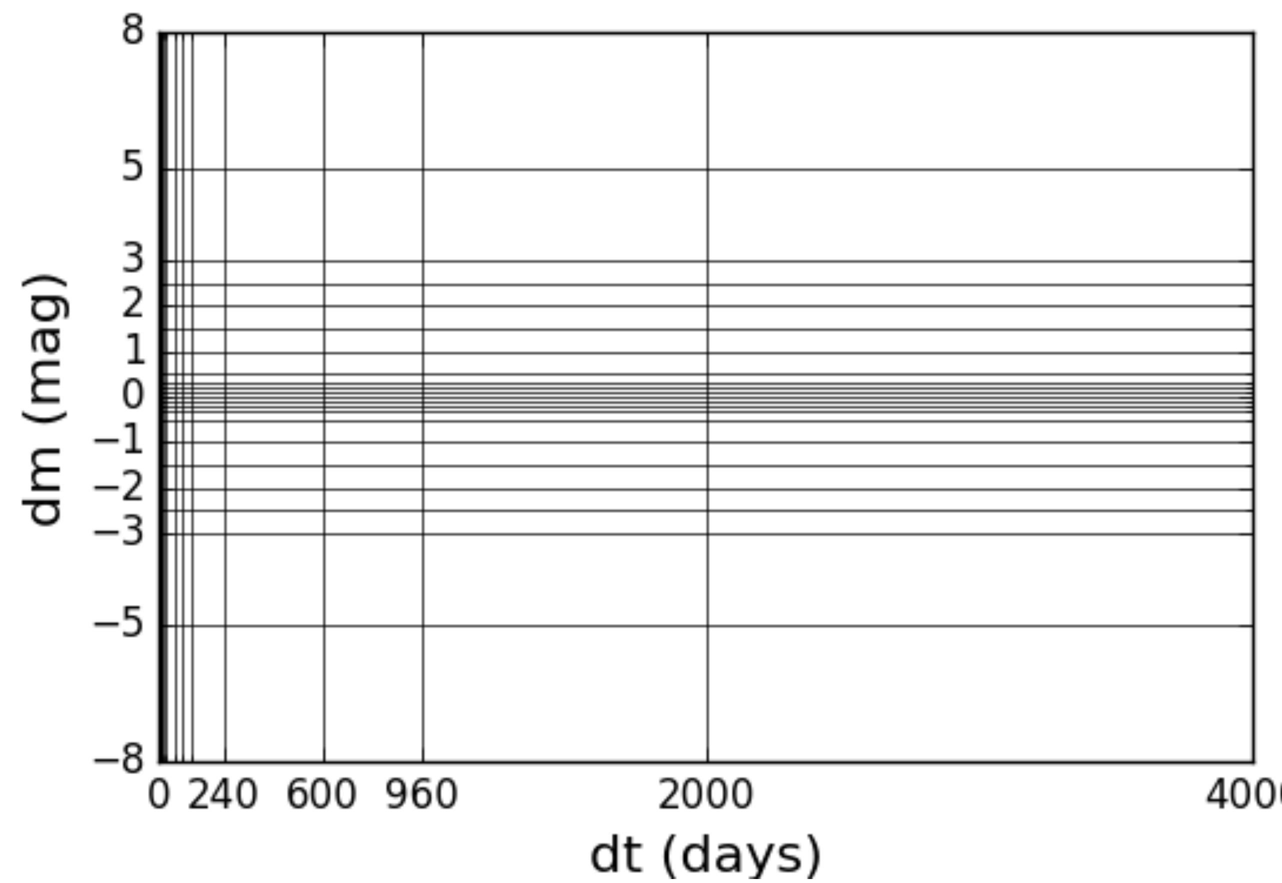


light curve with n points

23 x 24
output grid

n * (n-1)/2 points

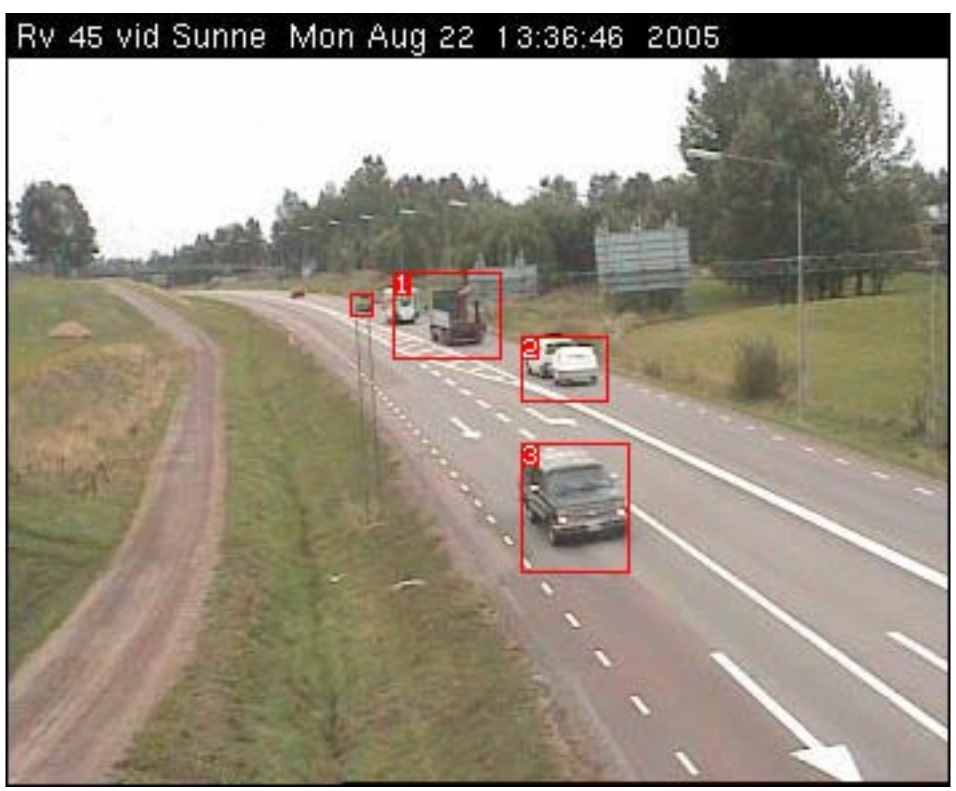




Area equalized pixels

Video Surveillance Analogy

non-convex robust PCA
Netrapalli et al., 2014



Andrew Kirillov

Each class is like a different road
Each individual object has/is perturbations over it

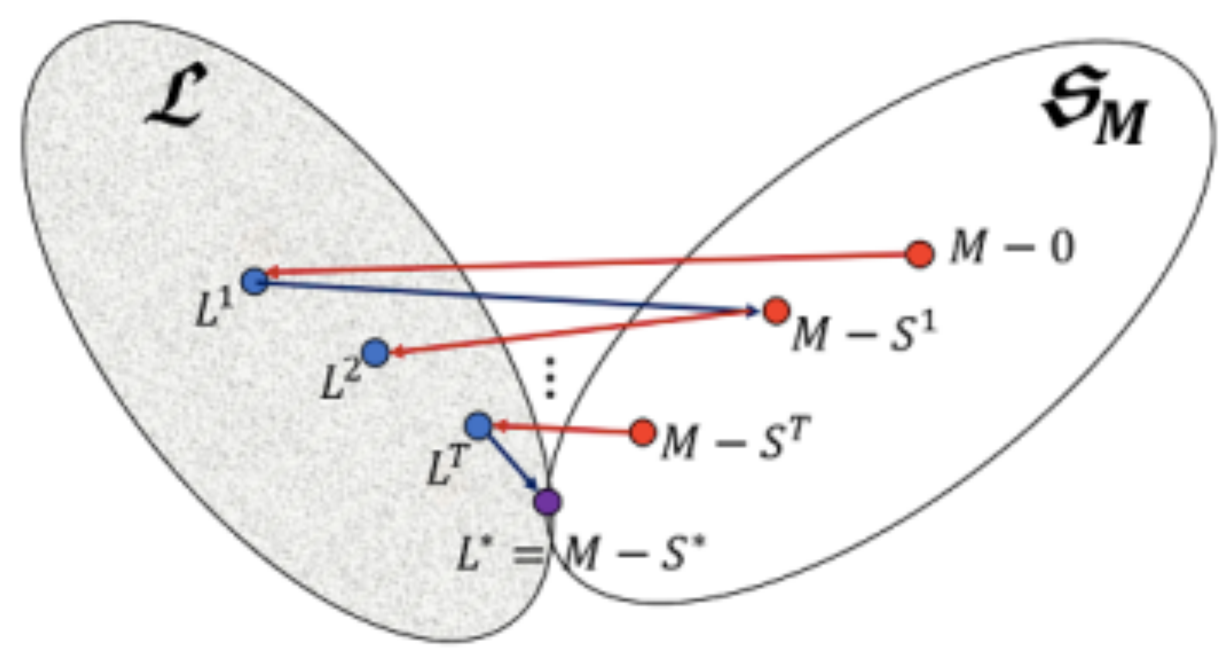
7 classes with at least 500 examples

dmdt–image = $b + ci + s$

- background (survey, cadence)
- class background
- individual object (specific)

$$\underset{L,S}{\mathbf{Min}} \|M - L - S\|_2$$

1. L lies in the set of low-rank matrices,
2. S lies in the set of sparse matrices.



non-convex robust PCA
Netrapalli et al., 2014

Diagnosing LIGO lockloss using auxiliary channels

Motivation

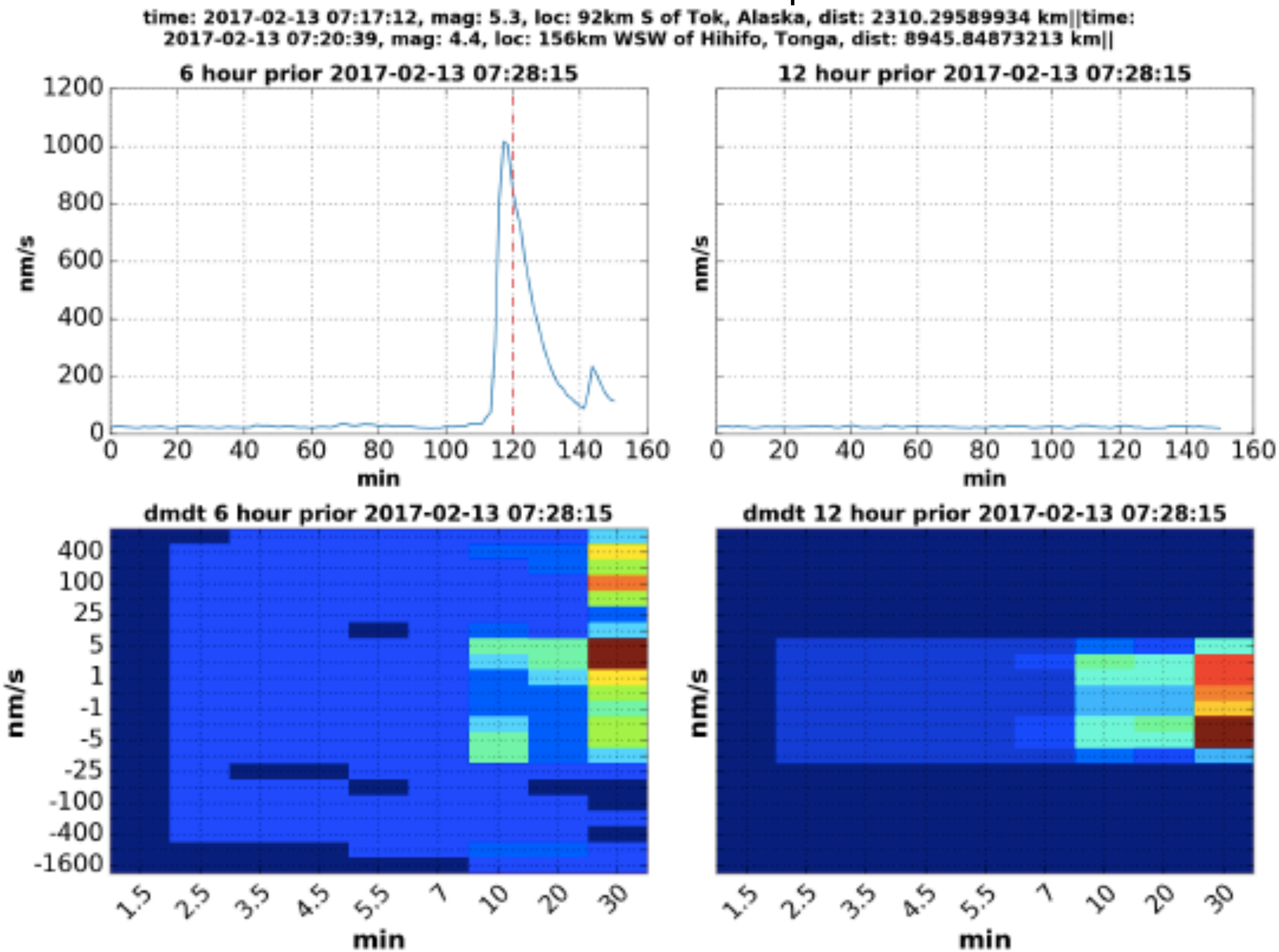
- Lockloss events due to environmental events lead to loss of observation time
- Monitor and diagnose lockloss events as they occur

Goals

- To find a minimal set of auxiliary channels that serve as good predictors for lockloss events
- Diagnosis of interferometer behavior leading to lockloss events

With Ayon Biswas and Jess McIver

Effect of earthquakes



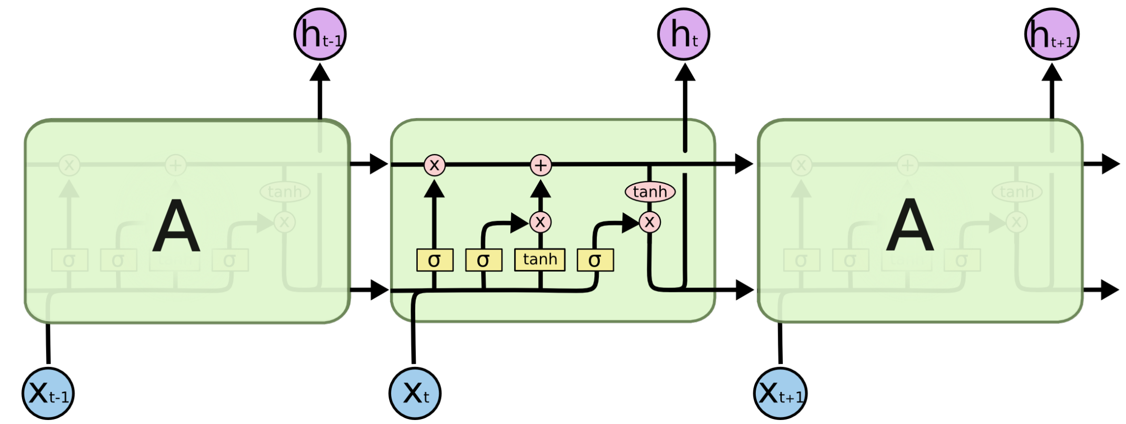
How to choose bins?

Histogram equalization in both axes?

(See the talk by Jess for details)

RNNs and delta-ts

using for multiple filters



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

$\boldsymbol{X}$: Input time series (2 variables);
$\boldsymbol{s}$: Timestamps for $\boldsymbol{X}$;

$\boldsymbol{M}$: Masking for $\boldsymbol{X}$;
$\boldsymbol{\Delta}$: Time interval for $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} 47 & 49 & NA & 40 & NA & 43 & 55 \\ NA & 15 & 14 & NA & NA & NA & 15 \end{bmatrix} \qquad \boldsymbol{M} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{s} = [\ 0 \quad 0.1 \quad 0.6 \quad 1.6 \quad 2.2 \quad 2.5 \quad 3.1] \qquad \boldsymbol{\Delta} = \begin{bmatrix} 0.0 & 0.1 & 0.5 & 1.5 & 0.6 & 0.9 & 0.6 \\ 0.0 & 0.1 & 0.5 & 1.0 & 1.6 & 1.9 & 2.5 \end{bmatrix}$$

Figure 2. An example of measurement vectors x_t, time stamps s_t, masking m_t, and time interval δ_t.

Che et al. 2018

SCIENTIFIC REPORTS | (2018) 8:6085 | DOI:10.1038/s41598-018-24271-9

Naul et al. (encoder-decoder)

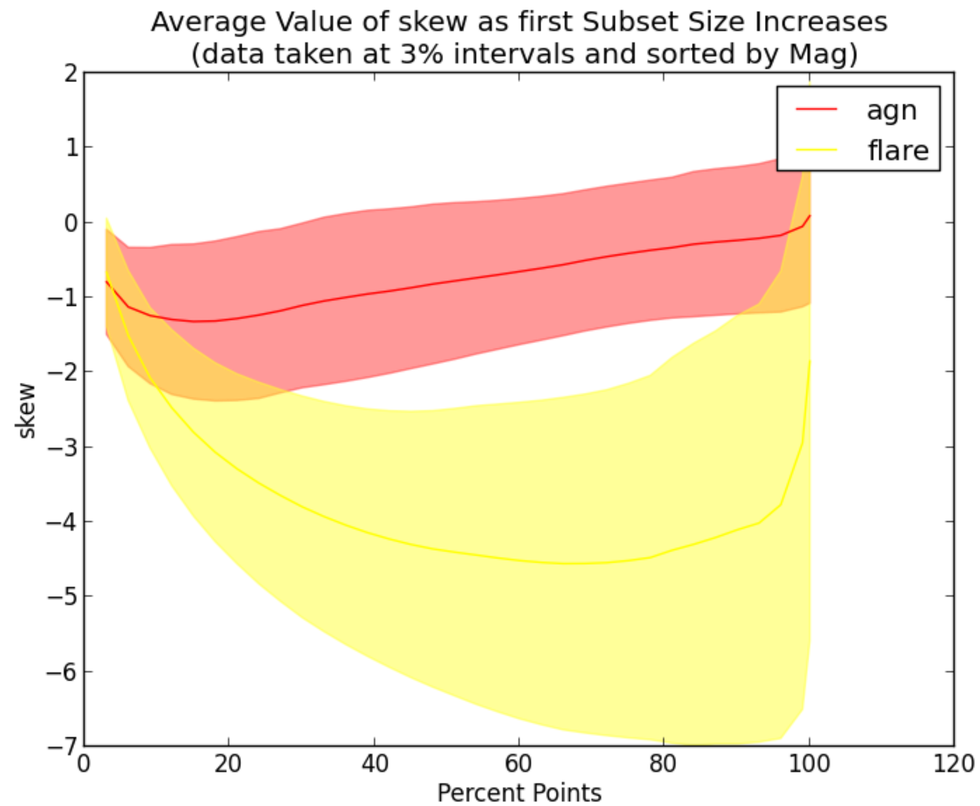
Best model: Static RNN with stitching, drop dt>120, input [dt, mag] (normalized)

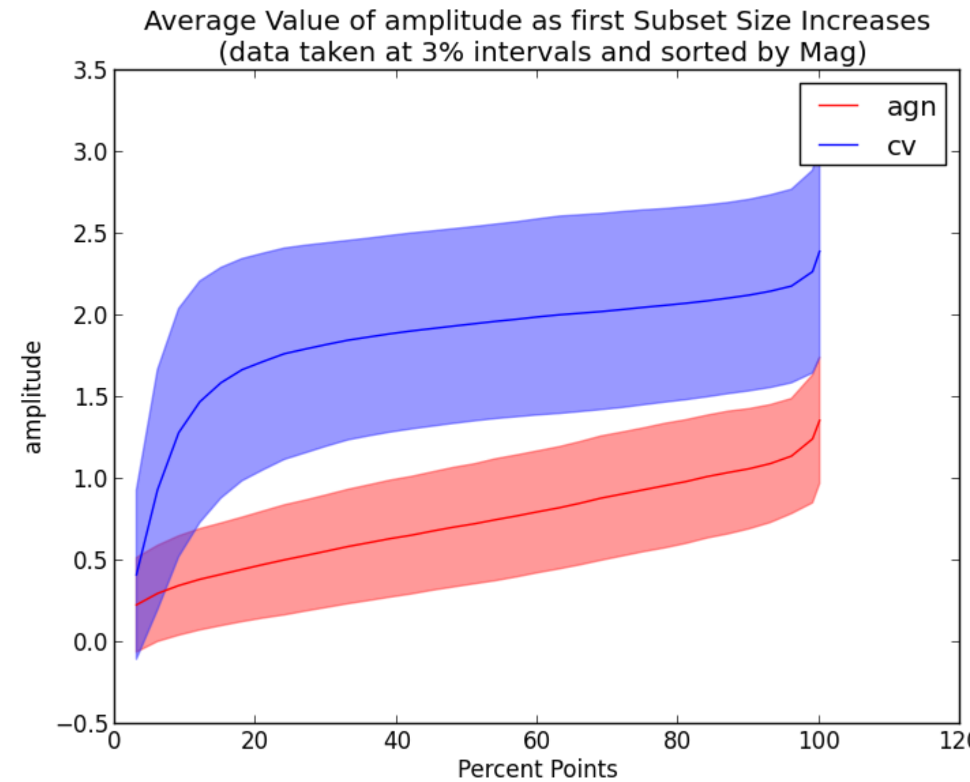
96% for easy classes

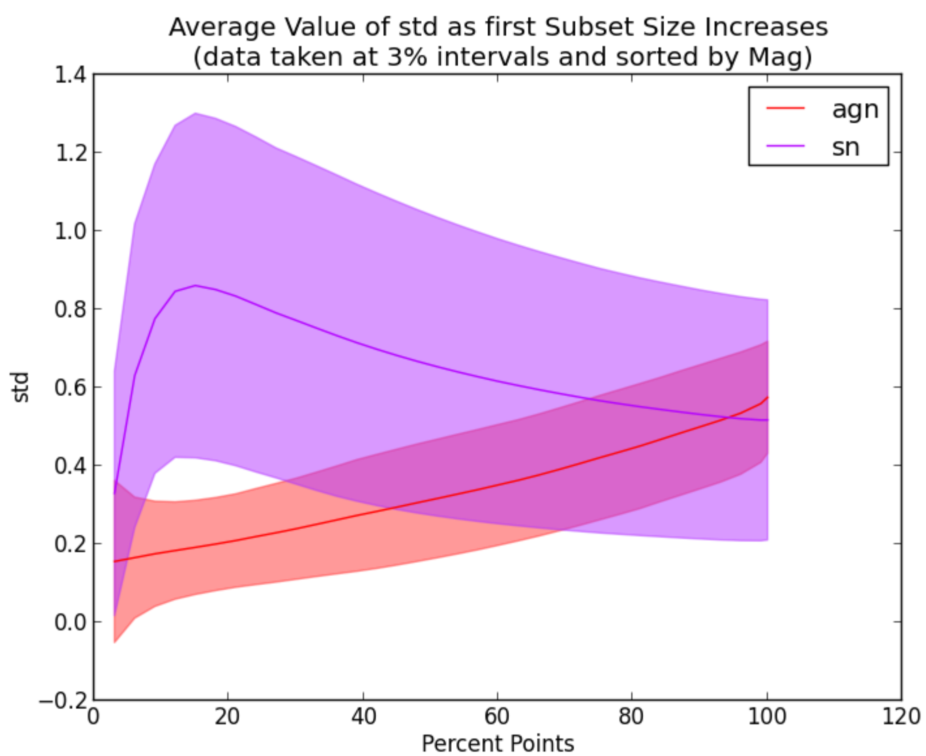
Should certain delta-ts be ignored?

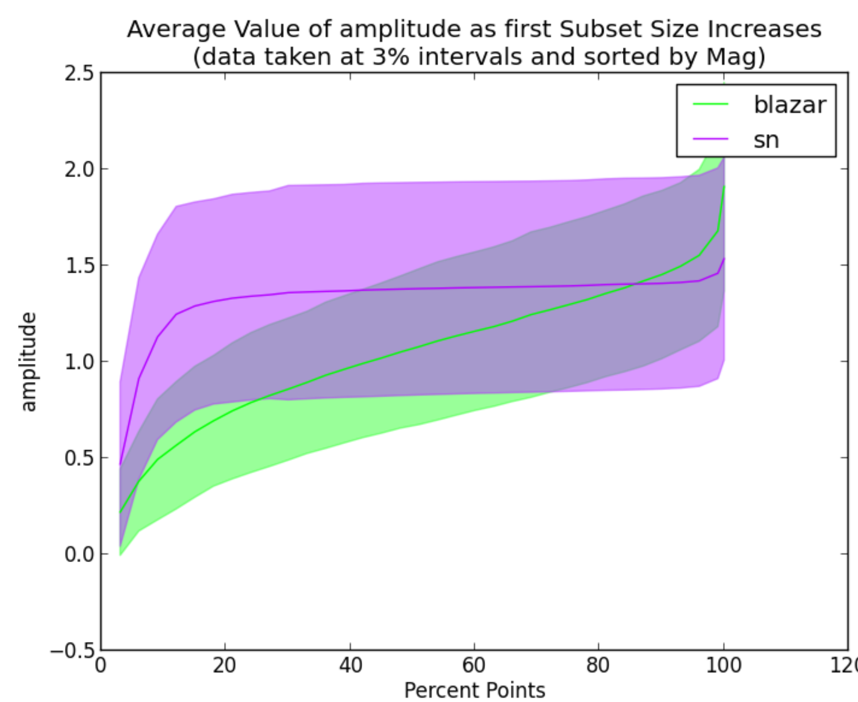
With Vinu Sankar

Using only fraction of points or, less may be more









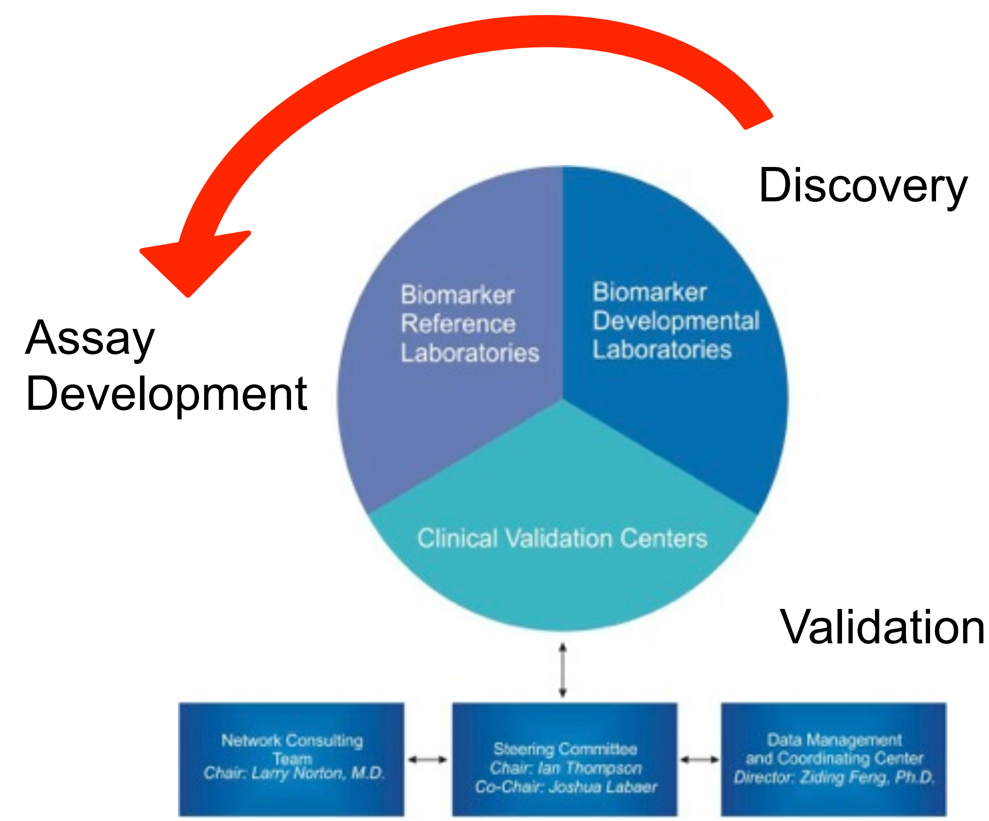
with Chengyi Lee

EDRN - Early Detection Research Network

Dan Crichton (PI), Luca Cinquini, David Liu, Heather Kincaid, Sean Keely, Kristen Anton, Maureen Colbert +++

- Early Detection Research Network (EDRN): 40+ institutions.
- Aim: Discovery of cancer biomarkers - early indicators of onset of disease
- NCI/NIH funded flagship program
  - Started in ~2000

Organizational Structure



Emphasis: Automation, Reproducibility, Scalability

MCL:Consortium for the Characterization of Screen Detected Lesions

Dan Crichton (PI), Luca Cinquini, David Liu, Heather Kincaid, Sean Keely, Kristen Anton, Maureen Colbert +++

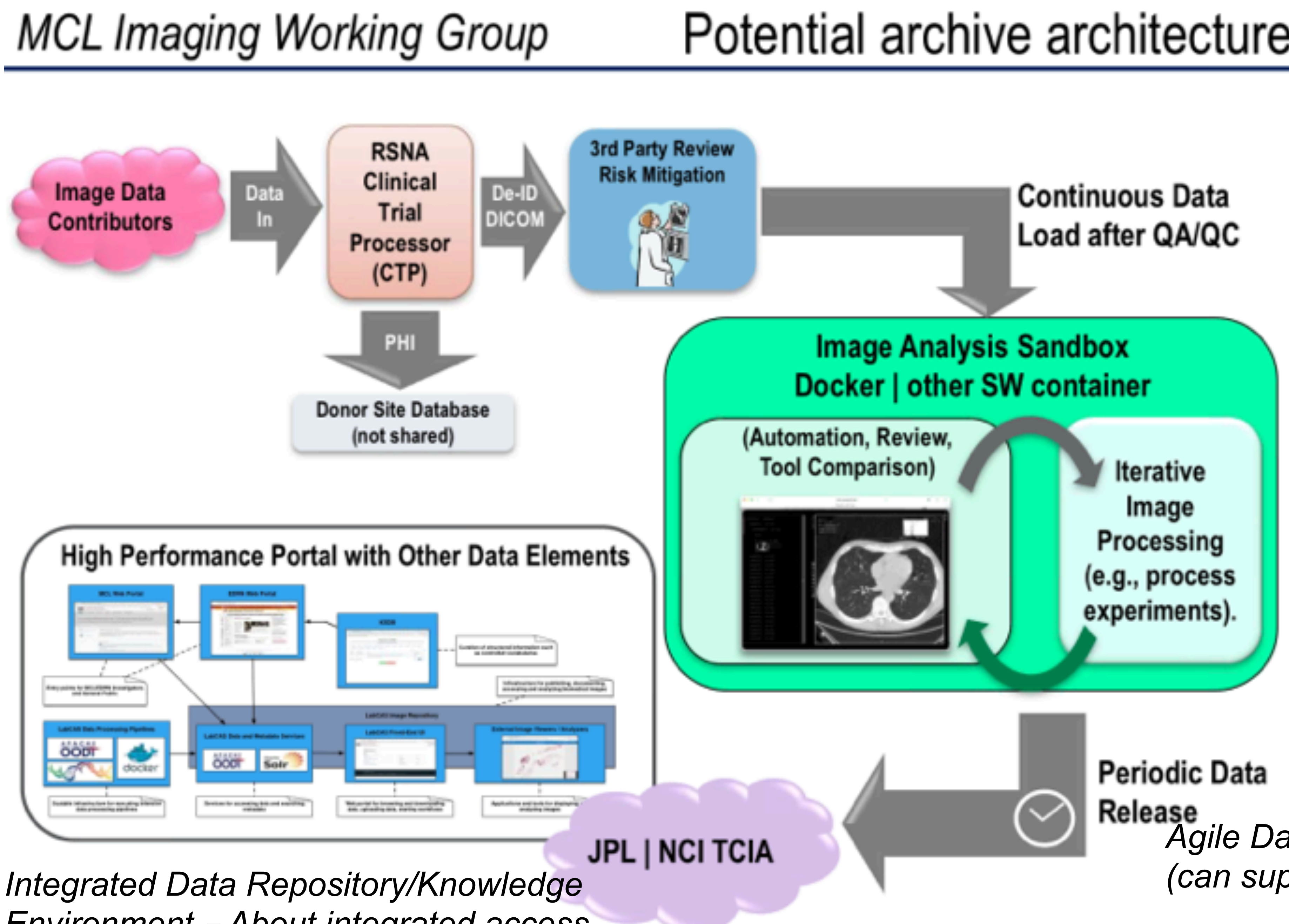


Image Types
- DICOM
- Pathology

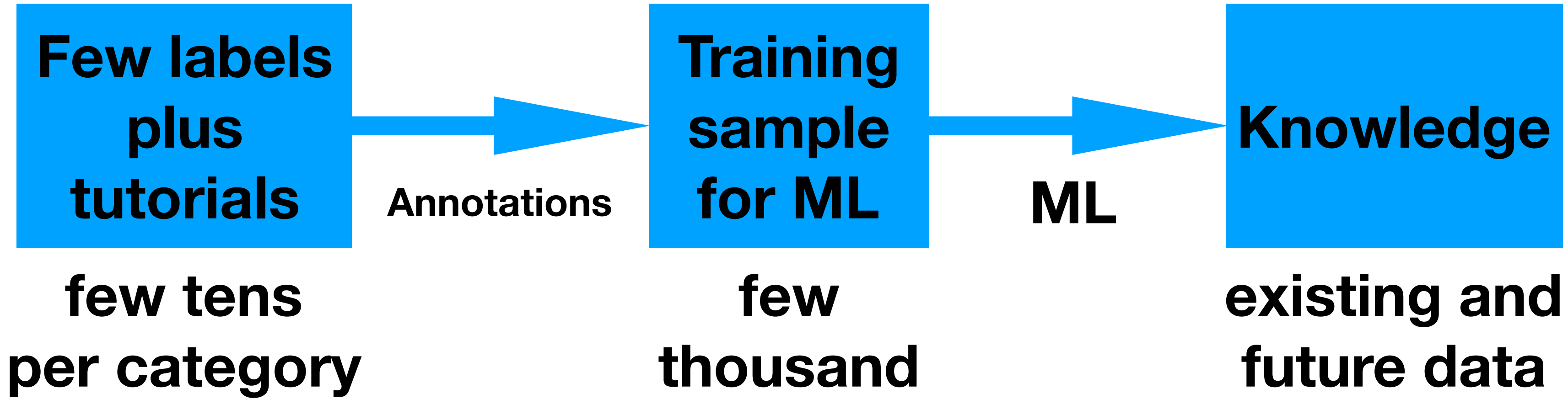
Organ Types
- Lung
- Breast
- Prostate
- etc

Agile Data Pipelines (can support as needed)

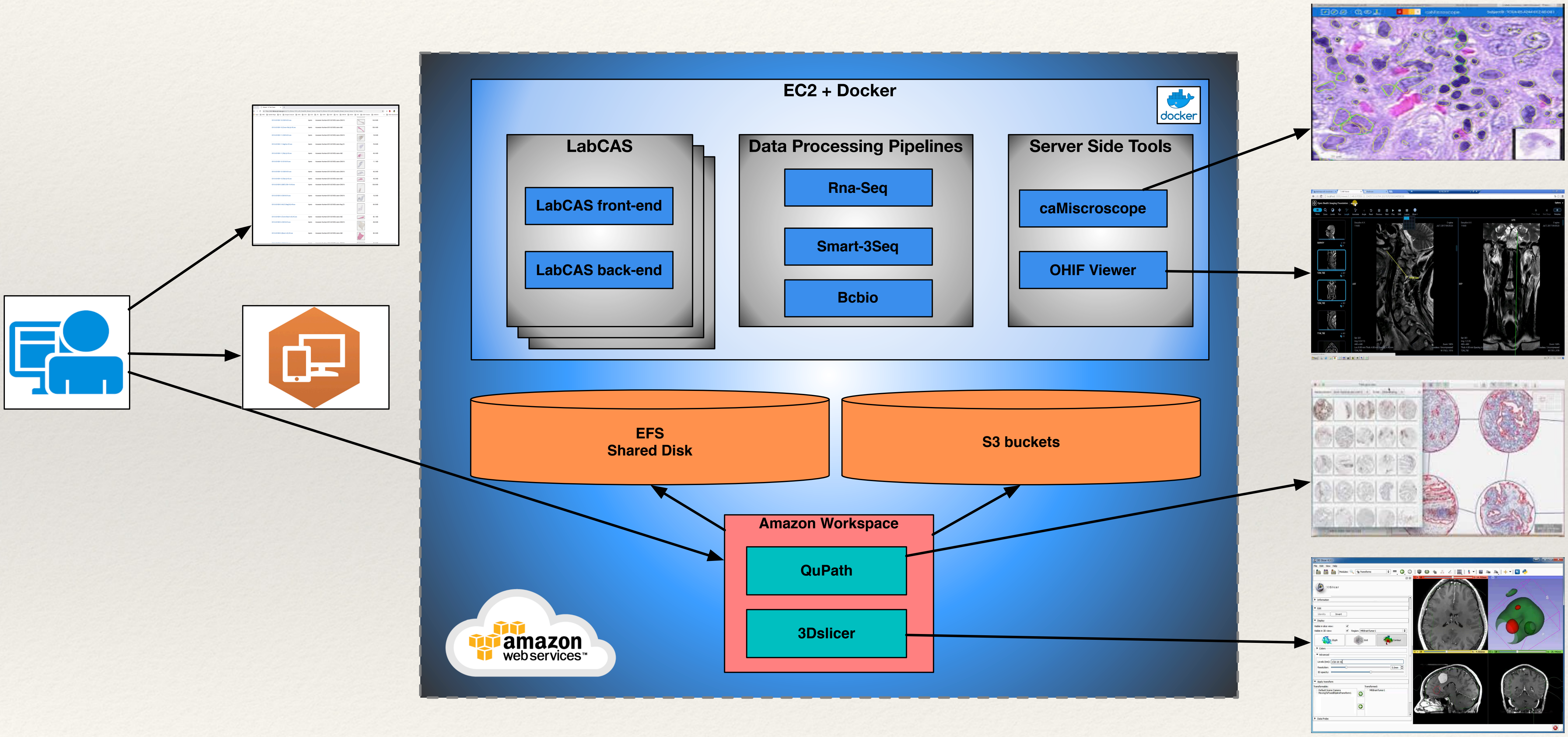
Integrated Data Repository/Knowledge Environment – About integrated access

Towards the Pre-cancer (Imaging) Atlas

- Create annotated images with nodules, cysts
  - attendant features
- Annotations done by trained personnel
  - Radiologists (**capturing + training**)
  - Citizen scientists (through tutorials)
- Use Machine Learning for large-scale classification

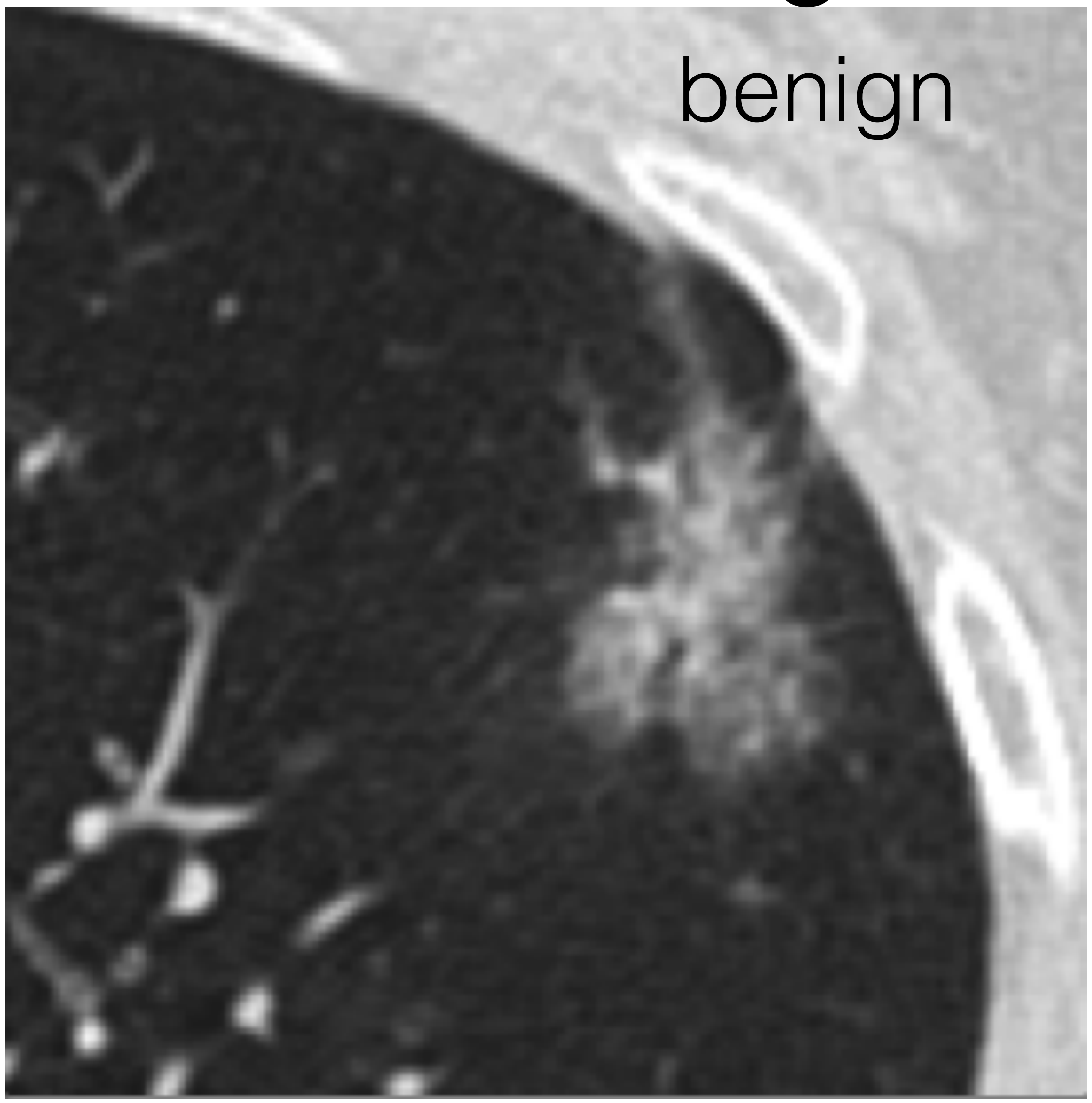


LabCAS Cloud Architecture



Deep Learning

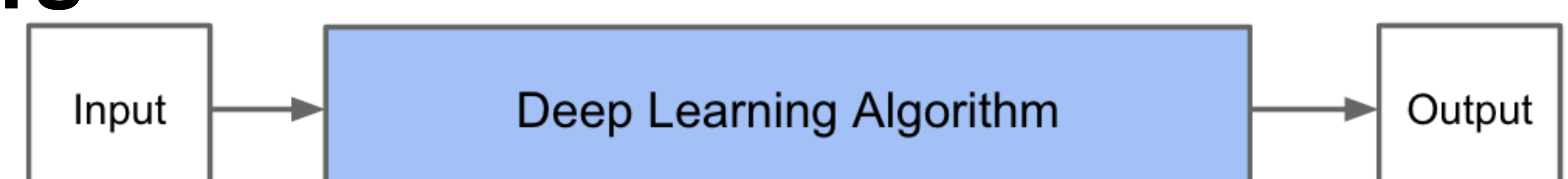
Lung Cancer Dataset (NLST)



Traditional Machine Learning Flow

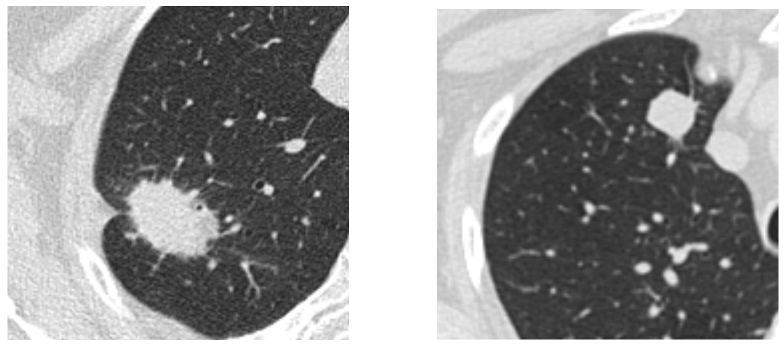
50000 Heavy smokers

Followed over years

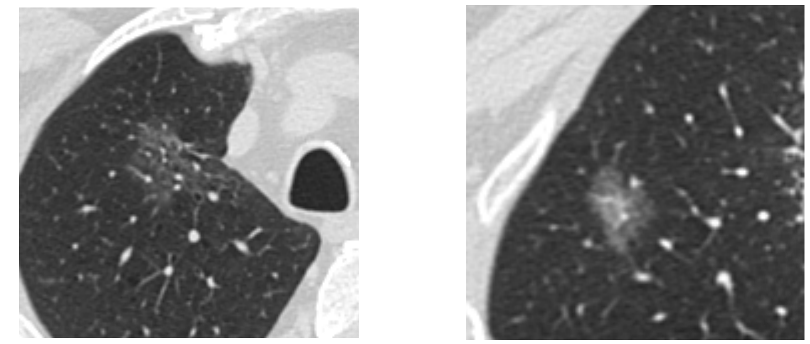


Deep Learning Flow

Consistency - solid

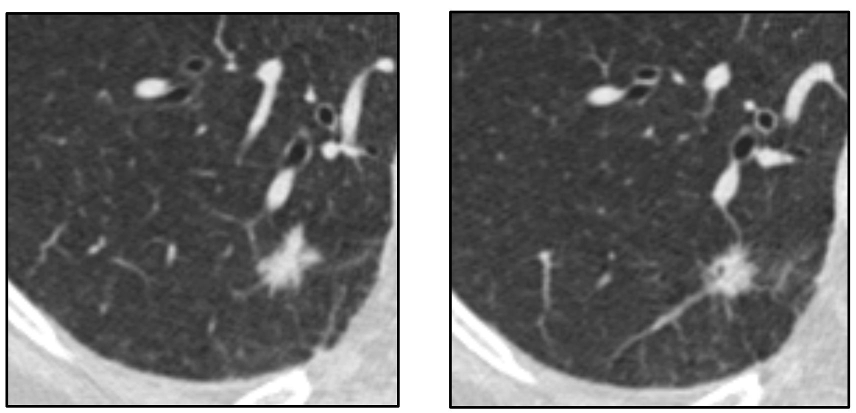


Consistency - pure GGN



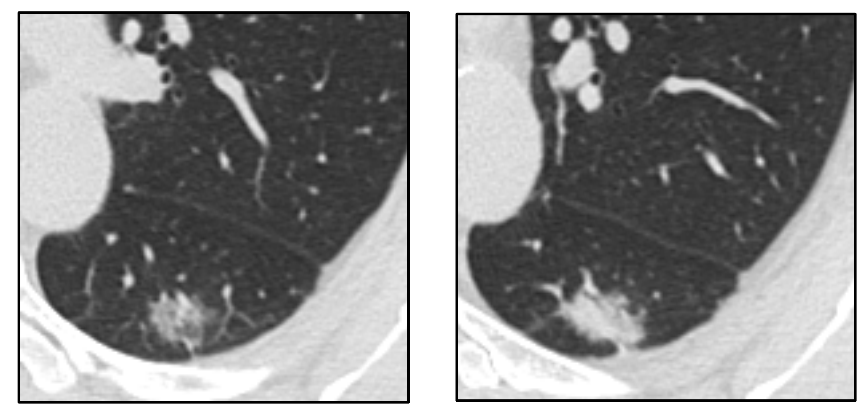
Solid or PSN

(Part Solid Nodule)

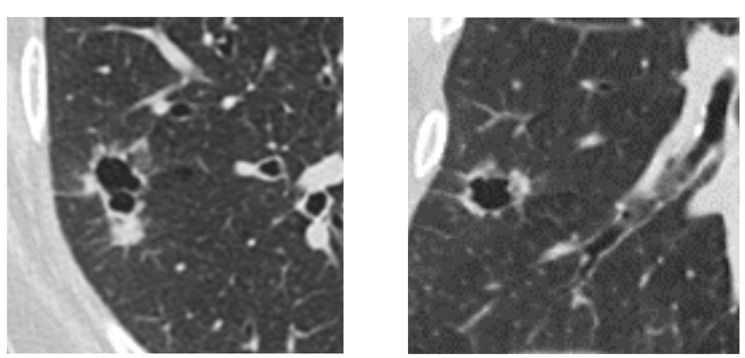


Diff. axial levels - PSN

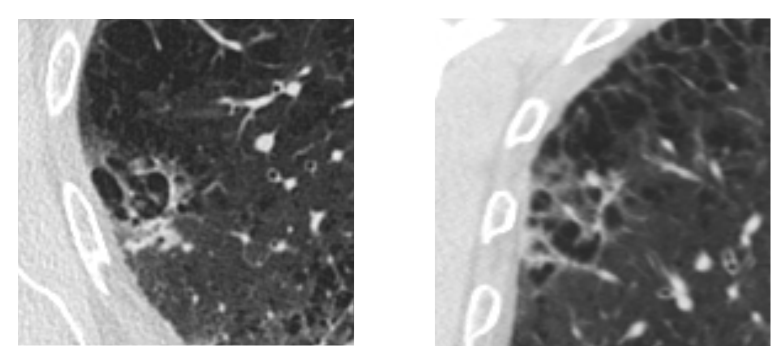
(by consensus)



Per-cystic or cystic



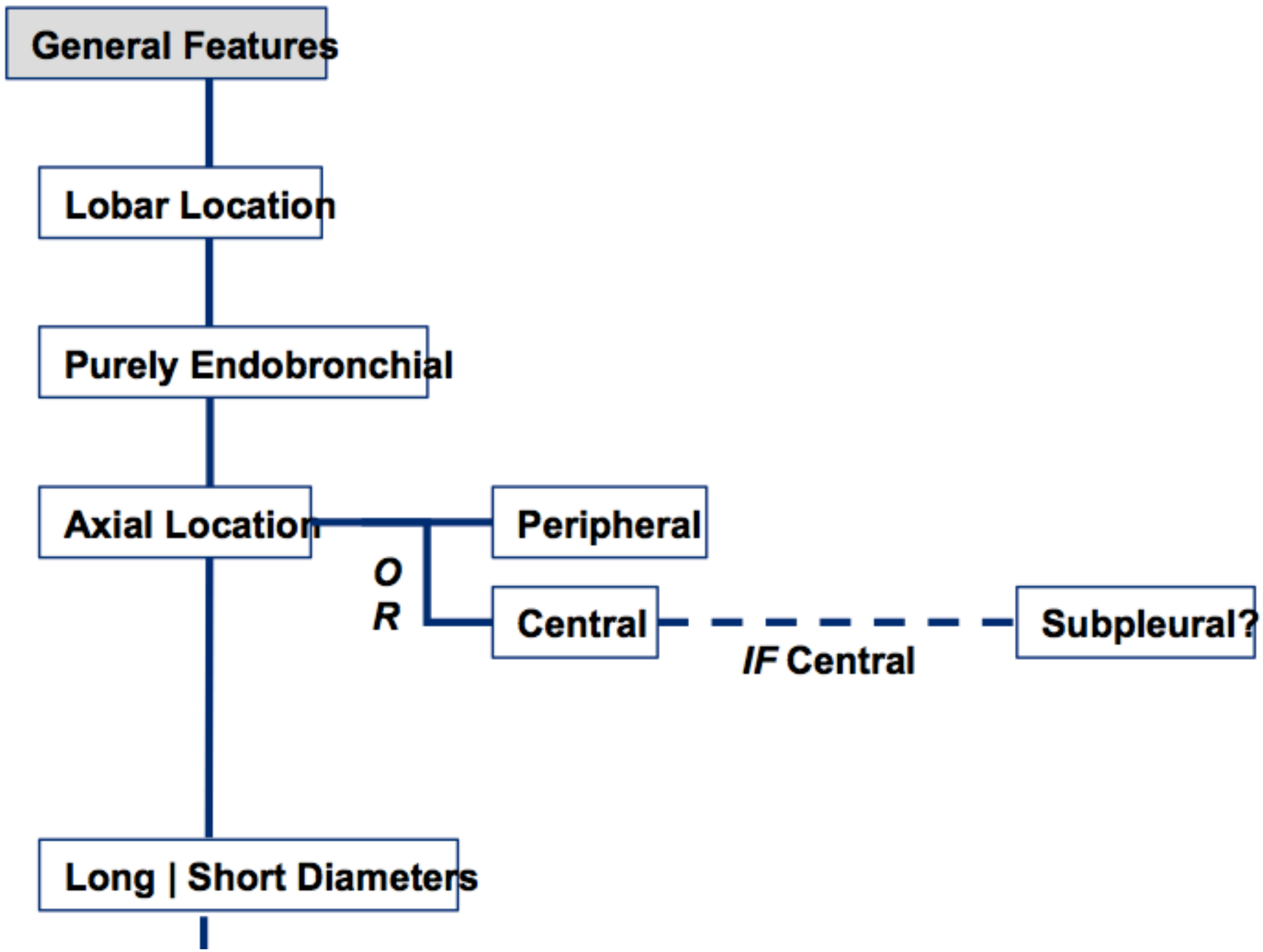
Axial **Coronal**



Axial **Coronal**

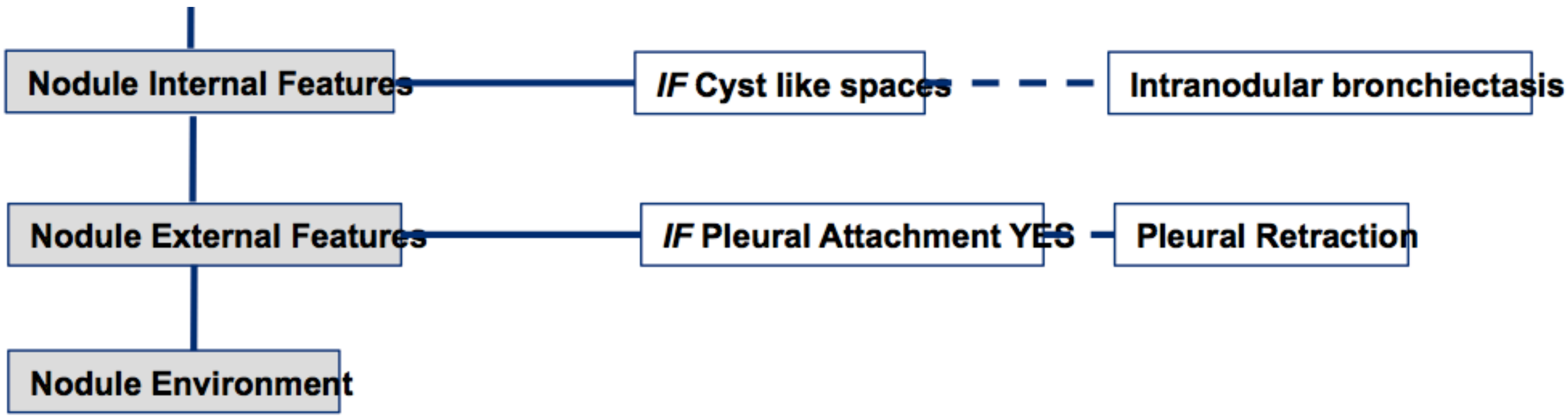
Branching Tree for Illustrated Lexicon

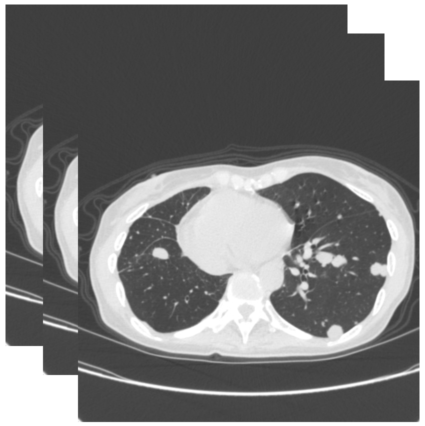
Deni Aberle et al.



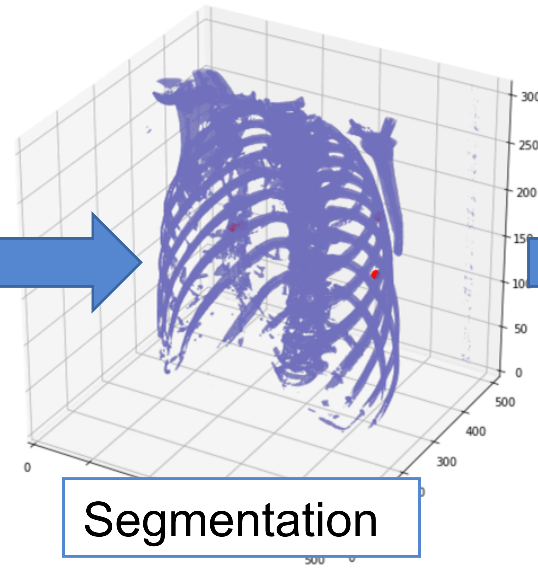
Branching Tree for Illustrated Lexicon

Deni Aberle et al.



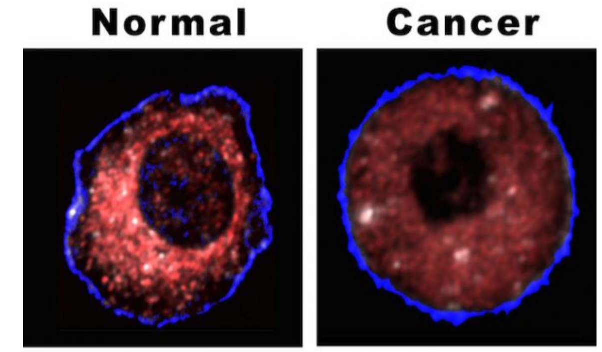


NLST Dicom Slices:
~1000 Patients



Segmentation

Extract bounding boxes of nodule



LUNA16/Kaggle1 trained model predicts cancer

GRT123

Fangzhou, L. (2017)

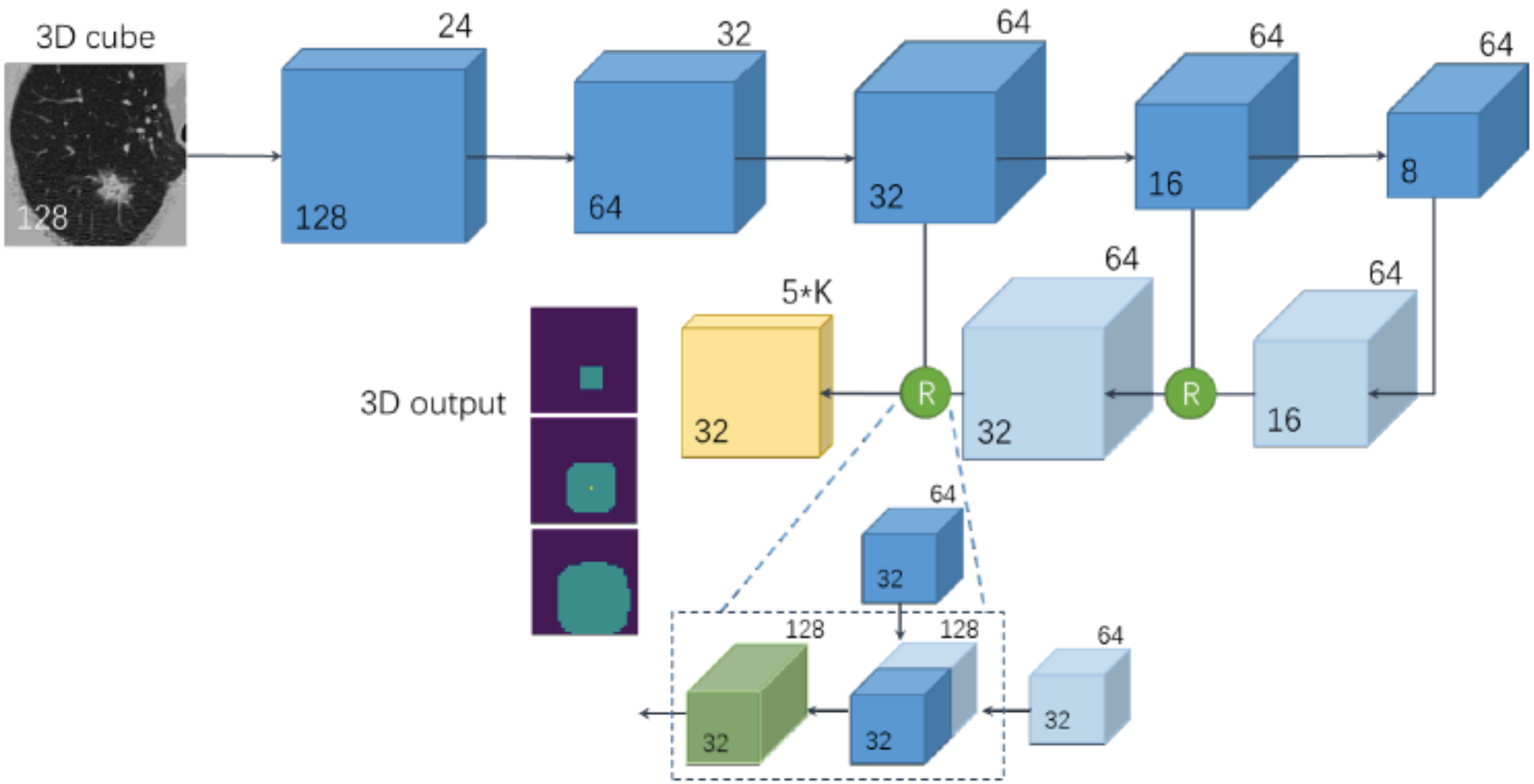
Domain adaptation and transfer learning

With David Liu

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

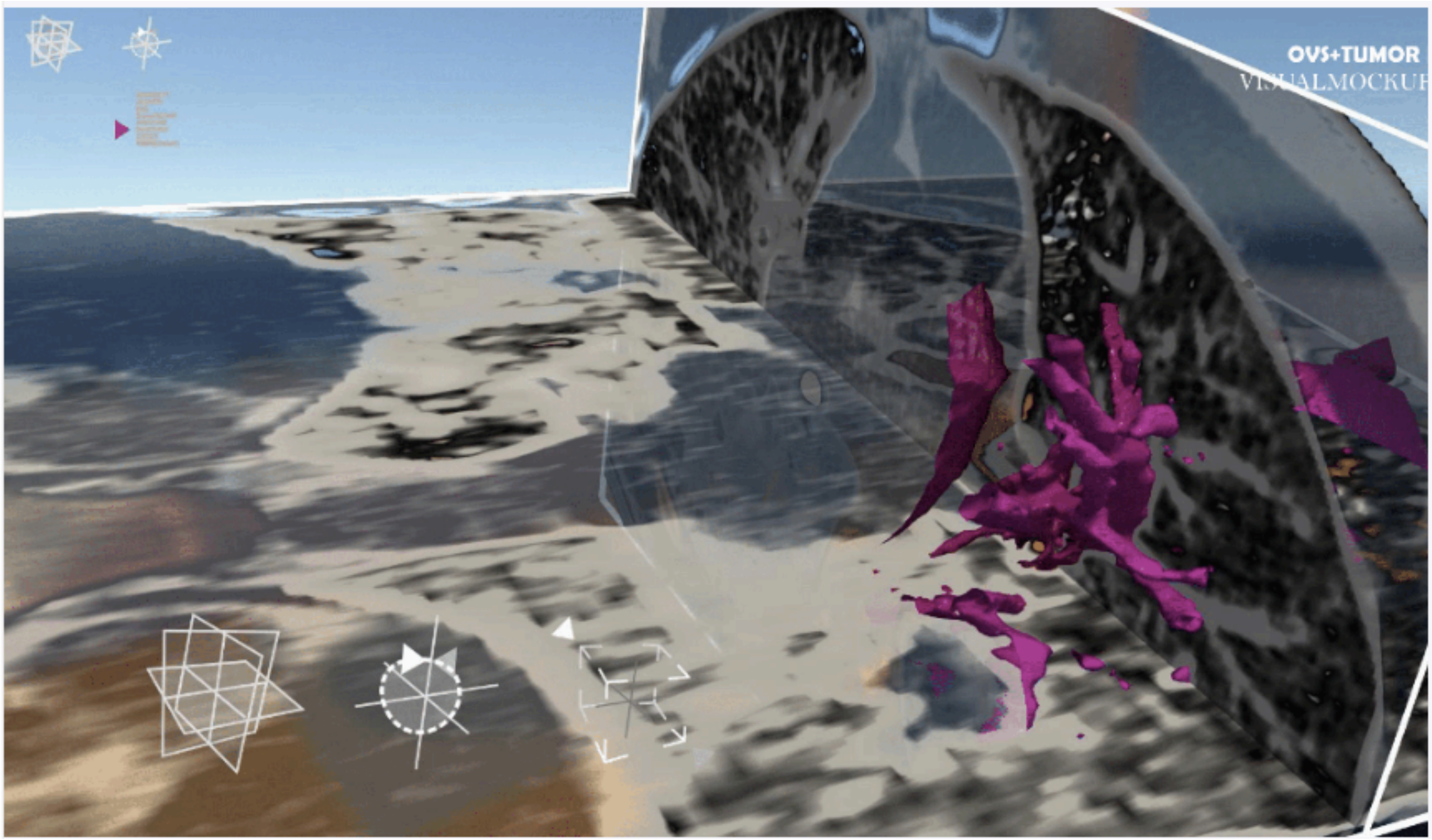
Explainability/Interpretability!

The GRT123 model for segmentation



16 CPUs

VR model with Santiago Lombeyda



SigGraph demo/talk coming up

Retrospective Cohort

U01CA189283: Duke 2005-2015. Predictive Biomarkers for interval cancers in high-risk women undergoing MRI screening n=1,421/year

Normal MRI
(82% n=1,165)

MRI in 12 months

Abnormal Screening MRI
(13% n= 181)

6 week MRI
No Biopsy
(5% n=71)

Abnormal Biopsy
(2% n= (11/5) 16)

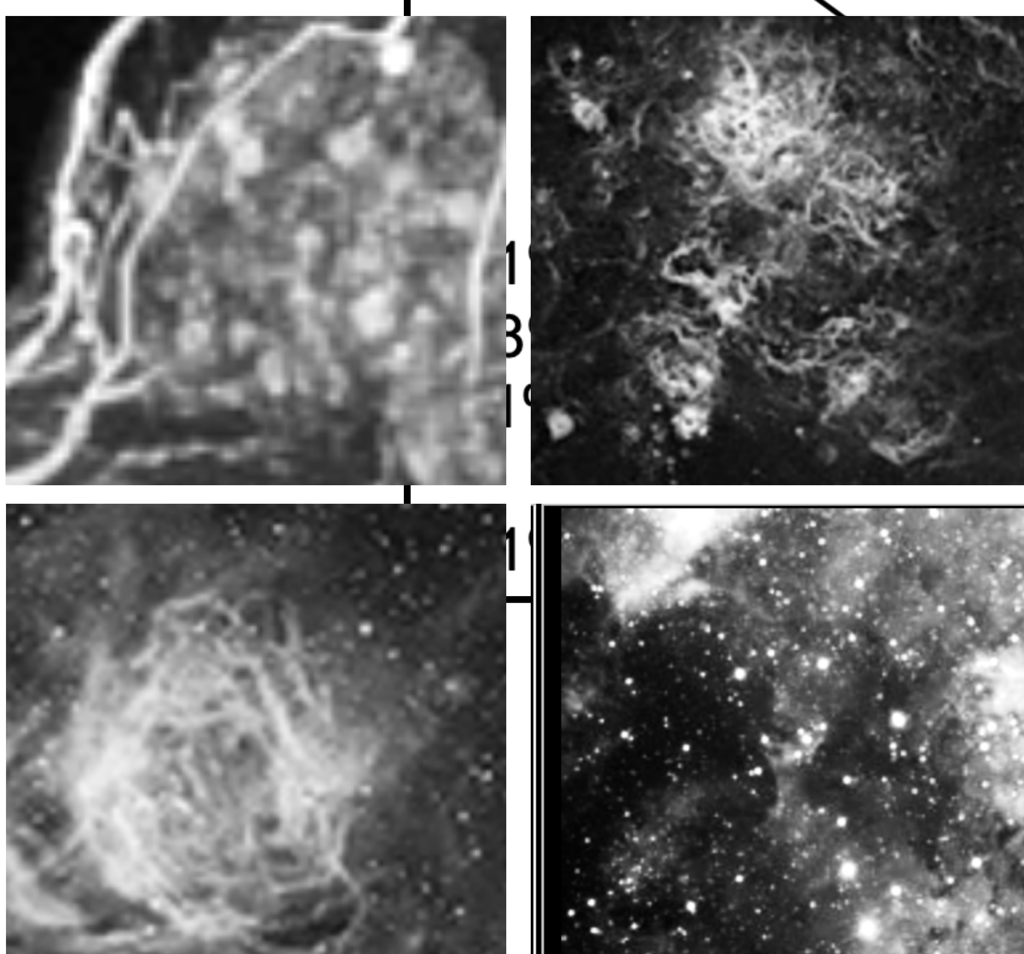
Standard-of-Care

Normal Biopsy
(6% n=94)

Follow up MRI 12 months

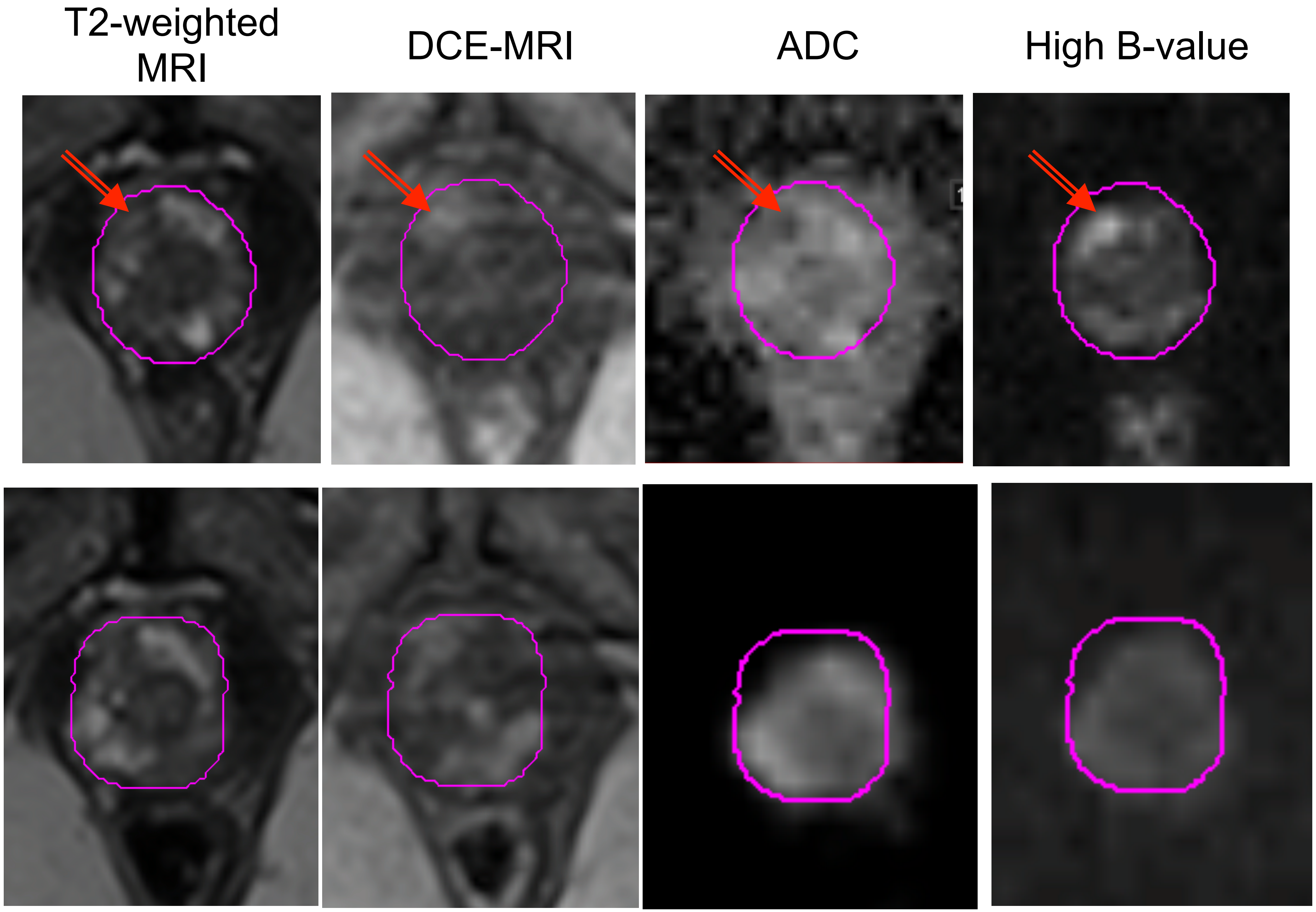
Normal f/u MRI
(5.4% n=86)

Interval Cancer
(0.7% n=8)



Victoria Seawaldt, COH, and team

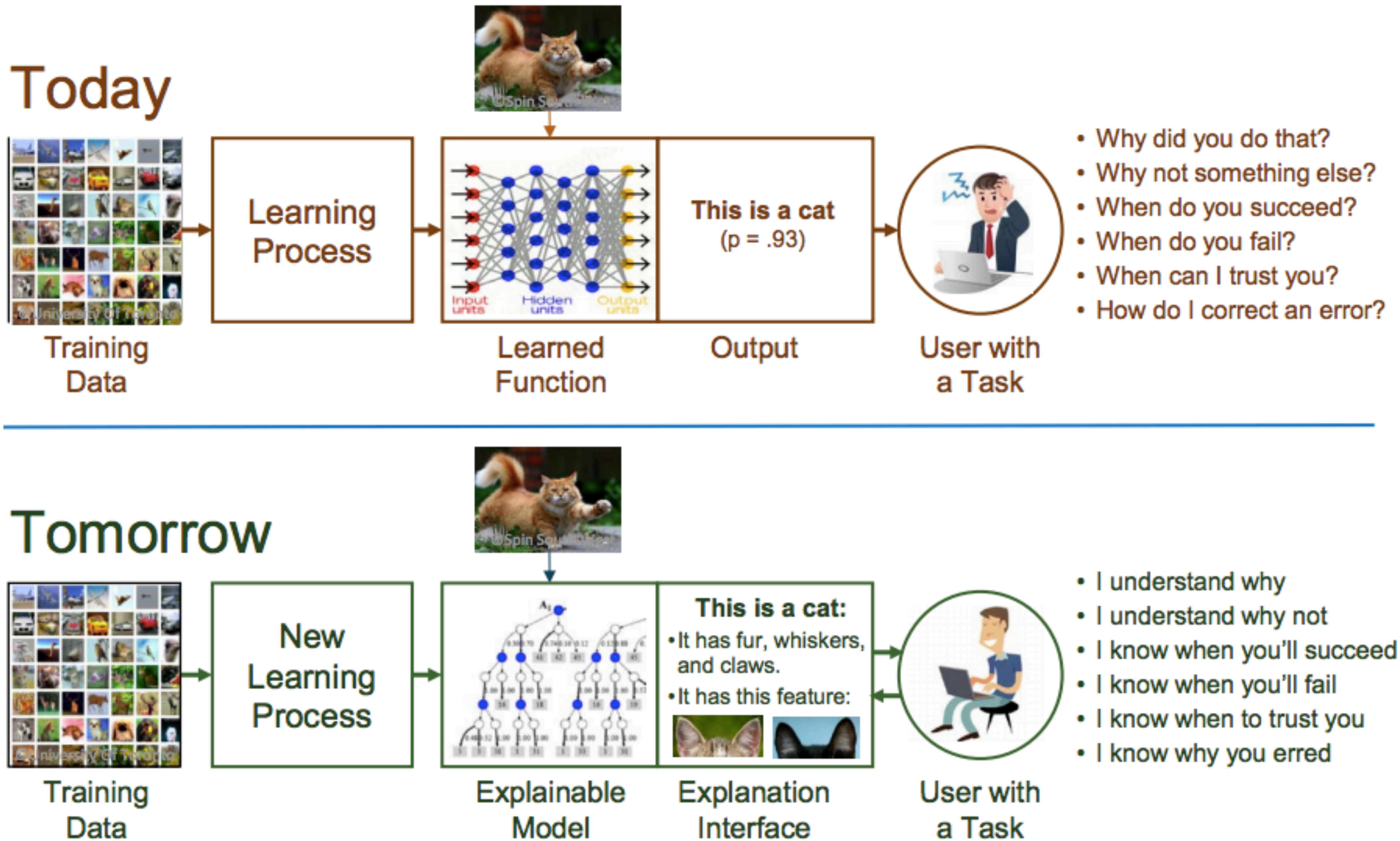
Prostate MRIs



T2-weighted imaging; Dynamic Contrast Enhanced (DCE-MRI); Apparent Diffusion Coefficient (ADC); High diffusion (B-value)

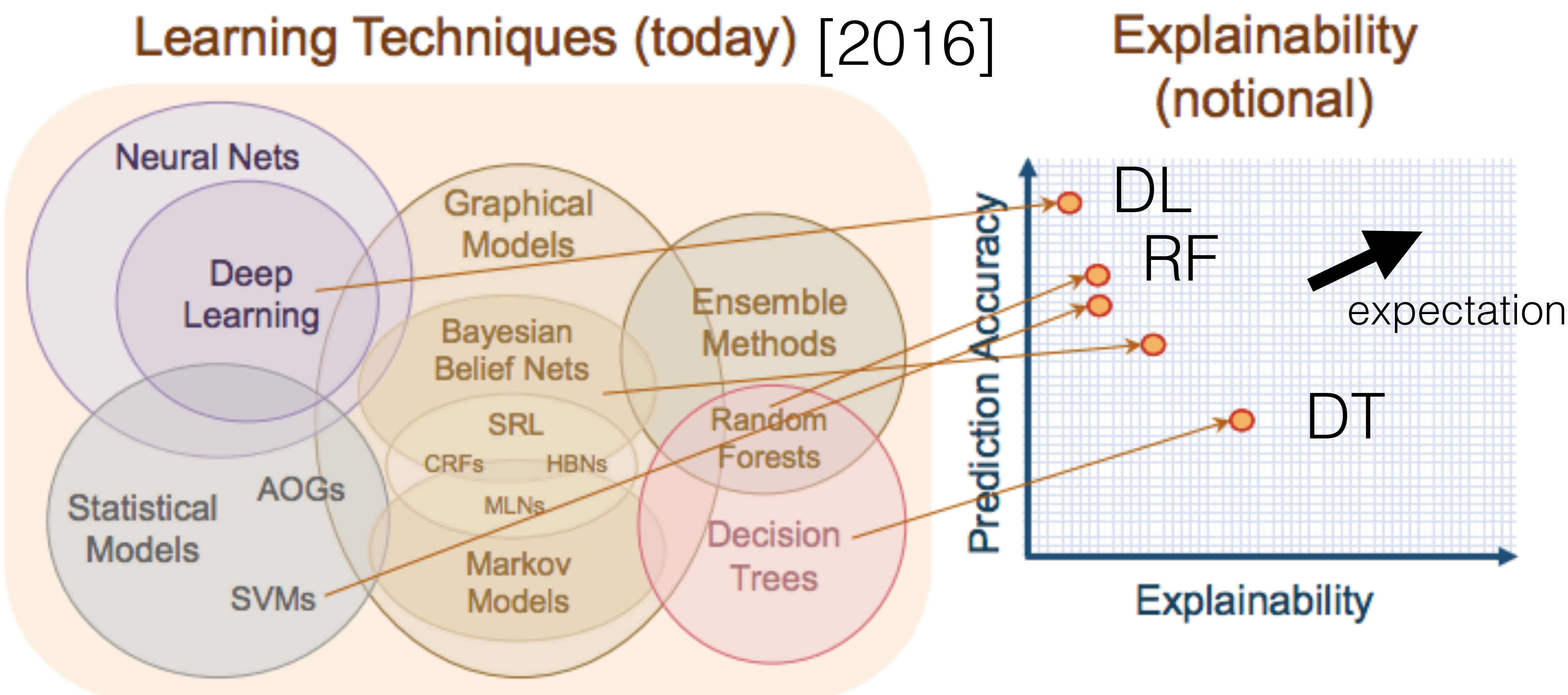
Radka Stoyanova
University of Miami

Interpretability



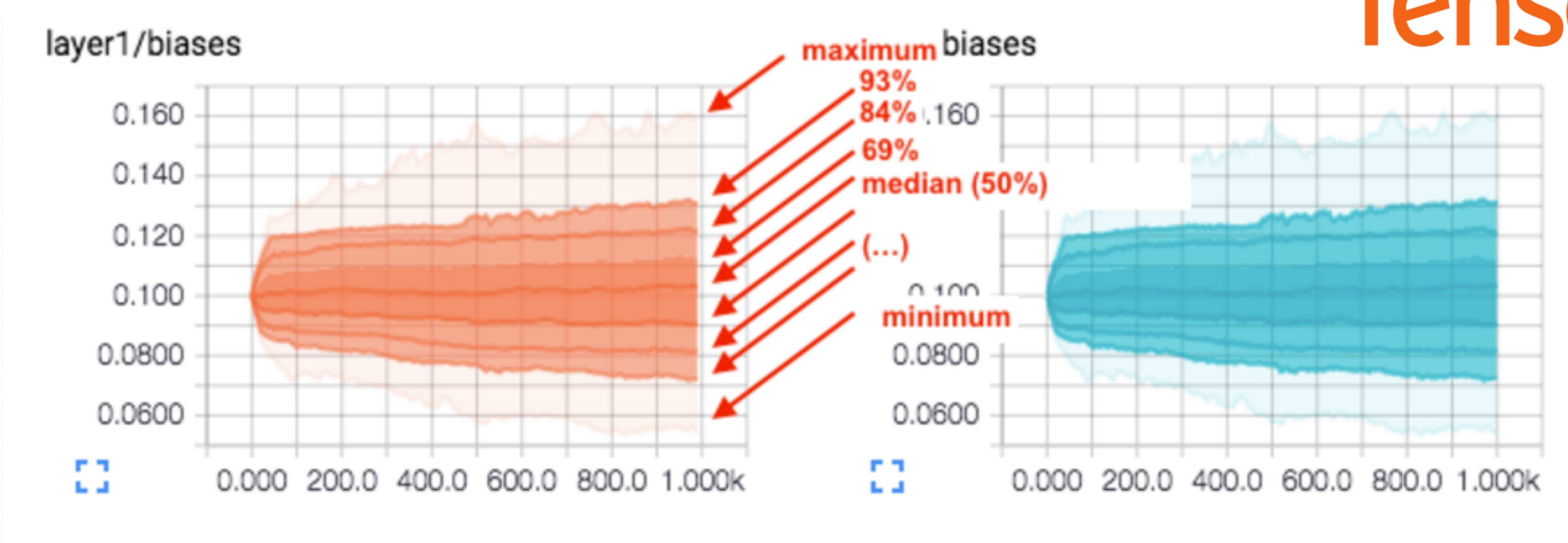
David Gunning (DARPA/I2O)

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

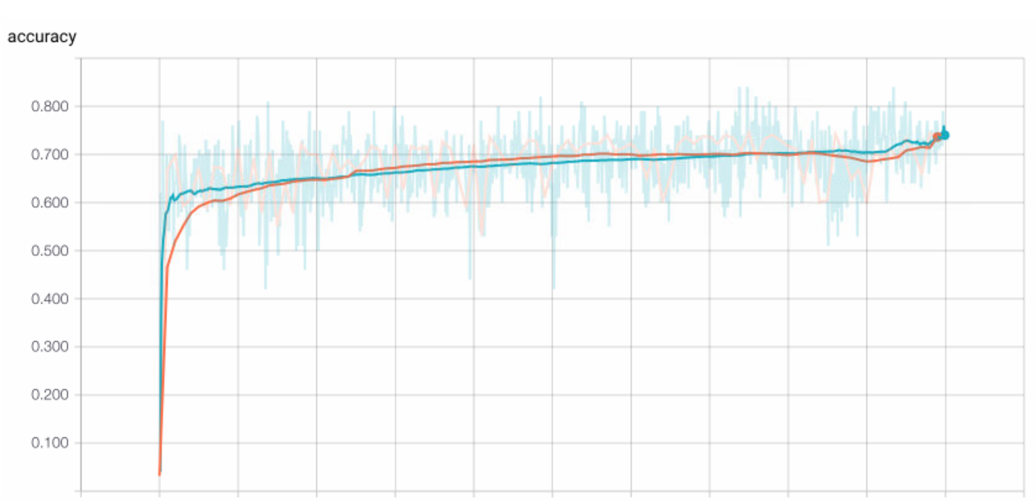


David Gunning (DARPA/I2O)

Distribution Summaries

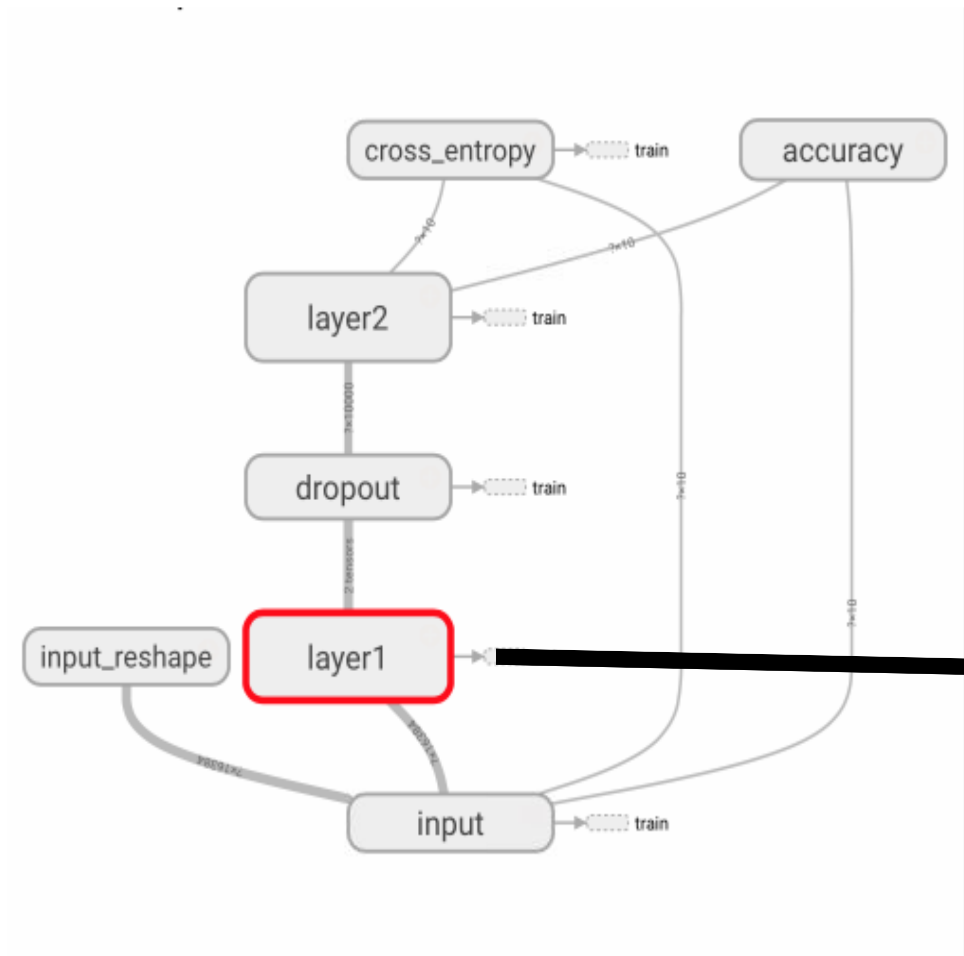


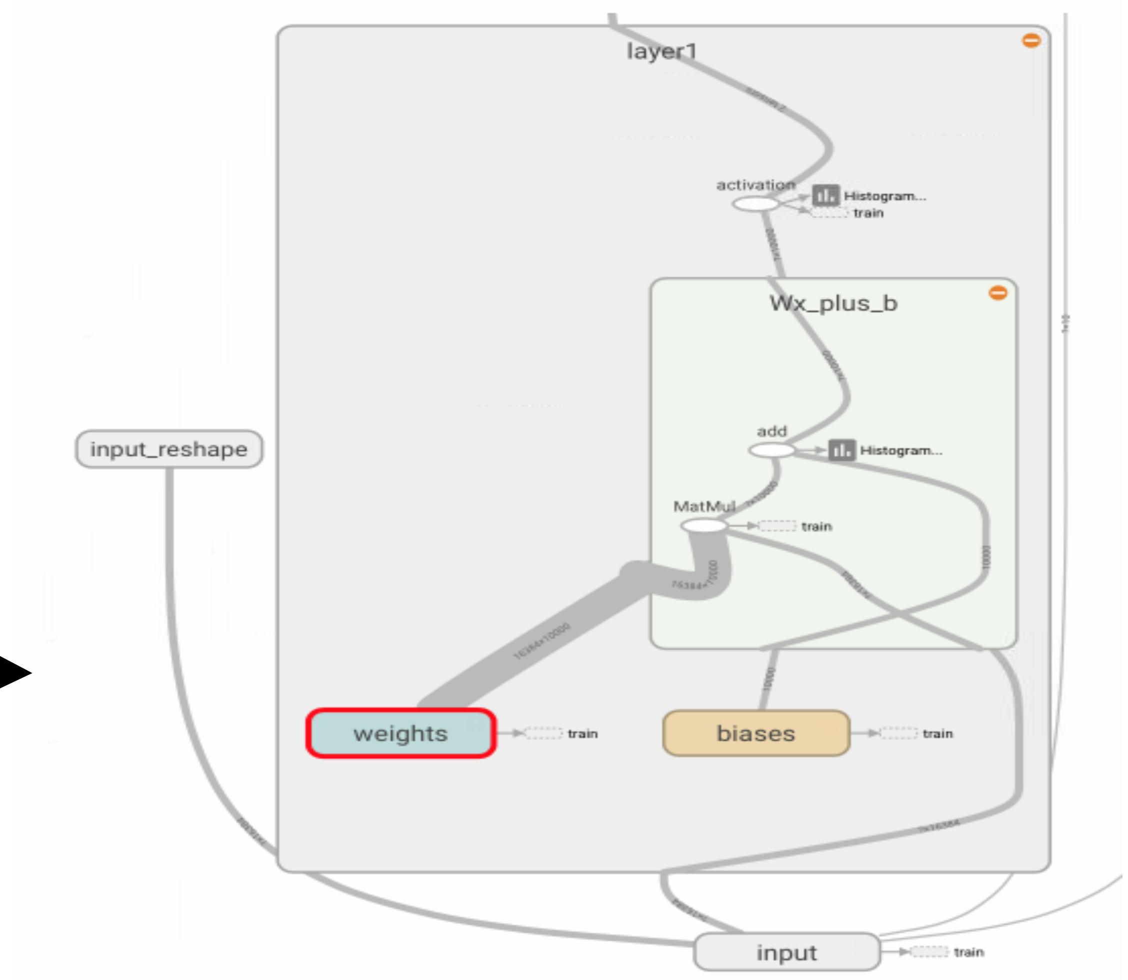
Percentile distributions over the data: max, 93, 84, 69, 50, 31, 16, 7, min



Interactivity

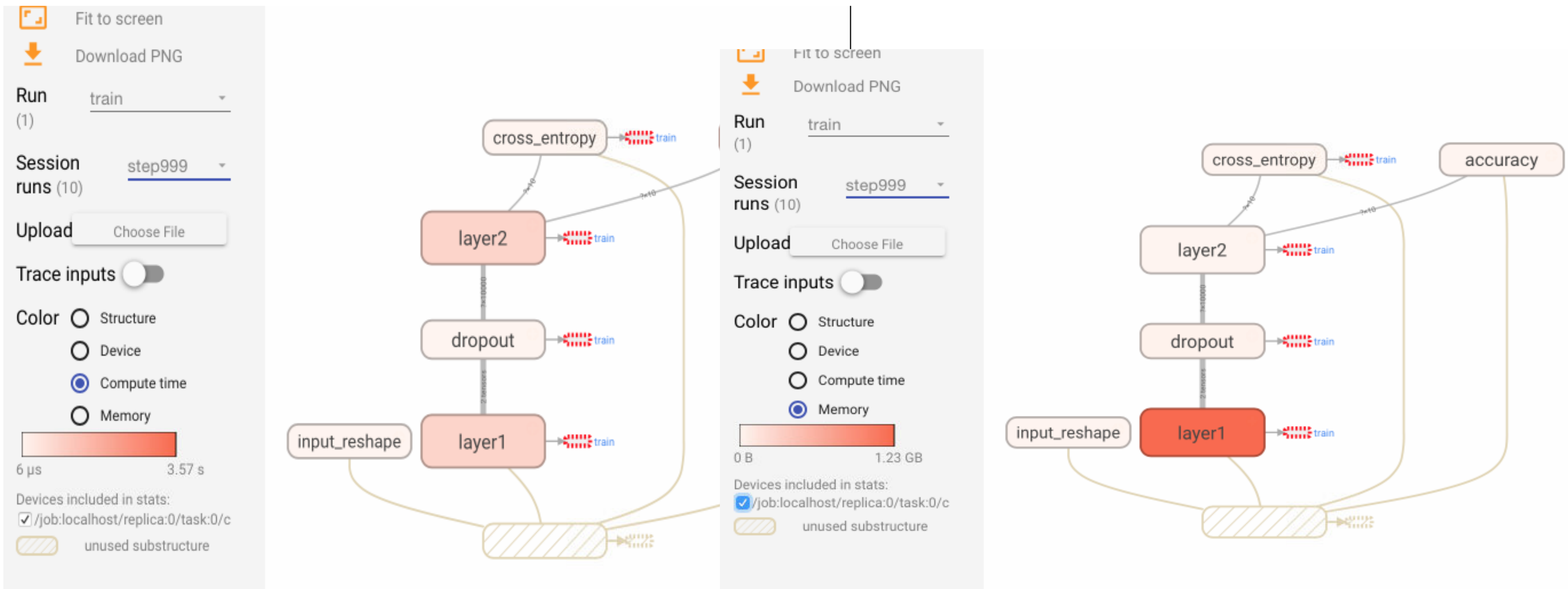
"Buttons" are portals to more details in the flow



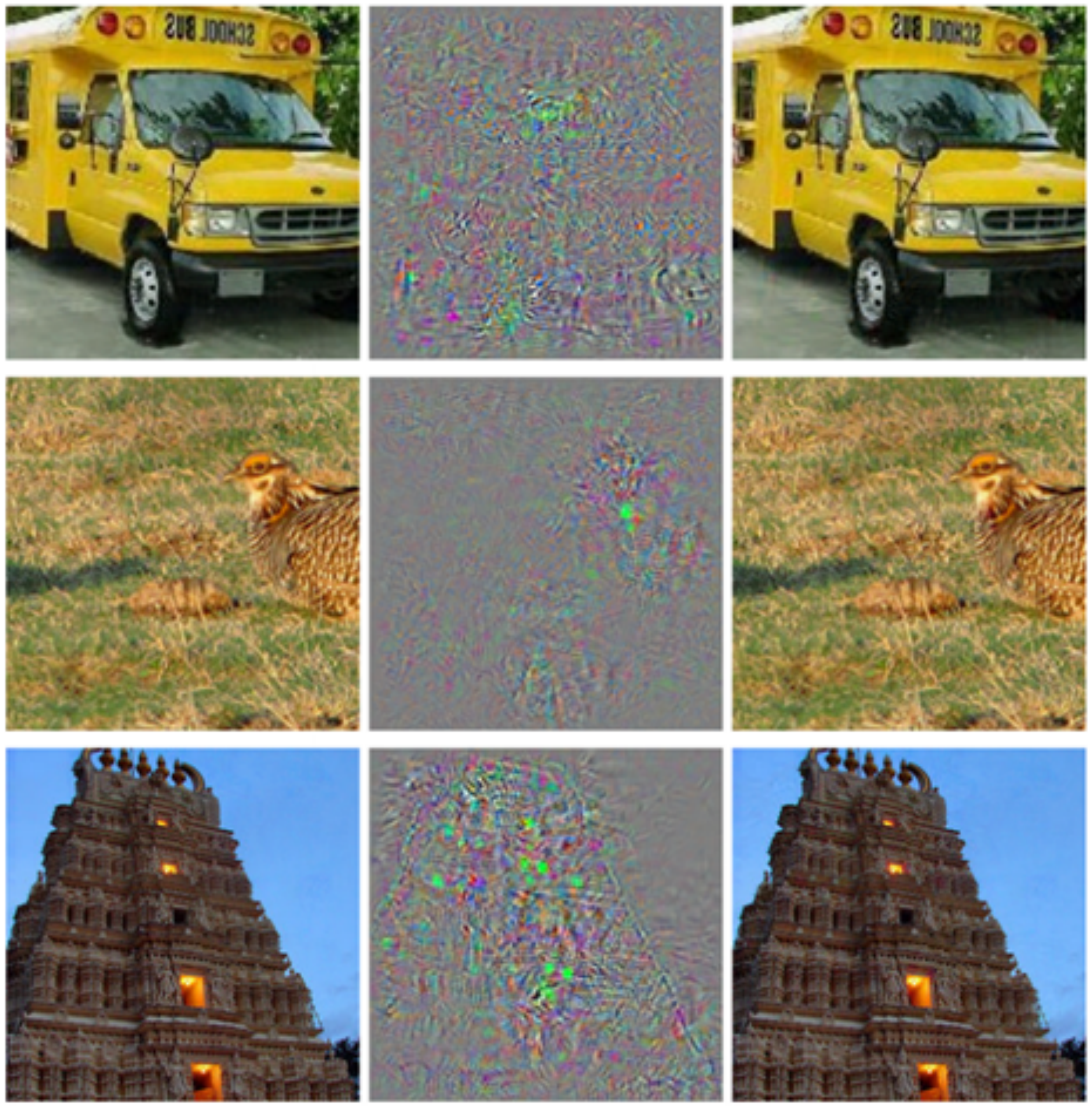


View options

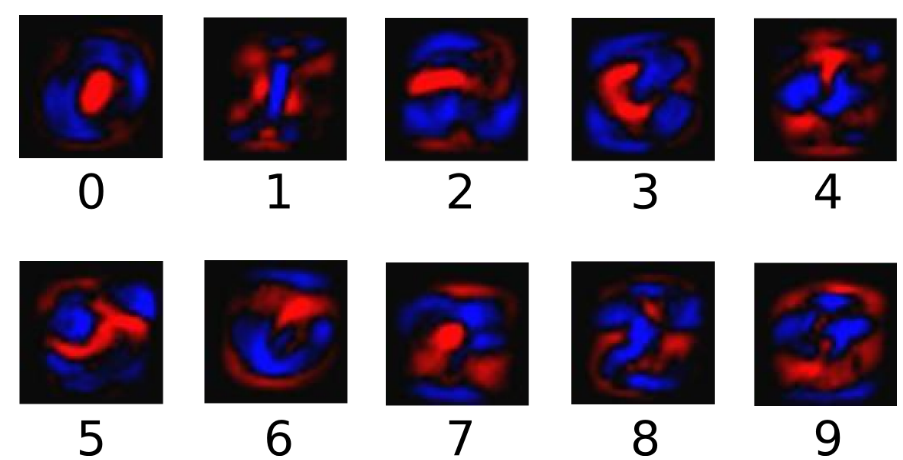
Device, computational time, memory for optimization, efficiency of computing



Including adversarial examples during training

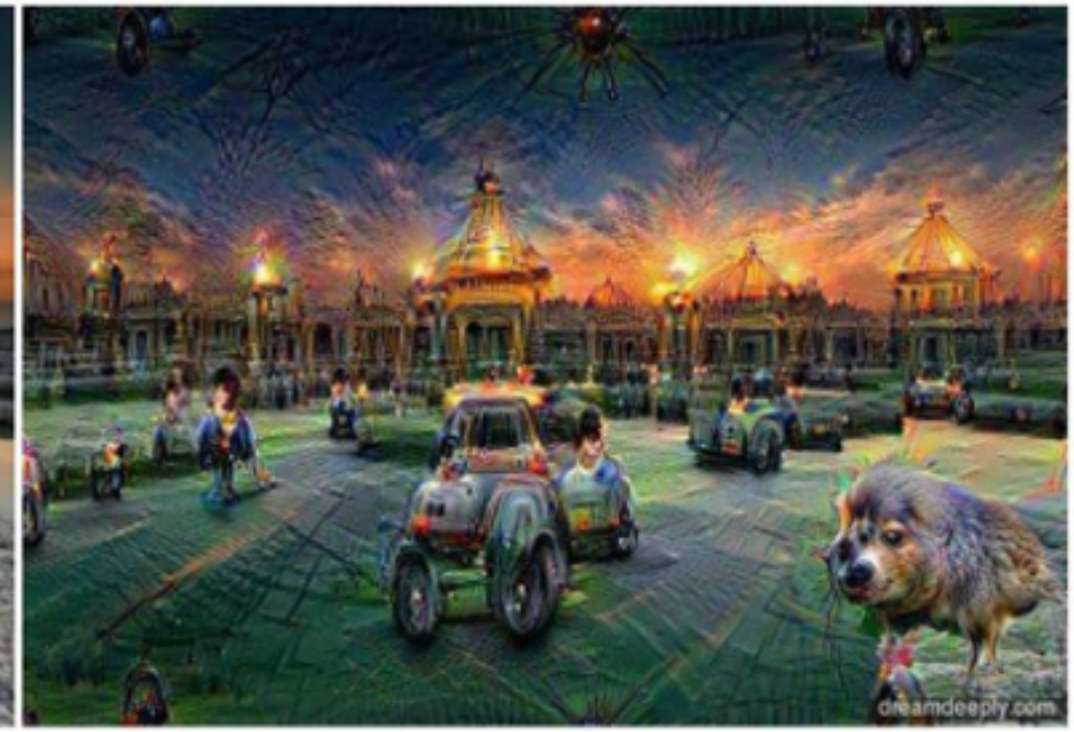


The images in the left most column are correctly classified examples. The middle column represents the distortion between the left and right images. The images in the right most column are predicted to be of the class ostrich! Even though the difference between the images on the left and right is imperceptible to humans, the ConvNet makes drastic errors in classification.



MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf



Before After fromthegrapevine.com

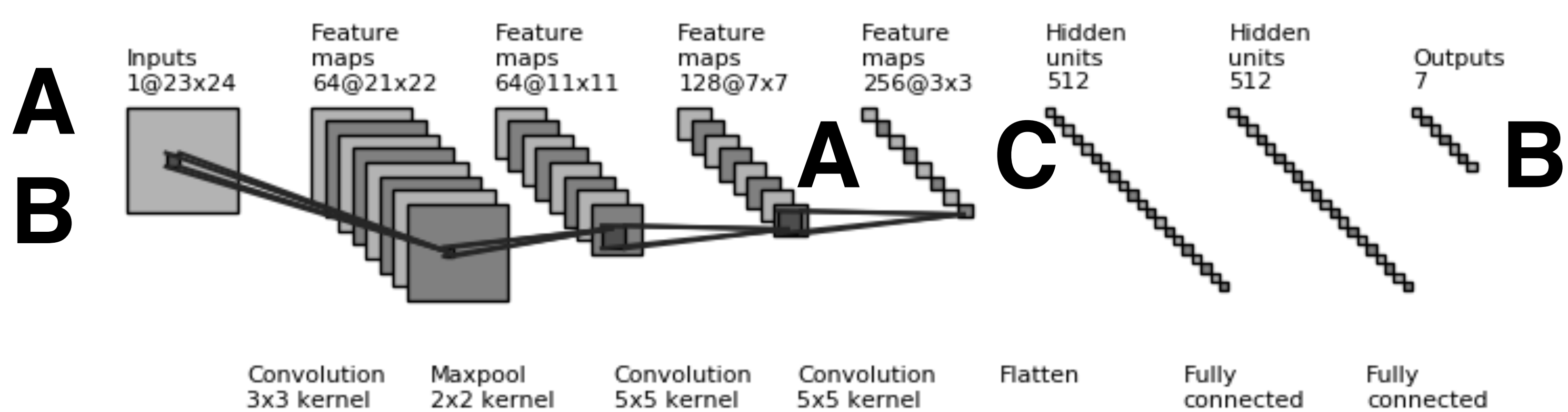
telegraph.co.uk

Style transfer

Deep Dreams Gatys et al. 2015

Visualization for interpretability

A. Activation Maximization
   - 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

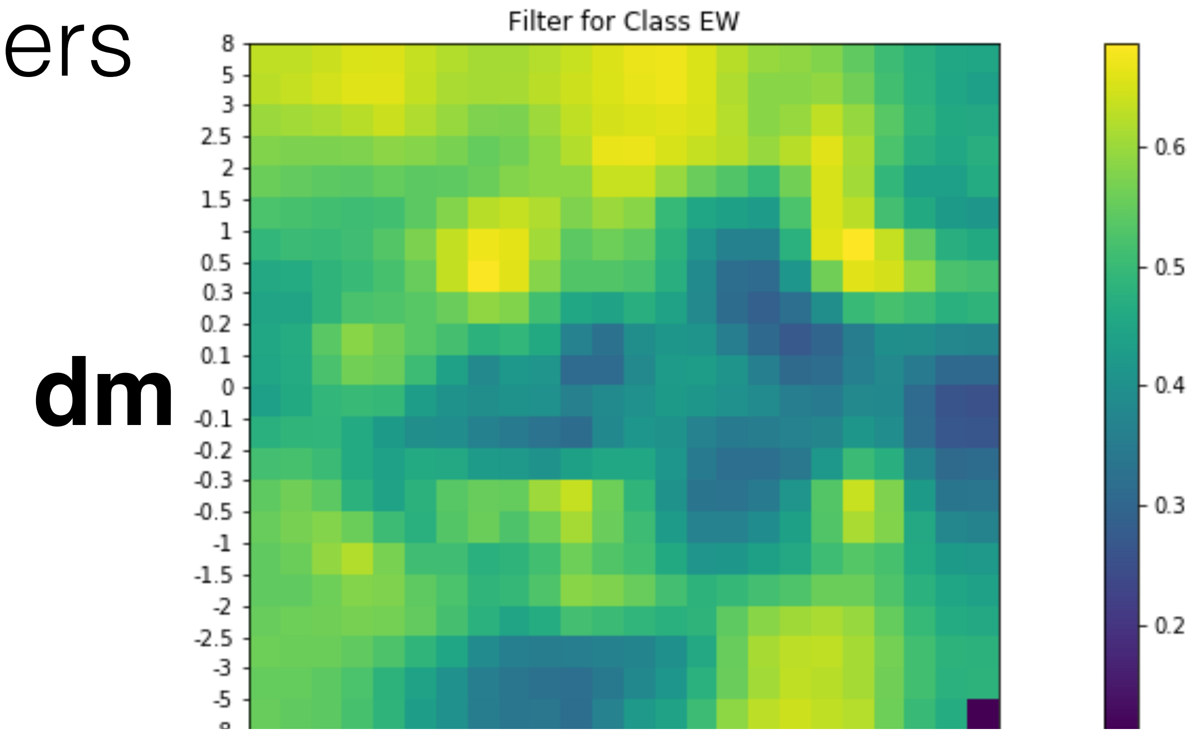


https://raghakot.github.io/keras-vis/

dmdt and activation maximization experiments

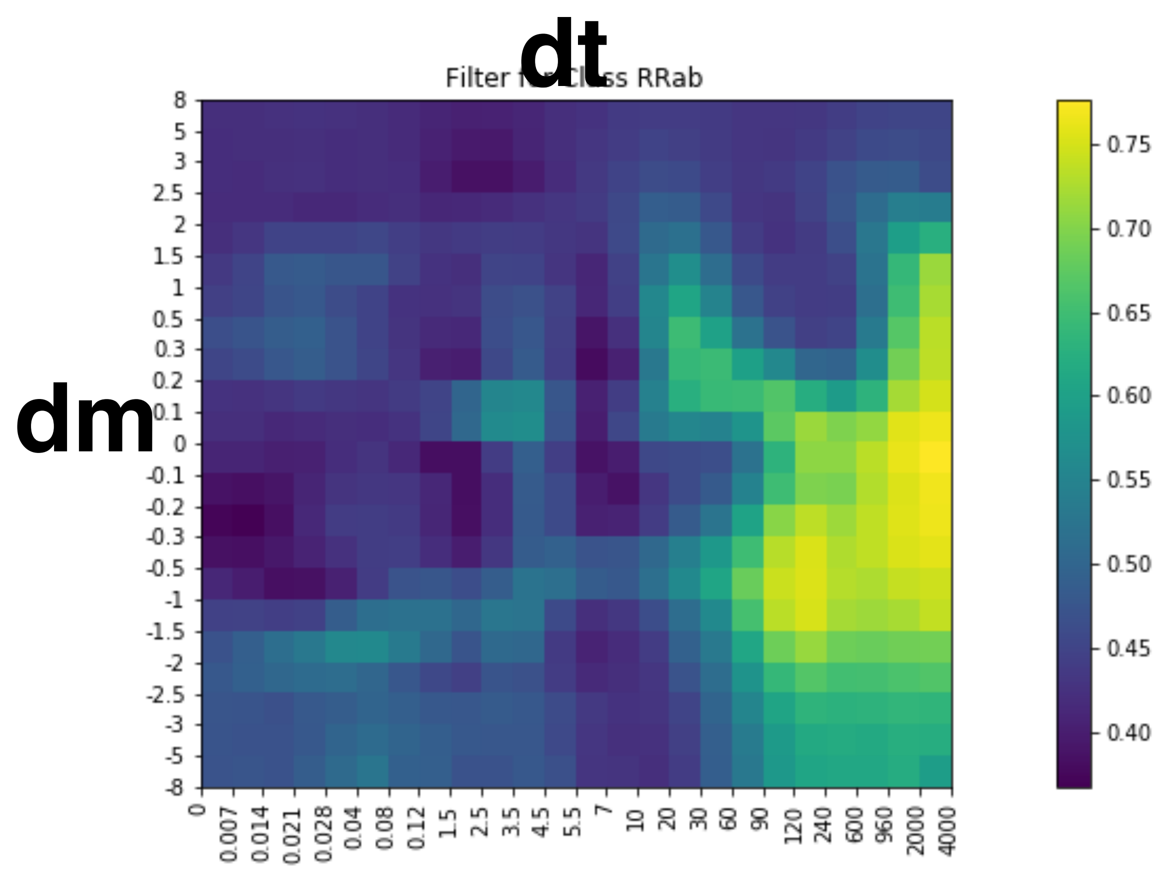
Activation of later filters

Attention more to what is not present!



EW

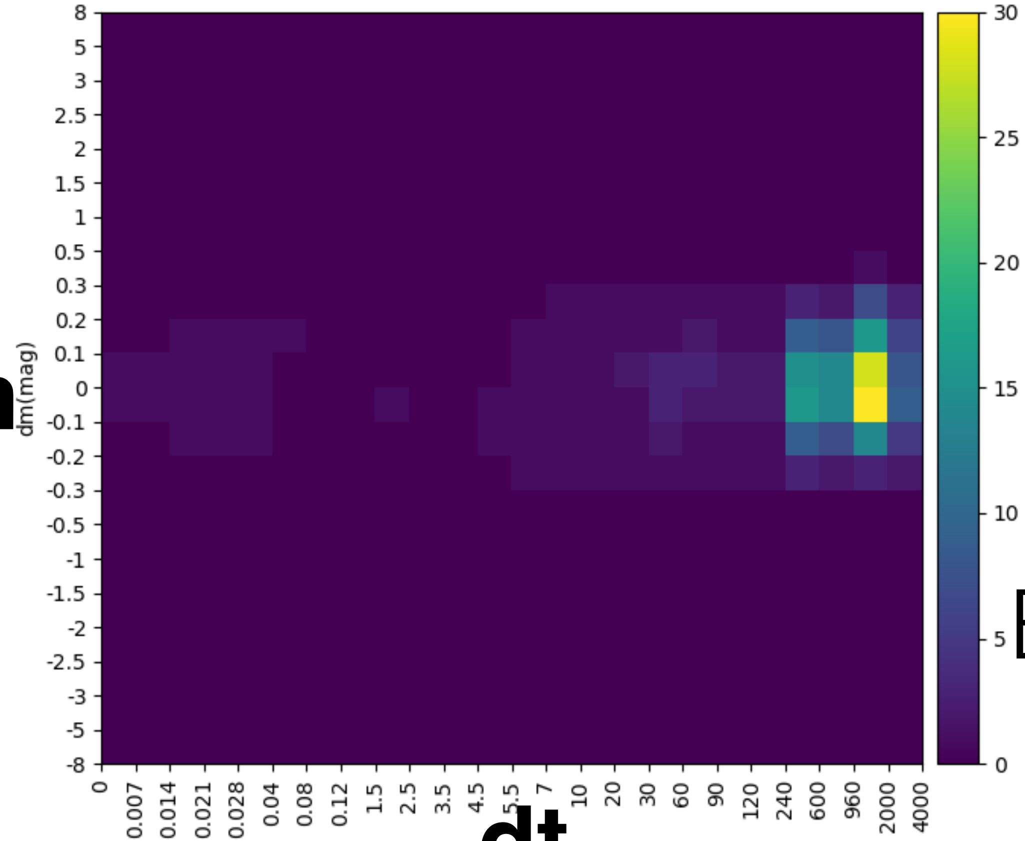
Attention more to longer durations



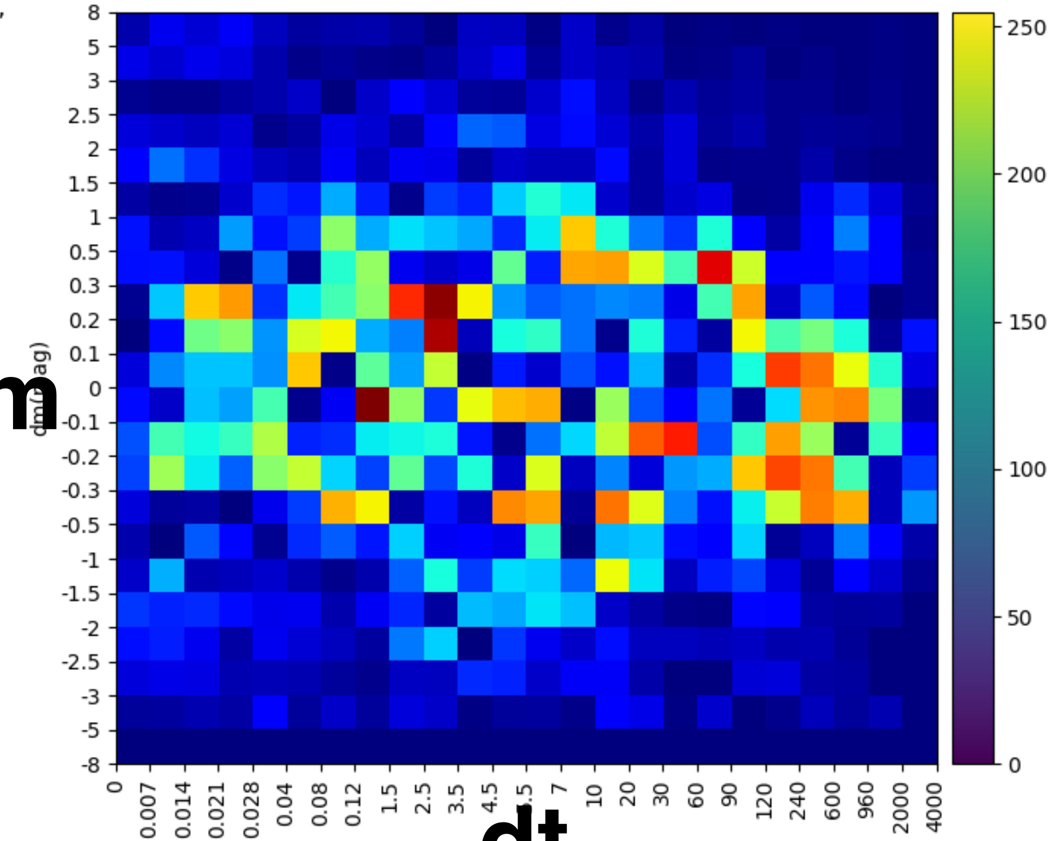
RRab

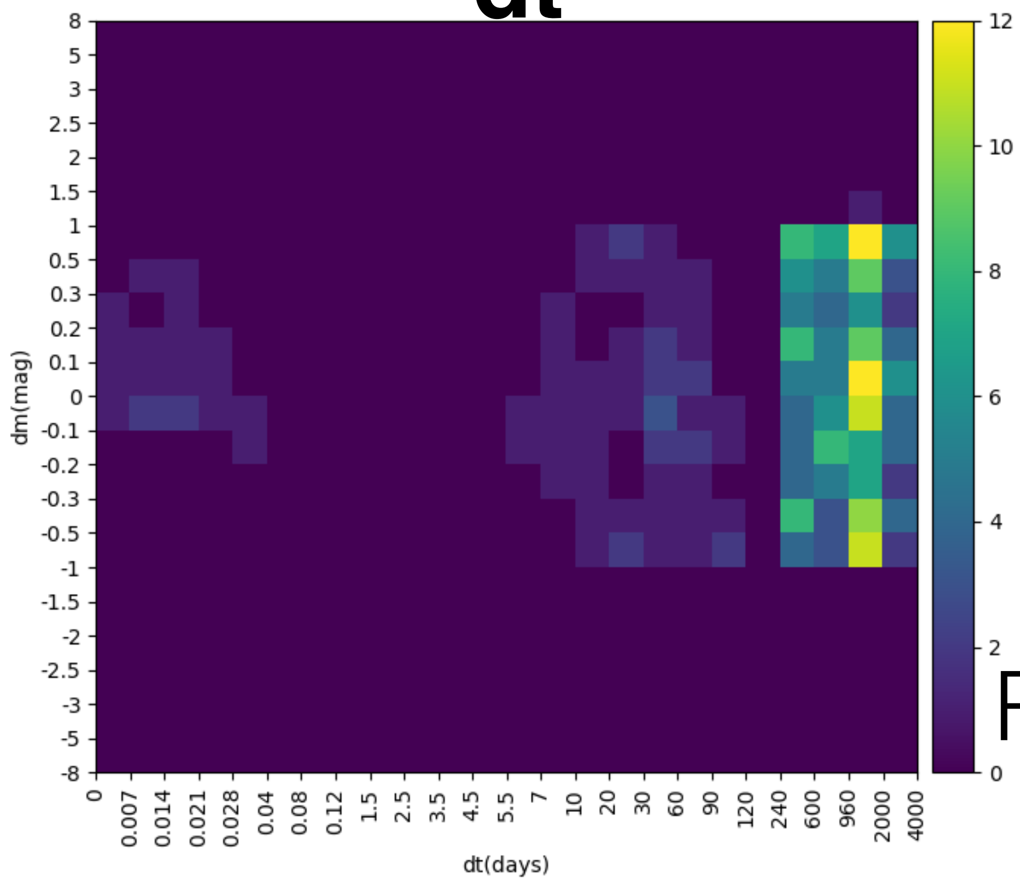
Meet Gandhi

Saliency (attention)

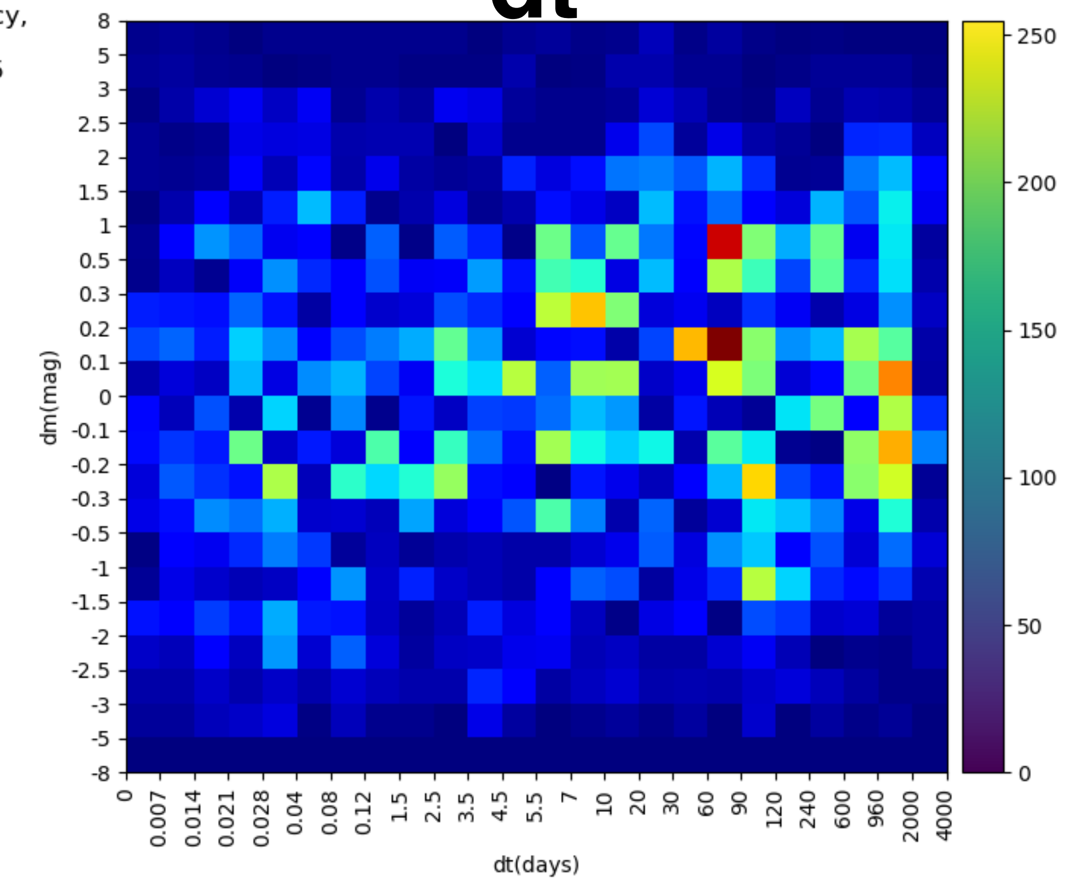


Class: EW, Saliency,
X_id: 1009
Pred_Prob: 0.9913

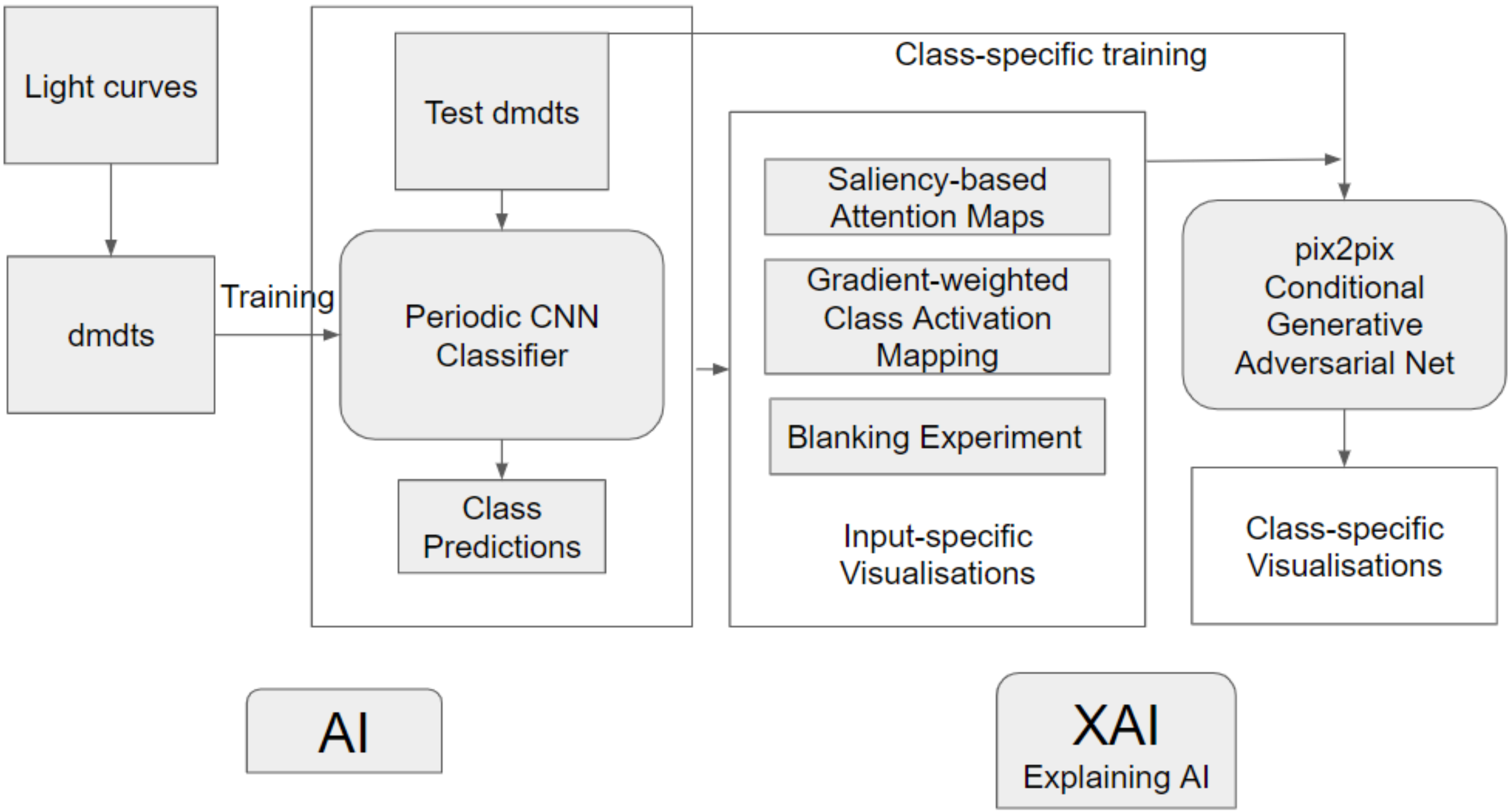




Class: RRab, Saliency,
X_id: 1067
Pred_Prob: 0.9795



Meet Gandhi



Light curves
dmdts
Training
Test dmdts
Periodic CNN Classifier
Class Predictions
Class-specific training
Saliency-based Attention Maps
Gradient-weighted Class Activation Mapping
Blanking Experiment
Input-specific Visualisations
pix2pix Conditional Generative Adversarial Net
Class-specific Visualisations
AI
XAI
Explaining AI

Meet Gandhi

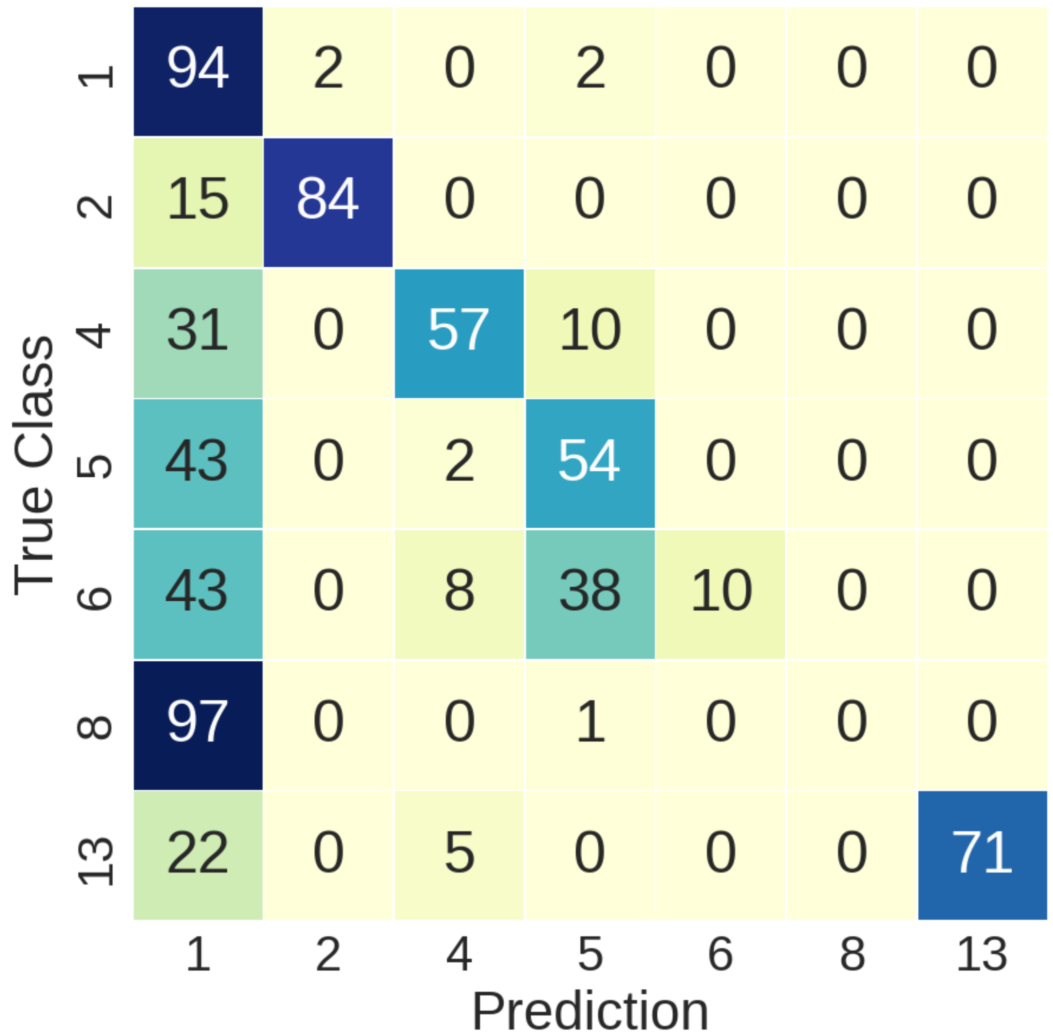
Summary

It has become easy to apply DI to astronomy data
There are many low-hanging fruit
Data massaging is required
So is formulating the problem correctly with domain knowledge

Applications to biology are also waiting to be exploited
Larger number of hurdles due to deidentification issues
Also of data fusion

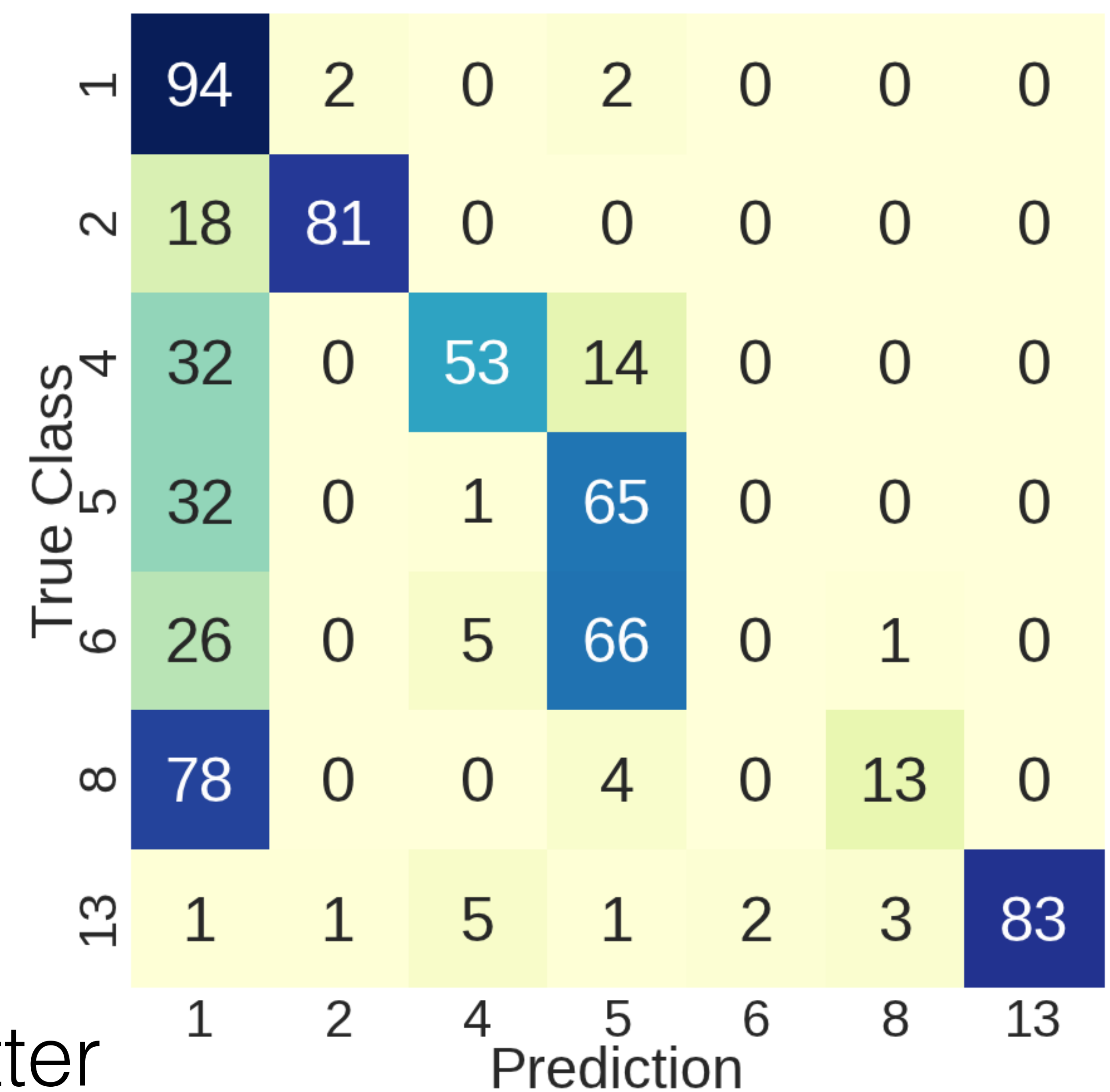
Interpretability and reproducibility are critical

Random Forest using standard features



no features
no dimensionality reduction
comparable results

Convolutional Network

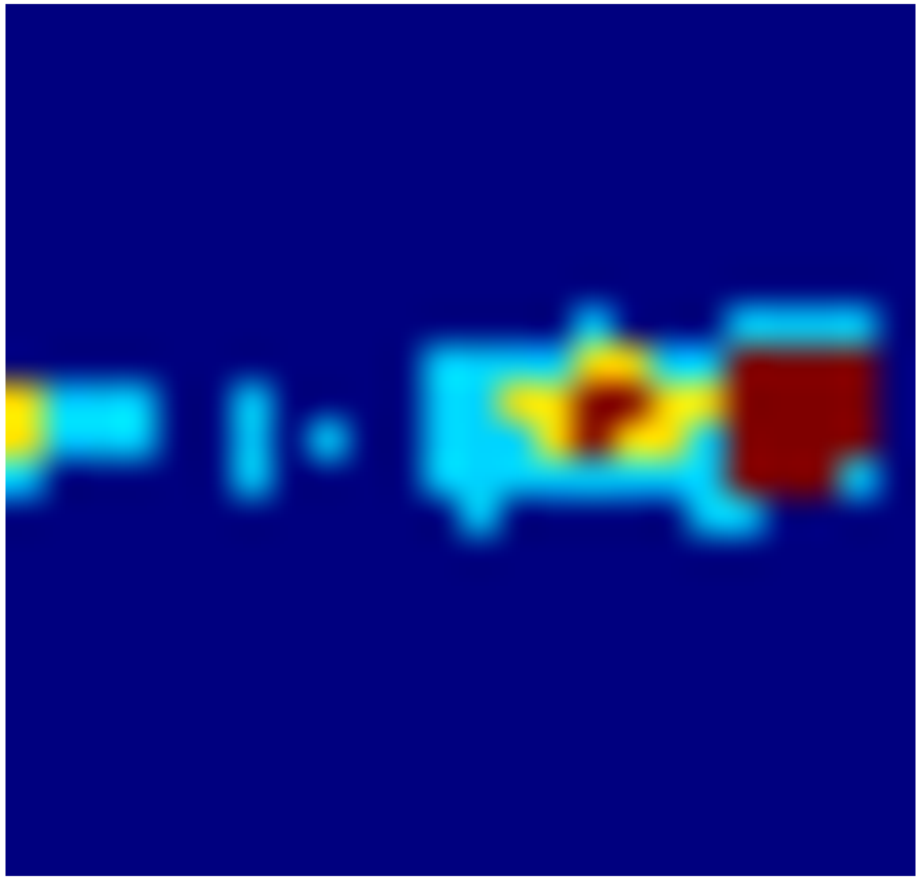


1	EW/EB
2	EA
4	RRab
5	RRc
6	RRd
8	RS CVn
13	LPV

Binary probabilities are better

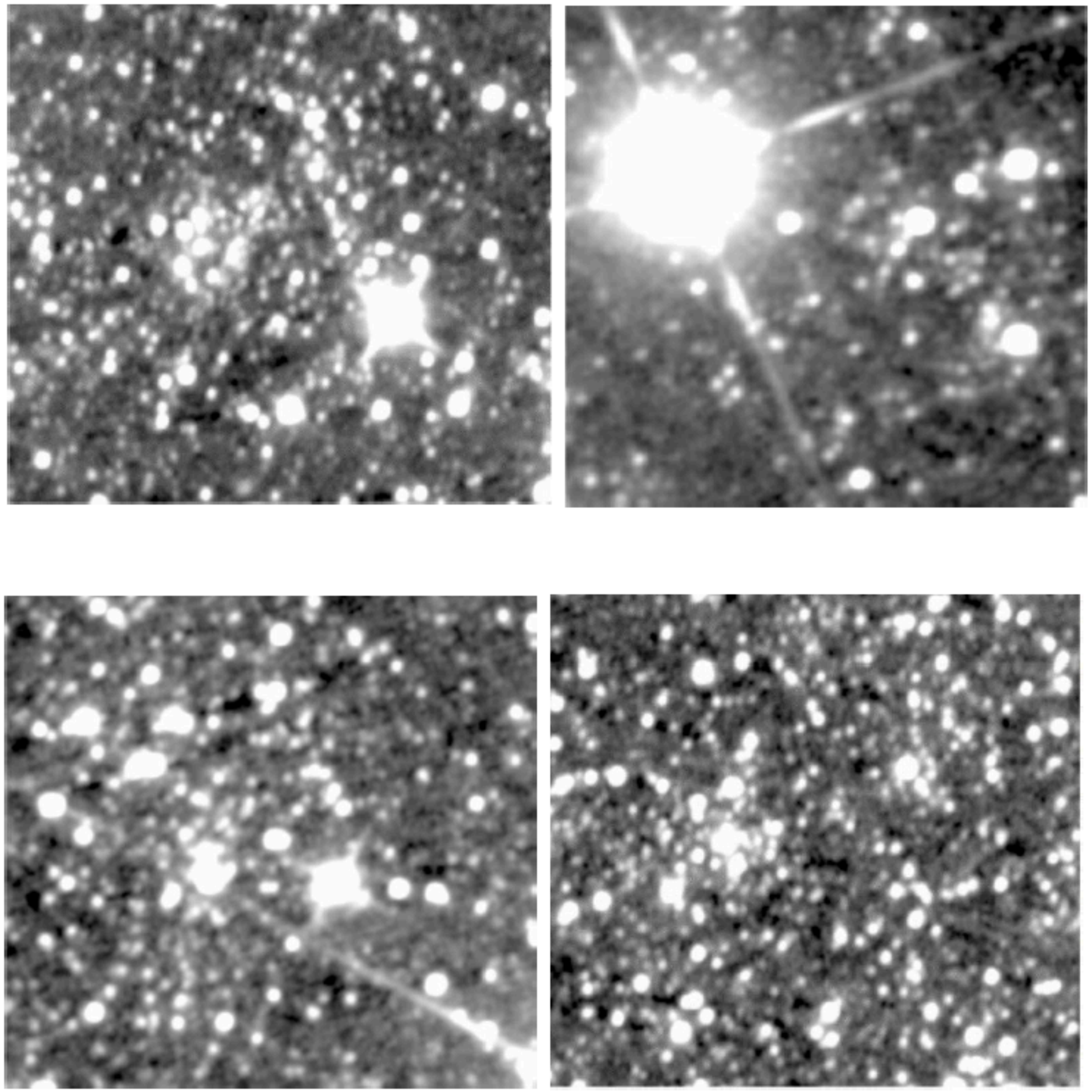
- (2) Astro background
  - Surveys needing classification, in real-time
  - fewer follow-up resources, finding rare events
- (1) traditional ML
  - feature extraction and dimensionality reduction
- Deep learning in astro
  - (1) streaks
  - (1) RB using triplets
  - (2) TransiNet
  - (1) GW
  - (1) Robopol
  - (1) Clusters of galaxies
  - (3) time series - dmdt
  - (3) RNNs
- (1) Good training set
- (5) Interpretability
- Deep Learning in Biomedicine
- (2) Issues with deidentification, holes, …
  - (3) Lung
  - (1) Pancreatic
  - (1) Breast

EW/EB separation?



Two separate backgrounds emerged for class 1

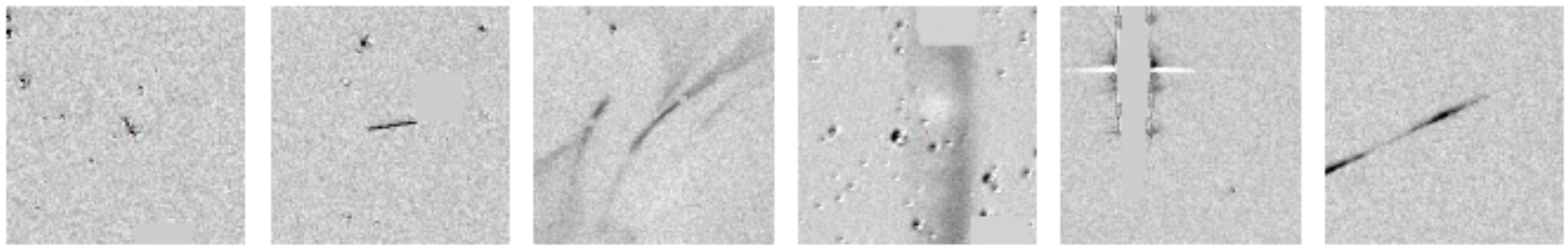
Detecting clusters of galaxies (in infrared)



Ouns El Harzli
Simona Mei
James Bartellt
SG Djorgovski

Identifying streaking asteroids

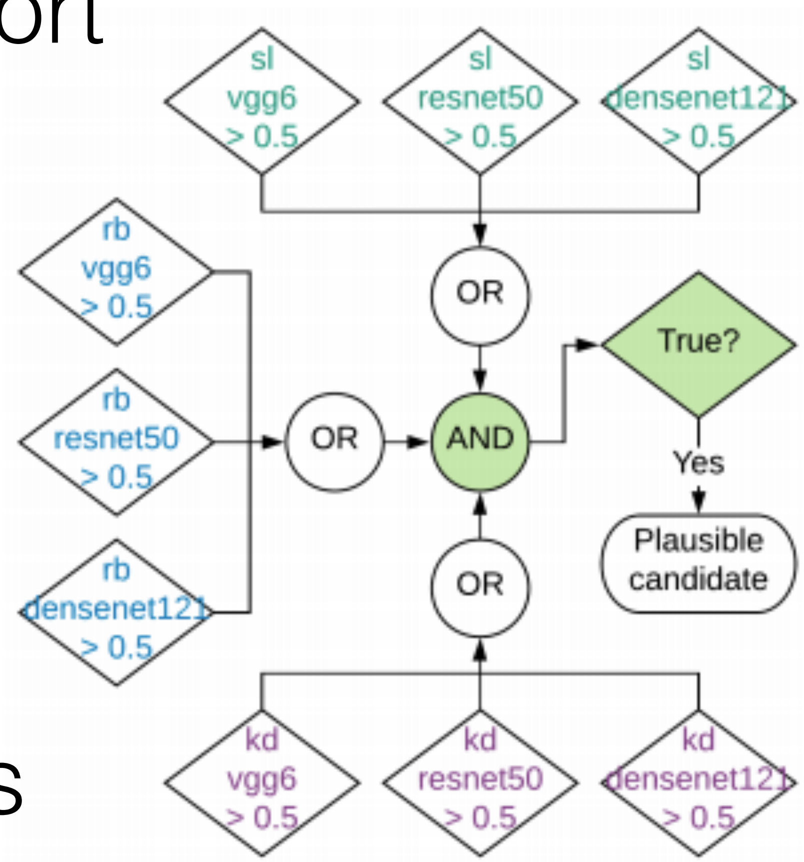
DeepStreaks: identifying FMOs in ZTF data 5



(a) Bad subtraction (b) Cosmic ray (c) "Dementor" (d) "Ghost" (e) Masked star (f) Satellite trail

Figure 4. Examples of different classes of bogus streak detections.

long/
short



real/
bogus

keep/
ditch

Figure 2. Decision logic used by **DeepStreaks** to identify plausible streaks. The problem is split into three simpler sub-problems, each solved by a dedicated group of classifiers assigning real vs. bogus ("rb"), short vs. long ("sl"), and keep vs. ditch ("kd") scores. At least one member of each group must output a score that passes a pre-defined threshold. See Section 2.1 for details.

combiner

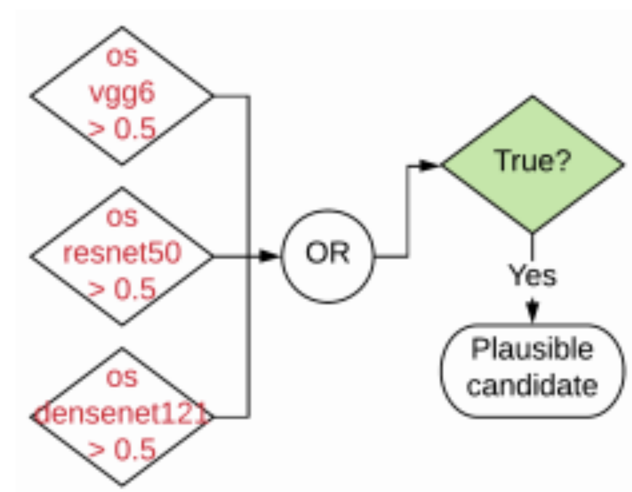
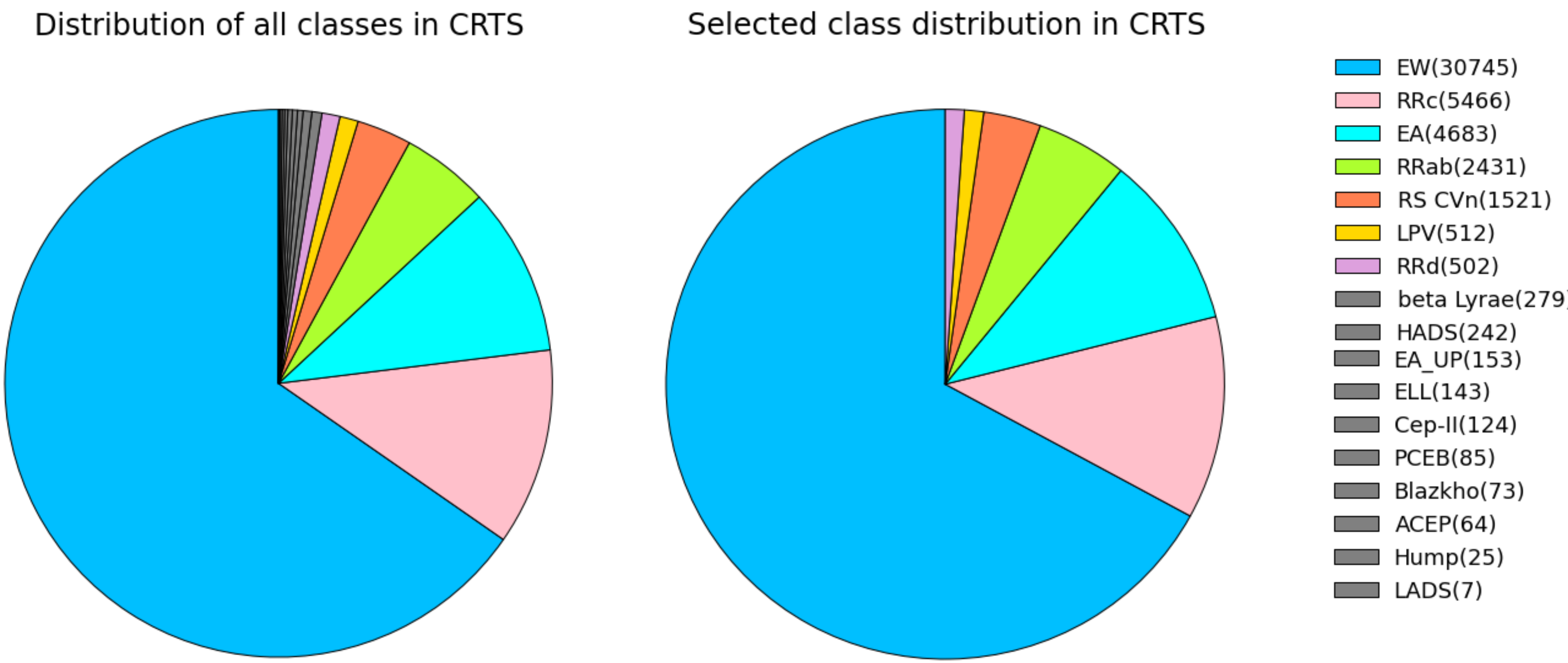


Figure 3. Decision logic to identify plausible streaks used in the one-shot ("os") classification approach. See Section 2.1 for details.

Duev, Mahabal, … 2019
arxiv:1904.05920

(See Dima's talk for details)

50K Periodic Variables from CRTS

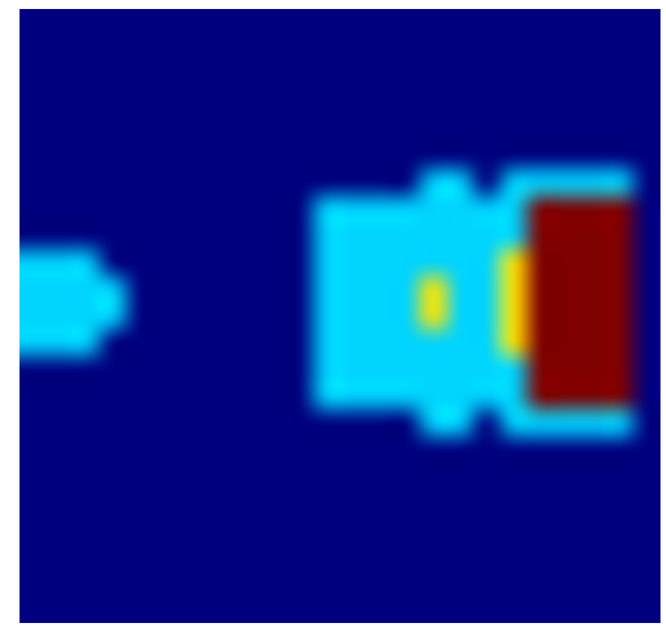


Drake et al. 2014

7 classes with at least 500 examples

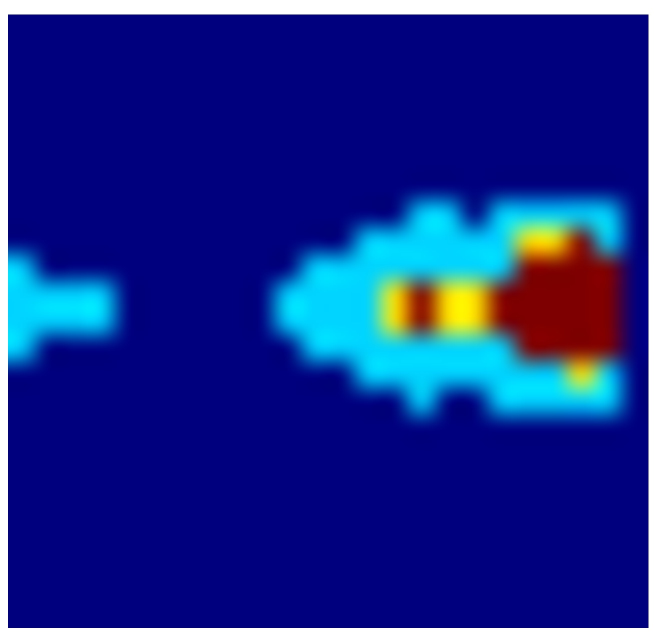
EW

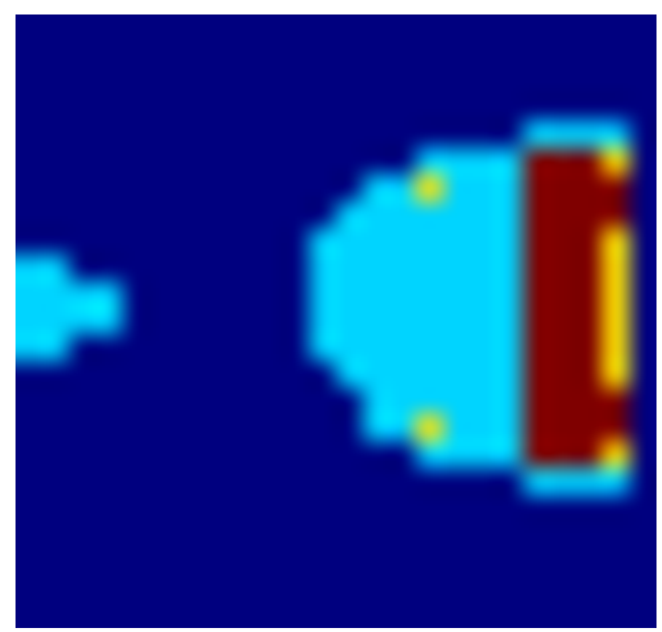
EA



RR

RS CVn





LPV

Kshiteej Sheth

medians