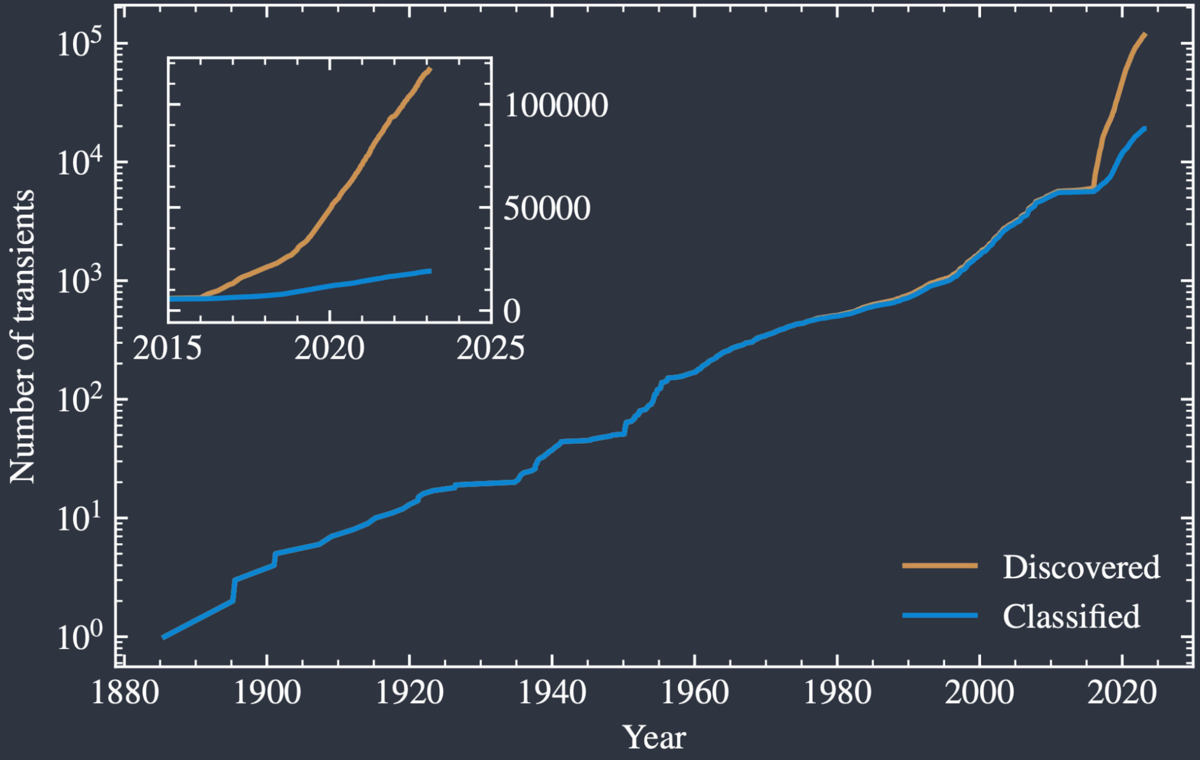
Sifting supernovae with deep-learned source classification

Tom Killestein
University of Turku

Growing survey capacity -> growing transient yields



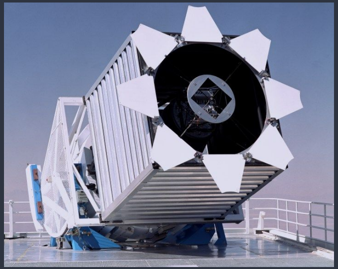
Massive step-changes in gradient from new surveys coming online.

128,000 transient discoveries as of this weekend

Rapidly-widening gap between discovery and classification

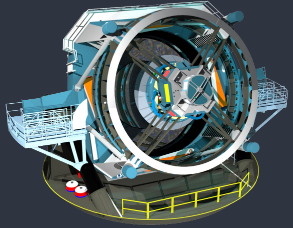
The deluge of data

> Astronomy is now among the most data-heavy sciences.



SDSS (1990s)
200GB / night

GOTO (now)
1TB / night



Rubin Observatory (2024-)
20TB / night

SKA (2030s)
160TB / second

> All of the above are reliant of machine learning to manage the data volumes involved!

Difference imaging -> generating transient candidates

Many algorithms, none perfect: complex PSF + background + norm matching

Intrinsically high false-positive rate, and many defects in process.

e.g. Alard and Lupton (1999), HOTPANTS (Becker 2006), ZOGY (Zackay+2016), SFFT (Hu+2021)

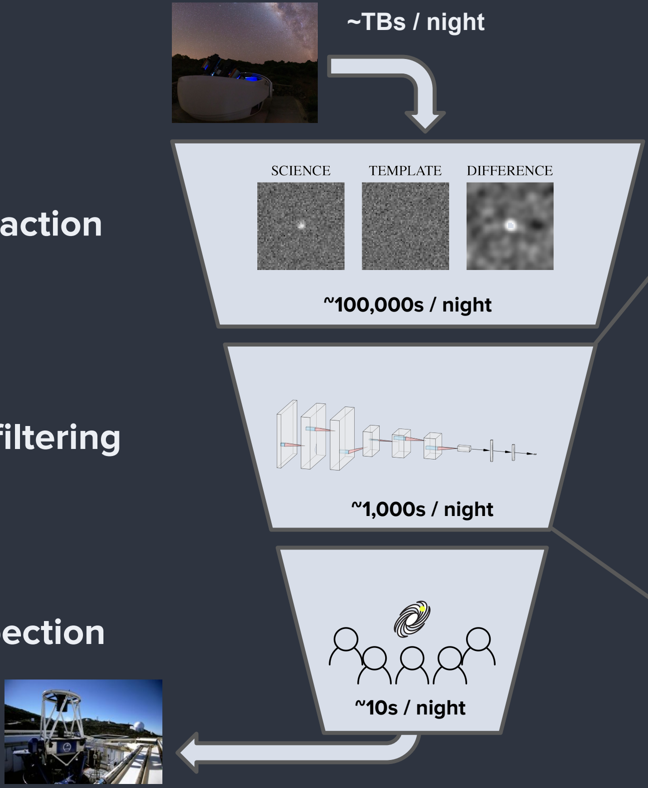
Time-domain astronomy: the real-bogus problem

How can we efficiently sift the deluge of data from current/upcoming transient surveys to maximise science yield?

Image subtraction

Real-bogus filtering

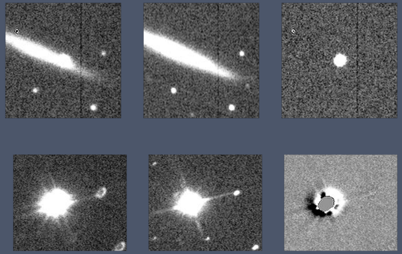
Human inspection



Real-bogus classification

Use ML/DL to automatically identify artifacts in difference images based on extracted properties.

Maximise completeness whilst simultaneously minimising false positives - avoid overwhelming human vetters!



SN2021hiz
(real, score 1)

Bad subtraction
(bogus, score 0)

see e.g. Bloom+12, Gieseke+17, Duev+19

Evolution of techniques

Blink
comparator,
eyeballing

Decision trees
(ESSENCE,
LOSS)

CNNs
(ZTF, ATLAS,
GOTO, et al.)

1940

1990

2000

2010

2020

Difference imaging
e.g. Alard
(1991)

Random forests
(iPTF, DES,)

Monolithic vs modular surveys: a range of challenges

Monolithic


e.g. PanSTARRS, ZTF, VRO

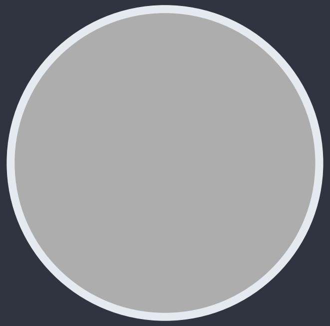
Data generated from one single, large, static instrument, yielding largely homogeneous dataset

Modular


e.g. ATLAS, GOTO, ASAS-SN

Data generated from multiple similar instruments: but each with subtly different characteristics (PSF, backgrounds)

Two approaches to dataset curation



Small, but pure (human-labelled) datasets

- Low label noise
- Low diversity of examples

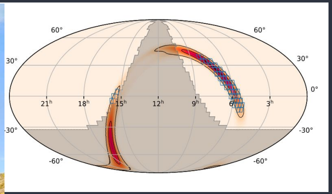
Large, but impure (algorithmic) datasets

- Higher label noise
- Higher diversity of examples

The Gravitational-wave Optical Transient Observer

Each telescope/node:
8x40cm astrographs, combined FoV of **40 square degrees** - reaches **L ~ 20.5 in 180s** enabling rapid and deep surveys of GW localisation regions in real-time.

Real-time response to GW/GRB/neutrino alerts, 2-3 day cadence over whole sky.



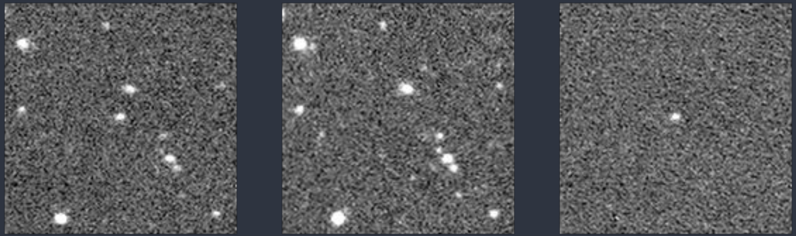
Rapidly fading (~hours) GW counterparts necessitates **real-time, accurate** source classification - only possible with high-performance ML/DL.

The significant data volume generated by GOTO (and other surveys) provides an ideal **data-rich proving ground** for novel ML/DL algorithms applied to a diverse range of data modalities.

gotorb: an algorithmic approach to SoTA RB classification

Real targets: Asteroids

Typical contaminant in transient searches, provide realistic difference image detections across a range of magnitudes, and intrinsically have instrumental PSF



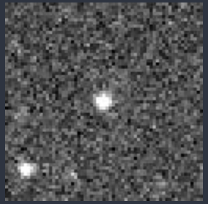
Bogus: randomly-sampled detections

Given the significant (99%) rate of artifacts in difference imaging, once removing asteroids and variable stars, we are left with a diverse range of negative targets

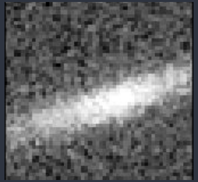
Generate a 400.000 example dataset in <24h with this method

Optimising for recovery of transients

“Data augmentation”: increasing the available training data by applying salient transformations.

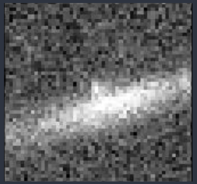


Randomly select minor planets in each image



Identify a galaxy near to the MP and extract an offset stamp

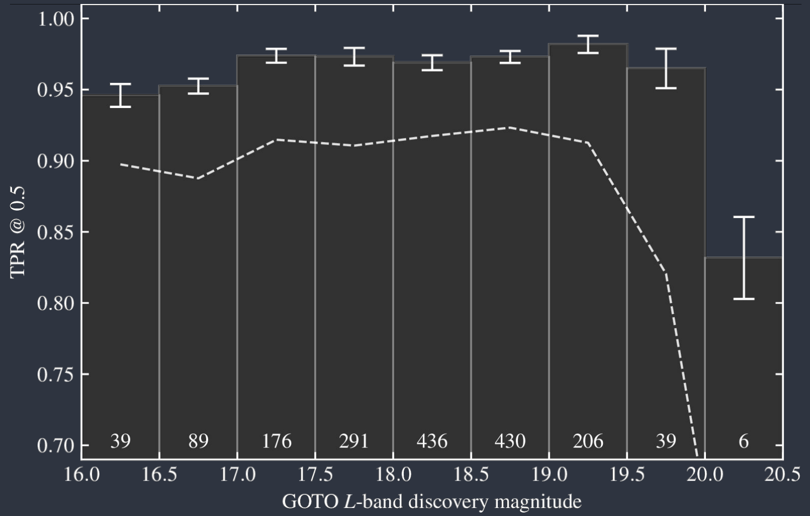
Realistic PSFs, fast, and configurable injection.



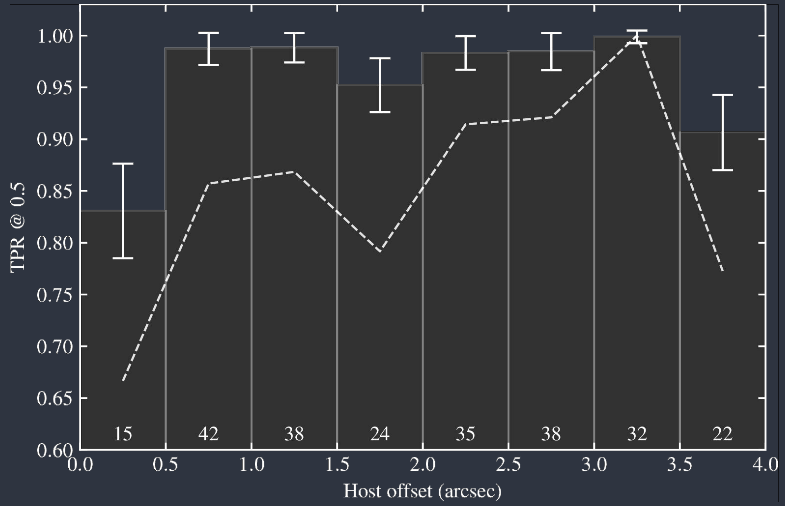
Sum galaxy stamp and minor planet to get a synthetic transient

Significantly improved recovery of transients

Test the classifier on >900 spectroscopically-confirmed transients recovered in the GOTO prototype phase — performance better with new instrumentation and pipeline upgrades.



Significant increase in recovery of faint transients - excellent prospects for kilonova recoveries in local Universe.



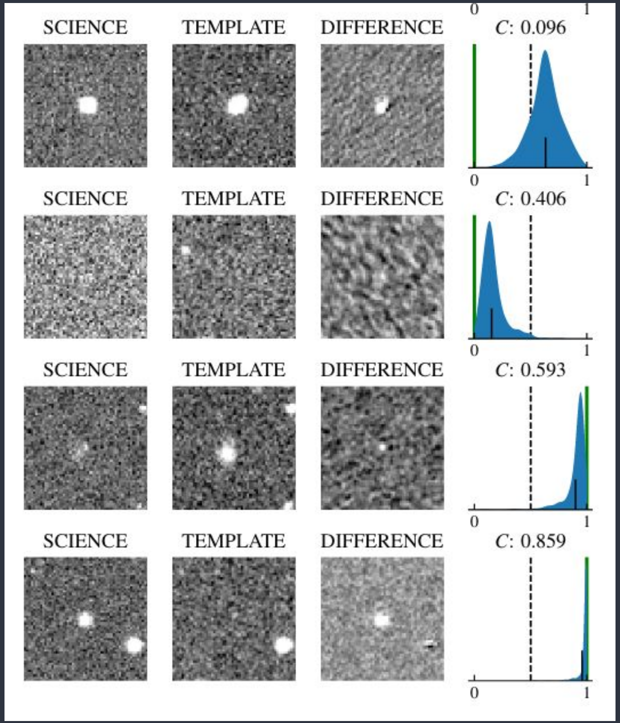
Uniform coverage of host-offset space leads to **boosted recovery of nuclear transients**

Uncertainty quantification: Bayesian NNs + others

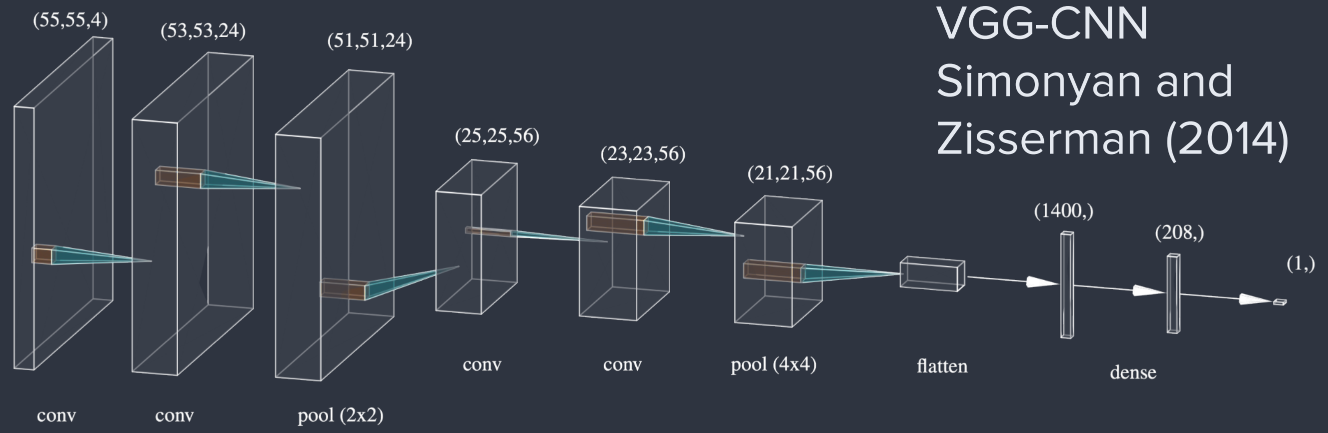
> Many techniques to estimate uncertainty - still a very open-ended area of study **(Charnock+2020)**

MCDropout, Bayes by Backprop, Variational Bayes, direct sampling

> Cast weights as parameters with prior distribution (usually Gaussian) and sample posterior distributions -> predictive uncertainty.



A commonality: simplistic architectures do the job!



VGG-CNN
Simonyan and
Zisserman (2014)

VGG16-likes

Duev+19
Killestein+21
Makhlouf+22
Takahashi+22

Either (or both!) astronomical images are comparatively simple, or RB classification is not a fundamentally taxing task.

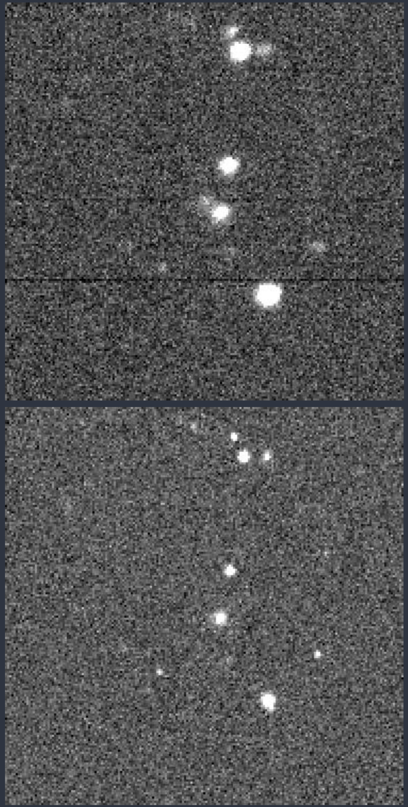
Most of task difficulty doesn't come from representative power 🤔

Just when you're done: concept drift and MLops

For all real-time inference, danger of **concept drift**: your models no longer operating on the dataset they were trained on.

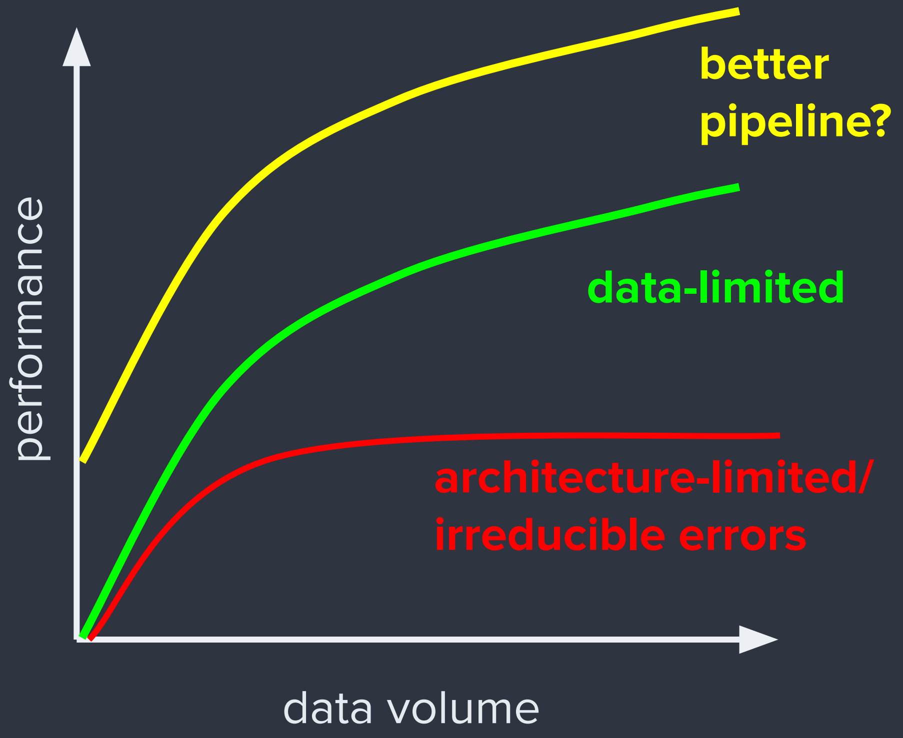
- **Real-world:** changing consumer preferences, seasonality, 'spikes', world events
- **Astronomical:** changes to underlying pipeline, degradation of optics, changes to survey strategy,

BUT: dynamic models at odds with things like well-defined selection functions in surveys. Challenging, both to identify, and mitigate - but this is part of the MLops lifecycle.



Future: where are we on the learning curve?

- Can we keep attaining ever-better performance with more data? Is there a benefit to scaling up?
- How much longer will we be able to keep using comparatively basic architectures?
- How much can e.g. new subtraction algorithms help with reducing false positives?



NB: issues with evaluation, both in terms of truncation of metrics, and dataset purity

Too many candidates!

Too many real transients!

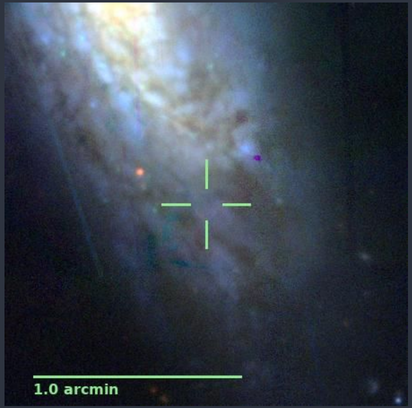
Contextual classification

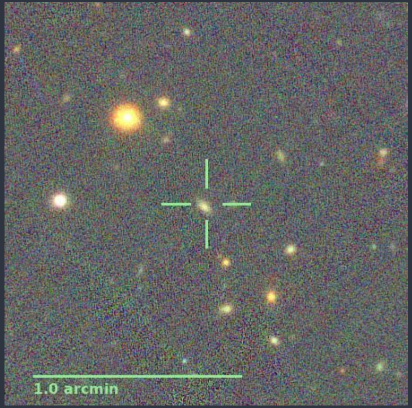
Learning what we can about a transient based on it's context

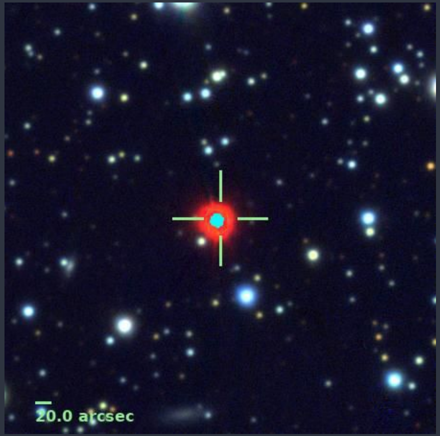
Galaxy association: redshift -> distance modulus -> absolute magnitude, physical offset

Coincident with active galaxy: likely AGN flare, nuclear transient

Underlying point source: M dwarf flare? CV? Distant/compact host?







How do we automate this process of inference, and provide maximally informative predictions to do our science with?

Challenges

Catalog information:

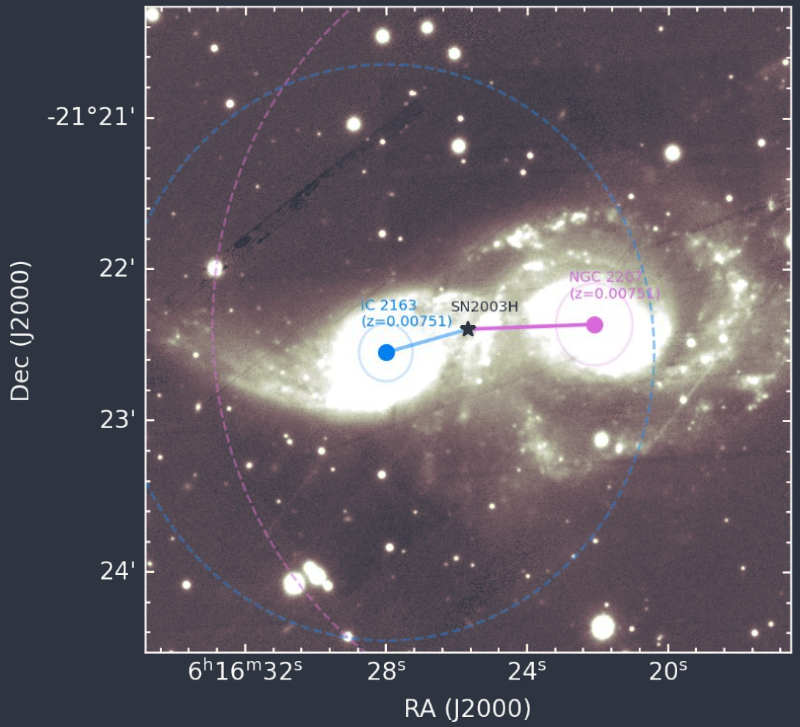
Incomplete, inhomogeneous, and sometimes incorrect.

Host association:

Often ambiguous, reliant on the catalog information often

Demands of real-time inference:

Any method needs to keep up with data volumes



Methodologies

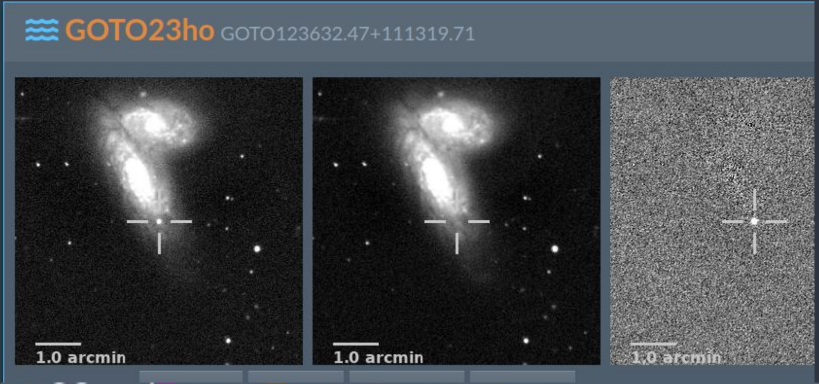
Create a ‘knowledge base’ of astronomical catalogs, make them queriable in real-time, and write logic for de-duplication, aggregation of information, priorities of catalog information, and a ‘decision tree’ for turning context into a classification.

Data-driven classification process – beginning to see the ‘features’ from this process be used as inputs to hybrid models.

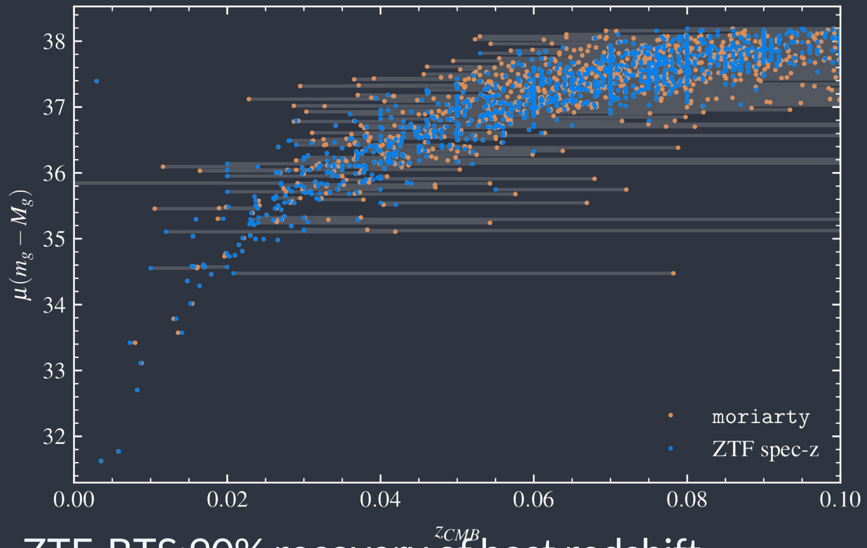
Gold standard in this area: Dave Young’s **sherlock** code.

moriarty

Enhanced host association algorithms, different technology stack.



Professor James Moriarty (2 weeks, 3 days ago) GOTO123632.47+111319.71 is likely associated with the B=11.32 mag galaxy NGC 4568 in the gladeplus_galaxies, ps1_stackobjectview_minimal catalogs (65.12" away). Probability of connection 98.829% -- host is at 18.4 Mpc (z=0.0041+/-0.0001), implying a transient absolute magnitude of -15.26 and sky-projected offset of 5.81 kpc. This transient is spatially coincident with the cluster Virgo cluster (20.0 Mpc)



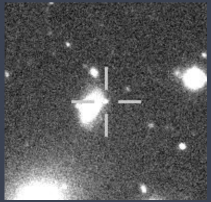
ZTF-BTS:90% recovery of host redshift, outperforms simple cone search on challenging gap transients of Dong+2021

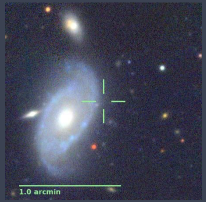
Public release next year, alongside paper

Documented in my PhD thesis

Unifying image and catalog-based data with hybrid architectures

Image-only:





+ Always available (discovery image)
- Limited by image quality/resolution
- Apparent type != actual type

Catalog-only:

```
ra,dec,gmag,...
32.042,12.321,19.042,...
31.987,12.021,18.428,...
32.321,12.124,21.421,...
```

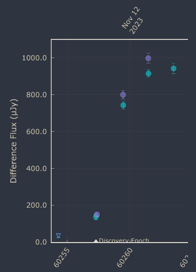
Salient information already extracted

+ Far deeper than survey images
- Incomplete (galaxy catalogs)
- Incorrect (misclassifications)
- Inhomogeneous (variable coverage)

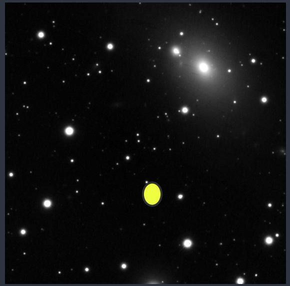
How to optimally combine these to get the benefits of both in a ‘sensible’ way?

Cross-modal datasets -> novel challenges

Just as humans perform their inferences based on all available information, our algorithms should too.



Rapid rise?



Remote from potential host?

Ca-rich transient?

Worth following up?

Multi-modal learning at forefront of current CS/ML literature: potentially promising directions given the constrained reasoning space of astrophysics problems.

Common limiting factor: dataset purity (Bayes error rate)

> Ambiguity between classes = source of irreducible error to classifier

e.g. distinguishing AGN activity vs TDEs

> Hierarchical schemes suffer from ‘dwindling data’ problem

> ‘Absolutely’ pure datasets tend to be small, thus challenging to apply data-hungry DL to

Citizen science for time-domain astrophysics

Generating large datasets by distributing work:

- High-quality labels via wisdom of the crowd
- Uncertainty quantification
- Outreach and public engagement

ZOONIVERSE

Many projects using this approach:

Transients: SN Hunters (PS), Superluminous Supernovae (Lasair/ZTF),

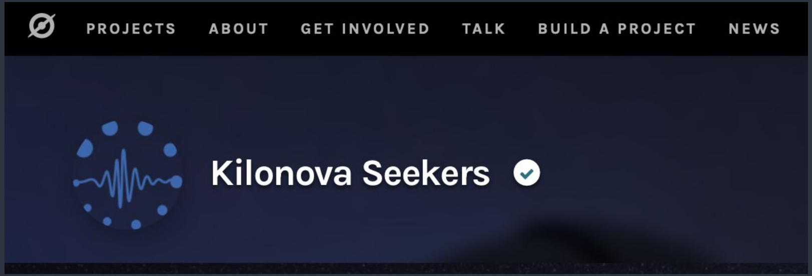
Variables: Gaia Vari, Citizen ASAS-SN

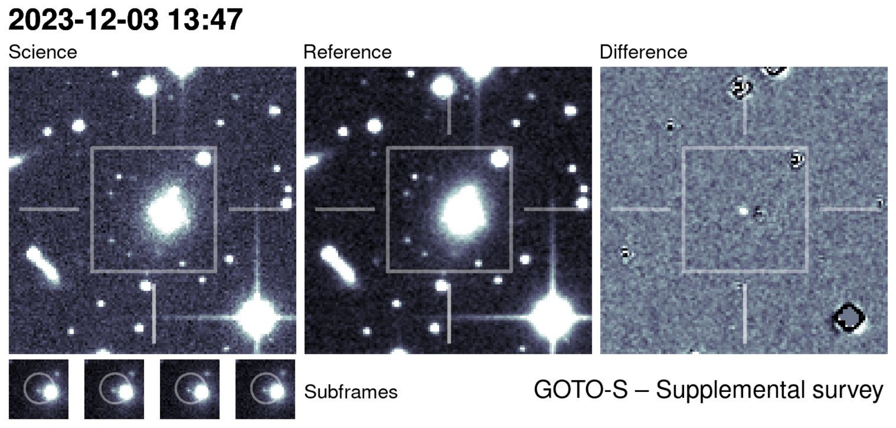
Kilonova Seekers

True real-time citizen science and public engagement – GOTO candidates at 3h latency, largely unfiltered.

1800 volunteers, 470k classifications, 19 transient discoveries on TNS

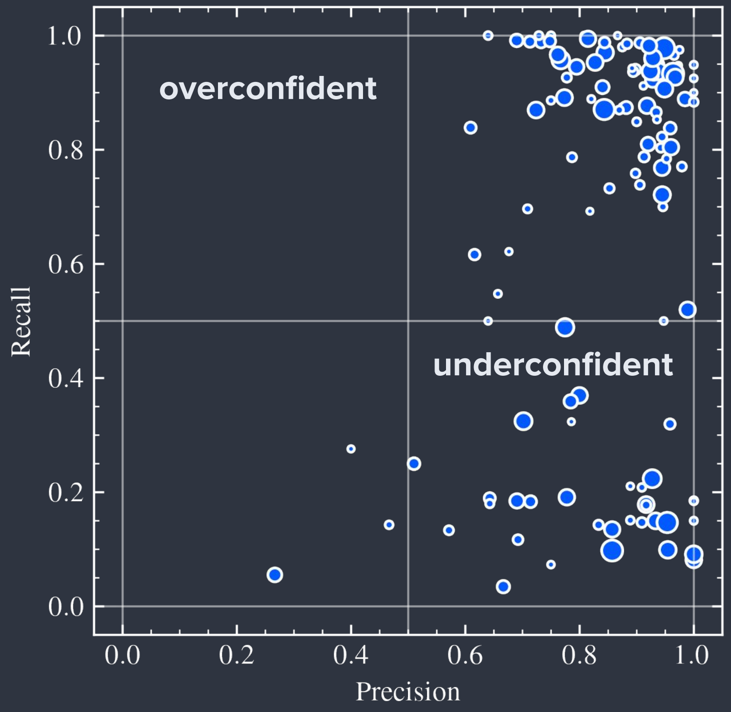
Active learning + multi-class workflows under development

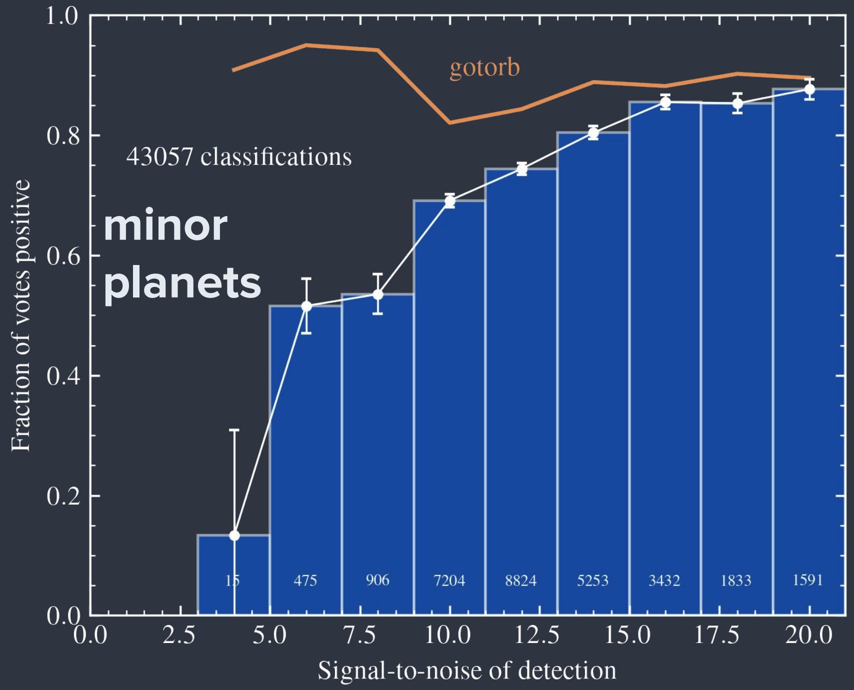




kilonova-seekers.org

Understanding the human factor better





Killestein, Kelsey, GOTO et al. in prep.

Brokers - aggregation and value added classifications

Lasair

Alerce, AMPEL, ANTARES, Babamul, Fink, Lasair, Pitt-Google

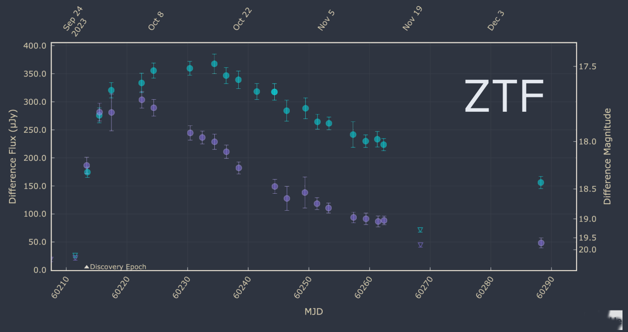
Currently focused on ZTF, but ready for VRO-LSST and all the science waiting there.

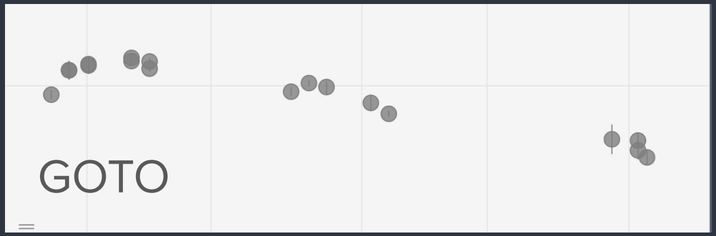
Rich opportunities for time-domain ML at the largest scales!

Deep-learned filters!
e.g. **Lasair** now has NEEDLE (X. Sheng), **Fink** T2 (Allam Jr.+2021.)

Looking ahead

Cross-survey learning(?)





Huge value in automatically combining data across multiple surveys: complementary constraints, colour evolution, augmenting cadence/depth.

But: N surveys = N unique data-streams to reconcile, also limited ‘overlapping’ training data.

Collaboration is key here.

Rubin Observatory

Remarkable data challenge: processing, storage, orchestration, syndication

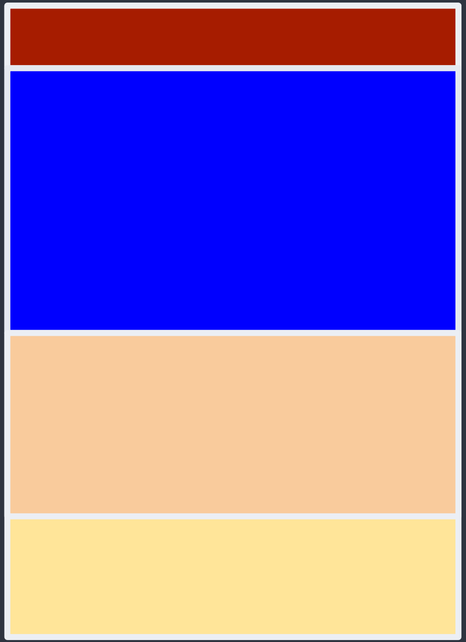
One of the largest datasets thus far in astronomy + volume effects + population magnitude distributions -> step change in discovery rate.

ML/DL is essential - not only for speed, but for accuracy.

Even at Rubin scale, CNNs probably good enough for RB classification

10 million alerts per night [1]

Assuming current-performance RB classifiers, similar difference imaging performance



100,000 false positives (until a second visit)

Solar system objects

Variable stars, AGN activity

Transients of interest

Many of these at 23rd mag with sparse cadence - beyond our current abilities!

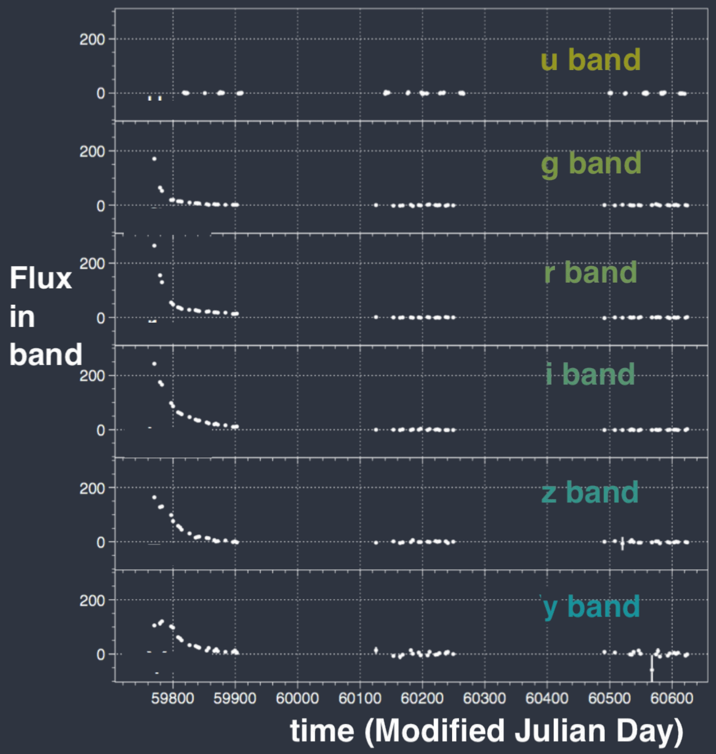
[1] https://www.lsst.org/scientists/keynumbers

The gap between simulation and reality

LSST Data Challenges preparing the community for the deluge of data (PLaSTICC, ELaSTICC)

Driving many innovative algorithms: *avocado* (Boone+2019)

What degree of resiliency to common artifacts do these models have? Domain shift? Ablation of cadence?



The importance of active learning

Finite labelling capacity will be stretched by unprecedented alert volume

Need to identify most valuable objects to label to rapidly augment classifiers trained on simulated data.

But: how (if at all) does active learning bias classifier training set?

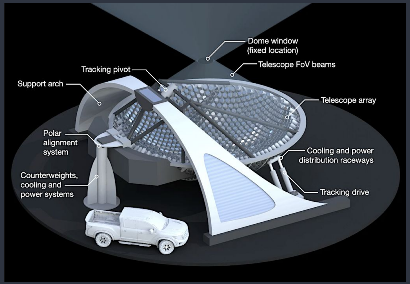
Another step change, closer to home?

Modular surveys acquiring new nodes

LS4

LAST

Argus Array

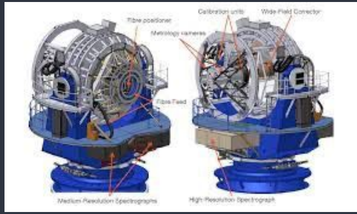


New planned surveys

Large-scale spectroscopic surveys

DESI

4MOST



thomas.killestein@utu.fi

tkilleste.in