The role of domain knowledge experts in the era of big data

Where the Earth Meets the Sky
remotely - 27 May 2021

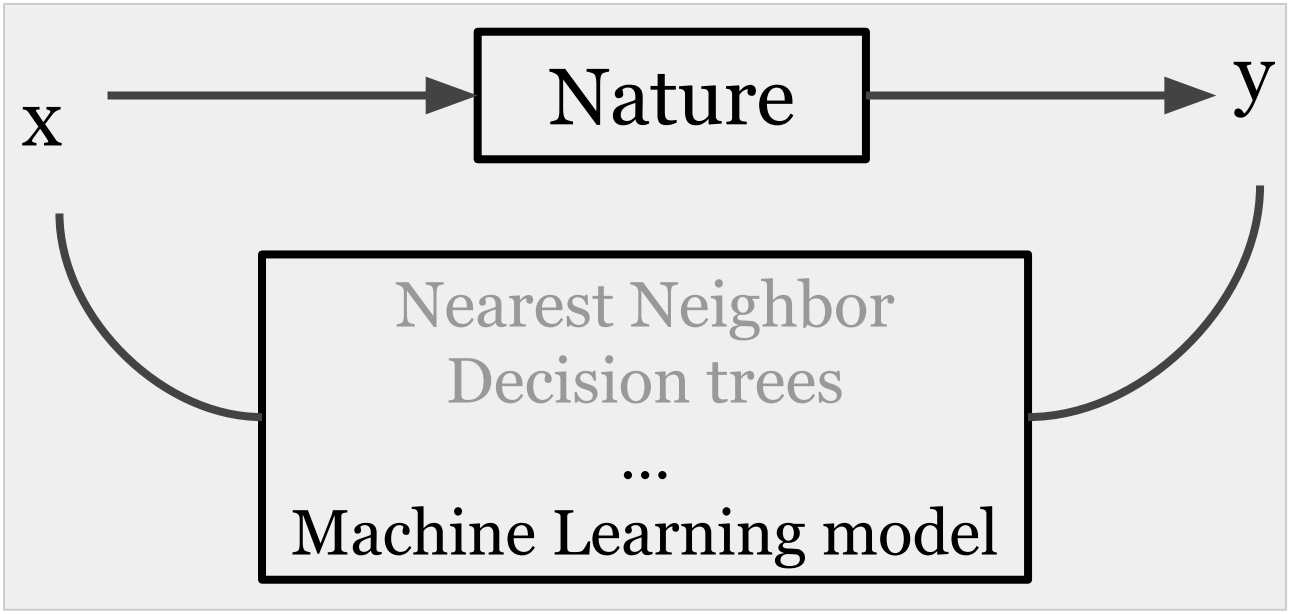
Emille E. O. Ishida
Laboratoire de Physique de Clermont - Université Clermont-Auvergne
Clermont Ferrand, France

I. Supervised learning

Hypothesis:

Physical modeling:

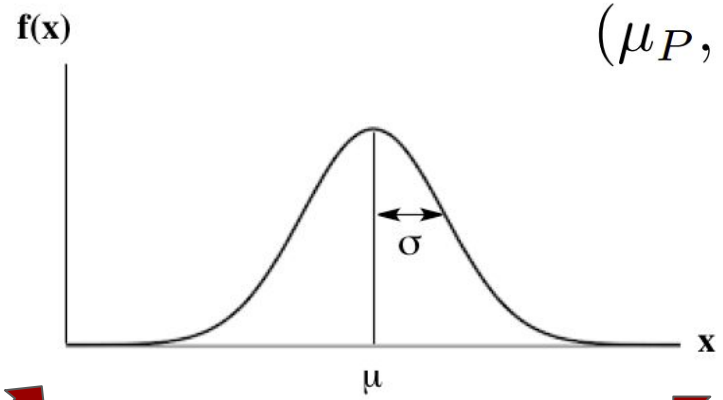
Algorithmic modeling:



Breiman, L., *Statistical Modeling: The Two Cultures*, Stat. Sci, Volume 16 (2001)

Representativeness

Probability distribution, *P*



(μ_P, σ_P)

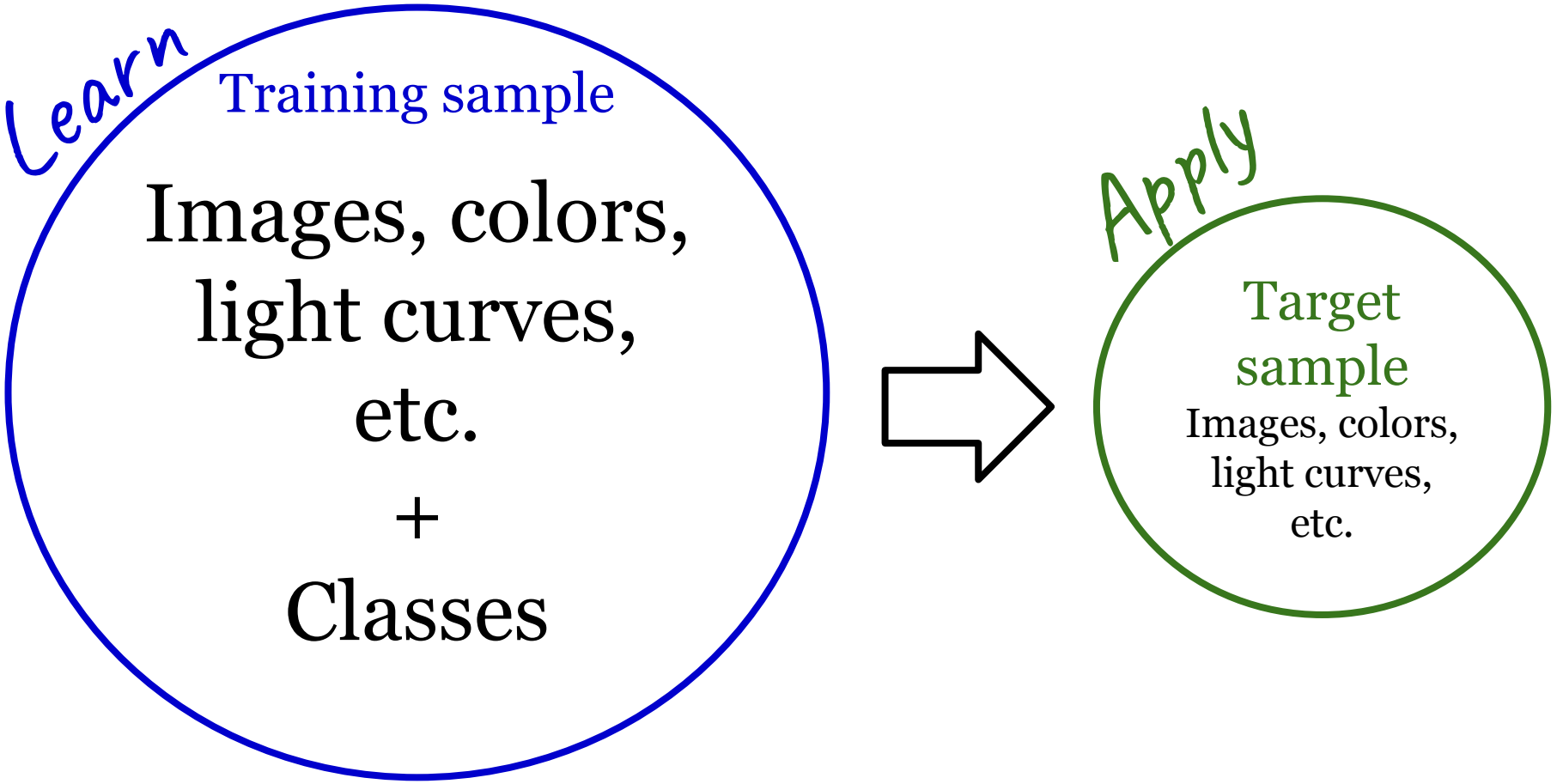
Sample, S_1

Sample, S_2

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

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

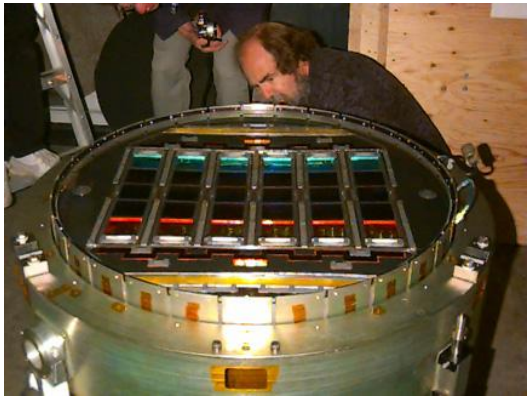
Ideal supervised learning situation



More likely situation in a real science data set

Features x Labels

Spectroscopy x Photometry



Exposure time 2 x 54s

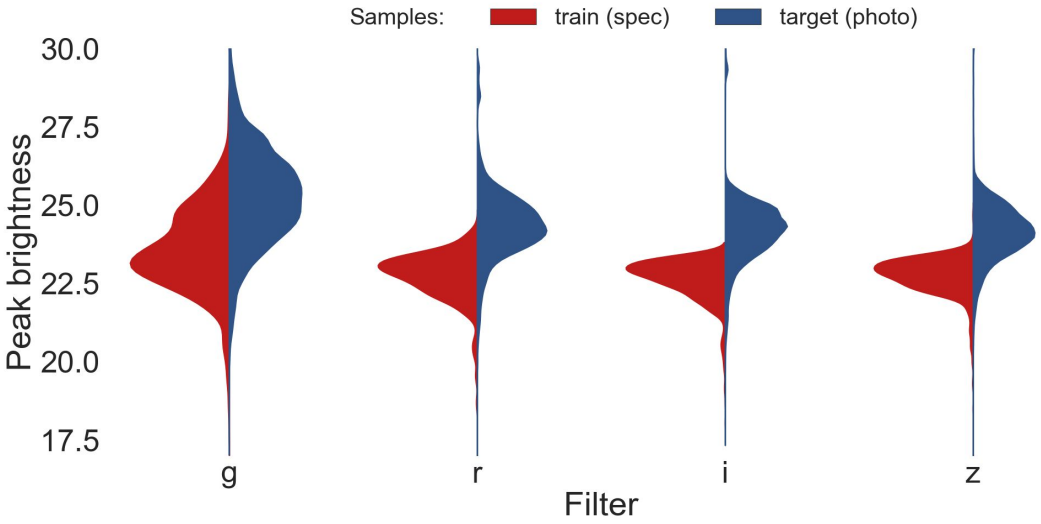
~ Integration at least 45 minutes

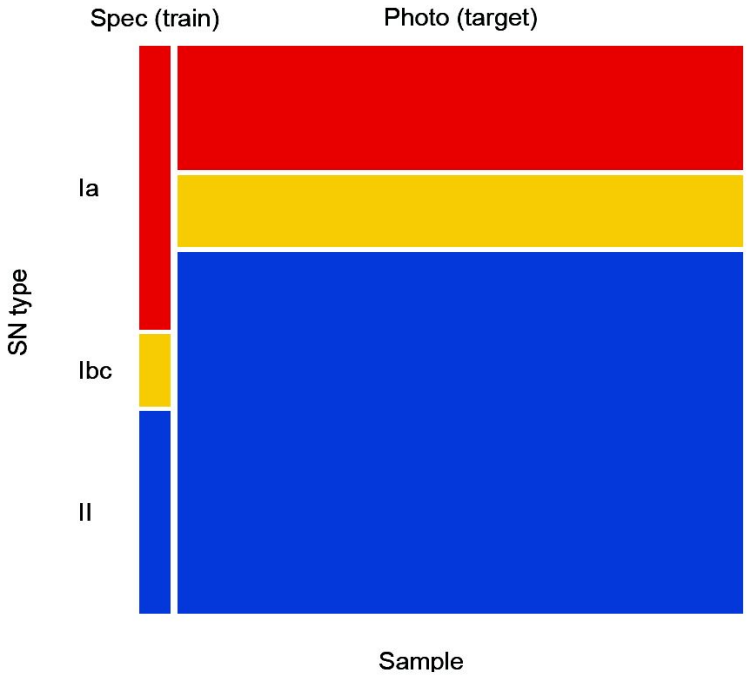
http://www.stsci.edu/~inr/bdpics/bd5.htm

://www.preposterousuniverse.com/blog/2009/10/06/practicality-and-the-universe/

In astro, training means spectra

Representativeness



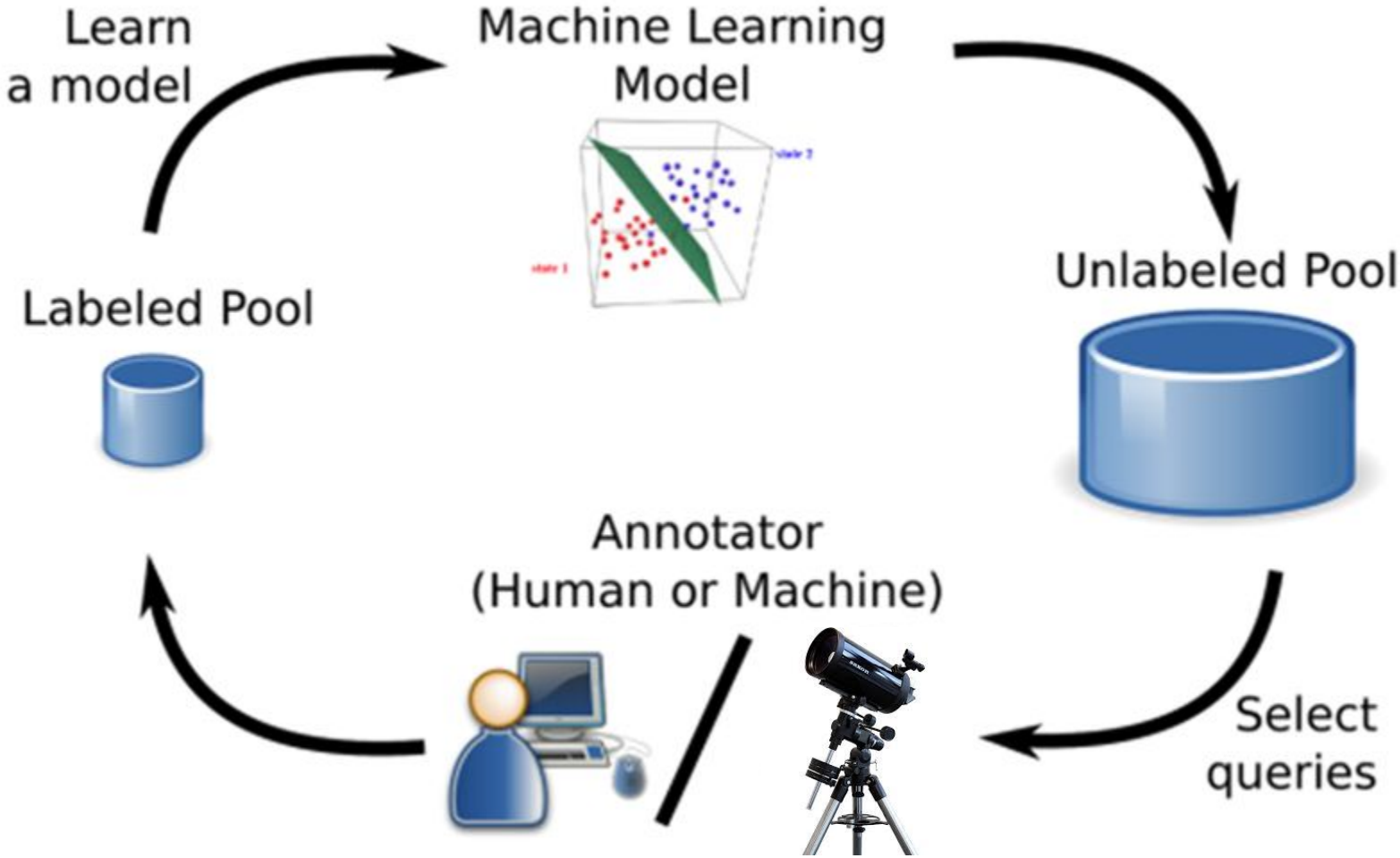


This is not something we can change ...

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

Active Learning

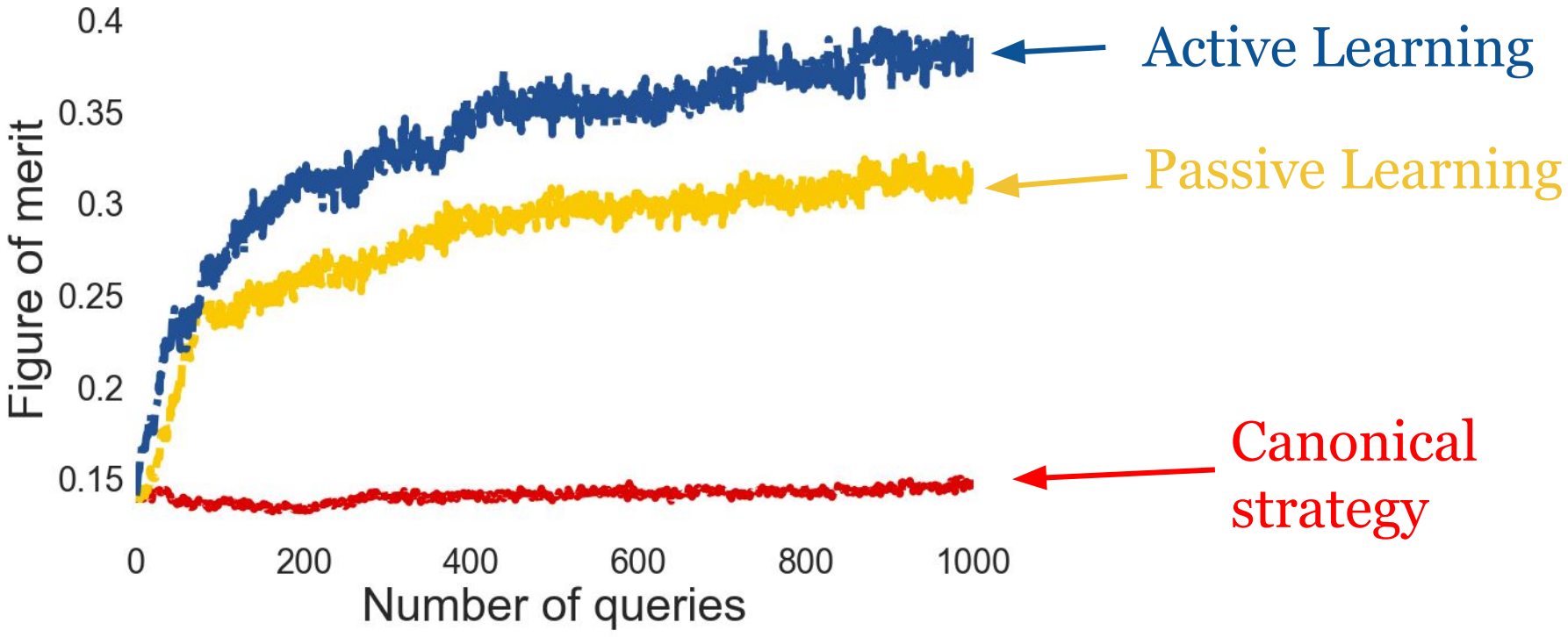
Optimal classification, minimum training



AL for SN classification

COIN

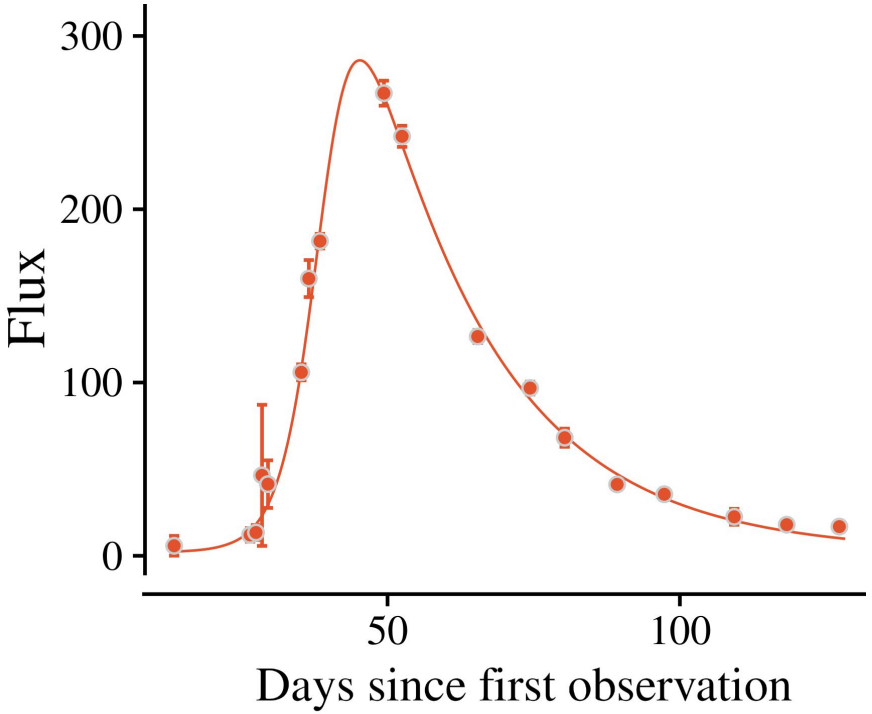
Static results



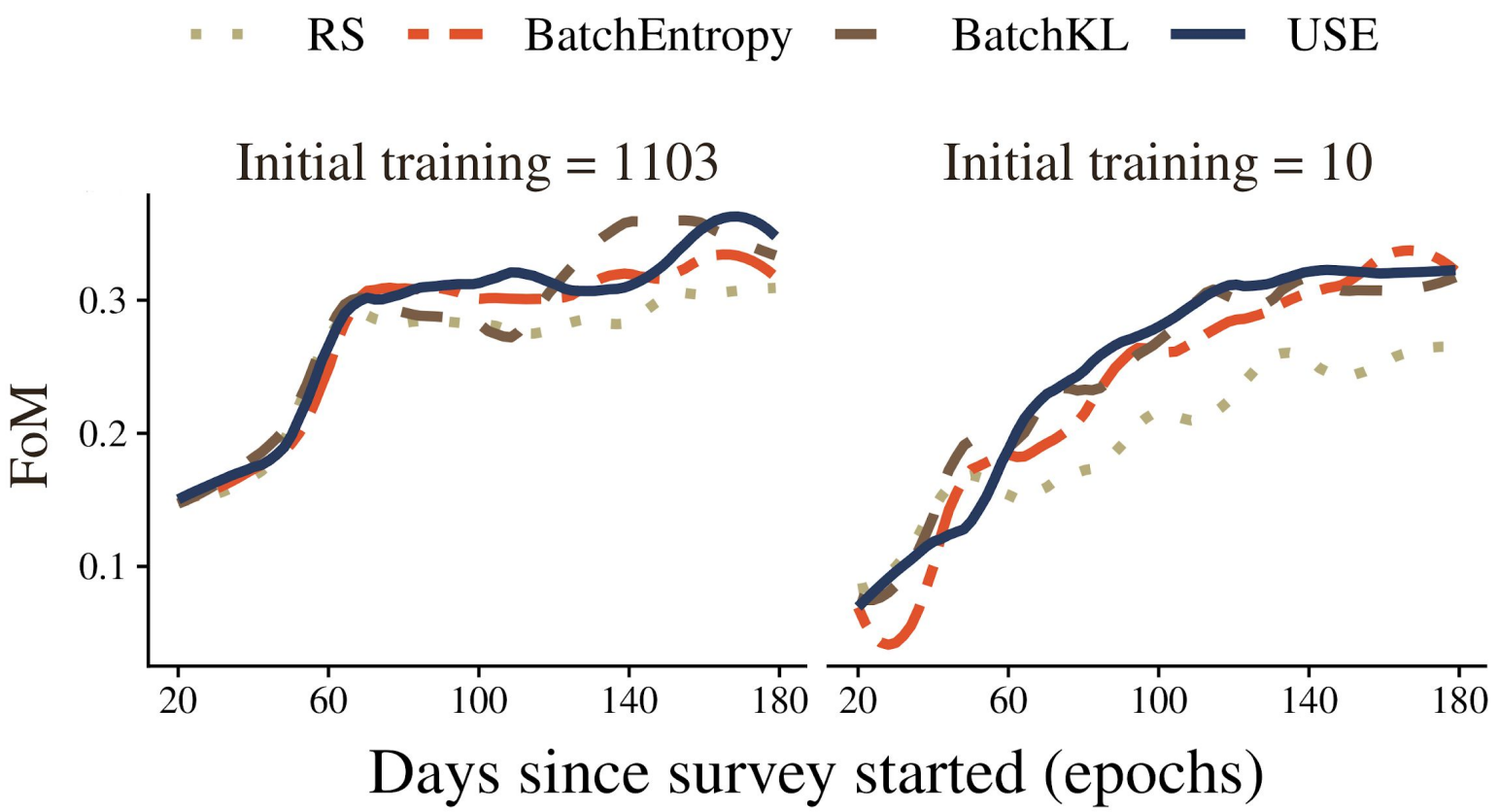
From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

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
- Evolving Costs
  - Observing costs for a given object changes as it evolves.

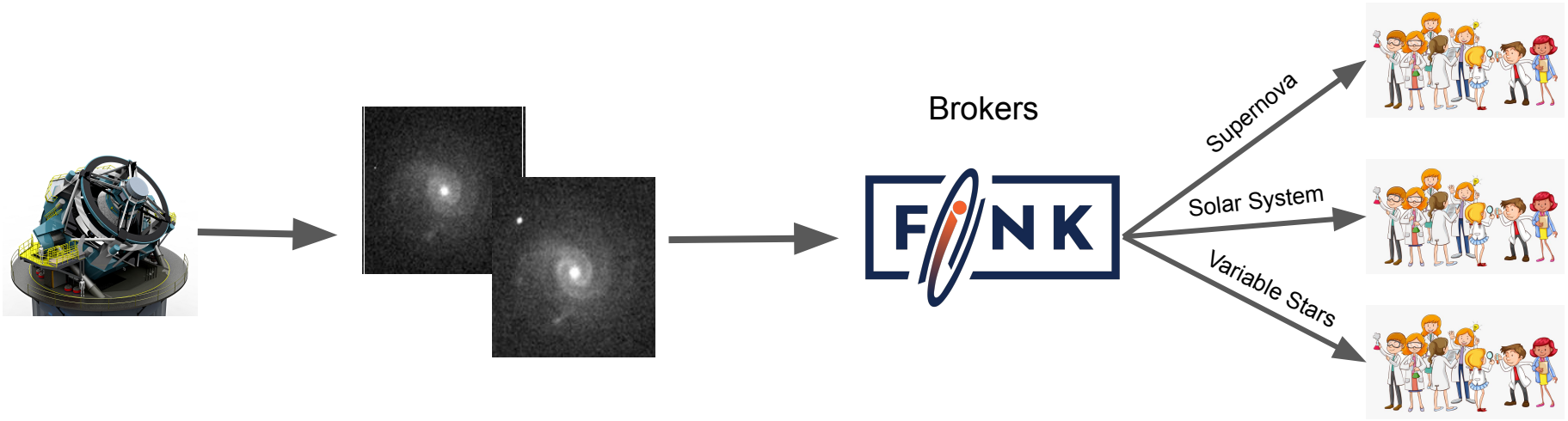


More is not always better

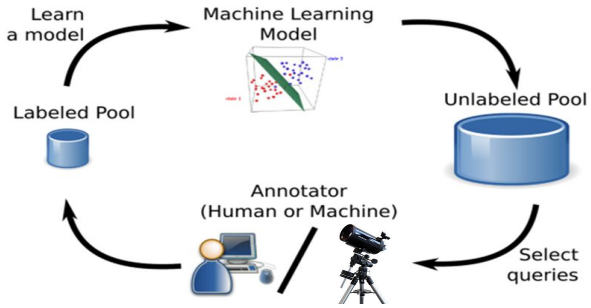


Kennamer, Ishida et al., 2020 - arXiv:astro-ph/2010.05941 - LSST-DESC and COIN, the RESSPECT team

Now, on real data



- Already available data
- Start with 20 randomly chosen objects (10 of each class)
- Run through 40 iterations



*Moller, Peloton, **Ishida** et al., - arXiv:astro-ph/2009.10185*

https://fink-broker.org/

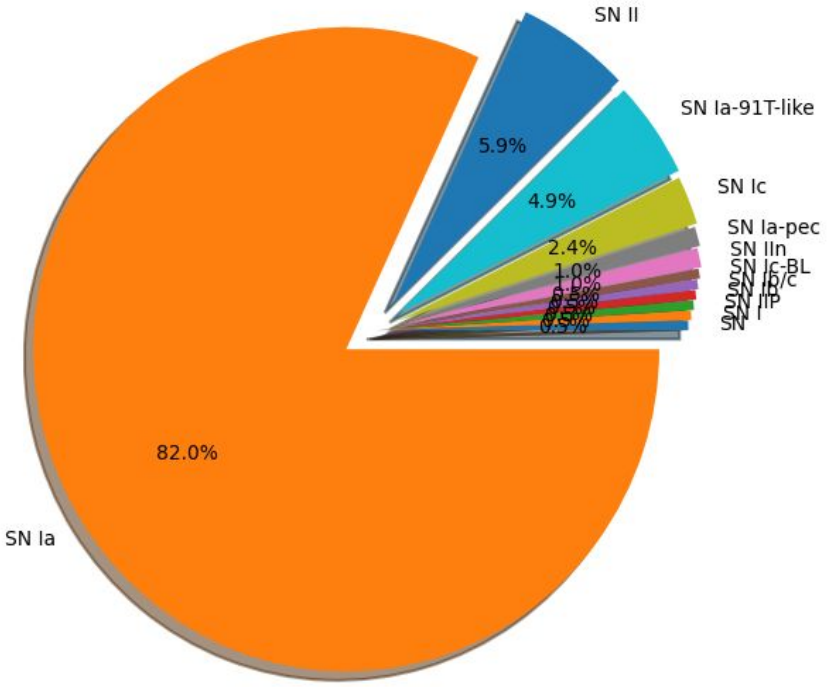
Early SN Ia classifier in Fink

Since 2020/11:

- Each night, we submit *Early SN Ia* candidates to TNS for spectroscopic follow-up
  - The submission is done via bot
  - No human inspection required

- As of 22/03/2021, 310 candidates were submitted (about 2/day)

- As of 22/03/2021, 205 candidates were classified as SN (about 1.5/day)



https://fink-broker.org/

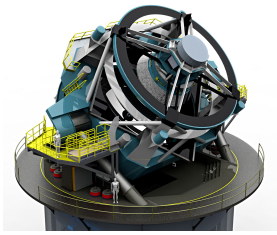
*Leoni, **Ishida** et al., in prep*

We should build recommendation systems for our targets of interest

- Training samples can be optimize to boost machine learning efficiency on your science case
- There are caveats in using machine learning and we should avoid off-the-shelf and black bloxes applications

~10 million transient candidates/night

Over a total life span of 10 years



https://www.lsst.org/

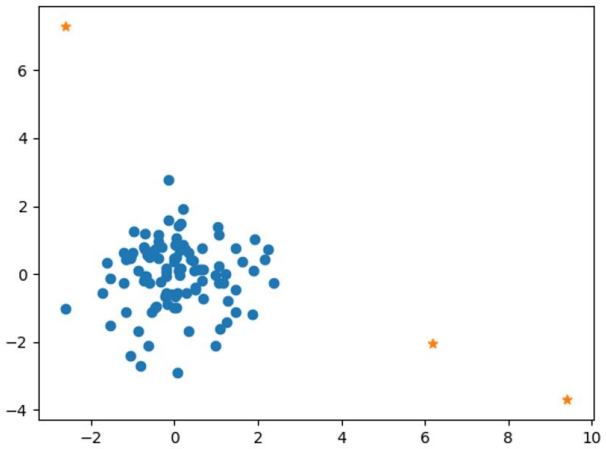
II. Unsupervised learning

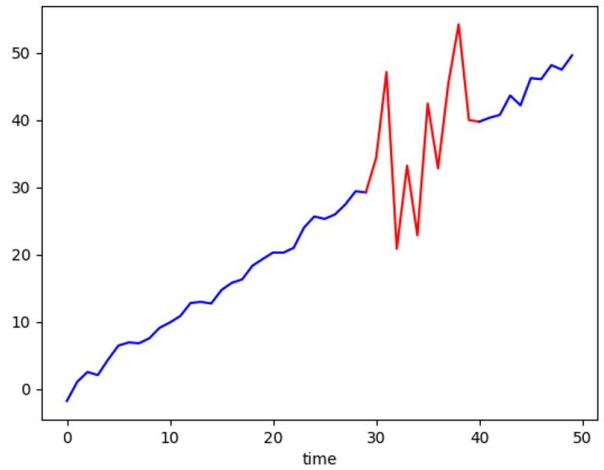
In algorithmic terms

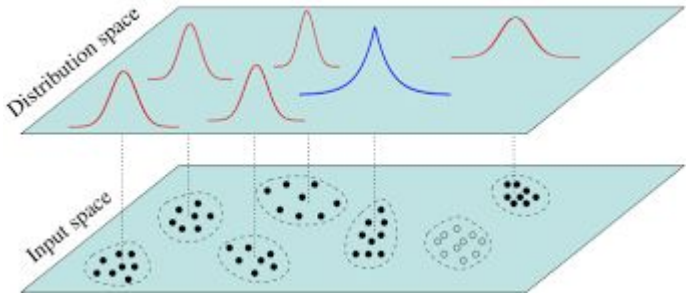
Anomaly Detection

"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







Philosophically,

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"

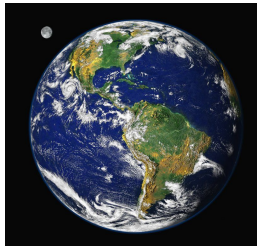
Hawkins, 1980

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)

Machine Learning only produces recommendations

Potentially interesting anomalies:

- Candidate 1
- Candidate 2
- Candidate 3
- …
- …
- ...
- ...

Expert analysis

Not interesting

Interesting

Very interesting!

Get more data

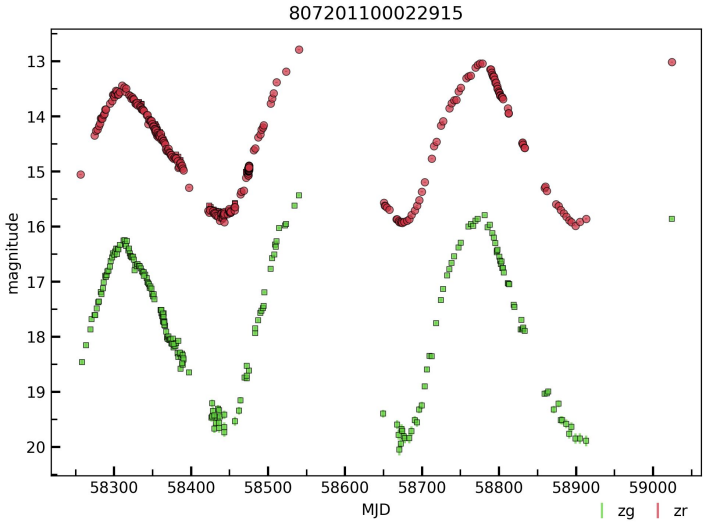
or

Publication

Observed data set

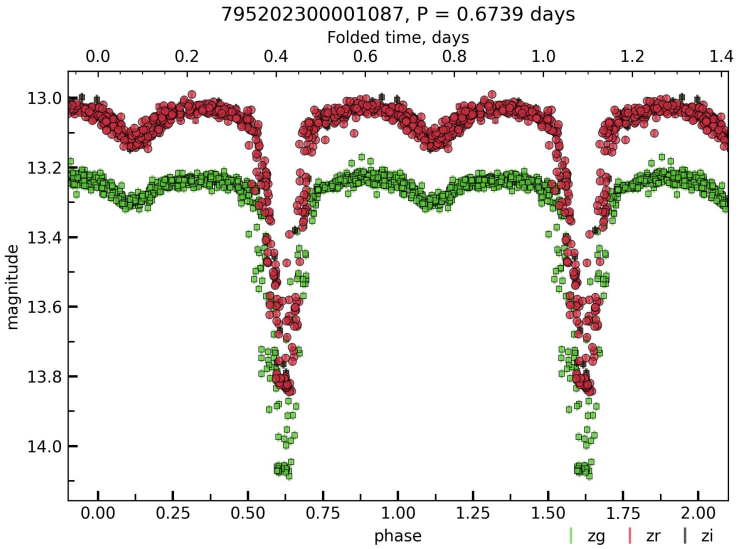
Anomaly detection algorithm

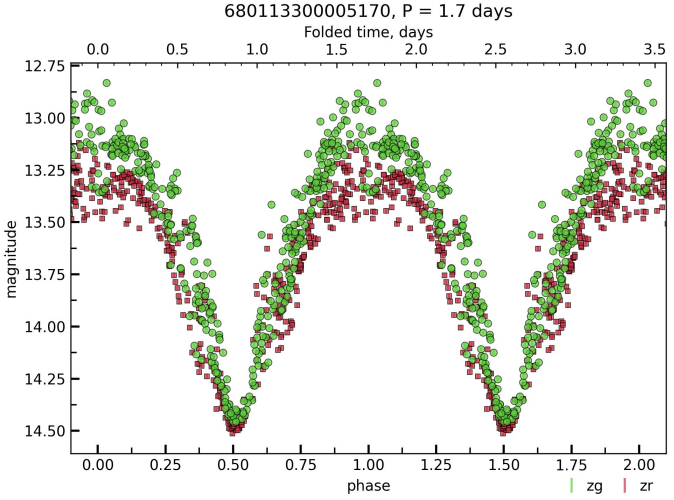
Zwicky Transient Facility DR3



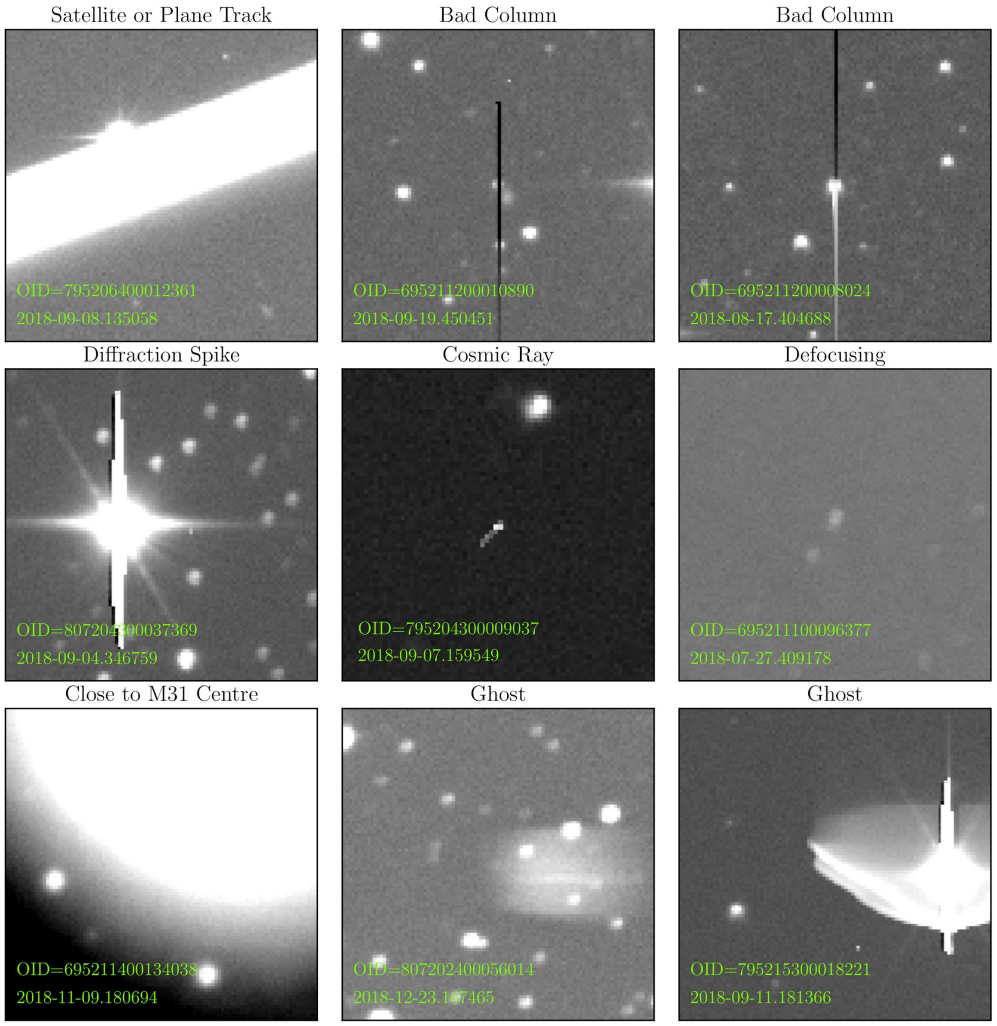
ZTF Data Release 3 was expected to contain stars and periodic variables (no transients)

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



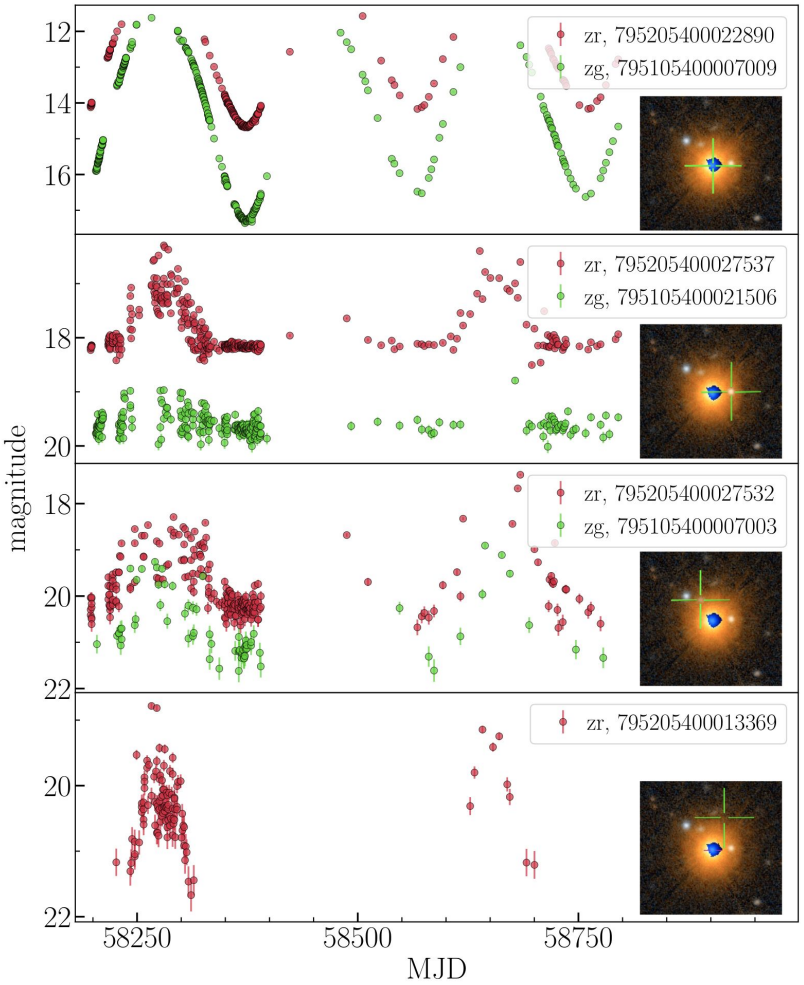


From ZTF DR3: Examples of artifacts



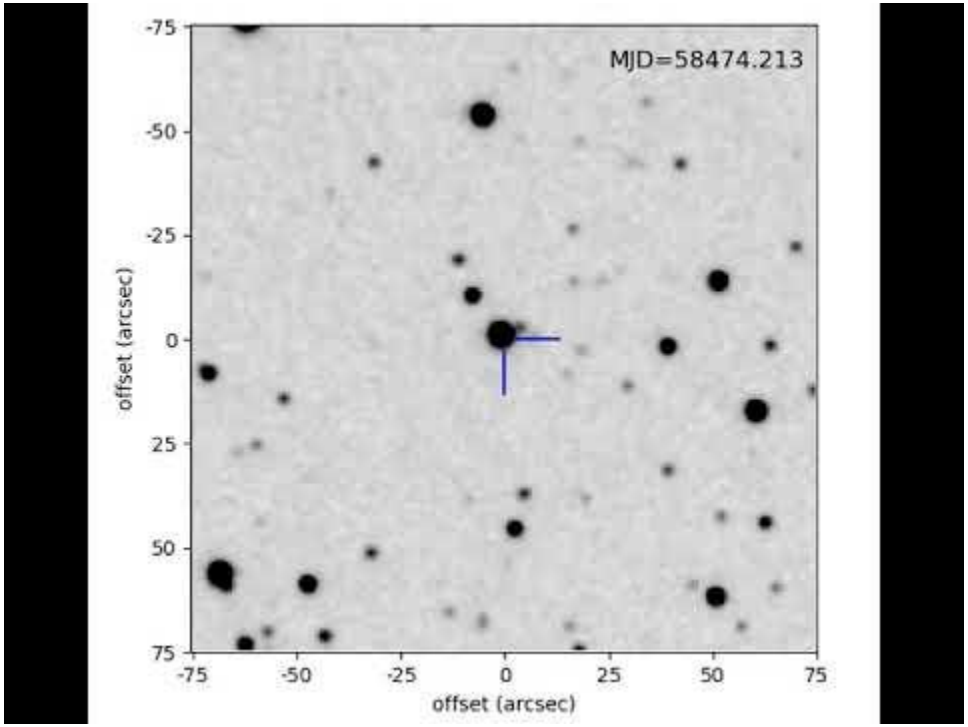
Malanchev et al., 2020 - MNRAS - https://arxiv.org/abs/2012.01419

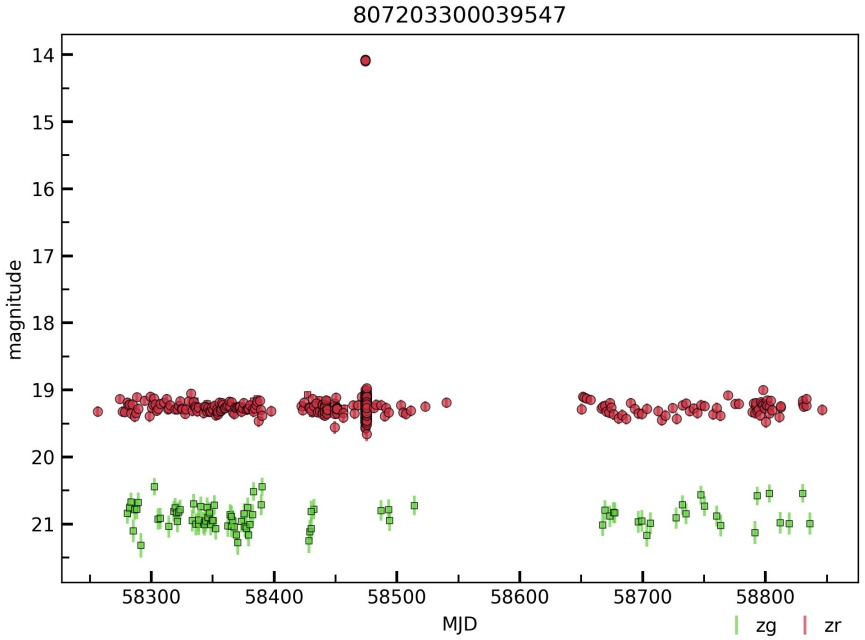
From ZTF DR3: IW Dra and its echoes



Malanchev et al., 2020 - MNRAS - https://arxiv.org/abs/2012.01419

From ZTF DR3: The Barcelona asteroid





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

Philosophical question:

What is a scientifically interesting anomaly?

Problem: Still high incidence of "non-important" anomalies *(68 % for ZTF DR3)*

Goal: Maximize the number of scientifically interesting anomalies shown to the expert

Strategy:

Incorporate human knowledge in the machine learning model

a. k.a. adaptive learning ...

The recommendation system can get better with time …

Machine Learning only produces recommendations

Potentially interesting anomalies:

- Candidate 1
- Candidate 2
- Candidate 3
- …
- …
- ...
- ...

Expert analysis

Not interesting

Interesting

Very interesting!

Get more data

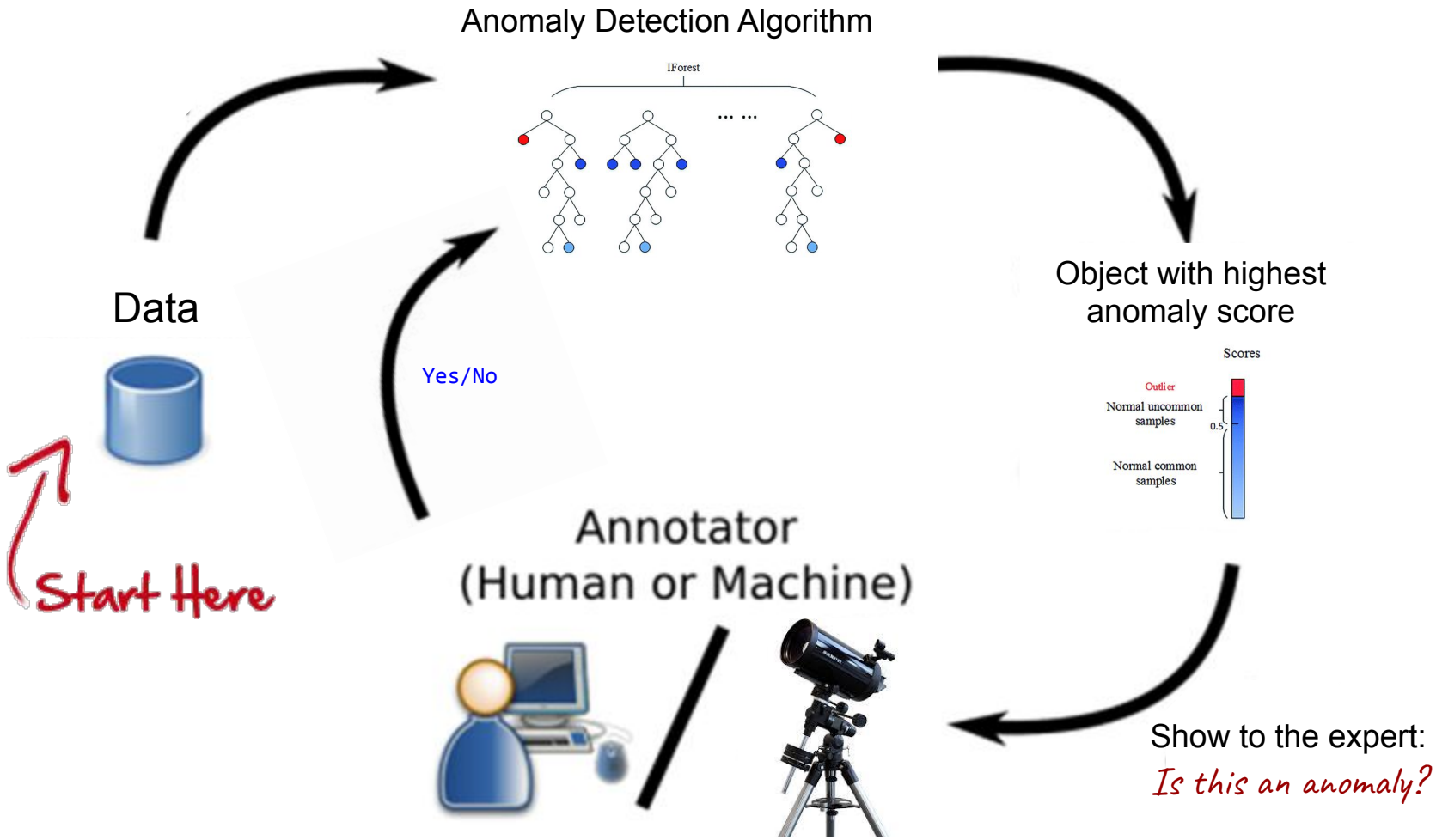
or

Publication

Observed data set

Anomaly detection algorithm

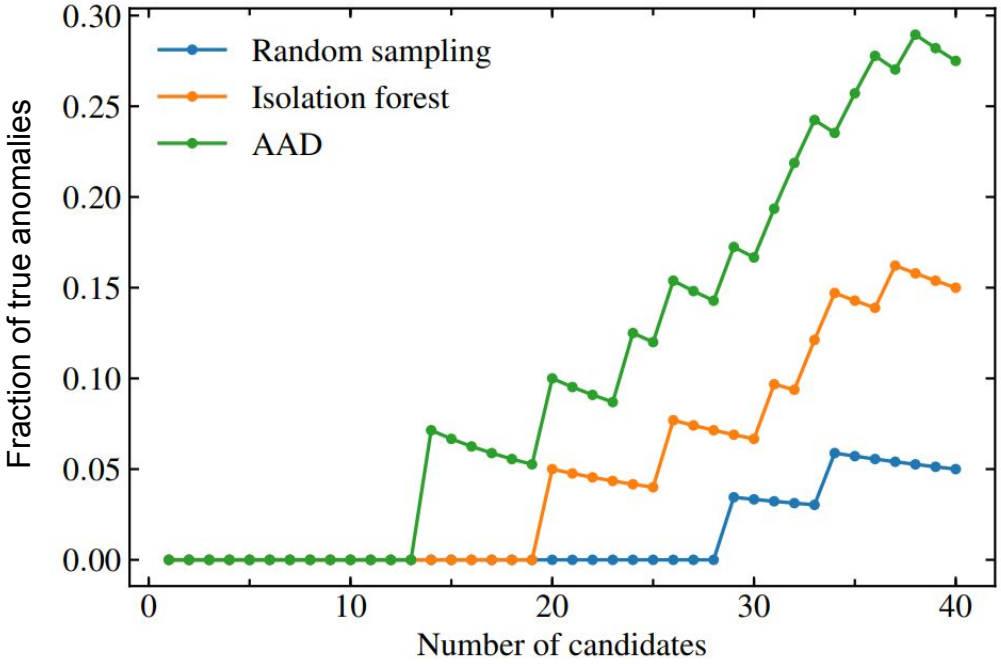
Active Anomaly Detection



Das, S., et al., 2017, in Workshop on Interactive Data Exploration and Analytics (IDEA'17), KDD workshop, arXiv:1708.09441

The Open Supernova Catalog

SNAD

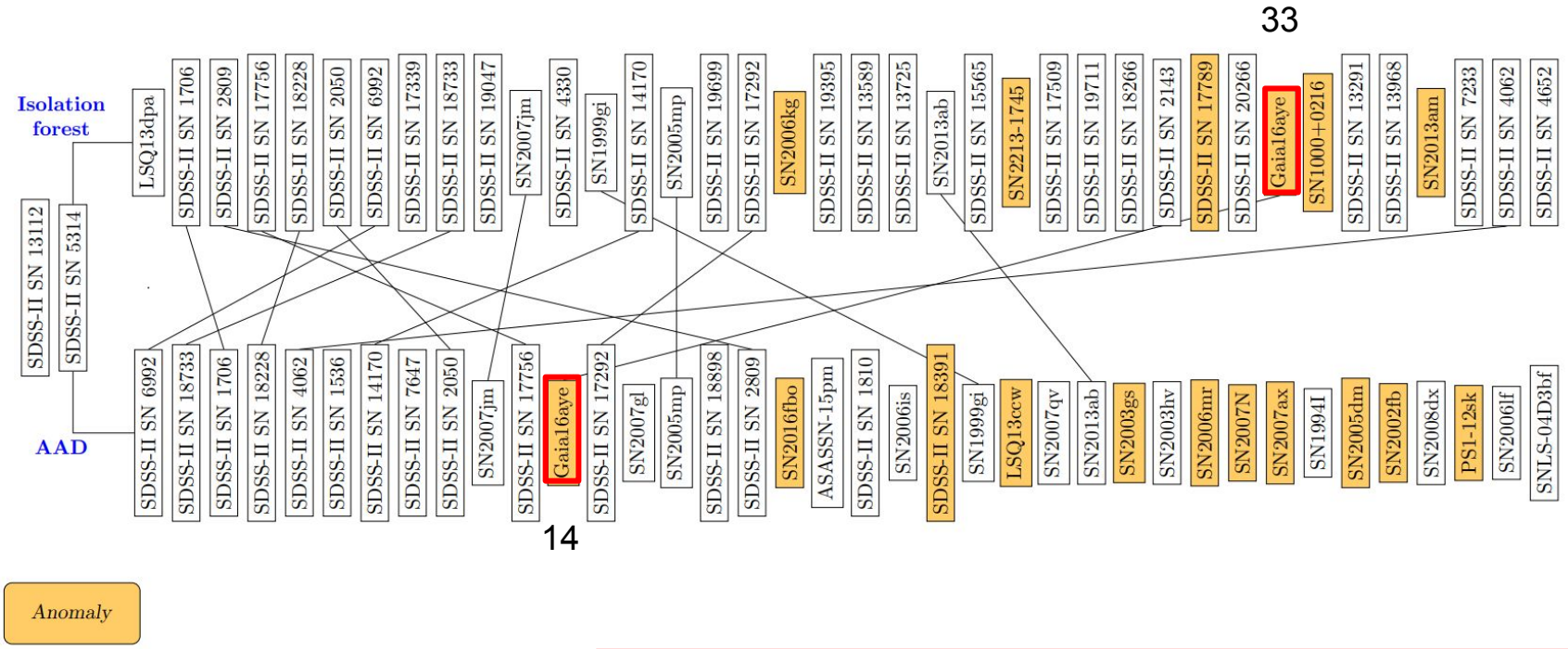


AAD was able to increase the incidence of true anomalies presented to the expert in 80%

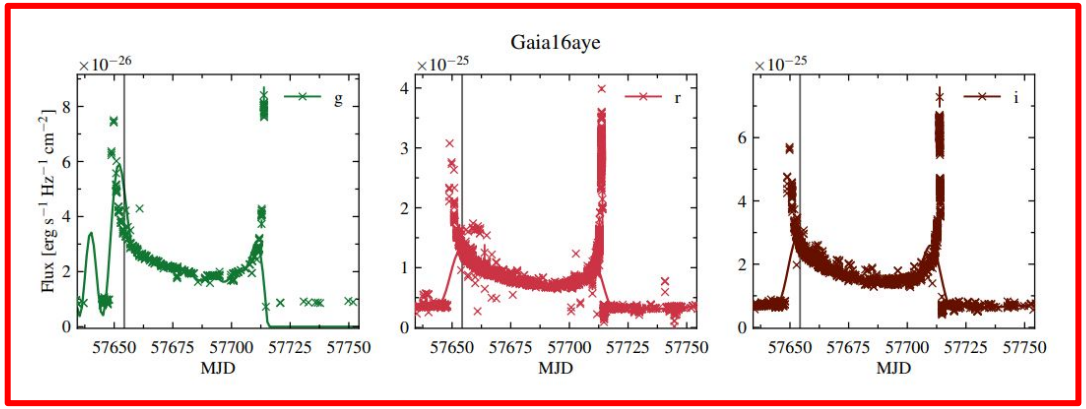
Ishida et al., 2019 - https://arxiv.org/pdf/1909.13260v1.pdf

https://snad.space/

AAD on real data: The Open Supernova Catalog



Fast identification of binary microlensing event



Ishida et al., 2019 - https://arxiv.org/pdf/1909.13260v1.pdf

SNAD

We should plan for the unknown

- In the era of big data, serendipitous discoveries will **not** happen
- Adaptive Learning strategies have the potential to allow the **discovery** with limited spectroscopic resources

PS: The SNAD team has just published a set of missed supernova in a variable star catalog found through an active anomaly detection technique:

https://snad.space/news/sn_tns/index.html

THANK YOU

https://cosmostatistics-initiative.org/

Extra slides

Experimental details

- Simulated data covering 180 observation days
- Data separated into four groups
  - Original training set - 1,103
  - 18,216 in pool set
  - 1,000 objects each in validation and test sets
- Assumed access to an 8m and 4m telescope for labeling
  - 6 hours per telescope on each night
- Pre-processed data with parametric fits (Bazin *et al.* 2009)
- Observing Costs calculated from brightness estimates of each objects and telescope properties

Active Learning details

- Ensemble of Random Forest Classifiers for query decisions
- **Four** Active Learning Strategies under **knapsack constraints**:
  - Random Sampling
  - Uncertainty Sampling
    - Entropy used to measure uncertainty
  - Batch Entropy
    - Measures a joint entropy over batches
    - Takes advantage of submodular properties of entropy
  - Batch KL-Divergence
    - Measures a Joint KL-Divergence/Mutual Information, equivalent to BatchBALD
    - Takes advantage of submodular properties of the KL-Divergence/Mutual Information