Human in the Loop:
Active Learning for Astronomy

Michigan Cosmology Summer School - part 2
8 June 2023 - Ann Arbor, USA

Emille E. O. Ishida
Laboratoire de Physique de Clermont, CNRS - Université Clermont-Auvergne
Clermont Ferrand, France

Acknowledgements

COIN

Alberto Krone-Martins (UCI)
Rafael S. de Souza (Hertfordshire)
and the entire Cosmostatistics Initiative (COIN)

Konstantin Malanchev (UIUC) SNAD
Maria Pruzhinskaya (LPC)
and everyone in the SNAD collaboration

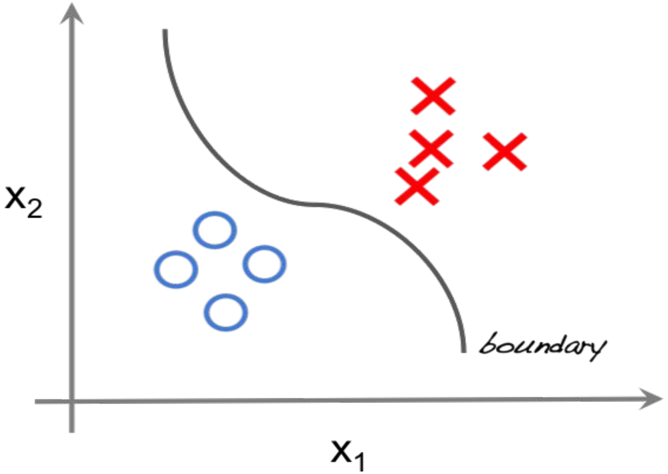
FINK

Anais Moller (Swinburne)
Julien Peloton (IJCLab)
and everyone in the Fink broker

LSST DESC & COIN RESSPECT

Alex I. Malz (CMU)
Mi Dai (JHU)
Kara Ponder
Amanda Wasserman (UIUC)
and all those working in the RESSPECT team

Supervised x Unsupervised

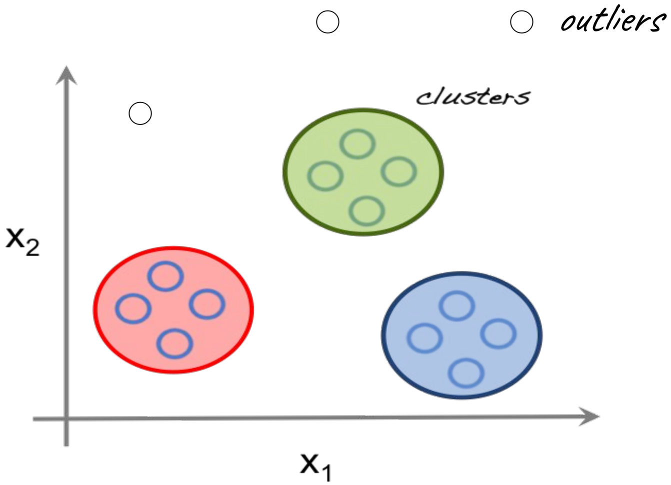


Training sample:

features + labels

Target sample:

features



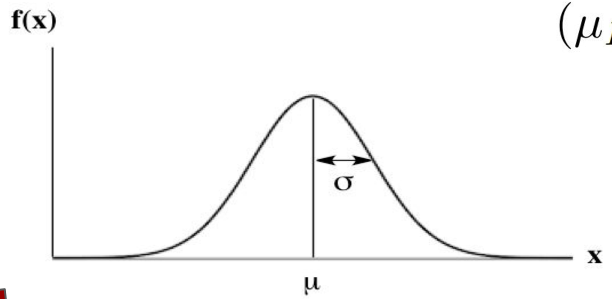
Data sample:

features

Representativeness

Probability distribution, *P*

(μ_P, σ_P)



This is why it works!

Training

$(\mu_{S_1}, \sigma_{S_1})$

Target

$(\mu_{S_2}, \sigma_{S_2})$

In astro, training means spectra…

Ideal Supervised learning situation

Learn

Training sample

Images, colors, light curves, etc.

+

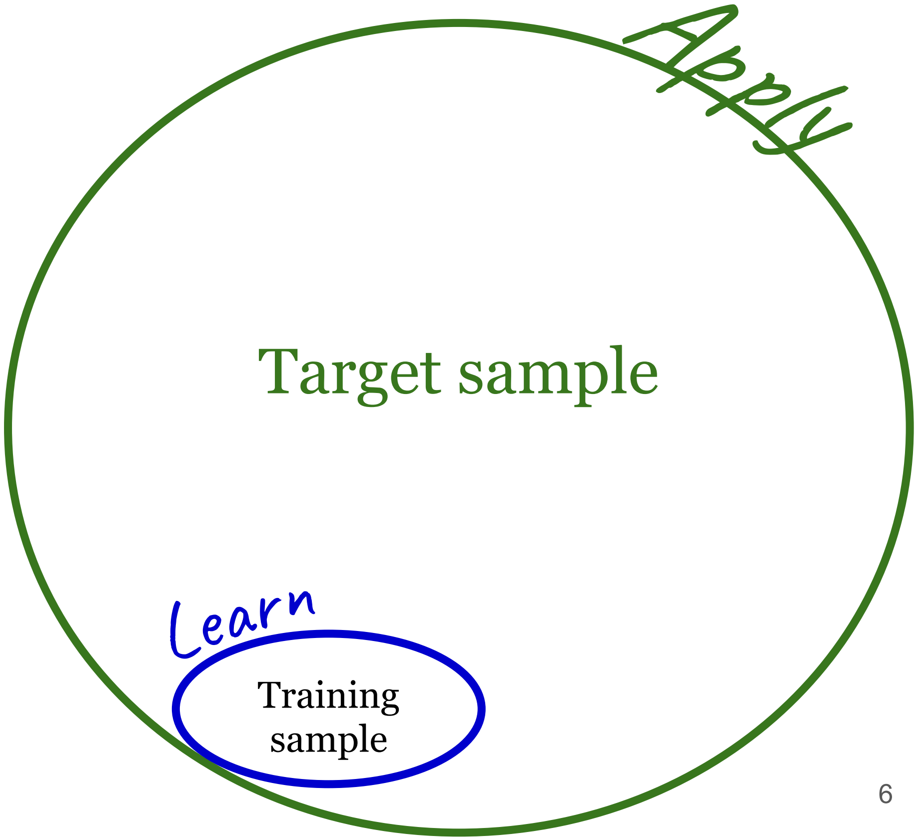
Classes

Apply

Target sample

Images, colors, light curves, etc.

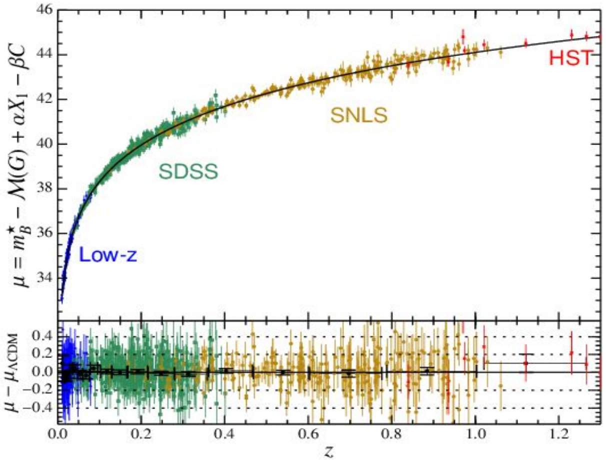
Real astro-learning situation



Type Ia supernova cosmology

Standard candles used to measure cosmological distances

https://supernova.eso.org/exhibition/images/1111_E_549779main_pia14095_full/

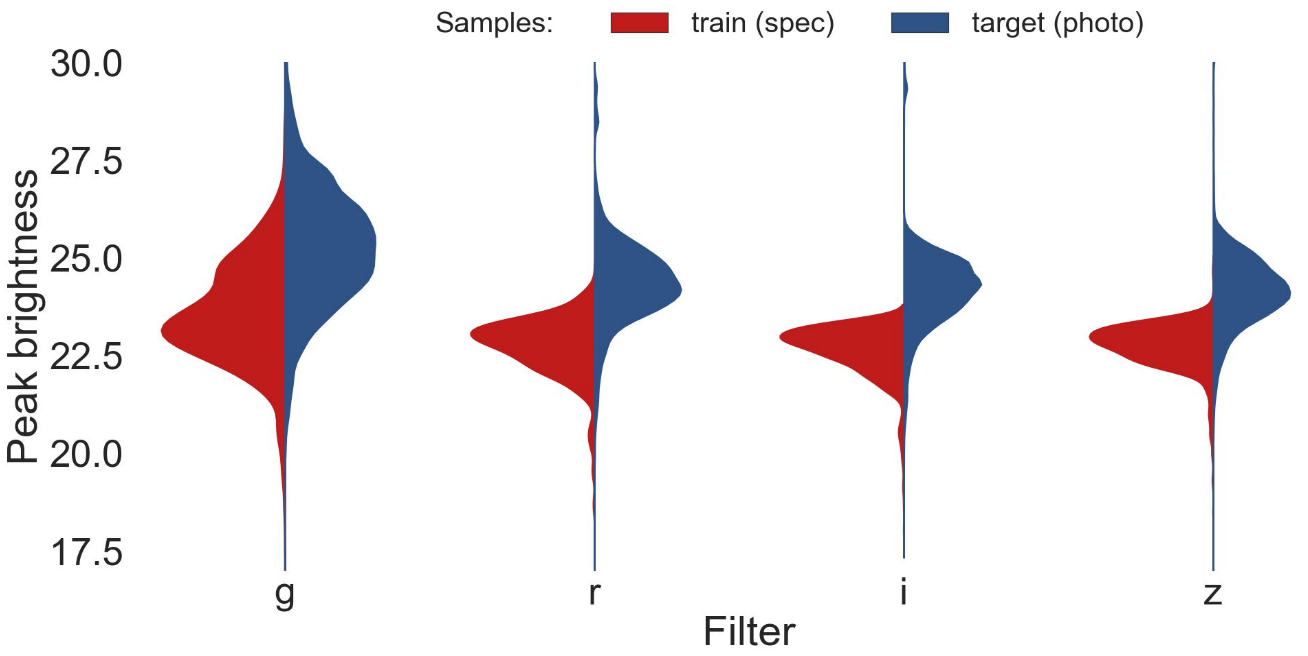


http://supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html

Ω

Λ

Real astro-learning situation



From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

Very common situation

Labels are often far too expensive!

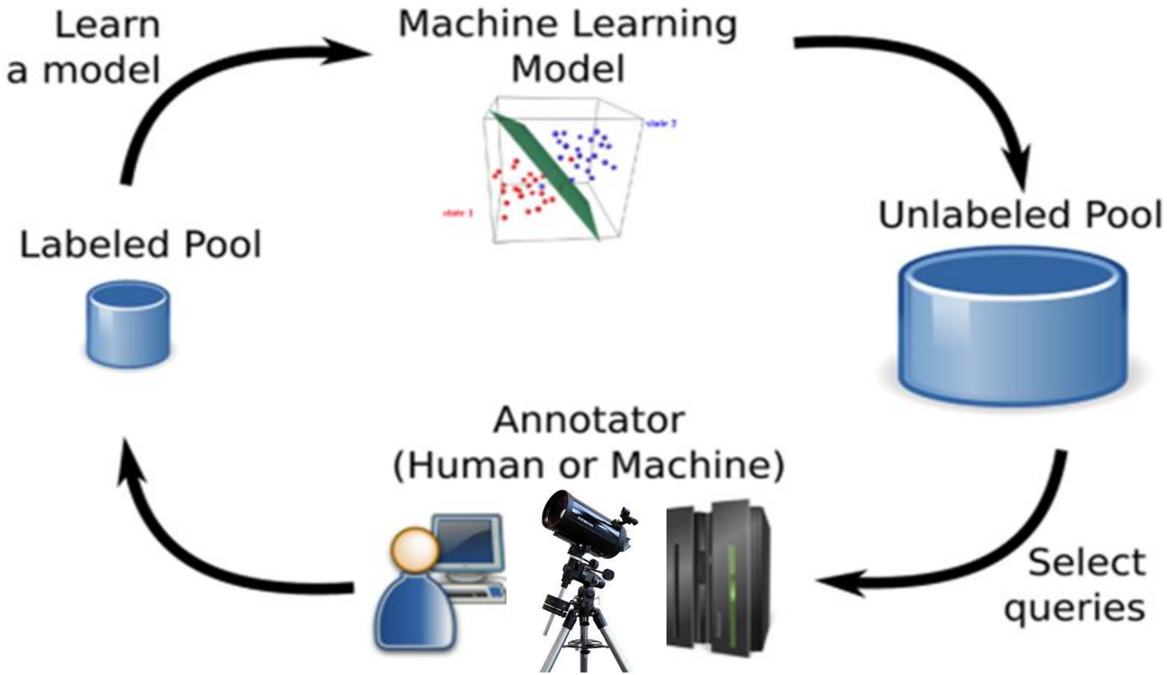
Given limited resources, we need recommendation systems!

35% OF AMAZON'S REVENUE ARE GENERATED BY IT'S RECOMMENDATION ENGINE.

75% OF USERS SELECT MOVIES BASED ON NETFLIX'S RECOMMENDATIONS.

Active Learning

Optimal classification, minimum training

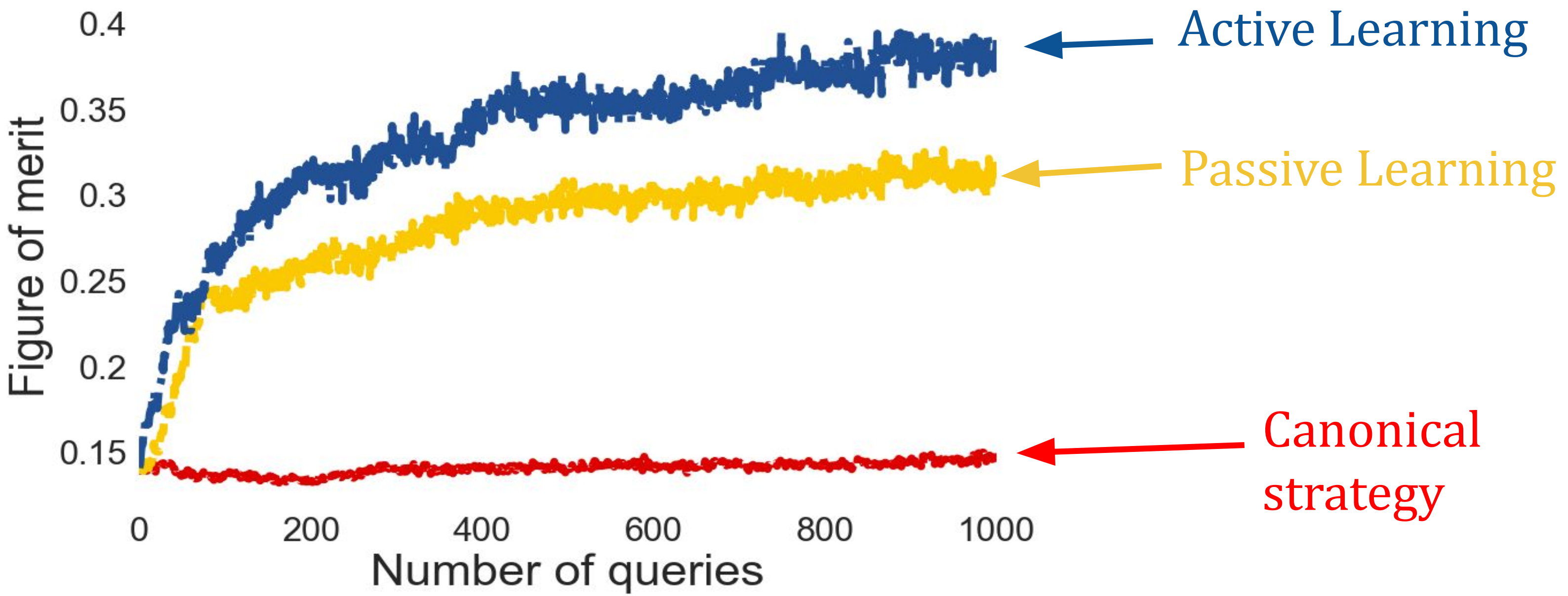


Plot modified from [*Chowdhury et al., 2021, SPIE Medical Imaging*]

COIN

AL for SN classification

Static results

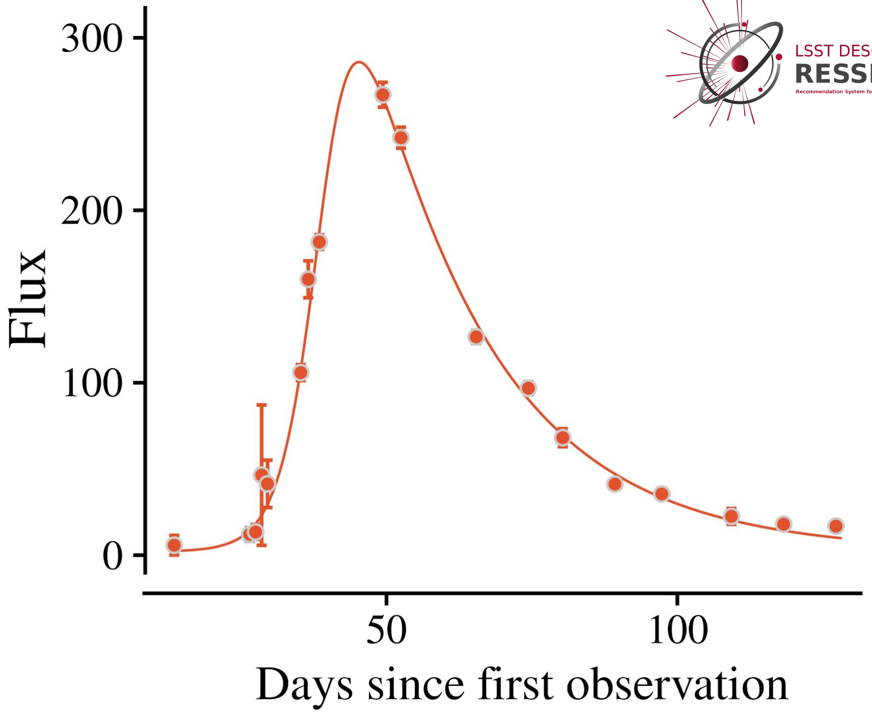


From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

If only it were that simple ...

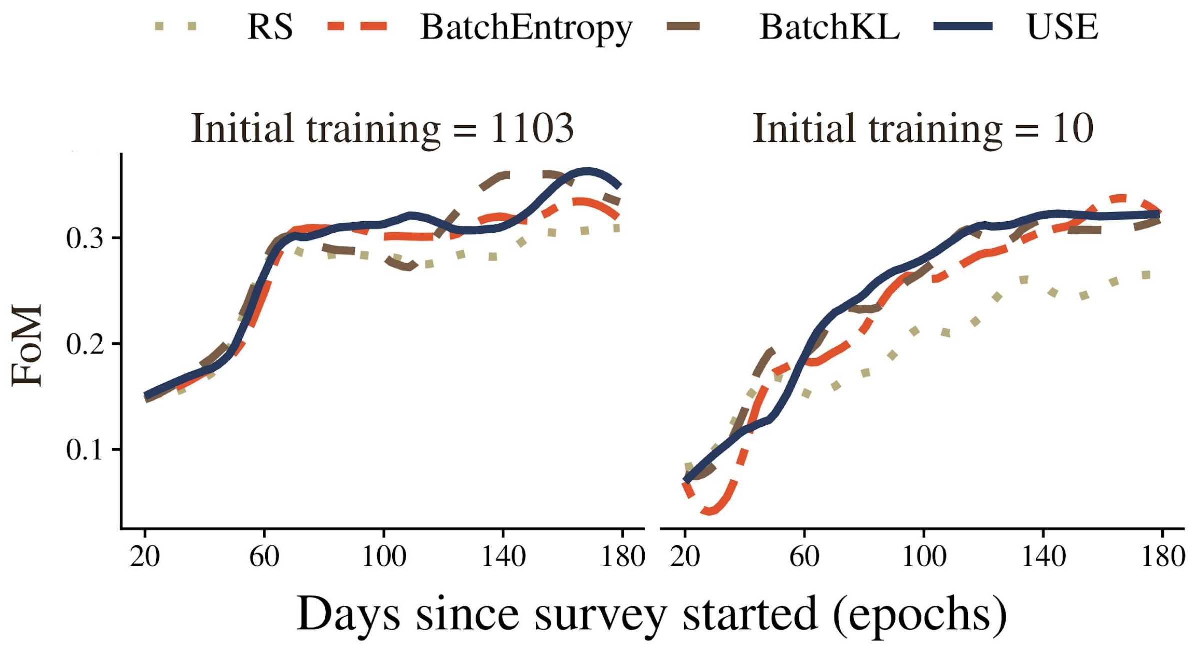
- Window of Opportunity for Labelling
- Evolving Samples
  - We must make query decisions before we can observe the full LC
- Multiple Instruments for labelling
- Evolving budget
  - Other people want to use the telescope
- Evolving Costs
  - Observing costs for a given object changes as it evolves.

LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up



[Kennamer, Ishida et al., 2020 - arXiv:astro-ph/2010.05941](https://arxiv.org/abs/2010.05941) - *The RESSPECT team: LSST-DESC and COIN, IEEE Symposium on Computational Intelligence for Astroinformatics, 2020, Canberra, Australia*

Start from scratch, do not overcomplicate



[*Kennamer, Ishida et al., 2020 - arXiv:astro-ph/2010.05941*]() *- The RESSPECT team: LSST-DESC and COIN, IEEE Symposium on Computational Intelligence for Astroinformatics, 2020, Canberra, Australia*

What about science?

Cosmology results from photometrically classified SN IA

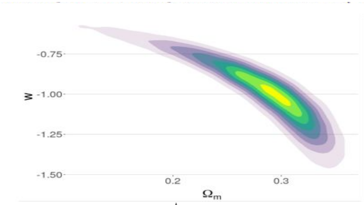


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

learning algorithm

+

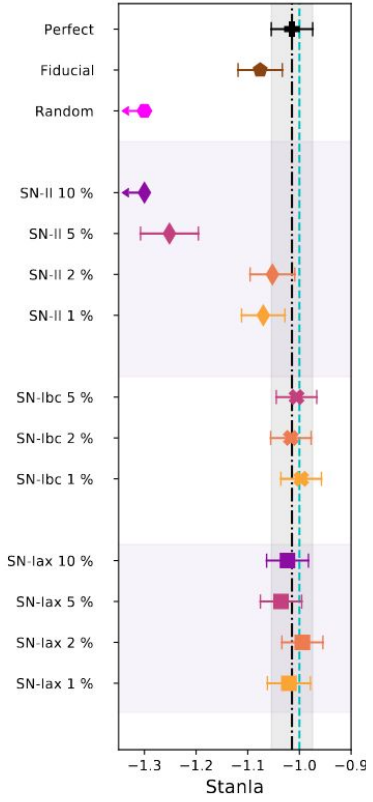
Training sample

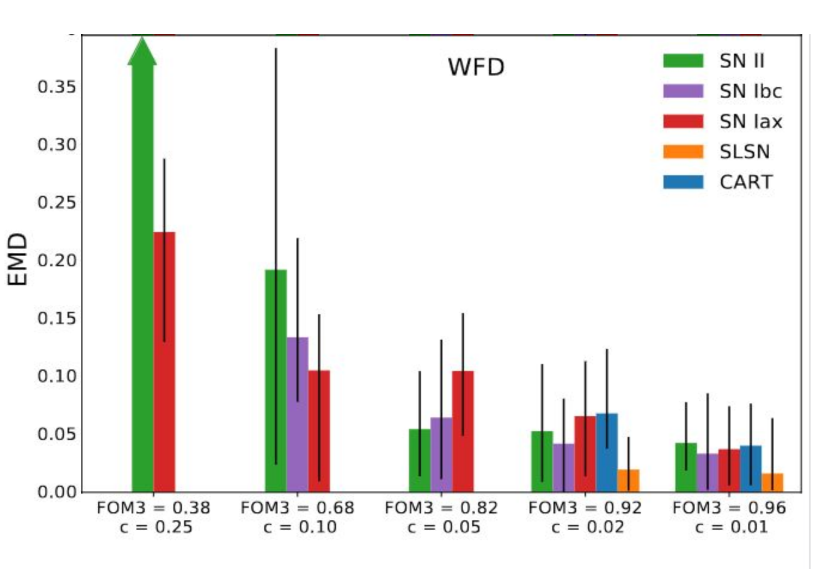
2. Impacts this

1. Different choices of this!

LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up

Good classification might not be enough

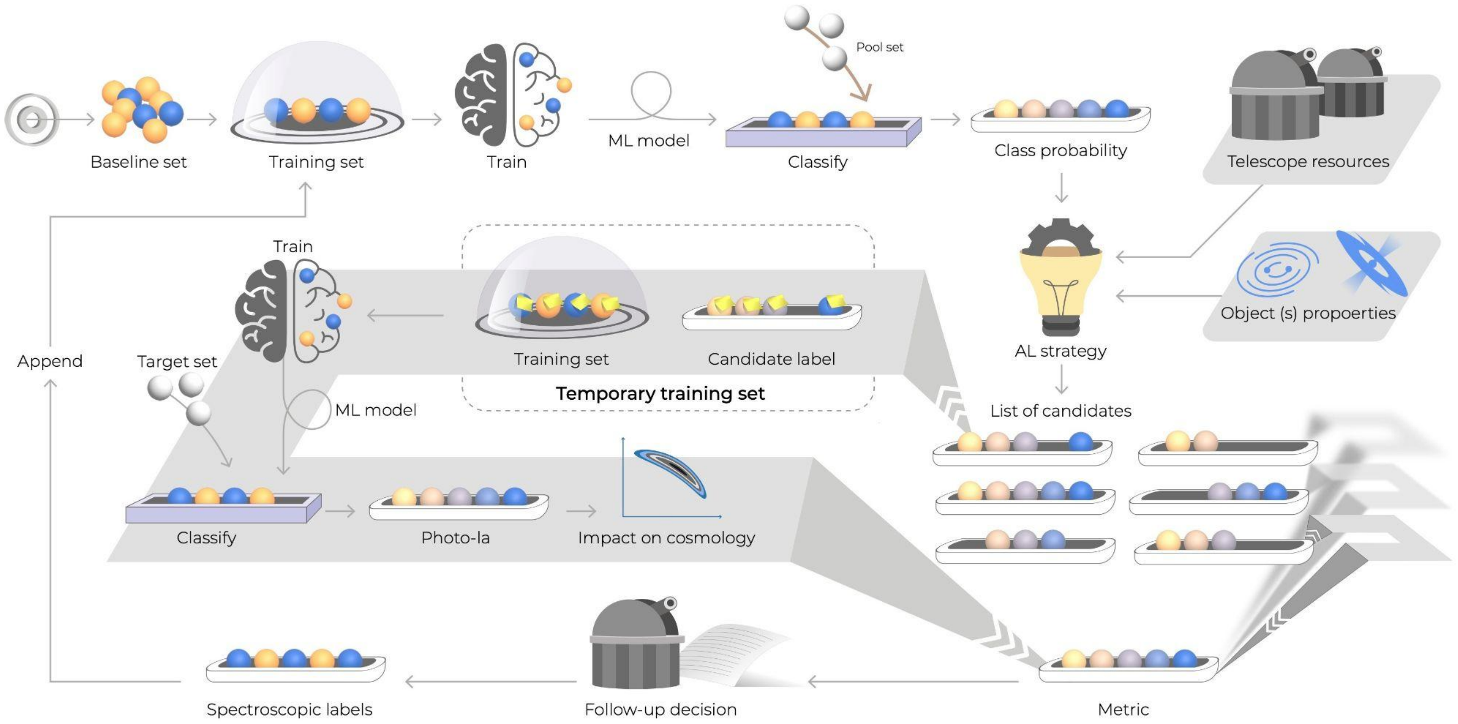




LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up

Malz et al., 2023 - arXiv:astro-ph/2305.14421 - The RESSPECT team: LSST-DESC and COIN, Are classification metrics good proxies for SN Ia cosmological constraining power? -- submitted to A&A

The RESSPECT workflow



https://github.com/COINtoolbox/RESSPECT

The difficult part is data treatment/gathering

- The power of machine learning is in its connection with domain knowledge
- There are caveats in using machine learning and we should avoid off-the-shelf and black bloxes applications
- ML for science must be personalized

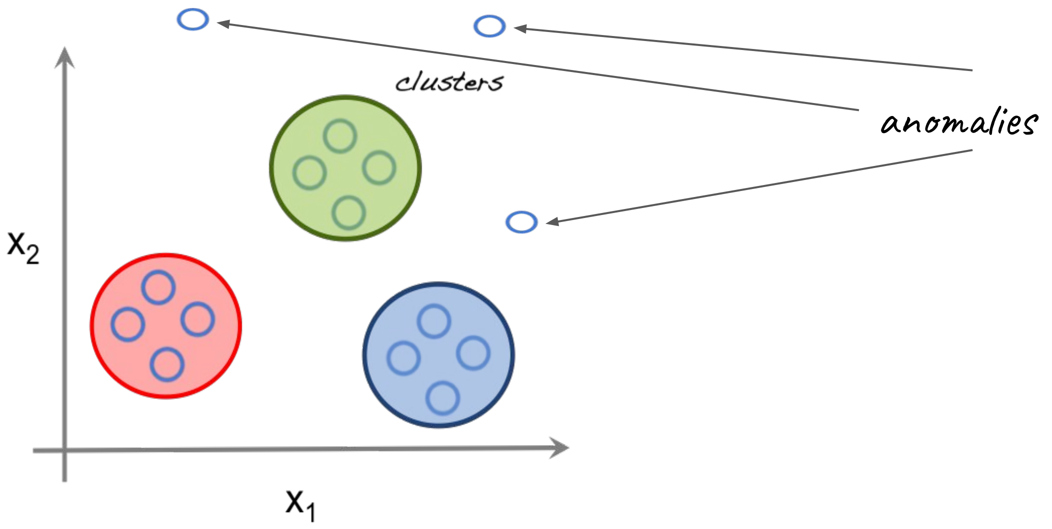
The beauty of an observational science

"… telescopes that merely achieve their stated science goals have probably failed to capture the most important scientific discoveries available to them."

Norris, R. (2017). Discovering the Unexpected in Astronomical Survey Data. Publications of the Astronomical Society of Australia, 34, E007. doi:10.1017/pasa.2016.63

Statistically,

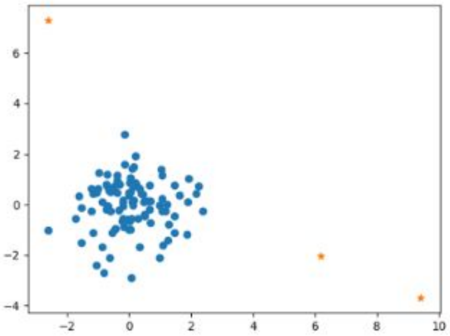
Anomaly Detection



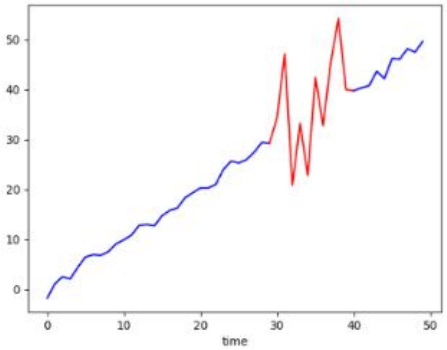
"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Anomaly Detection

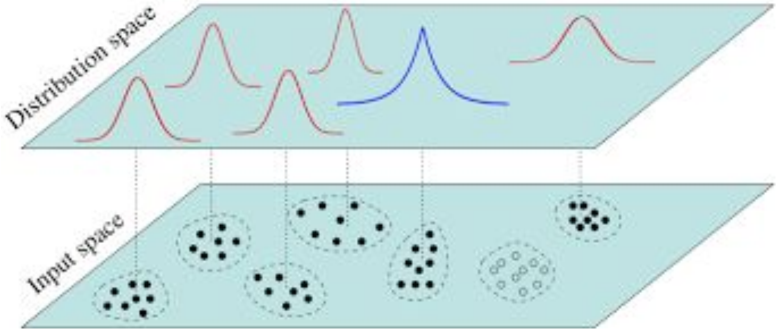
"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"
Hawkins, 1980



isolation



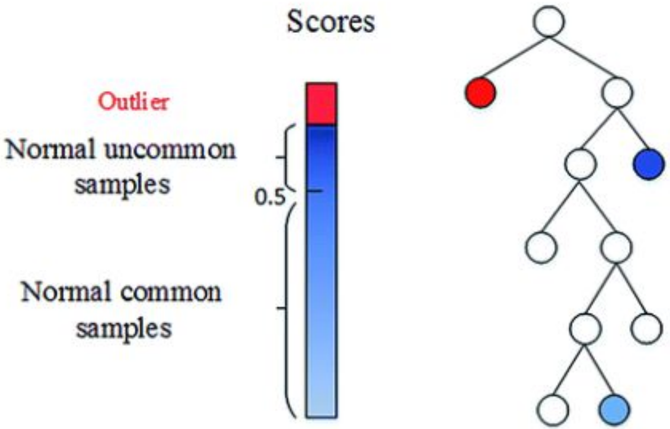
behaviour

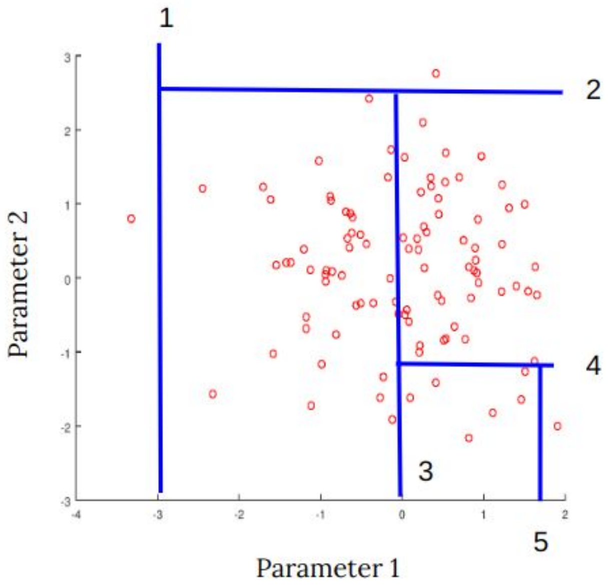


Plot from Muandet and Scholkopf, 2013 - arXiv:stat.ML/1303.0309

group

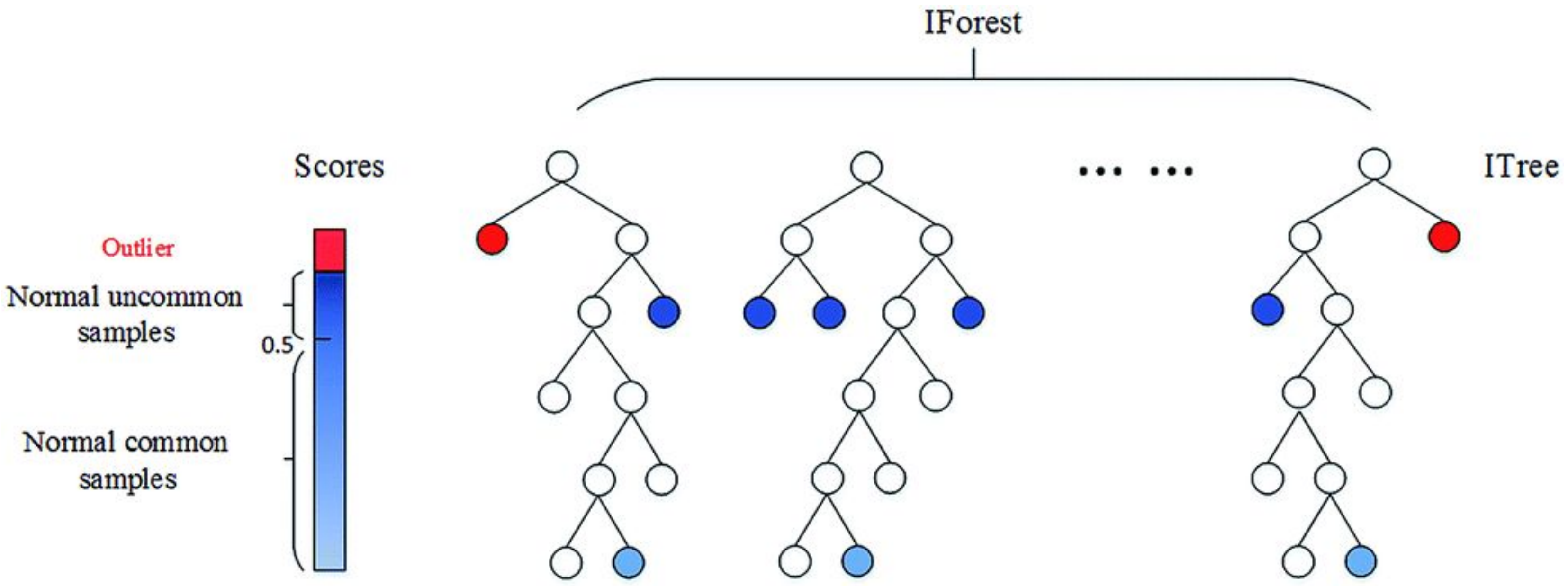
Isolation tree





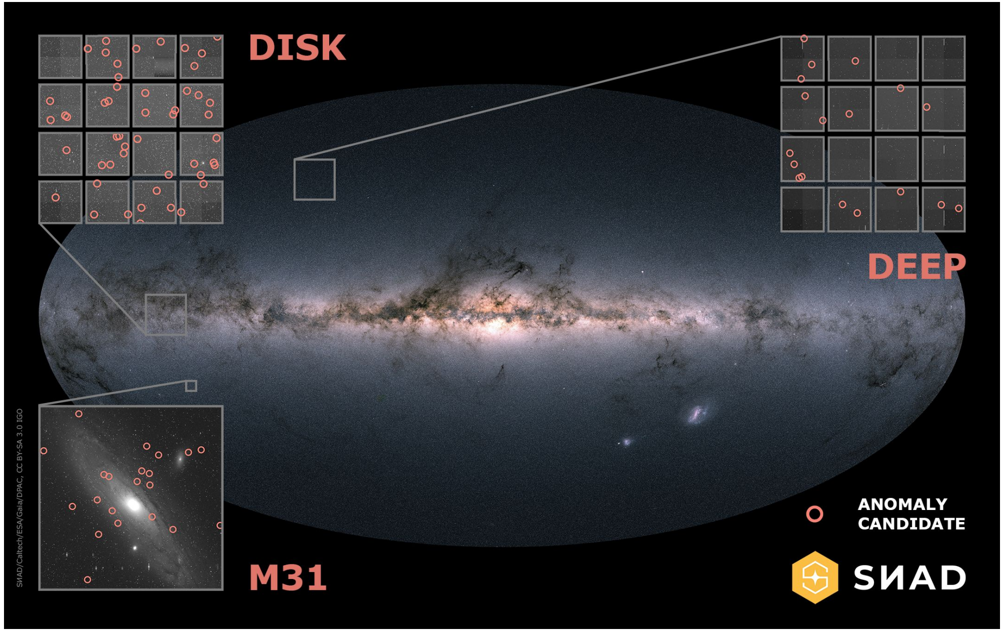
Plot from https://donghwa-kim.github.io/iforest.html

Isolation forest



Plot from https://donghwa-kim.github.io/iforest.html

Zwicky Transient Facility DR3



- Survey currently in operation, telescope in California
- 3 fields from Dara Release 3 (DR3)

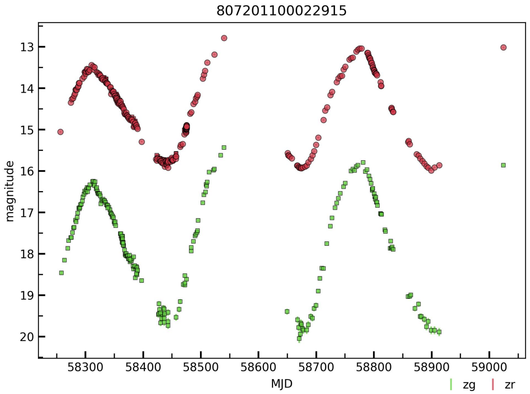
  After selection cuts and feature extraction, ***2.25 million objects***

Malanchev et al., 2021 - MNRAS - https://arxiv.org/abs/2012.01419

Figure by Maria Pruzhinskaya

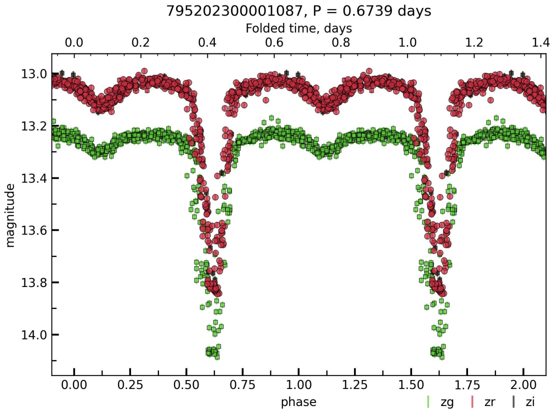
SNAD

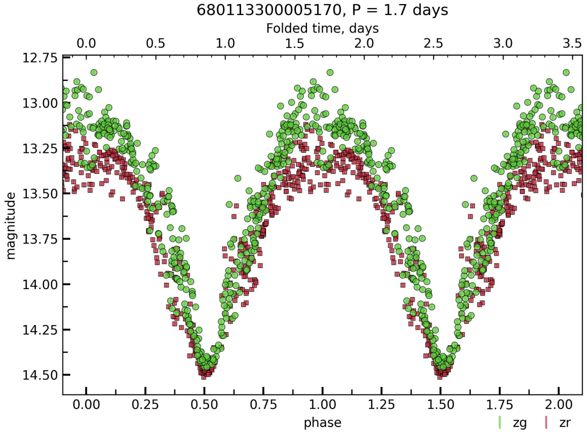
Zwicky Transient Facility DR3



expected to contain stars and periodic variables (no transients)

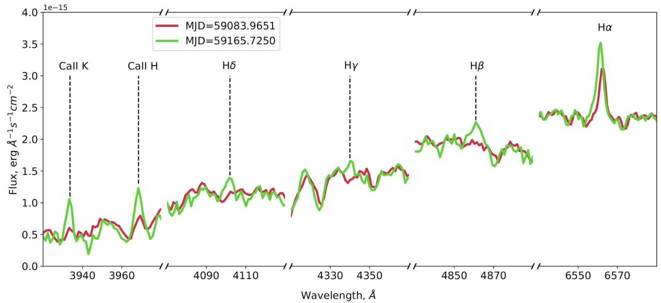
Visualization generated with the SNAD ZTF viewer: https://ztf.snad.space/



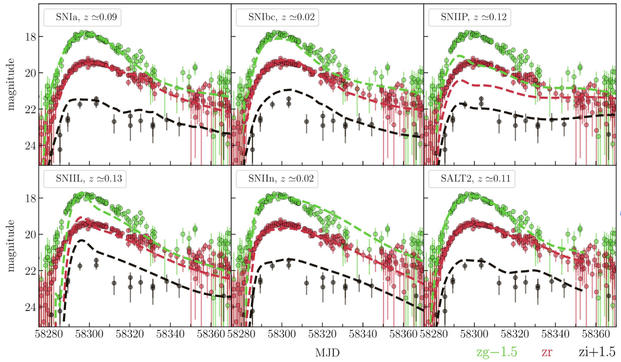


Zwicky Transient Facility DR3

- Feature extraction
- Anomaly detection algorithms:
    - *Isolation Forest*
    - *Local Outlier Factor*
    - *Gaussian Mixture Model*
    - *One-Class Support Vector Machine*
- Initial data: 2.25 million objects
- Expert analysis: 277 objects



- 1 RS Canum Venaticorum star
- 1 red dwarf flare
- 4 Supernova candidates



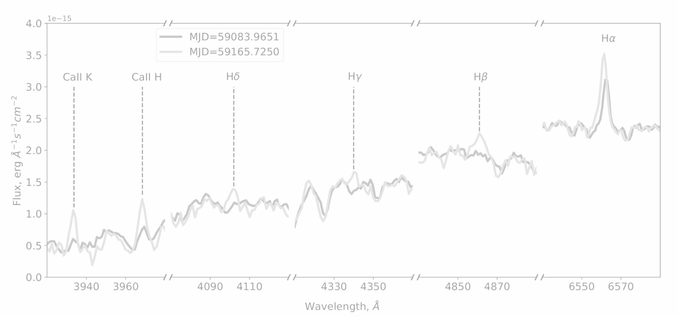
Results:

- 68 % (188) - artifacts, bogus
- 24 % (66) - previously cataloged
- 8 % (23) - discoveries

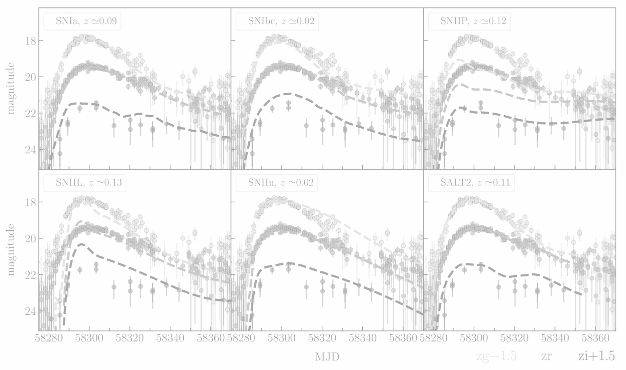
Malanchev et al., 2021 - MNRAS - https://arxiv.org/abs/2012.01419

Zwicky Transient Facility DR3

- Feature extraction
- Anomaly detection algorithms:
  - *Isolation Forest*
  - *Local Outlier Factor*
  - *Gaussian Mixture Model*
  - *One-Class Support Vector Machine*
- Initial data: 2.25 million objects
- Expert analysis: 277 objects



- 1 RS Canum Venaticorum star
- 1 red dwarf flare
- 4 Supernova candidates



Results:

- **68 % (188) - artifacts, bogus**
- 24 % (66) - previously cataloged
- 8 % (23) - discoveries

Malanchev et al., 2021 - MNRAS - https://arxiv.org/abs/2012.01419

It is about Discovery

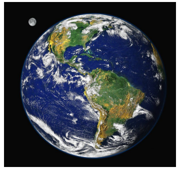
"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Stages of discovery in astronomy:

- Detection
- Interpretation
- Understanding
- Acceptance

Which mechanism?
Is it something we are familiar with but fail to proper model or recognise?
Is it something we have never seen before?

Is there something new for us to learn?

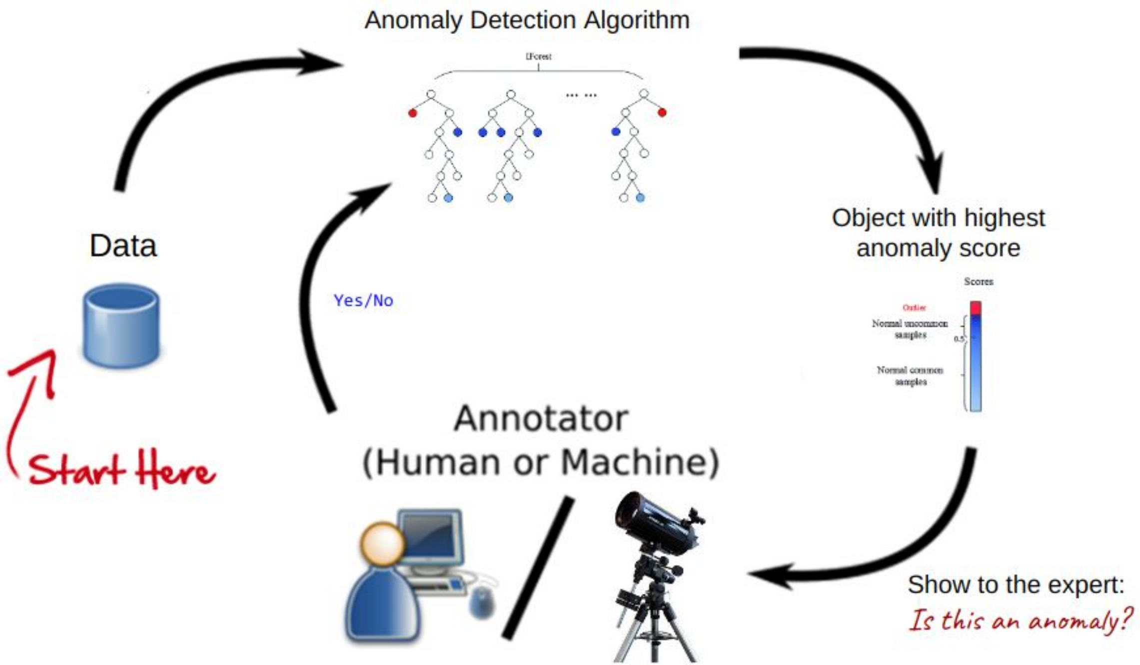


In order to identify the unusual we need to have a clear ideal of what is usual …

.. and that is a social construct. It changes and adapts with time!

Discovery and Classification in Astronomy - by Steven Dick - Cambridge University Press (2013)

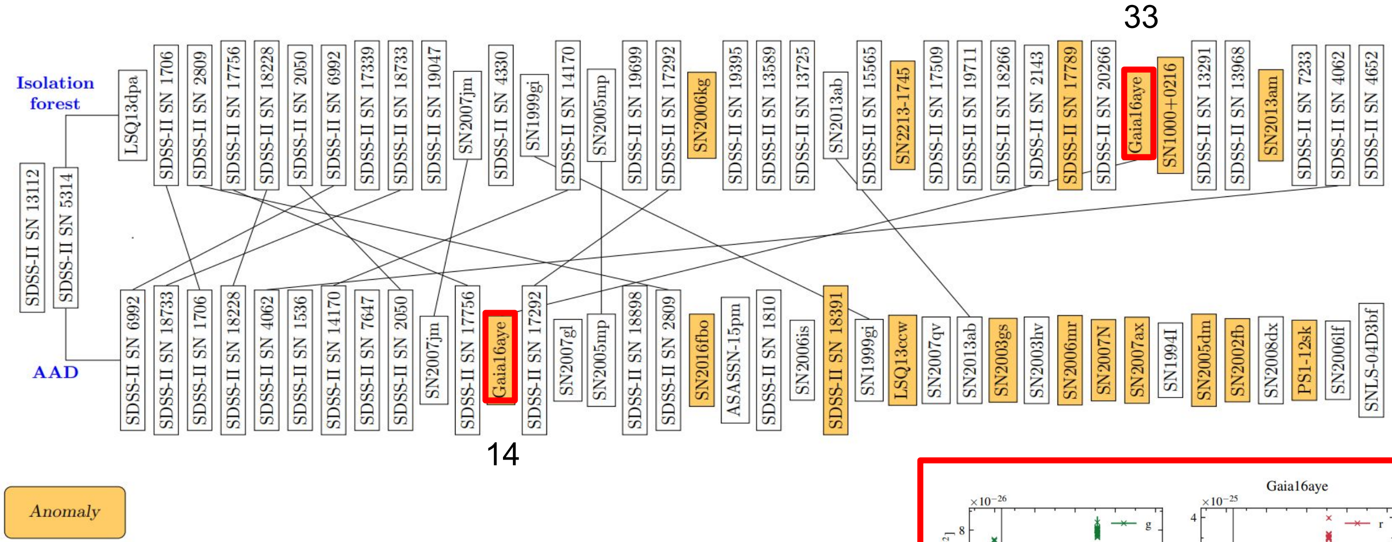
Active Anomaly Detection

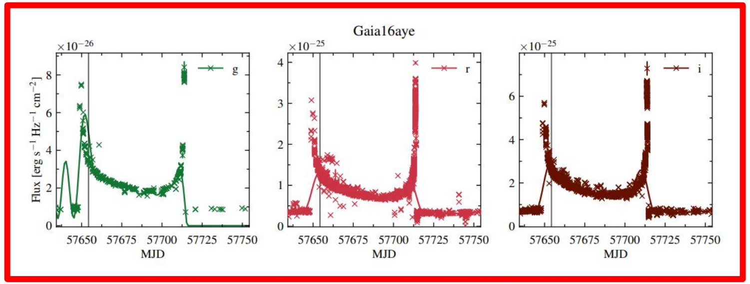


Plot modified from [Chowdhury et al., 2021, SPIE Medical Imaging]
Algorithm from Das, S., et al., 2017, in Workshop on Interactive Data Exploration and Analytics (IDEA'17), KDD workshop, [arXiv:cs.LG/1708.09441]
Try the SNAD implementation: [https://coniferest.readthedocs.io/en/latest/tutorial.html]

AAD on real data: The Open Supernova Catalog

SNAD

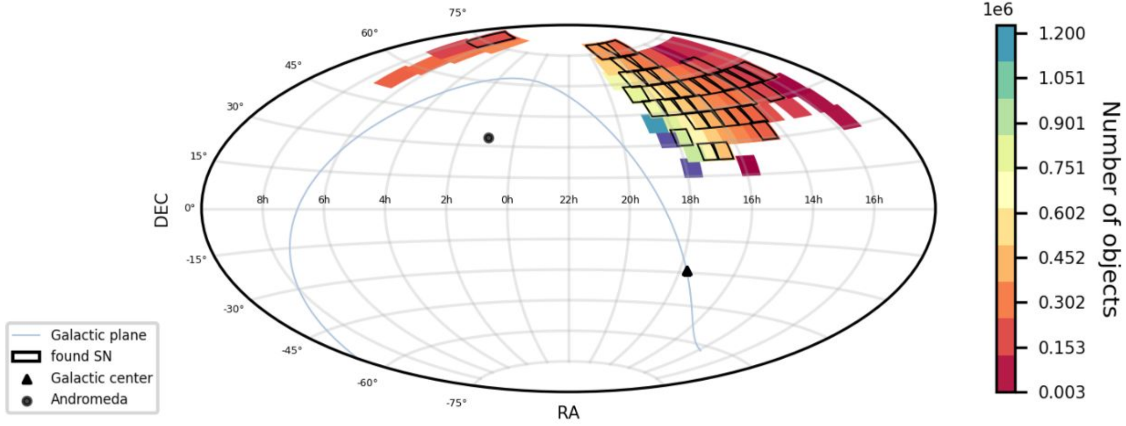




Fast identification of binary microlensing event

Ishida et al., 2021, Astronomy & Astrophysics, Volume 650, id.A195, 9 pp, arXiv:astro-ph/1909.13260

AAD on real data: ZTF data releases



- March - December/2018
- 70 fields
- 30 objects/field
- Total 2100 objects inspected

Found:

- 100 SN-like objects
  - 46 already catalogued
  - 54 newly discovered
- The SNAD catalog: https://snad.space/catalog/

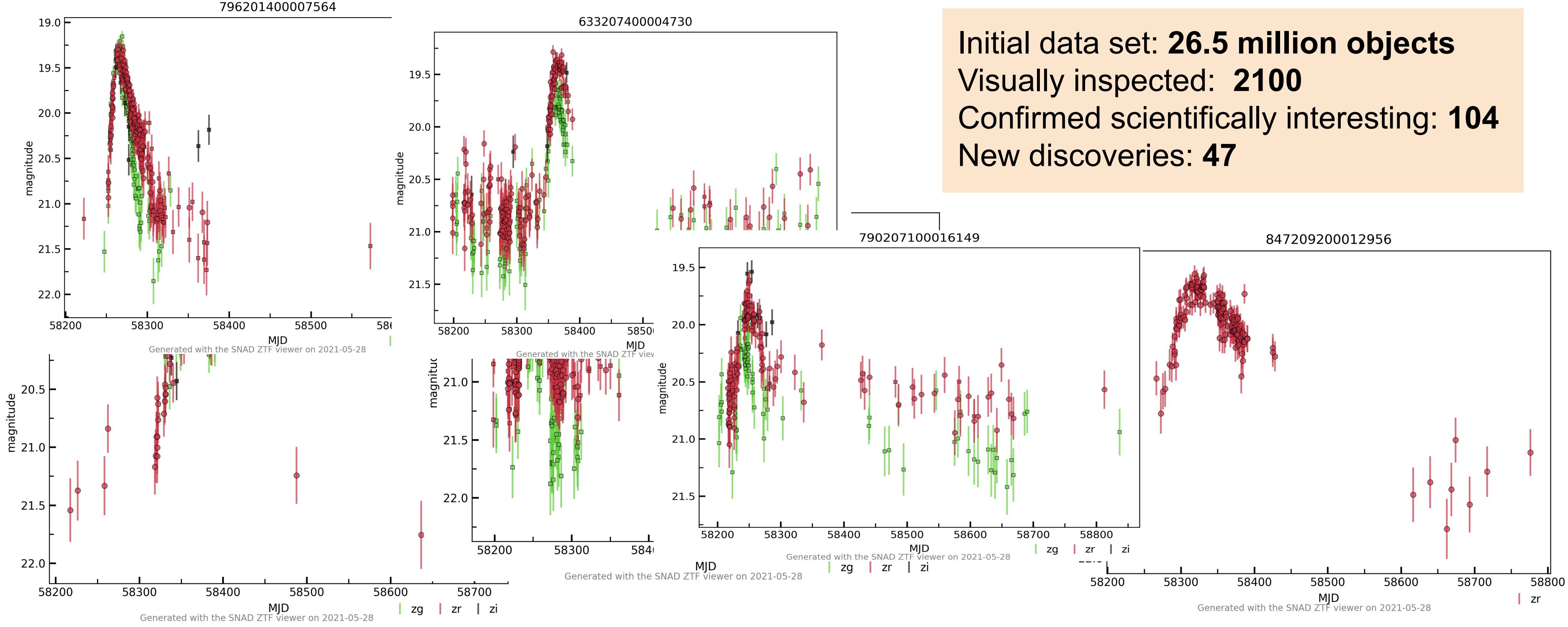
Pruzhinskaya et al., 2023, A&A, 672, id.A111, [*arXiv:astro-ph/2208.09053*](arXiv:astro-ph/2208.09053)

AAD on real data: ZTF DR3

"There are no new supernova-like objects in ZTF DR"

Basically everyone to whom we mentioned we were looking for them.

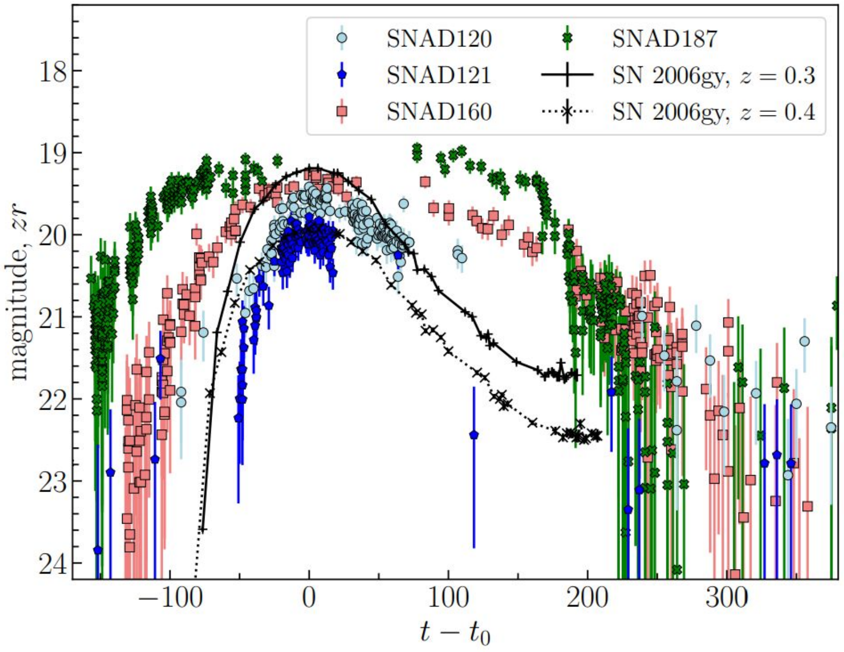
Initial data set: **26.5 million objects**
Visually inspected: **2100**
Confirmed scientifically interesting: **104**
New discoveries: **47**

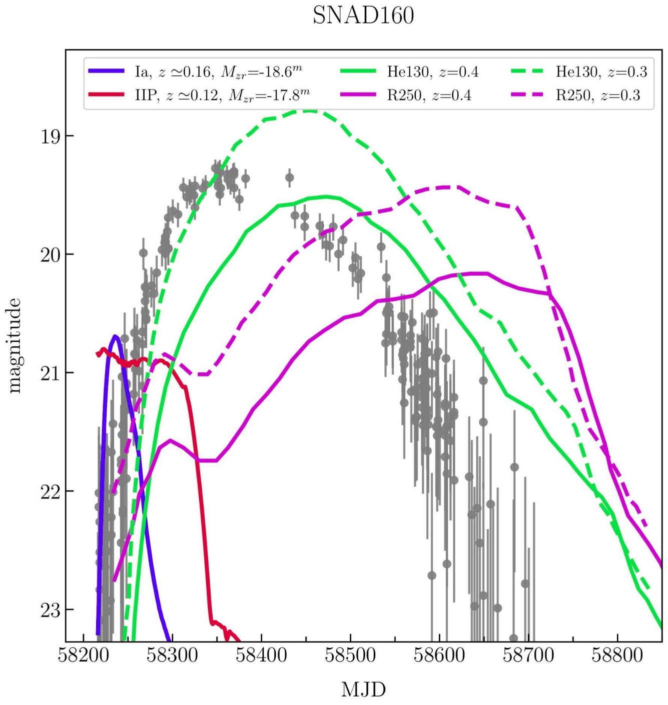


Catalog of lost transients: https://snad.space/catalog/

Pruzhinskaya et al., 2023, A&A - [arXiv:astro-ph/2208.09053](https://arxiv.org/abs/2208.09053)

Interesting SLSN candidates





Pruzhinskaya et al., 2023, A&A 672, A111 (2023), [arXiv:astro-ph/2208.09053]()

AAD is very expensive

Algorithm 2 Active Anomaly Discovery (AAD)

Input: Dataset $\mathbf{H}$, budget B
Initialize the weights $\mathbf{w}^{(0)} = \{\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}\}$
Set $t = 0$
Set $\mathbf{H}_A = \mathbf{H}_N = \emptyset$
while $t \leq B$ **do**
  $t = t + 1$
  Set $\mathbf{a} = \mathbf{H} \cdot \mathbf{w}$ (i.e., $\mathbf{a}$ is the vector of anomaly scores)
  Let $\mathbf{z}_i$ = instance with highest anomaly score (where $i = \arg\max_i(a_i)$)
  Get feedback {*'anomaly'*/*'nominal'*} on $\mathbf{z}_i$
  if $\mathbf{z}_i$ is *anomaly* **then**
    $\mathbf{H}_A = \{\mathbf{z}_i\} \cup \mathbf{H}_A$
  else
    $\mathbf{H}_N = \{\mathbf{z}_i\} \cup \mathbf{H}_N$
  end if
  $\mathbf{w}^{(t)}$ = compute new weights; normalize $\|\mathbf{w}^{(t)}\| = 1$
end while

$$
l(q, \boldsymbol{w}; z_i, y_i) = \begin{cases} 0 \quad \text{if} & \boldsymbol{w} \cdot \boldsymbol{z}_i \geq q \quad \text{and} \quad y_i = \text{anomaly} \\ 0 \quad \text{if} & \boldsymbol{w} \cdot \boldsymbol{z}_i < q \quad \text{and} \quad y_i = \text{normal} \\ q - \boldsymbol{w} \cdot \boldsymbol{z}_i \quad \text{if} & \boldsymbol{w} \cdot \boldsymbol{z}_i < q \quad \text{and} \quad y_i = \text{anomaly} \\ \boldsymbol{w} \cdot \boldsymbol{z}_i - q \quad \text{if} & \boldsymbol{w} \cdot \boldsymbol{z}_i \geq q \quad \text{and} \quad y_i = \text{normal} \end{cases},
$$

$$
\begin{aligned}
\mathbf{w}^{(t)} = \arg\min_{\mathbf{w}, \xi} \; & \frac{C_A}{|\mathbf{H}_A|} \left(\sum_{\mathbf{z}_i \in \mathbf{H}_A} \ell(\hat{q}_\tau(\mathbf{w}^{(t-1)}), \mathbf{w}; (\mathbf{z}_i, y_i)) \right) \\
& + \frac{1}{|\mathbf{H}_N|} \left(\sum_{\mathbf{z}_i \in \mathbf{H}_N} \ell(\hat{q}_\tau(\mathbf{w}^{(t-1)}), \mathbf{w}; (\mathbf{z}_i, y_i)) \right) \\
& + \frac{C_\xi}{|\mathbf{H}_A|} \left(\sum_{\mathbf{z}_i \in \mathbf{H}_A} \ell(\mathbf{z}_\tau^{(t-1)} \cdot \mathbf{w}, \mathbf{w}; (\mathbf{z}_i, y_i)) \right) \\
& + \frac{C_\xi}{|\mathbf{H}_N|} \left(\sum_{\mathbf{z}_i \in \mathbf{H}_N} \ell(\mathbf{z}_\tau^{(t-1)} \cdot \mathbf{w}, \mathbf{w}; (\mathbf{z}_i, y_i)) \right) \\
& + \|\mathbf{w} - \mathbf{w}_p\|^2 \qquad (2)
\end{aligned}
$$

Das, S., Wong, W-K., Dietterich, T., Fern, A. and Emmott, A. (2016). *Incorporating Expert Feedback into Active Anomaly Discovery*

SNAD

Aiming at bigger data

- Requires optimization
- Smooth incorporation of expert knowledge

SNAD

Pine Forest

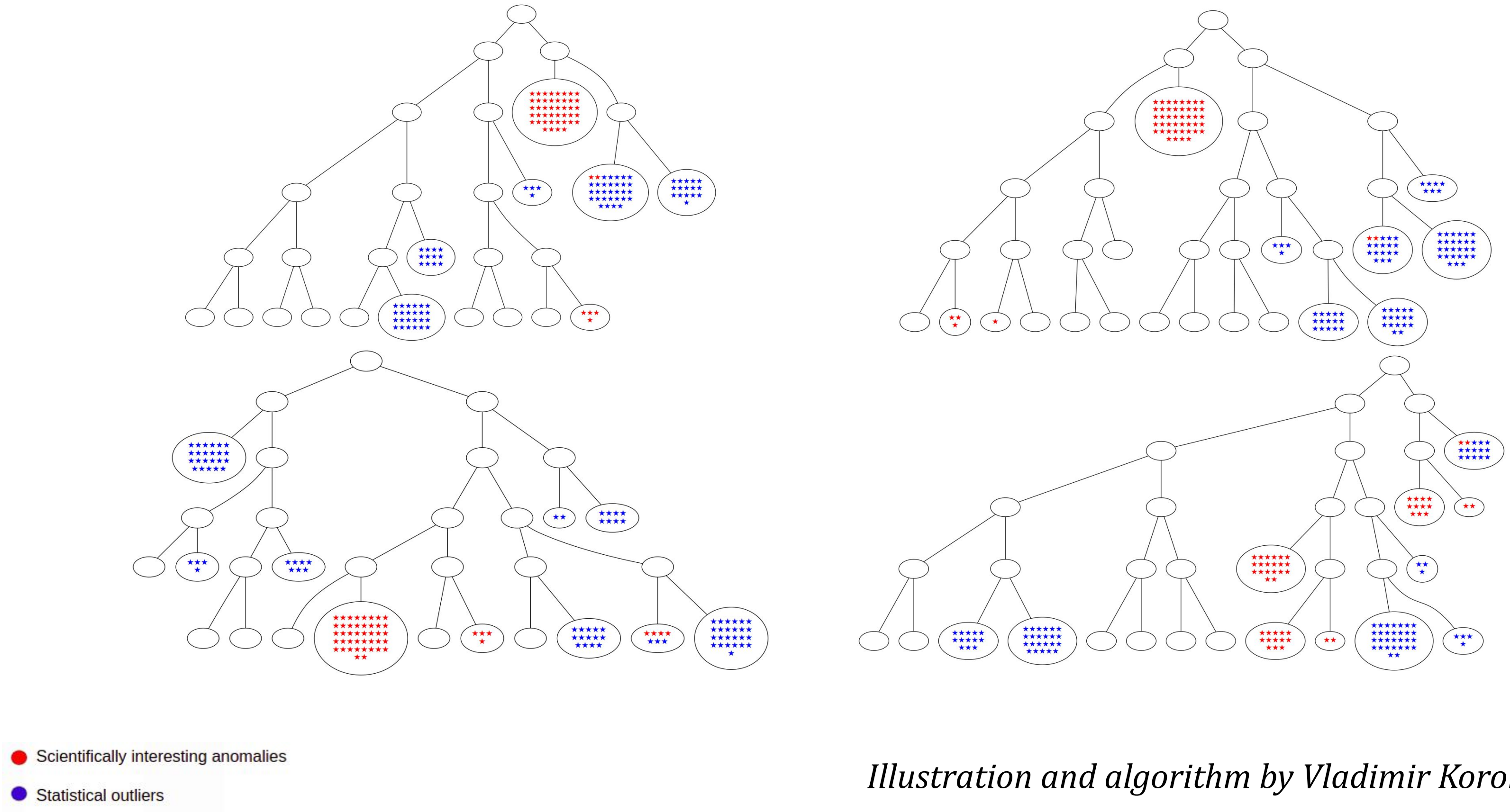
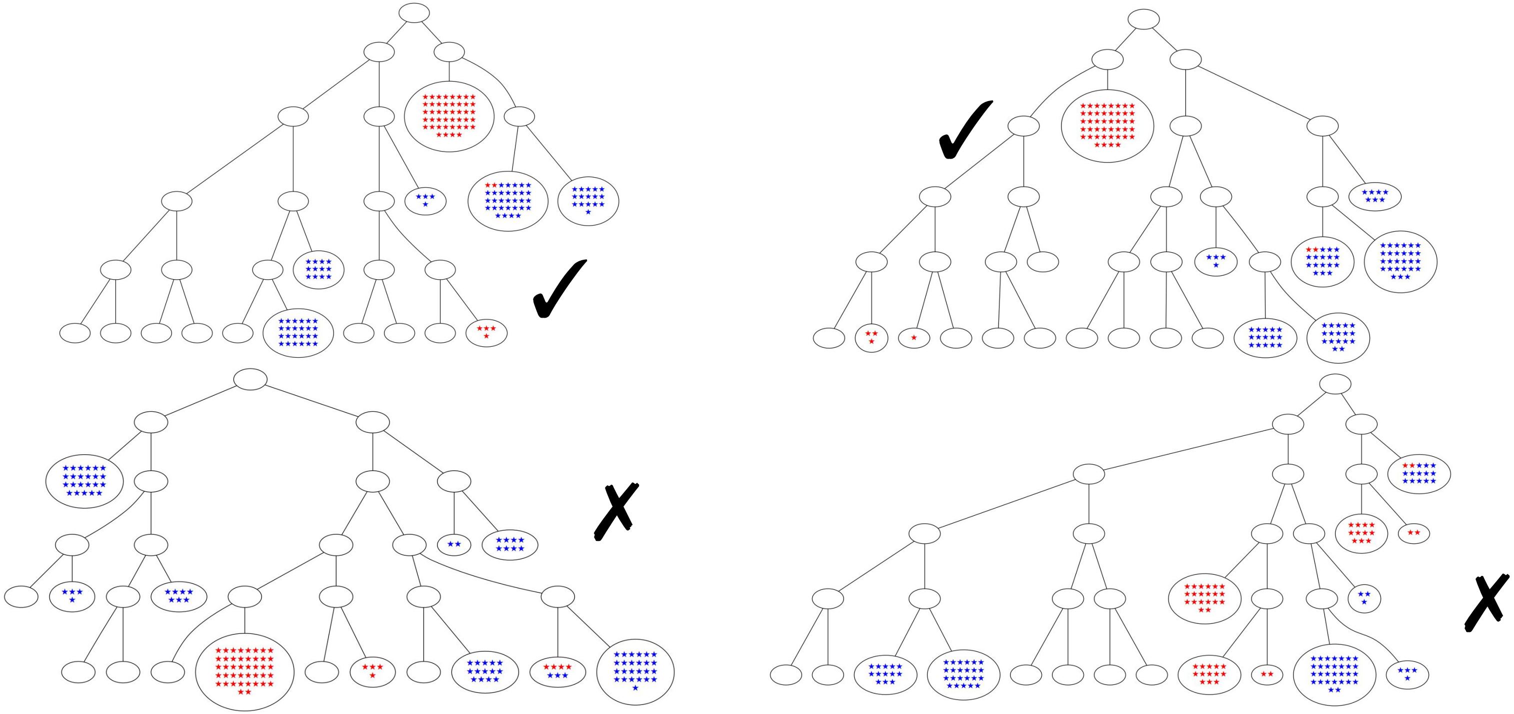


Illustration and algorithm by Vladimir Korolev

SNAD

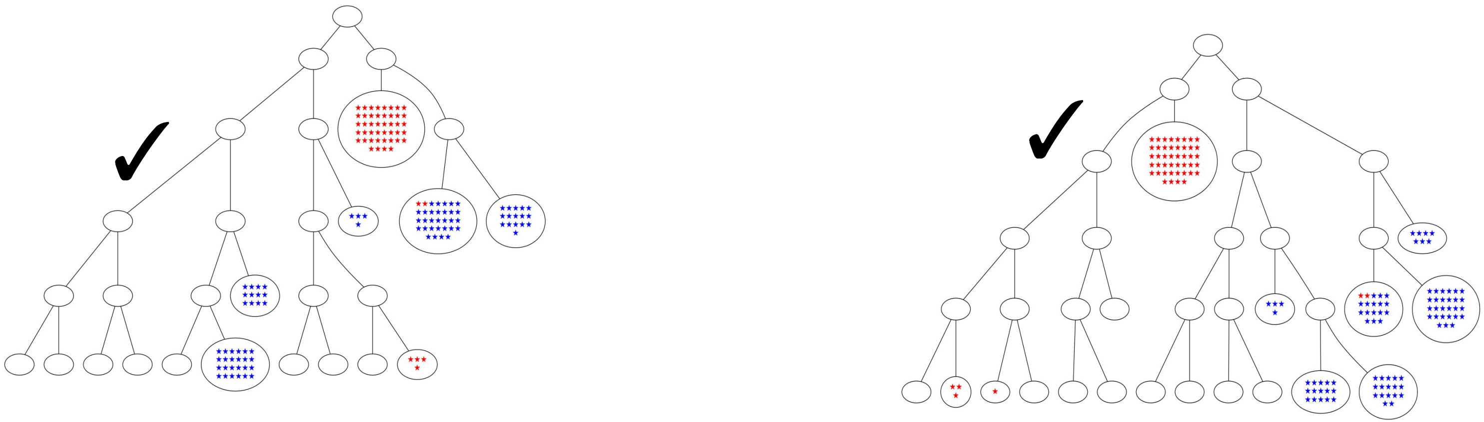
Pine Forest



- Scientifically interesting anomalies
- Statistical outliers

Illustration by Vladimir Korolev

Pine Forest



Generate new random trees
and filter again

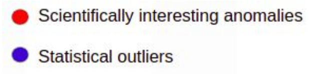


Illustration and algorithm by Vladimir Korolev

Break

Active Anomaly Detection tutorial

Remember to: File → Save a copy to drive

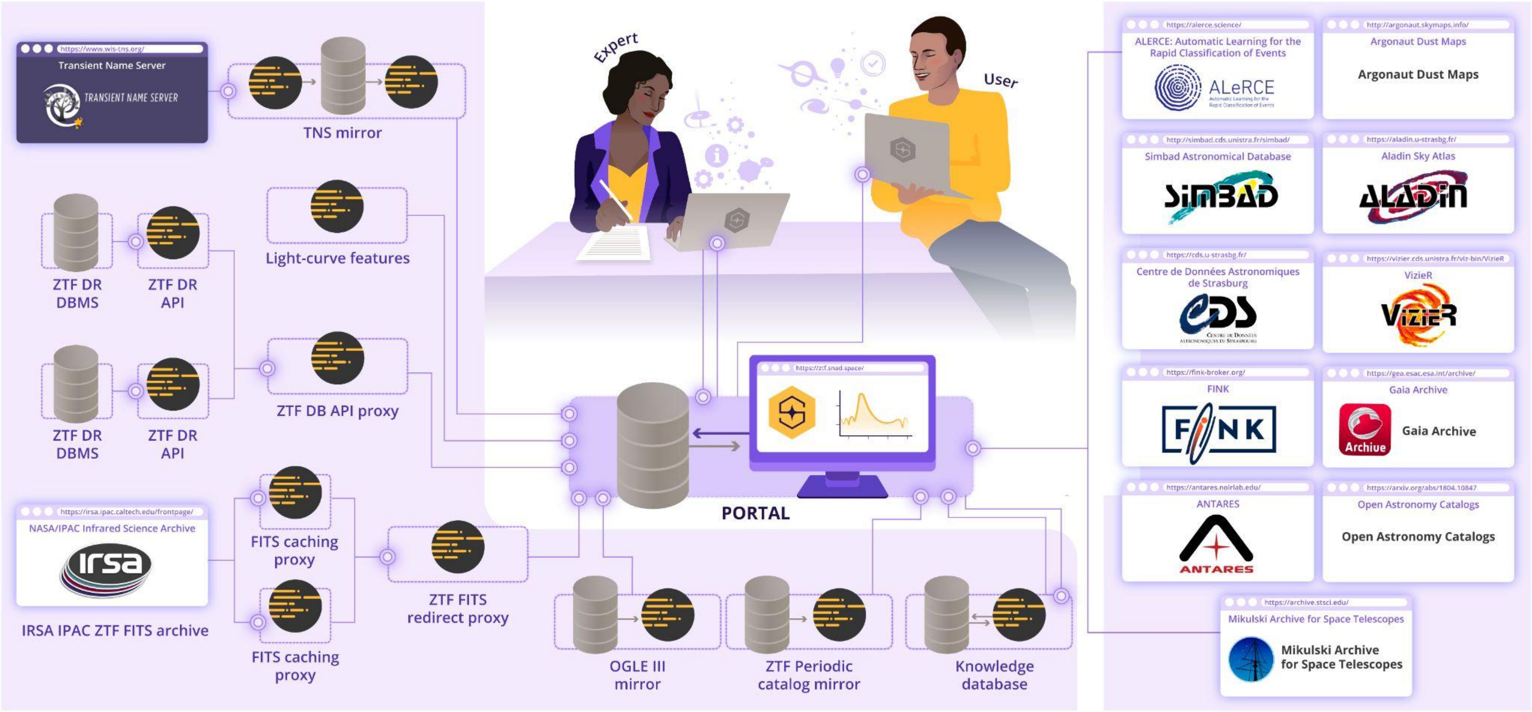
https://colab.research.google.com/drive/1LvC_a8QE7Q5MECL5uOAdze_xGgRSS4bm?usp=sharing

Active Anomaly Detection tutorial

- The coniferest package: https://coniferest.readthedocs.io/en/latest/tutorial.html

- To use with your own data:
  - **Features** (called data in the tutorial) ⇒ 2D pandas dataframe
    1 line per object, 1 column per feature
    - If you are using light curves, you might want to check the `light_curve` package for optimized feature extraction:
      https://github.com/light-curve/light-curve
  - **Metadata** ⇒ 2D array
    1 line per object: it may contain any information that will identify the candidate so you can make a decision.

The SNAD viewer



Malanchev et al., 2023, PASP, Volume 135, Issue 1044, id.024503, 18 pp, [arXiv:astro-ph/2211.07605]

The knowledge database

SNAD160 — 821207100004043

☐ artefact ☐ column ☐ bright_star ☐ cosmic ☐ defocusing ☐ ghost ☐ M31 ☐ spike ☐ track ☐ frame_edge
☐ VAR ☐ transient ☐ AGN ☐ QSO ☐ STAR ☐ Galaxy
☑ SN ☐ SNIa ☐ CCSN ☐ SLSN
☐ Eclipsing ☐ EA ☐ EB ☐ EW
☐ Pulsating ☐ CEP ☐ DCEP ☐ L ☐ LPV ☐ M ☐ RR ☐ RRAB ☐ RSG ☐ SR
☐ DSCT
☐ Cataclysmic ☐ AM ☐ N ☐ UG ☐ UGSS ☐ UGZ
☐ Eruptive ☐ INS ☐ SDOR ☐ TTS ☐ YSO ☐ M_dwarf_flare
☐ Rotating ☐ BY ☐ RSCVn
☑ uncertain ☐ non-catalogued ☐ 1-point ☐ TNS_candidate

Point tag name to see its description. See instructions and tag editor here

SNAD160

SUBMIT

RESET

Tags	Description	Changed by	Changed at
SN, uncertain	SNAD160	maria	2021-11-11T11:07:27.753Z
AGN, SN, uncertain	SNAD160	maria	2021-11-09T21:42:36.314Z
AGN, SN, uncertain, non-catalogued		maria	2021-11-08T13:33:49.098Z

More than 2000 objects already tagged by experts

Malanchev et al., 2023, PASP, Volume 135, Issue 1044, id.024503, 18 pp, arXiv:astro-ph/2211.07605

Explore the boundaries of your knowledge

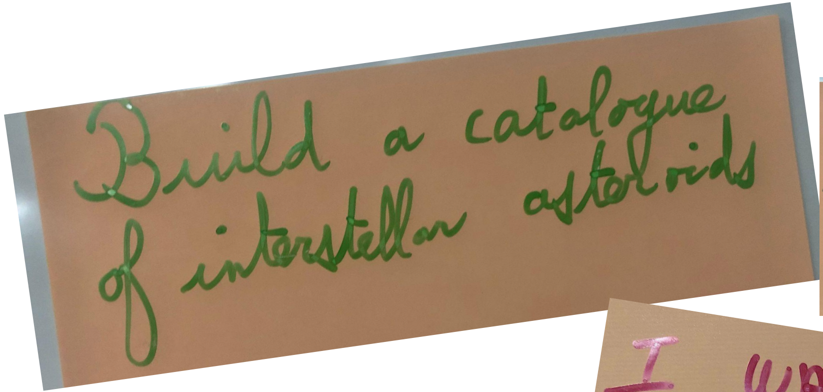
- In the era of Rubin, serendipitous discoveries will not happen
- Domain experts **must be included** in the development of new techniques **from the first stages**. They should supervise the first prototypes.

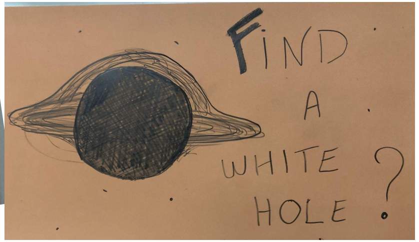
Explore the boundaries of your knowledge

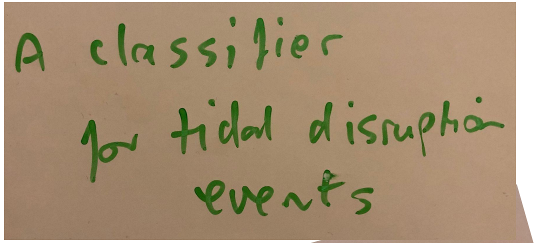
- In the era of Rubin, serendipitous discoveries will not happen
- Domain experts **must be included** in the development of new techniques **from the first stages**. They should supervise the first prototypes.

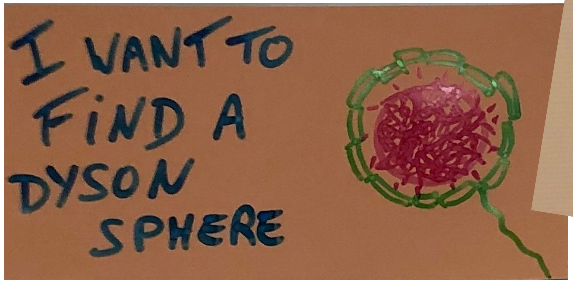
It is crucial to know what you are looking for

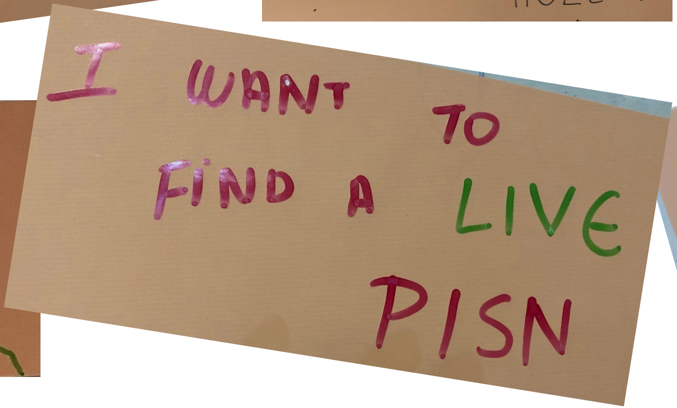
#FinkDreamShots

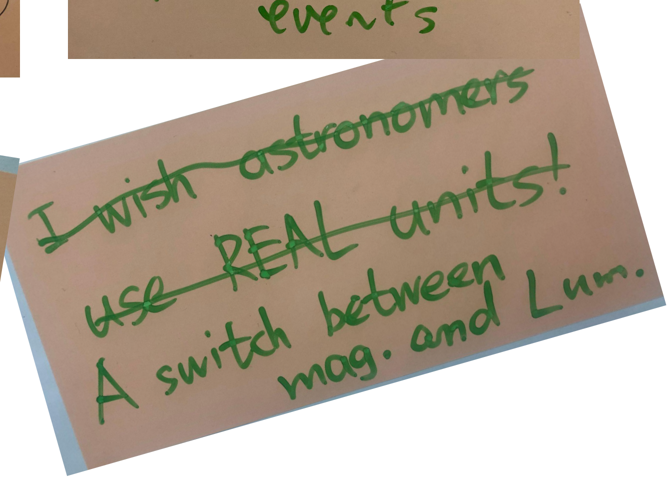












From OzFink 2023 - Melbourne, Australia - https://www.ozgrav.org/ozfink-workshop-2023.html

What do you want to see?

THANK YOU