Human in the loop:

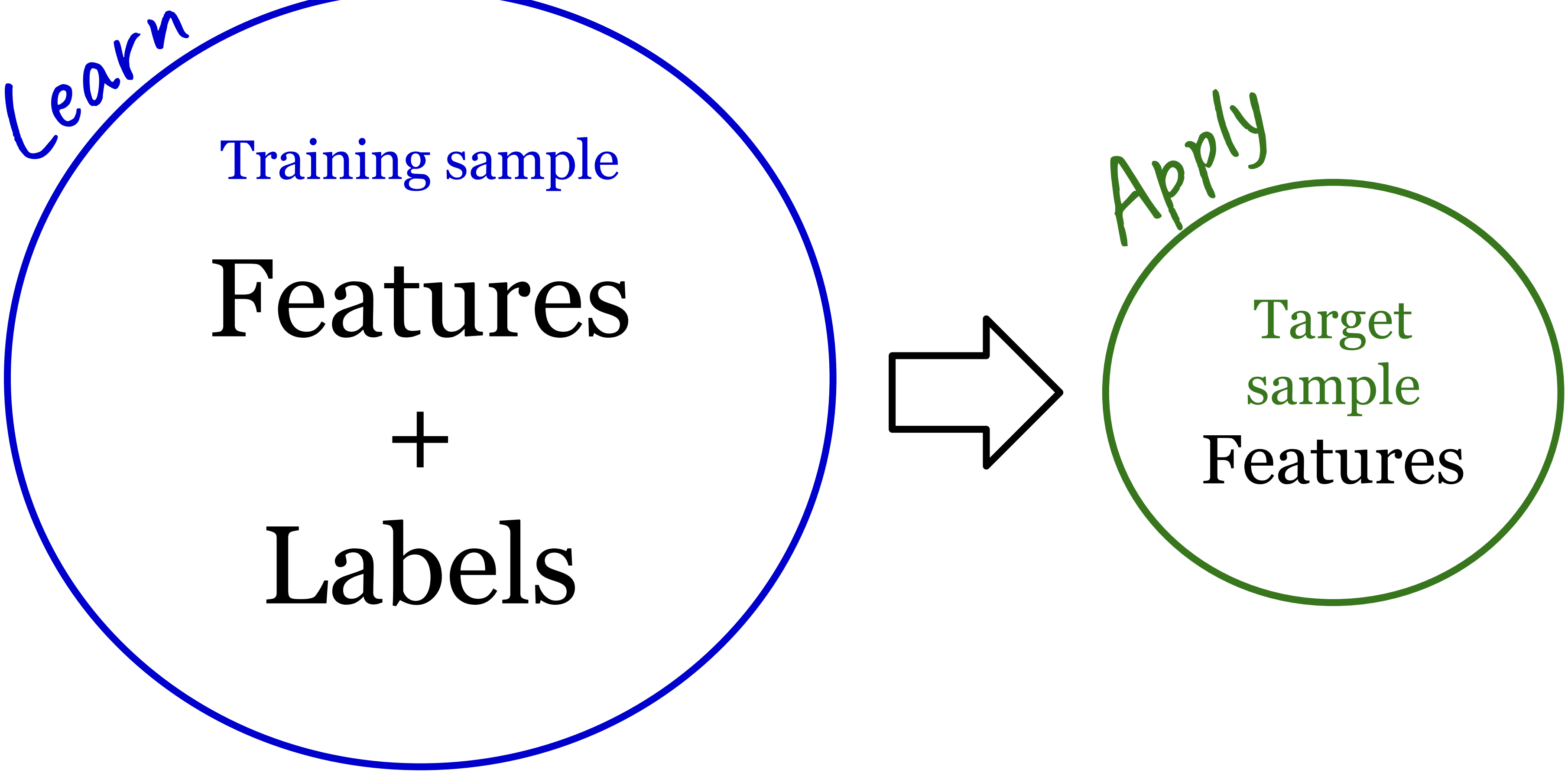
Active Learning in Astronomy

Artificial Intelligence in Astronomy - ESO, Germany
22 June 2019

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

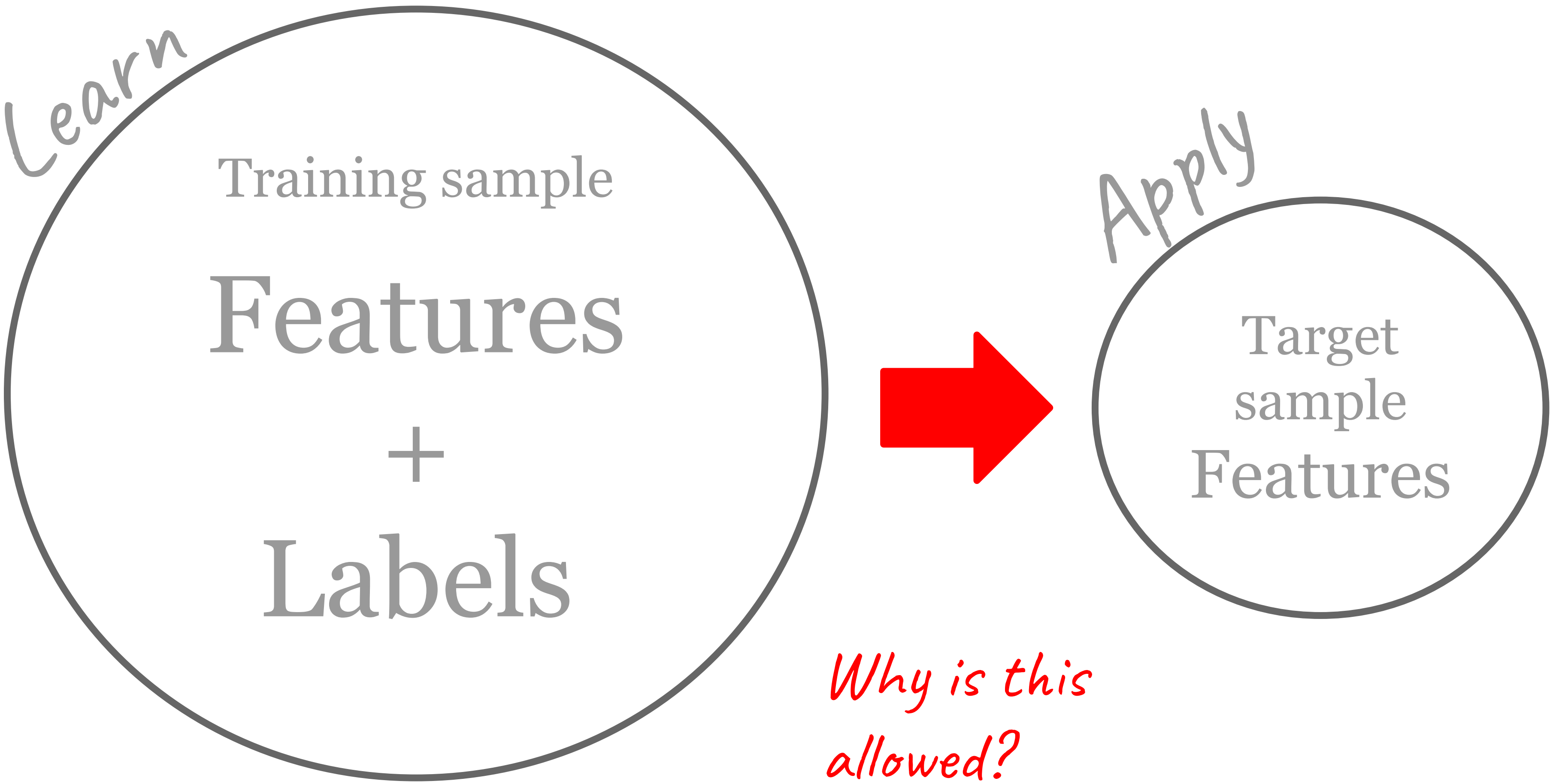
Supervised Learning

Ideal data scenario



Supervised Learning

Learn by example



Why is this allowed?

Supervised ML model

data training, target

χ set of all samples, x

Y set of possible labels, y

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Hypothesis: Training is representative of target

Data generation model:

$x_i \sim P_{\chi}$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x \sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Supervised ML model

data training, target

$\mathcal{X}$ set of all samples, x

Y set of possible labels, y

Machine Learning algorithm

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Hypothesis: Training is representative of target

Data generation model:

$x_i \sim P_{\mathcal{X}}$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Supervised ML model

data training, target

χ set of all samples, x

Y set of possible labels, y

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Hypothesis: Training is representative of target

Data generation model:

$x_i \sim P_\chi$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

How often does your data fulfill these requirements?

Ideal Supervised learning situation

In astronomy

Learn

Training sample

Images, colors, light curves, etc.

+

Classes

(spectra)

Apply

Target sample

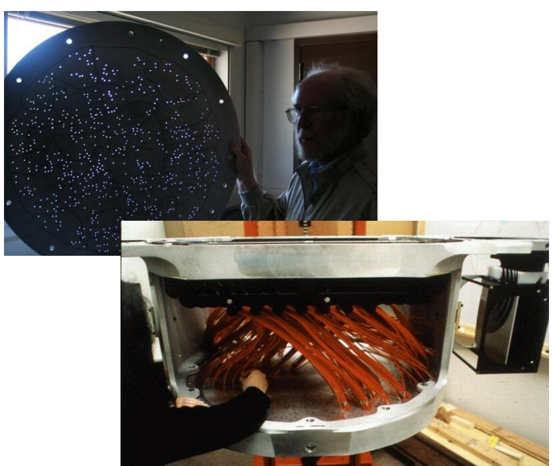
Images, colors, light curves, etc.

Astro: supervised learning situation

In astronomy, labels ⇒ spectra

Similar examples

Labels are often far too expensive!



Similar examples

Labels are often far too expensive!

amazon

35% of Amazon's revenue are generated by it's recommendation engine.

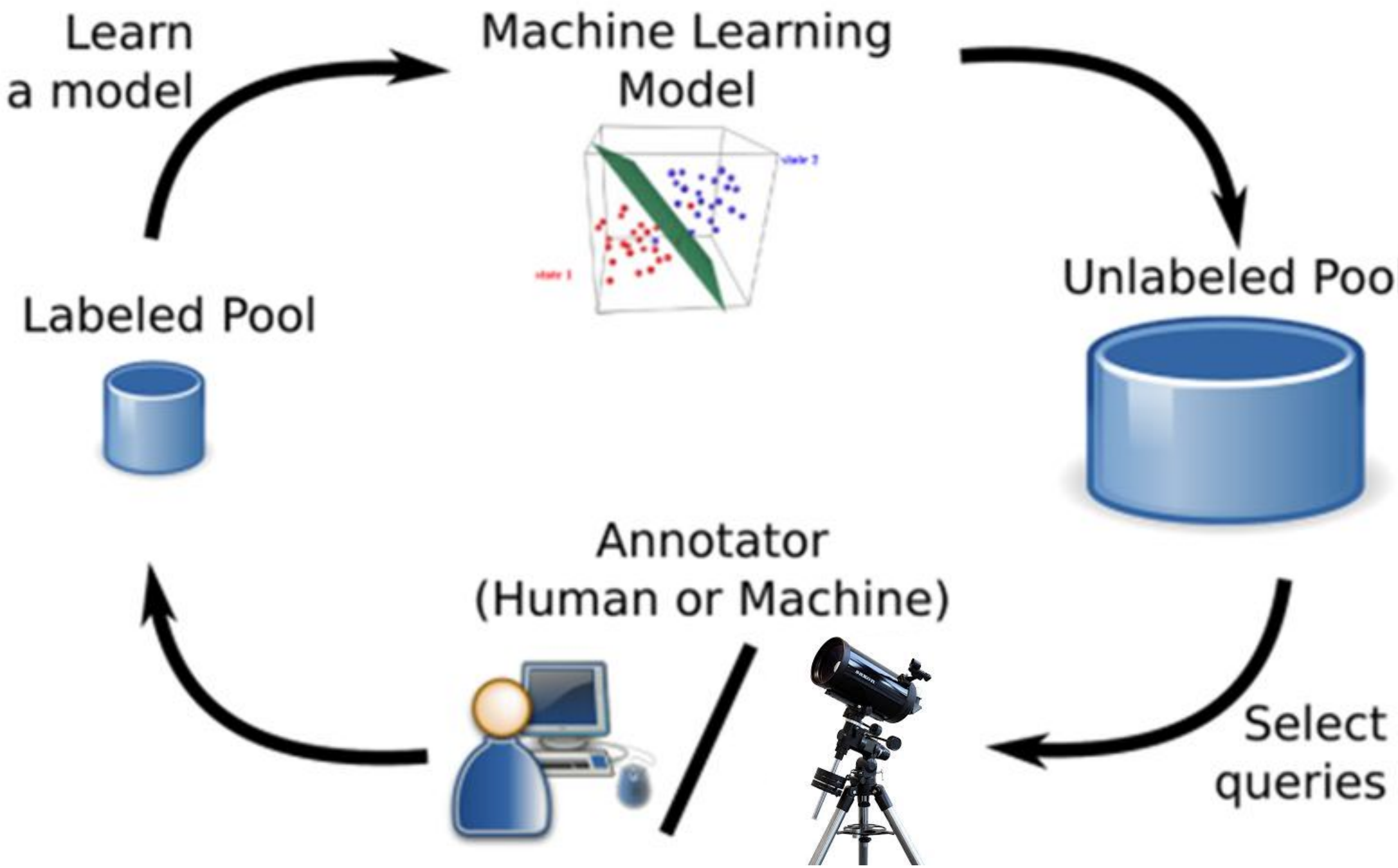
NETFLIX

75% of users select movies based on Netflix's recommendations.

Given limited resources, we need recommendation systems!

Active Learning

Optimal classification, minimum training



Optimal Experiment Design

In Statistics literature

$$PQ_{data,f}(x) \equiv P_{x\sim data}(h_{train}(x) \neq f(x) \mid previous\ results)$$

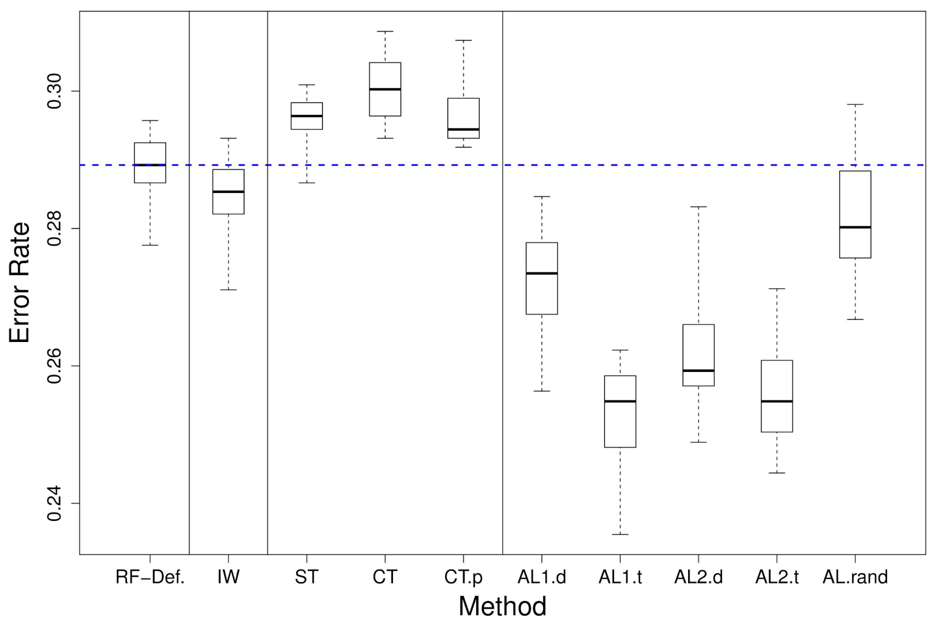
- Pool based
- Generative
- Sequential

Estimation of stellar population parameters

Simulated catalogs - Solorio et al, 2005

Classification of variable stars

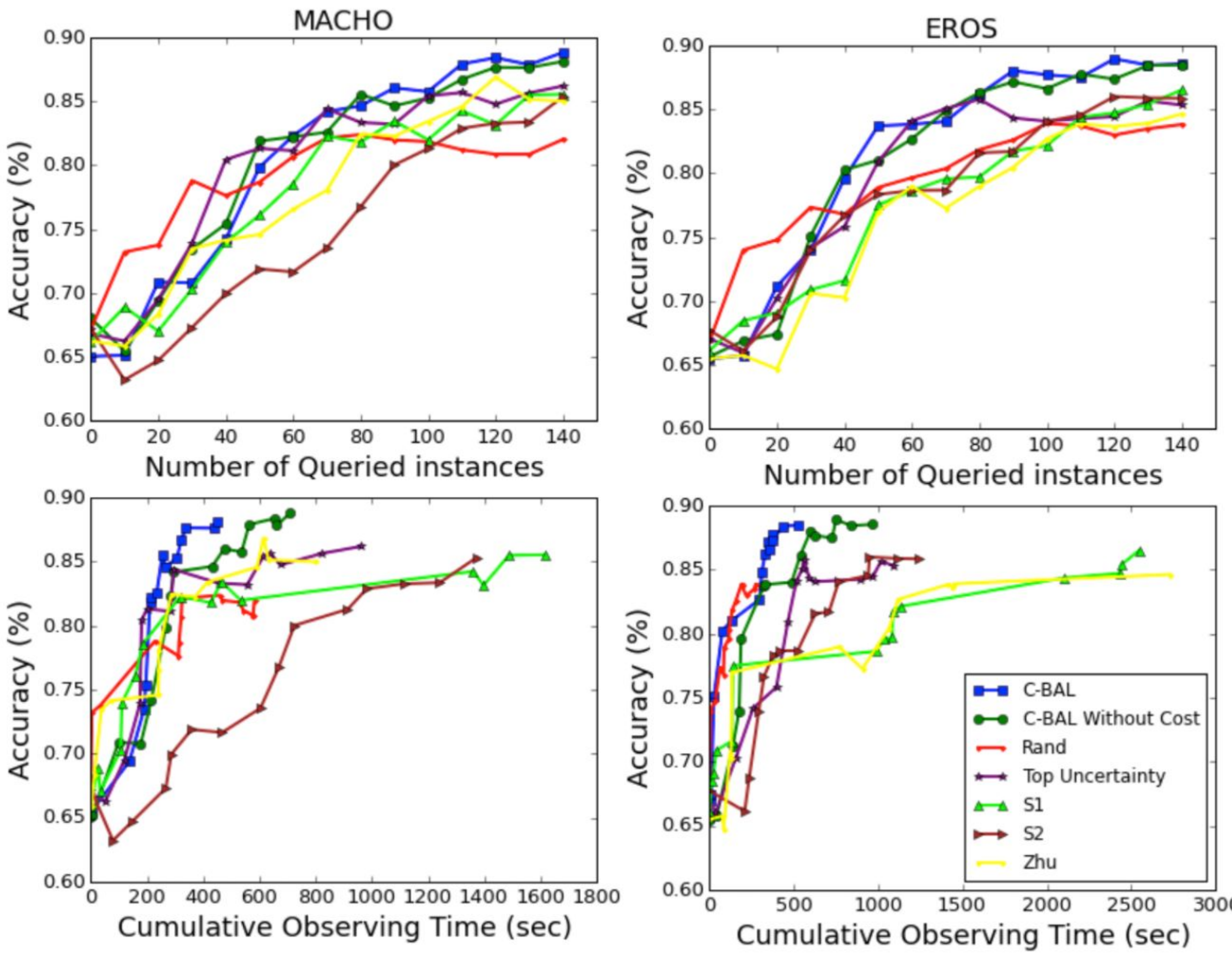
Real data, expected error change - *Richards et al., 2012*



Active Learning in Astronomy:

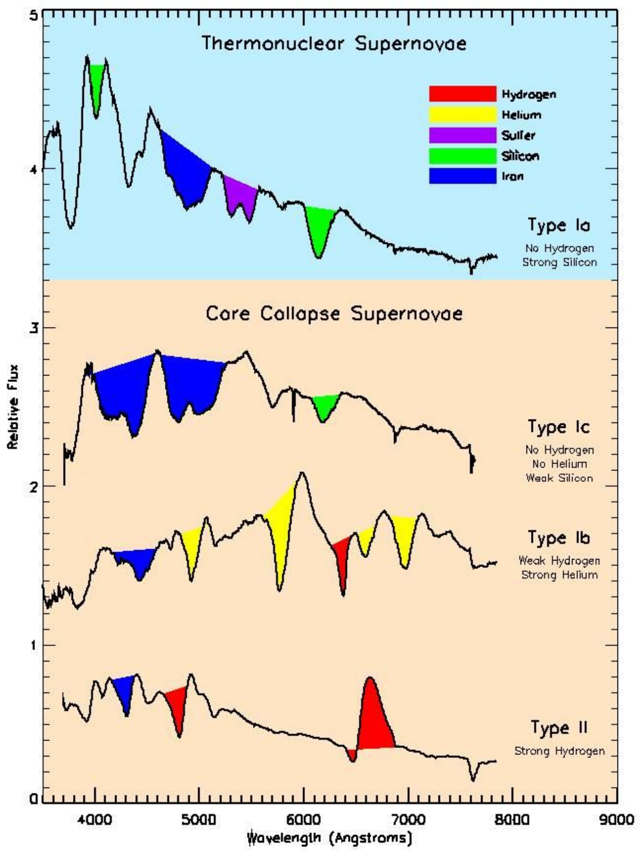
Choosing where to point the telescope

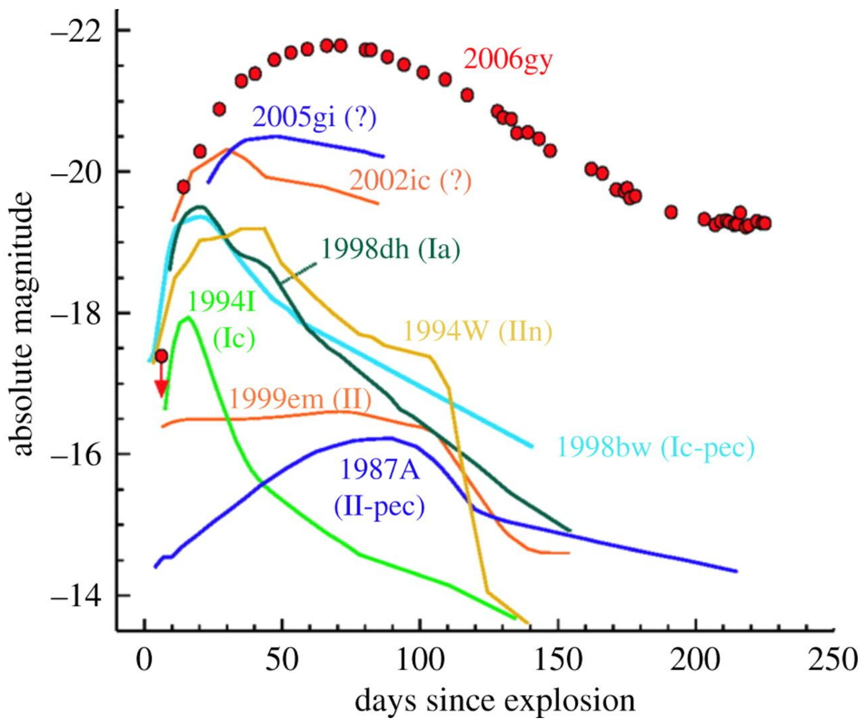
Real catalog - cost sensitive - *Xia et al., 2016*



Example of application to astro:

Supernova photometric classification

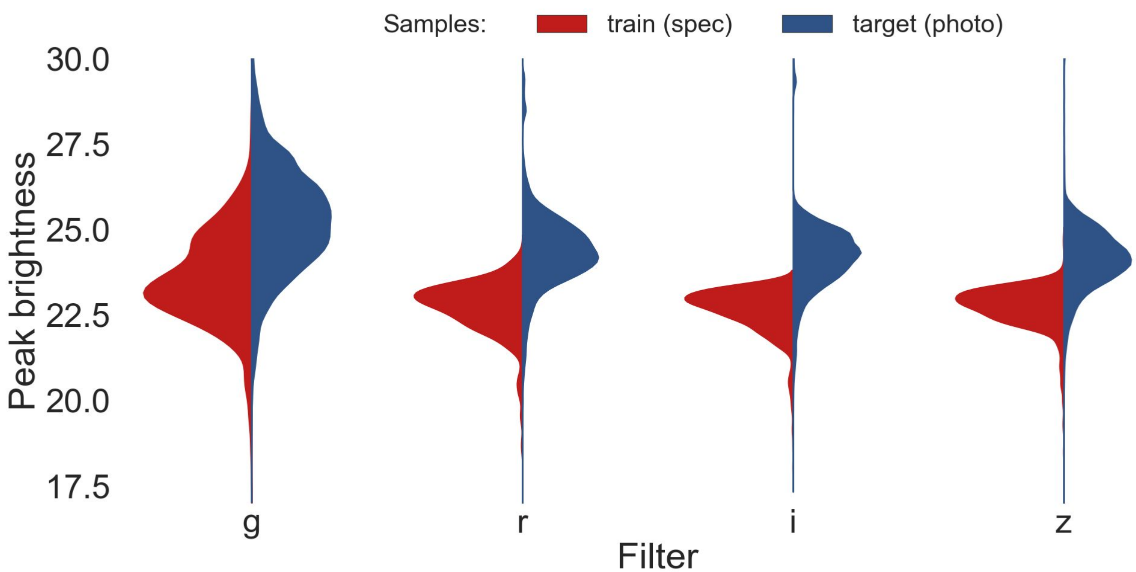


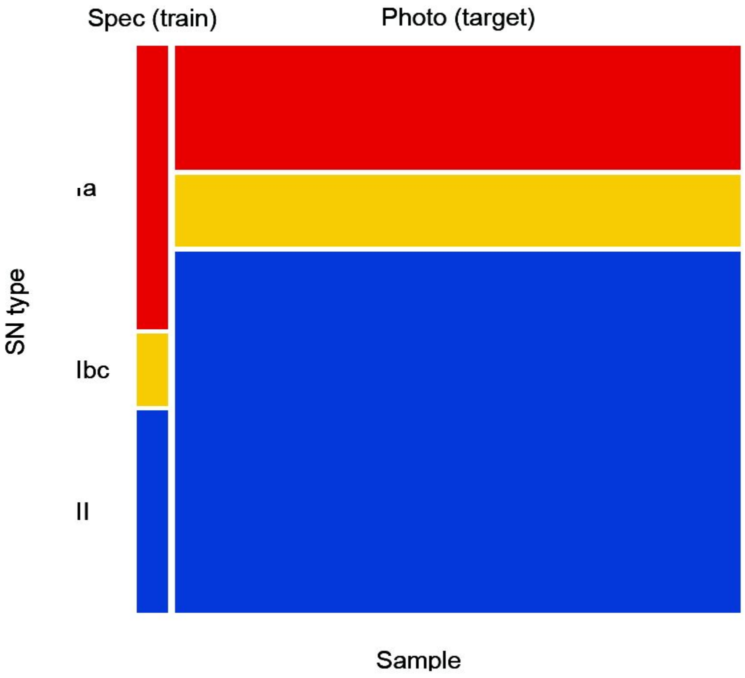


For more on SN classification see
Anais Moller's talk this afternoon!

http://rsta.royalsocietypublishing.org/content/370/1960/774

Representativeness





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

AL for Supernova classification

A strategy

Parametric Fit
5 parameters

$$f(t) = A\frac{e^{-(t-t_0)/\tau_f}}{1 + e^{(t-t_0)/\tau_r}} + B,$$

Random Forest

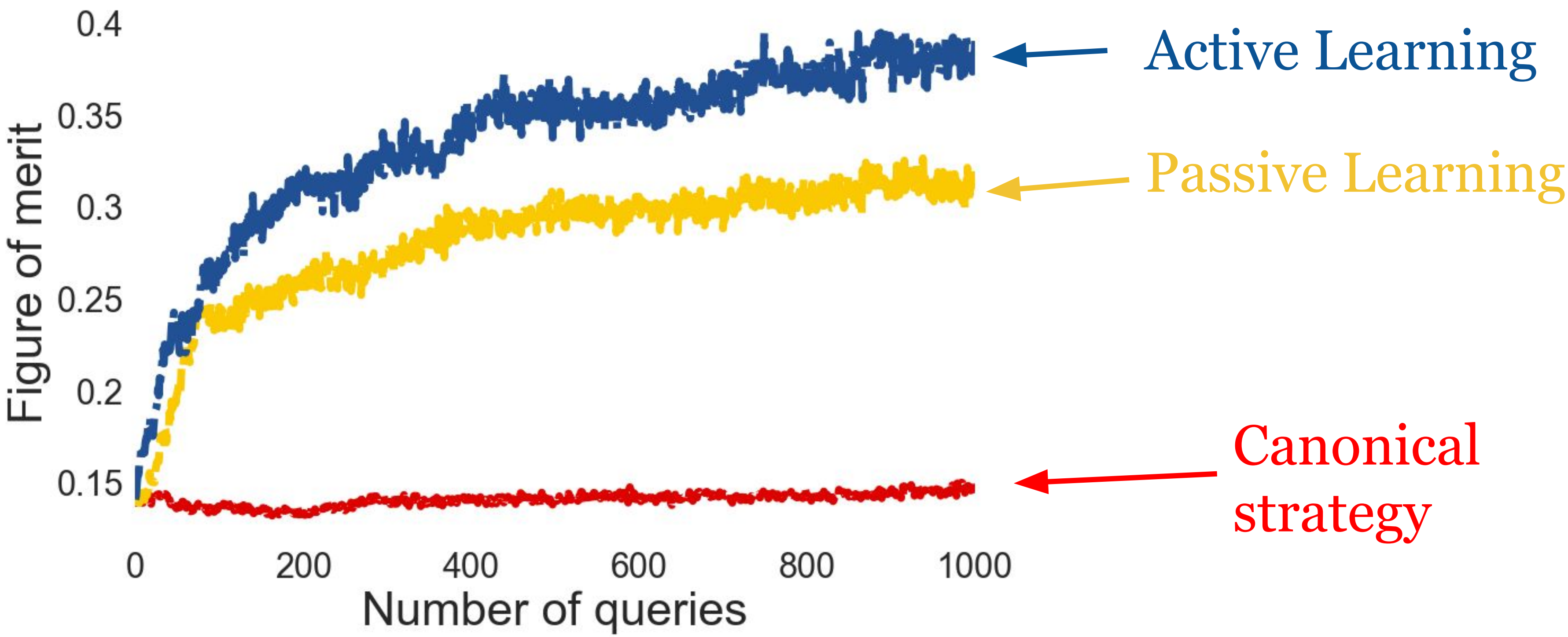
Photo-class uncertainty

Query the SN with highest uncertainty

LC+labels
LC

AL for SN classification

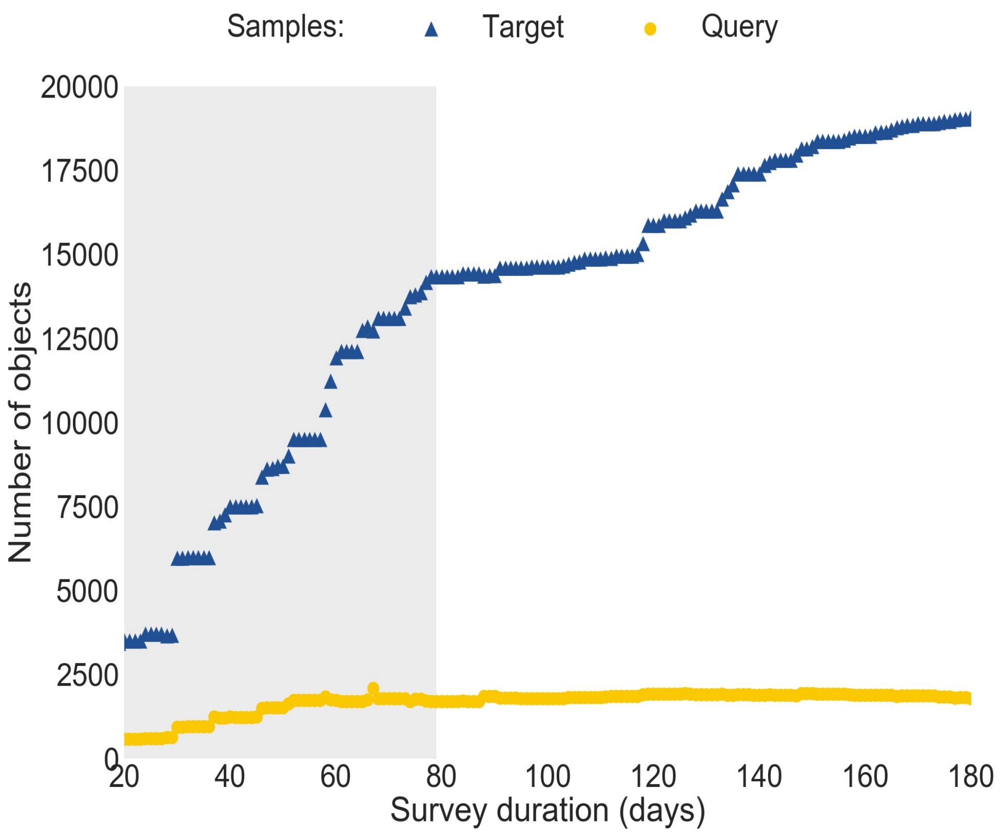
Static results



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

Complications: SNe are transients

Not everything is available for labelling

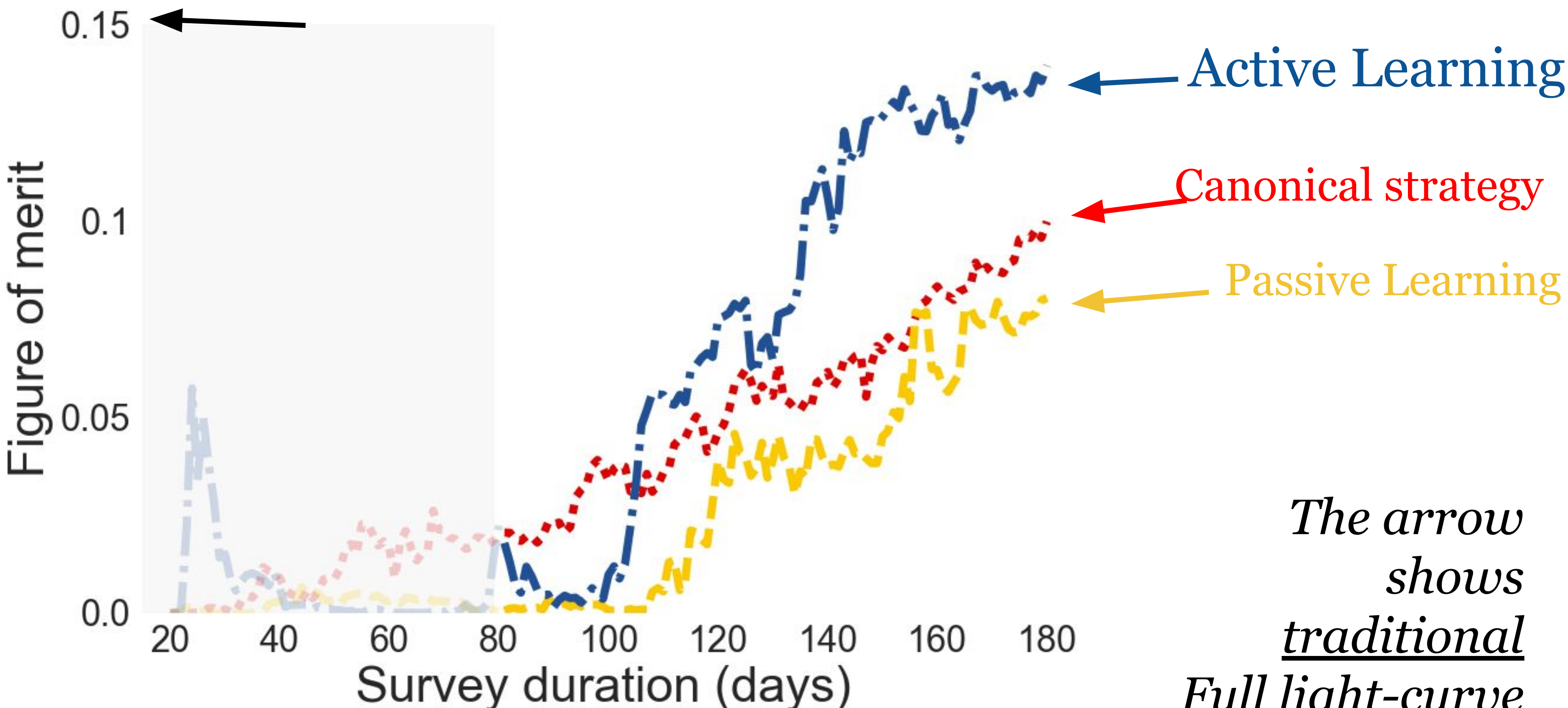


1. Feature extraction done daily **with available observed epochs until then.**

2. Query sample is also re-defined daily: objects with **r-mag < 24**

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

Do we even need a training set?



The arrow shows <u>traditional</u> Full light-curve results with full SNPCC spec 1103 spectra

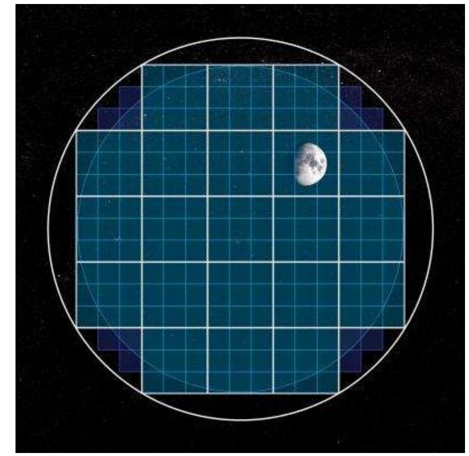
For more on The Cosmostatistics Initiative see Rafael de Souza's talk on Friday!

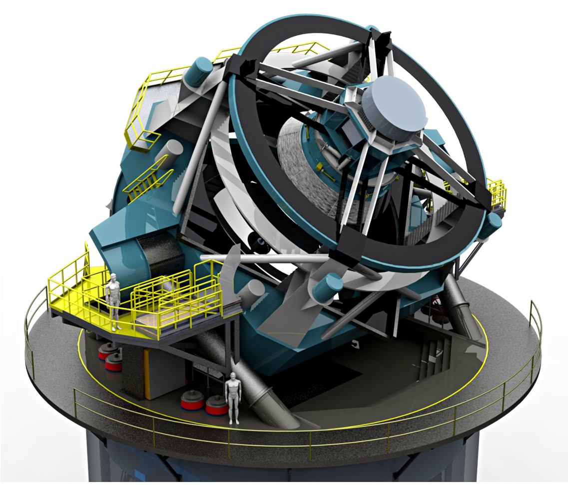
From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

The Large Synoptic Survey Telescope

Photometric obs:
~minute

Spectroscopic obs:
>= 1 hour (e.g. SDSS)
Multi-fiber spec.
Pointing is not trivial





Camera: 3**.2 Giga** pixels and 1.65m

Primary mirror: **8.4m**

Field of view: **3.5 deg,** 40x full moon

Data production :**15 TB/night**

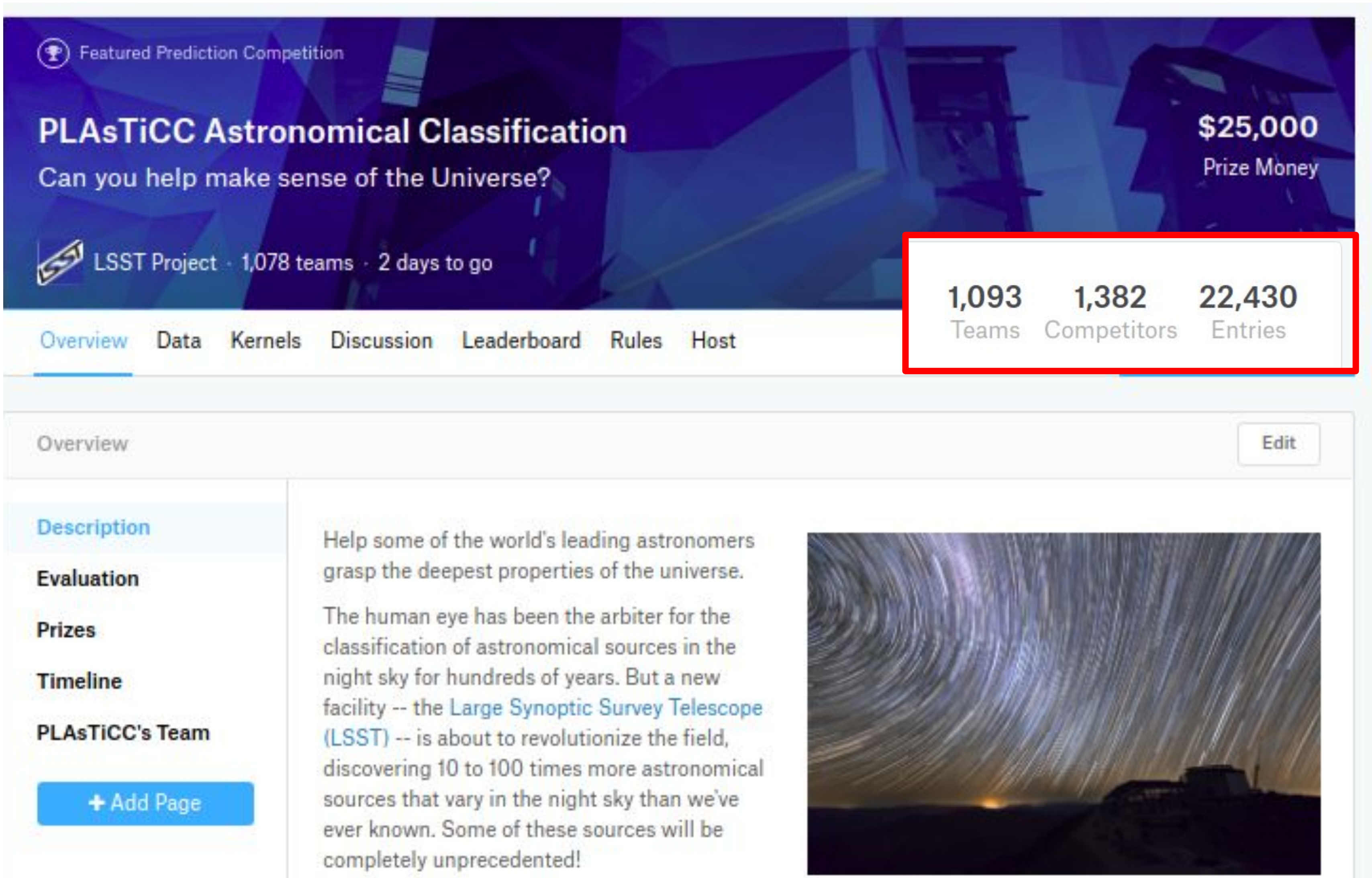
(3yr LSST=internet today)

~10 million alerts/night

30.000 type Ia SN/yr (today ~1000)

Expected ~ **1000 spectra/yr** (~ 3%)

https://www.kaggle.com/c/PLAsTiCC-2018

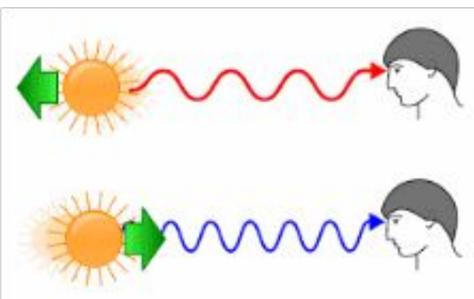


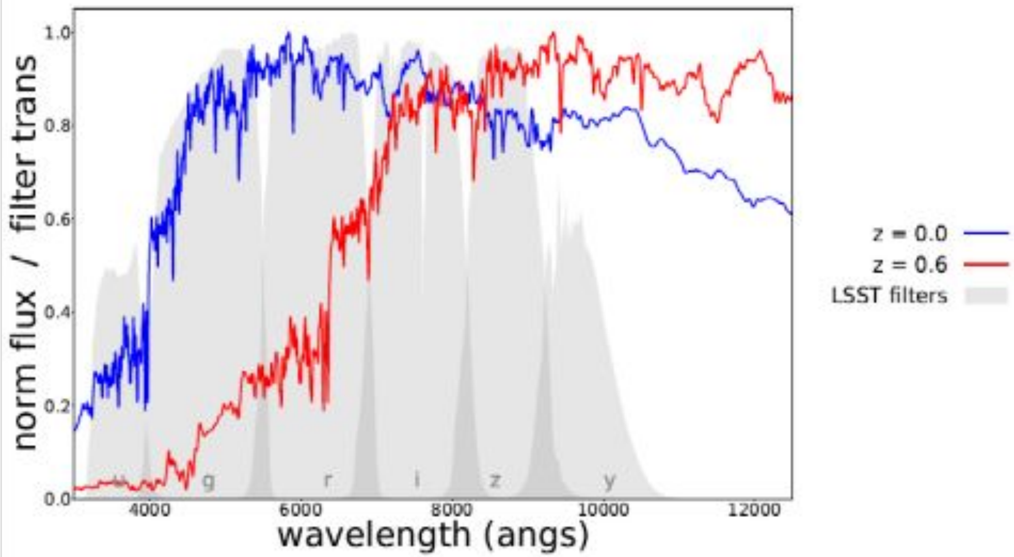
For more on PLAsTiCC see Mi Dai's talk on Wednesday!

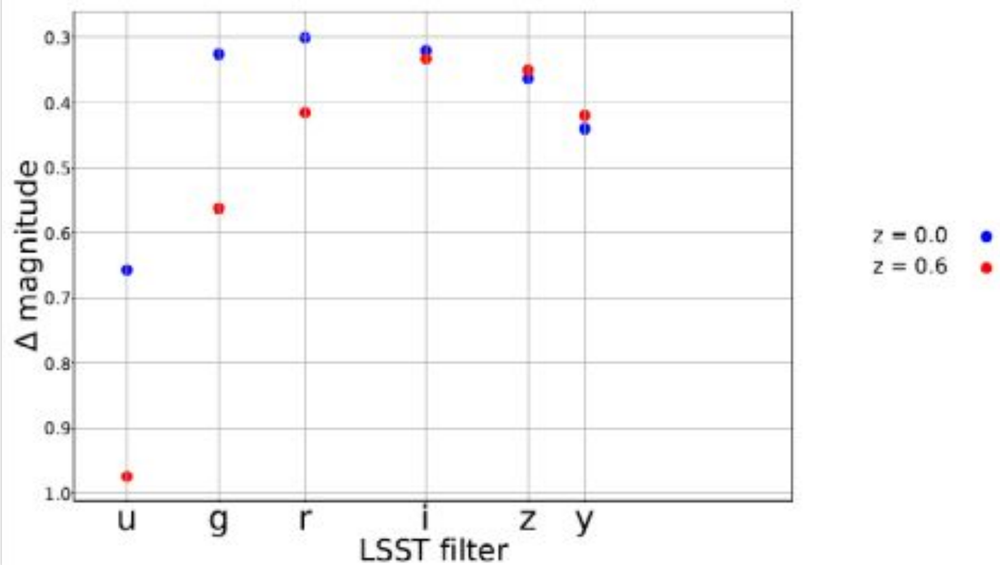
Example of application to astro:

Photometric redshift

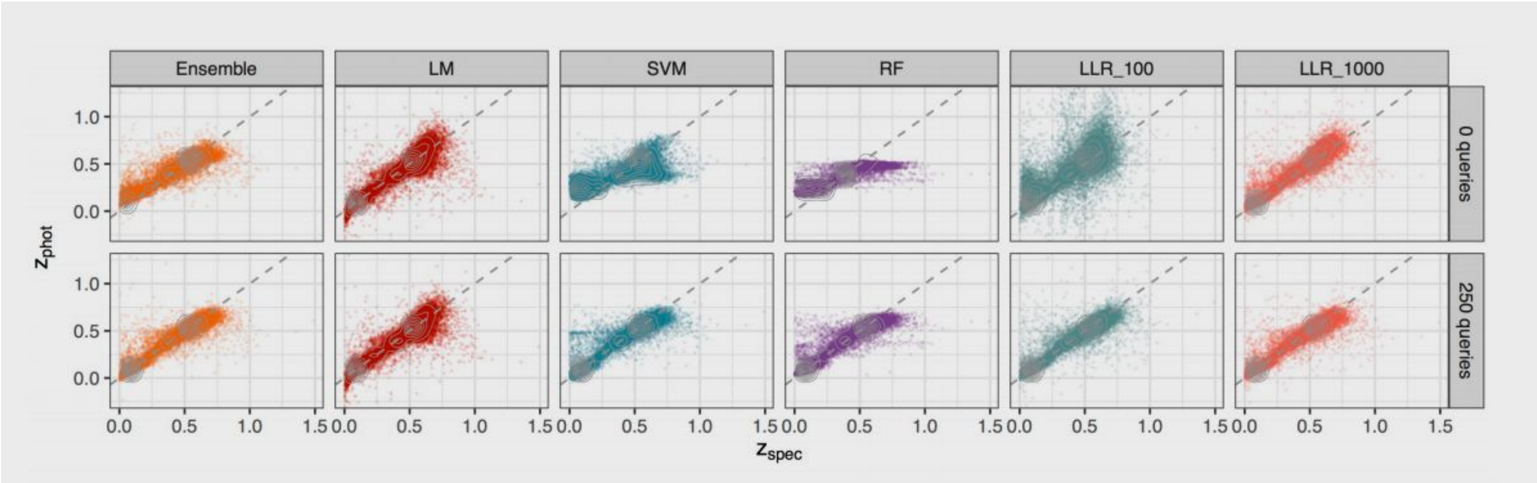
Idea







AL for Photo-Z



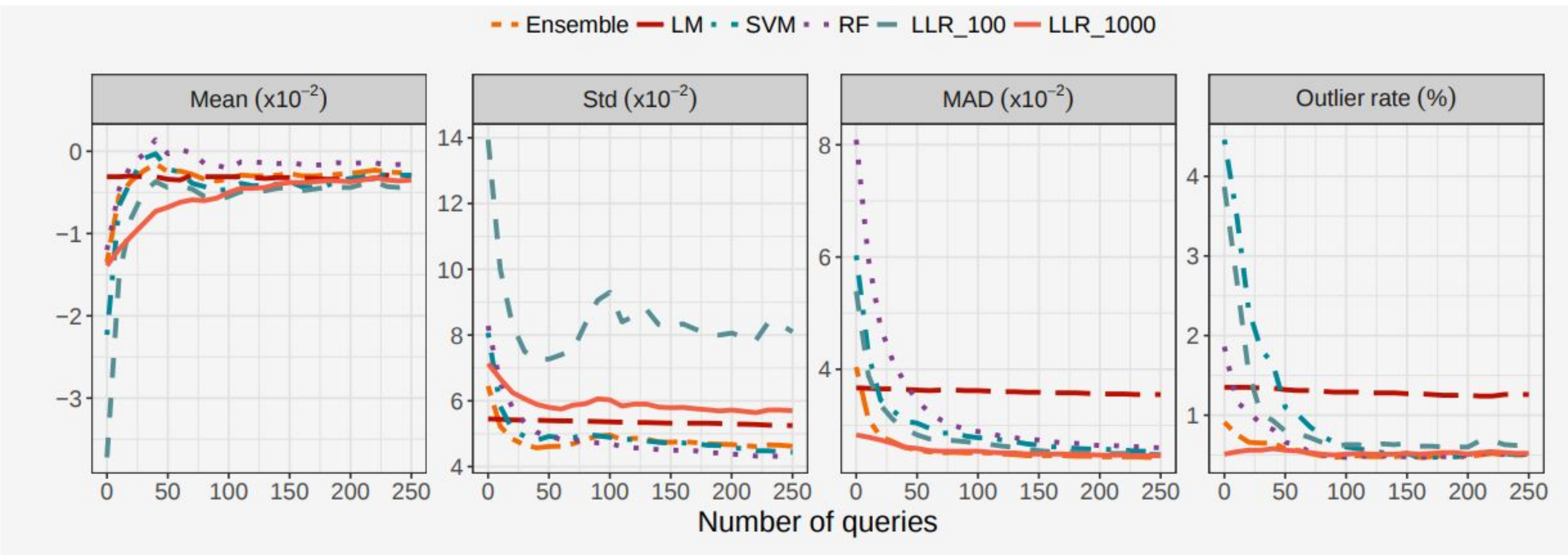
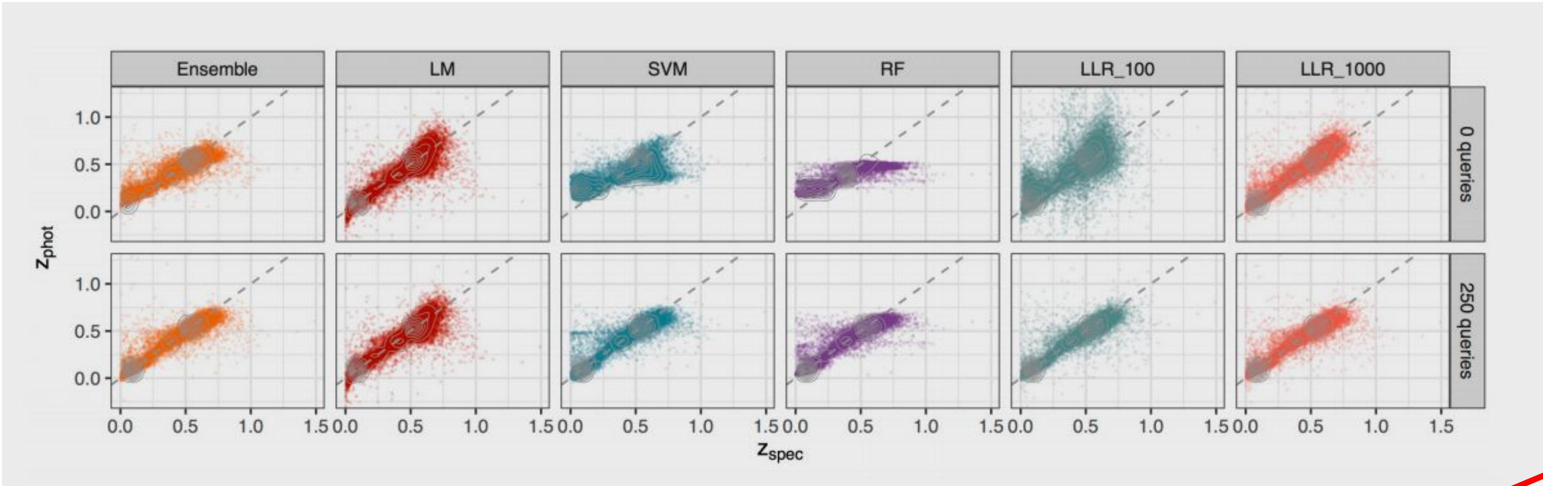
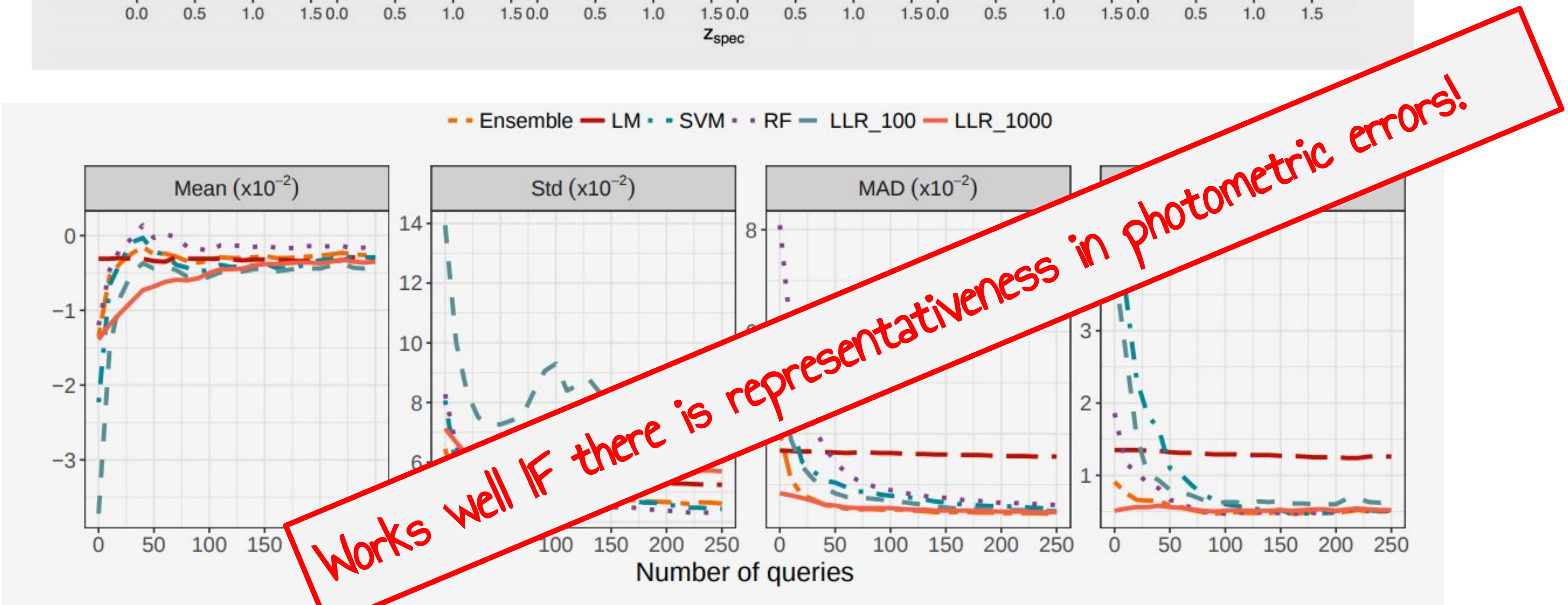


Figure 4. An assessment of the performance of the ensemble model and its constituent models using active learning. Performance diagnostics are shown as a function of the number of queries.

*Vilalta, **Ishida** et al., 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*

AL for Photo-Z





Works well IF there is representativeness in photometric errors!

Figure 4. An assessment of the performance of the ensemble model and its constituent models using active learning. Performance diagnostics are shown as a function of the number of queries.

*Vilalta, **Ishida** et al., 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*

Take home message 1:

Astronomy needs
optimized samples and
algorithms for
Machine Learning
applications

This means

Interdisciplinarity is the key

Text-book machine learning methods must be adapted to the peculiarities of astronomical data

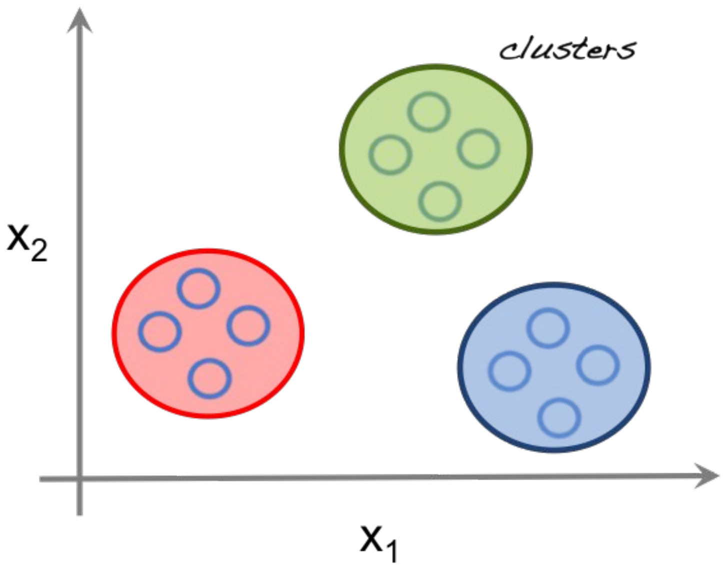
Summary: Supervised Learning

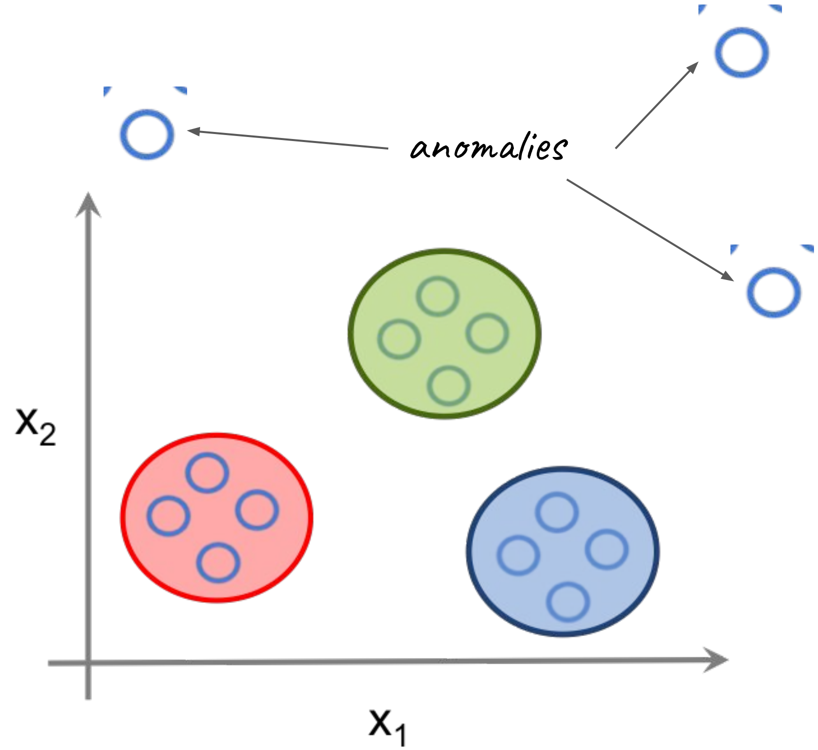
*"**How** do we optimize machine learning results with a minimum number of labeled training instances?"*

Adaptive Learning

<u>designed</u> for astronomical data

Clustering and Anomaly Detection



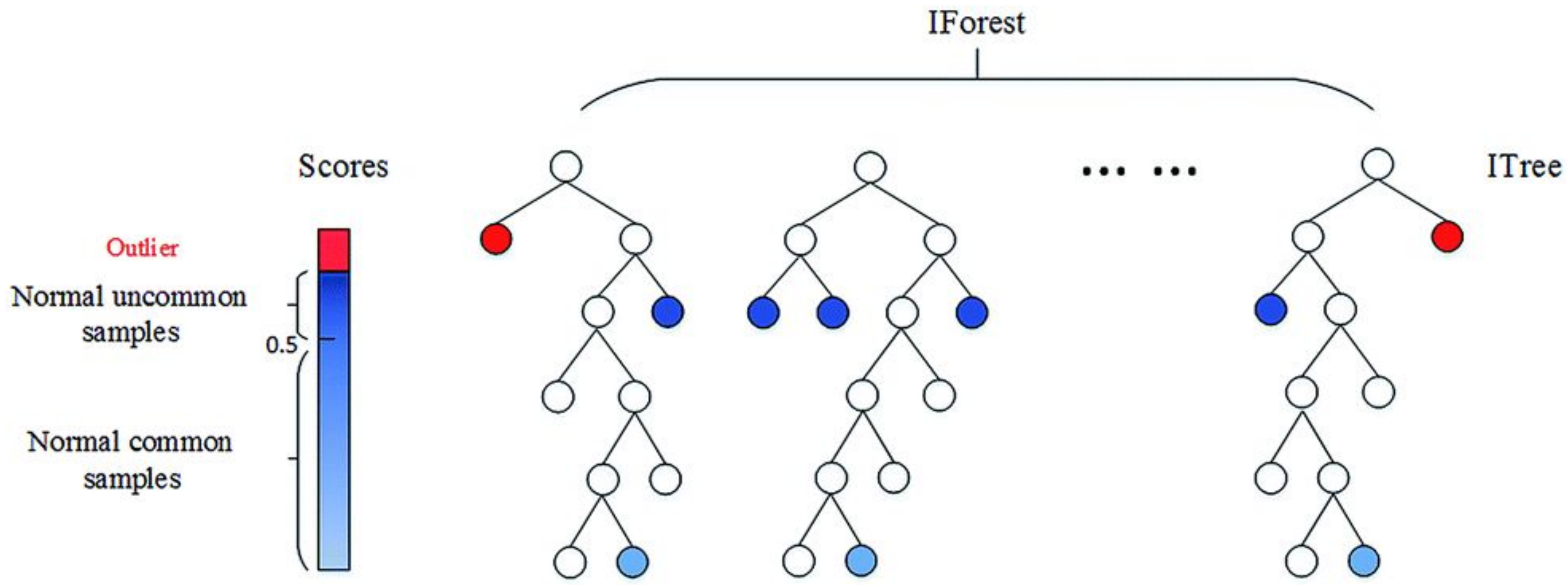


"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Hawkins, 1980

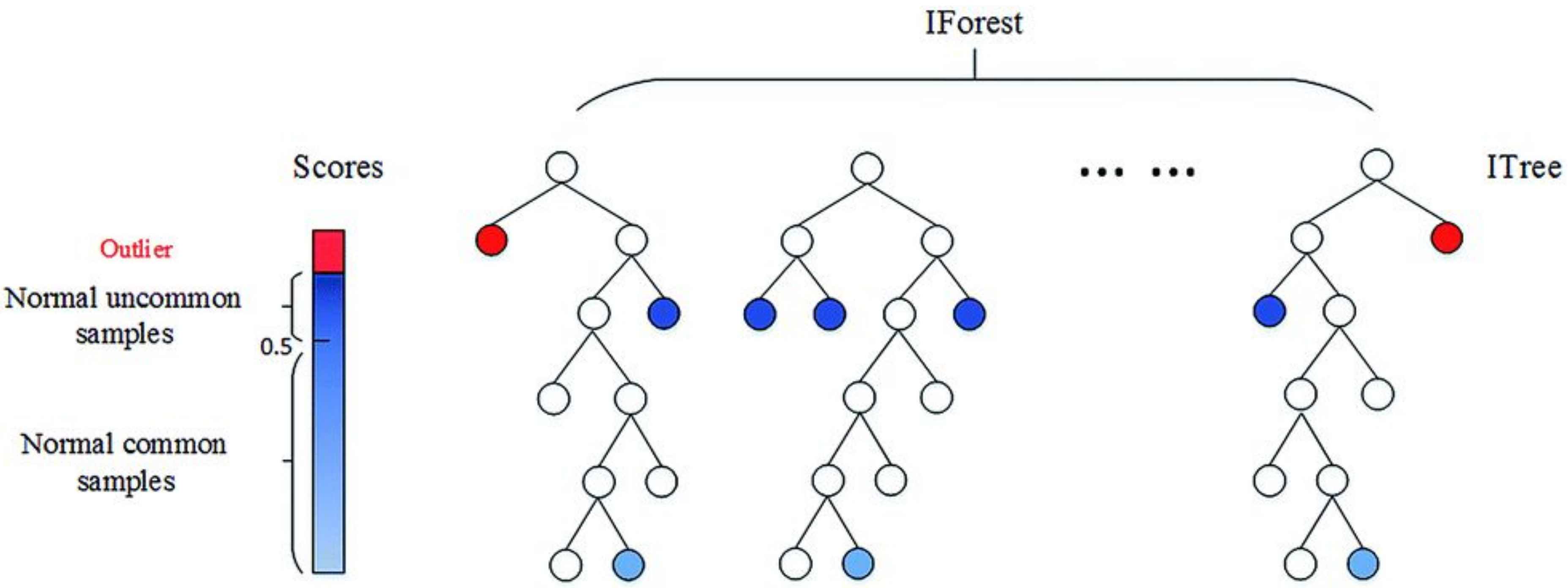
For more on unsupervised methods see Alberto Krone-Martins' talk on Tuesday and Dalya Baron on Wednesday!

Isolation Forest



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

Isolation Forest



Problem: high occurrence of false positives!

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

Active Anomaly Detection

A strategy

If yes: check next obj in the anomaly score board

If no: update hyperparameters to accommodate the new information

Isolation Forest

List of anomalies

Spec. check object with highest anomaly score

Das, S., Wong, W-K., Dietterich, T., Fern, A. and Emmott, A. (2016). *Incorporating Expert Feedback into Active Anomaly Discovery* in the Proceedings of the IEEE International Conference on Data Mining

Does this solve the problem completely?

No, it is just the best you can do!

Does this solve the problem completely?

No, it is just the best you can do!

Is this adaptable to the upcoming generation of large scale surveys?

We have to check!

The LSST alert stream

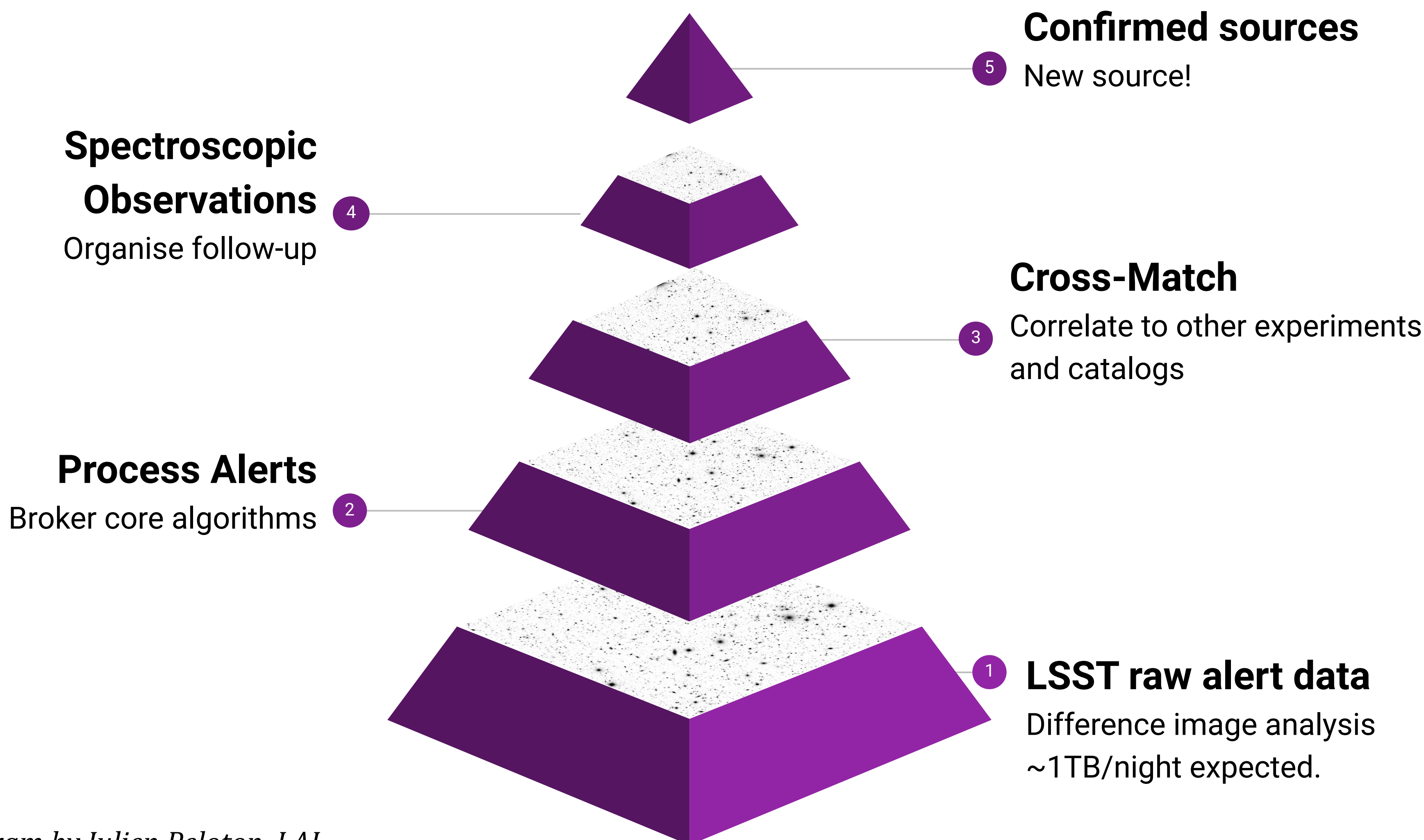


Diagram by Julien Peloton, LAL

Fink: a community broker based on Active Learning, BNN and Spark

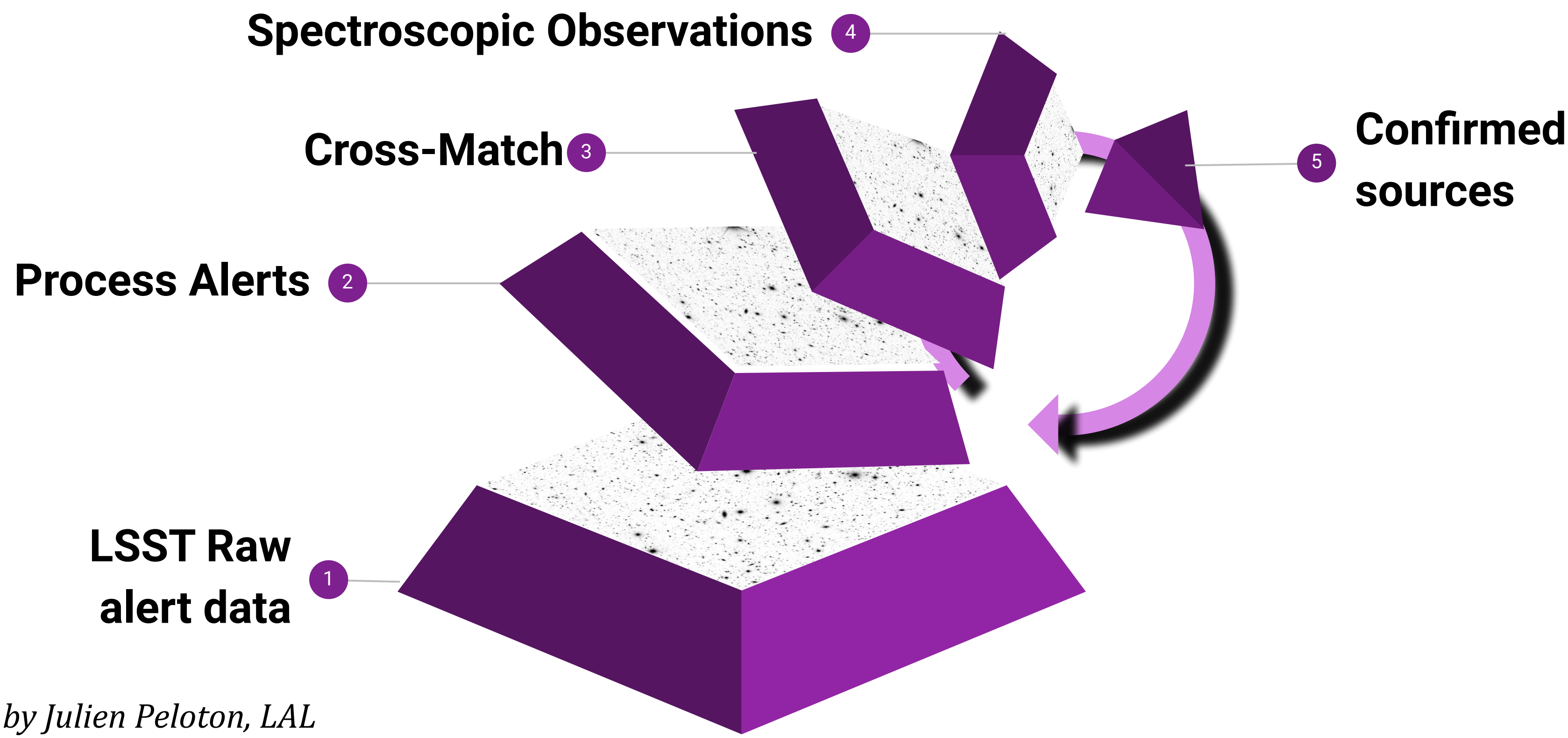


Diagram by Julien Peloton, LAL

https://fink-broker.readthedocs.io/en/latest/

Take home message 2:

Serendipitous discoveries will only get more **difficult** with the next generation of large scale surveys

This means

We need to plan for the unknown

Adaptable algorithms are one possible way to systematically search for new physics -
we should think of/try others

THANK YOU

Extra slides

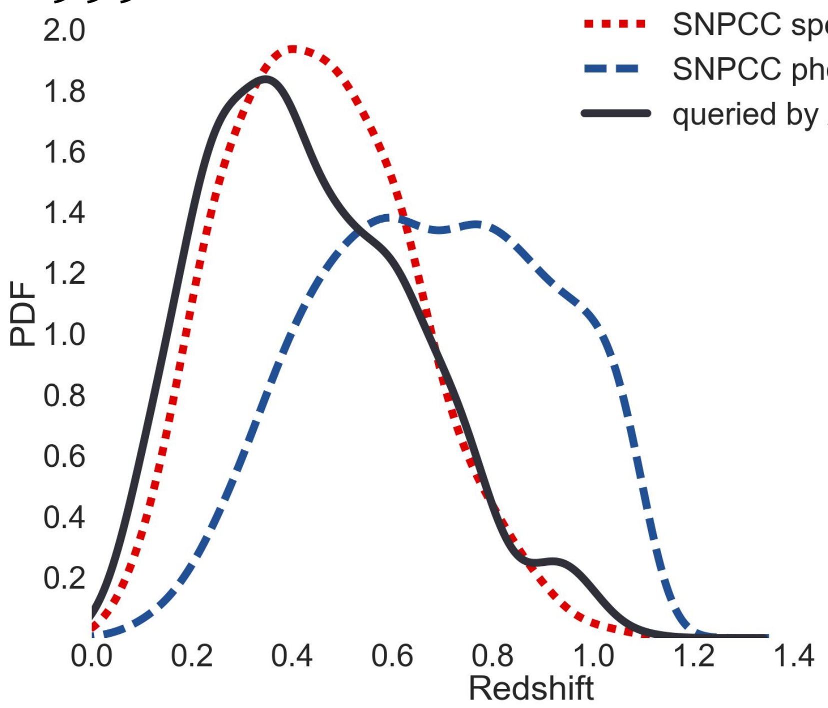
The queried sample

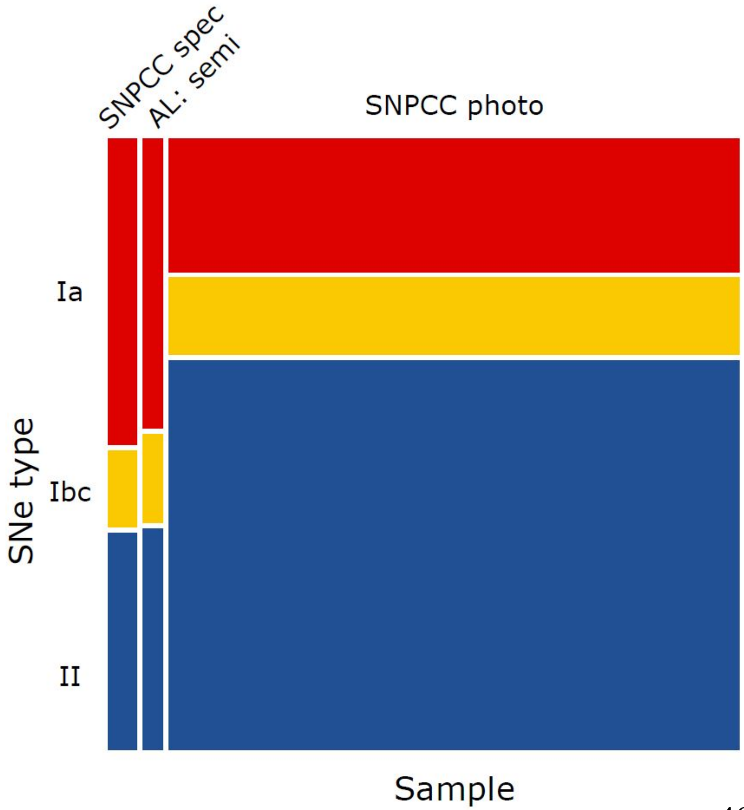
Partial LC, no training, time domain, batch

SNPCC spec:
1103 objects

Queried sample:
800 objects

Telescope time:
Queried/spec =
0.999

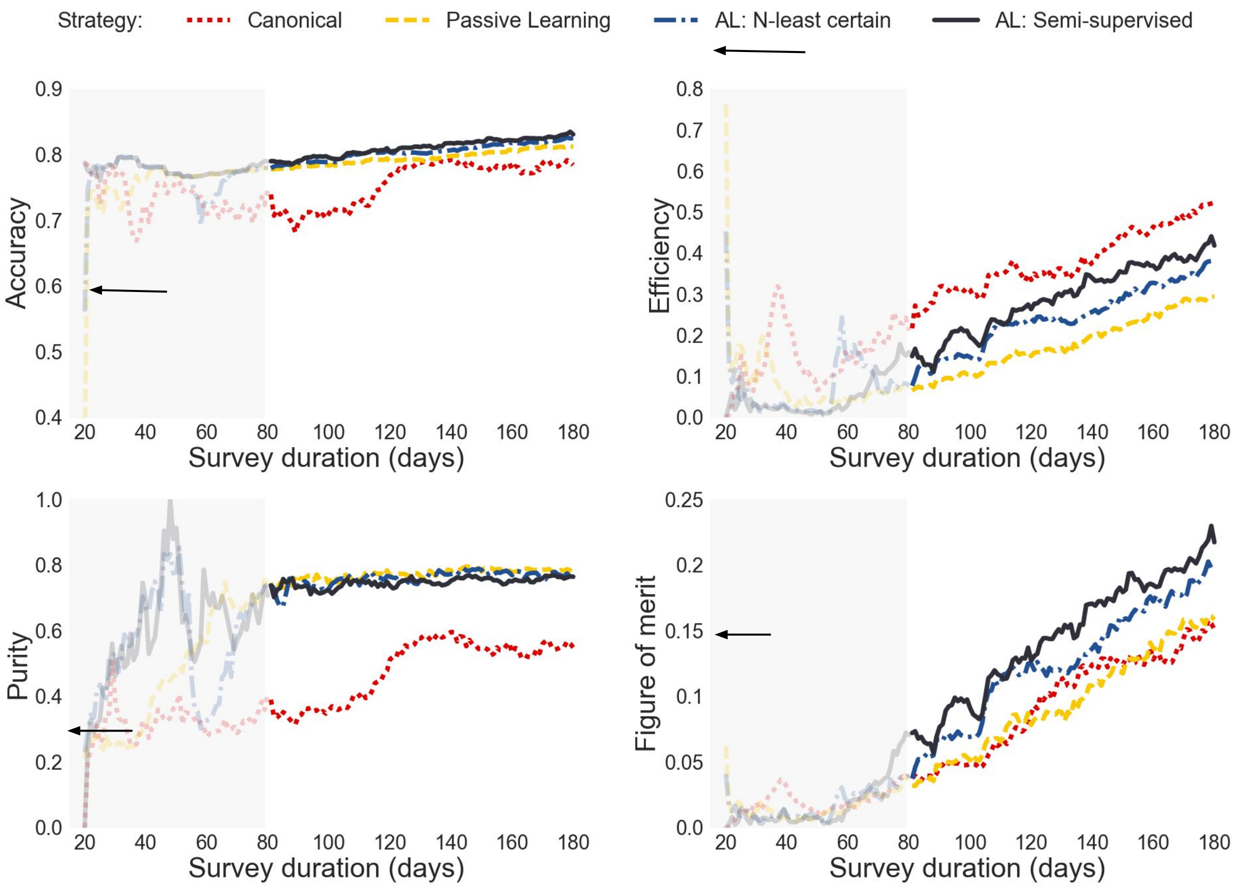




From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

Batch Mode

Partial LC, no initial training, time domain



The arrow shows <u>traditional</u> Full light-curve results with full SNPCC spec

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

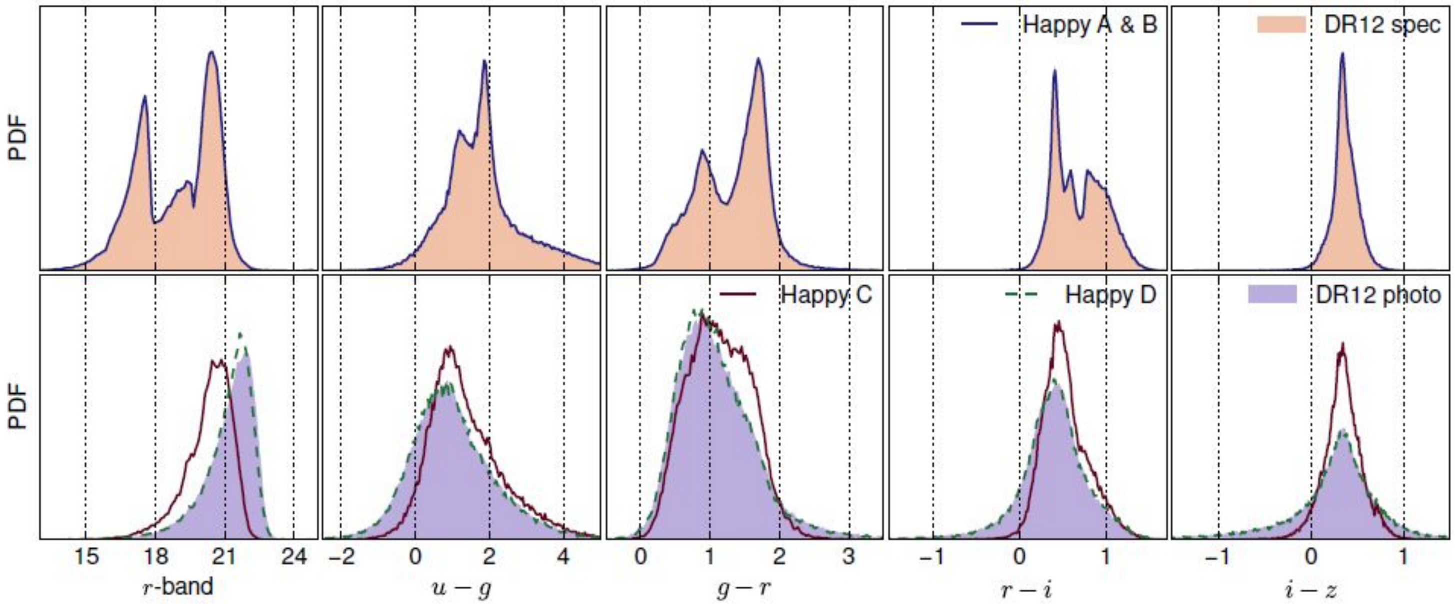
Happy catalogue

The effect of coverage + photometric errors

Photometric redshifts

Photometry from SDSS
Spec-z from many different surveys leads to larger photometric errors and consequently wide domain in r-band and color

From CRP#3 - Beck et al., astro-ph:1701.08748, MNRAS 2017



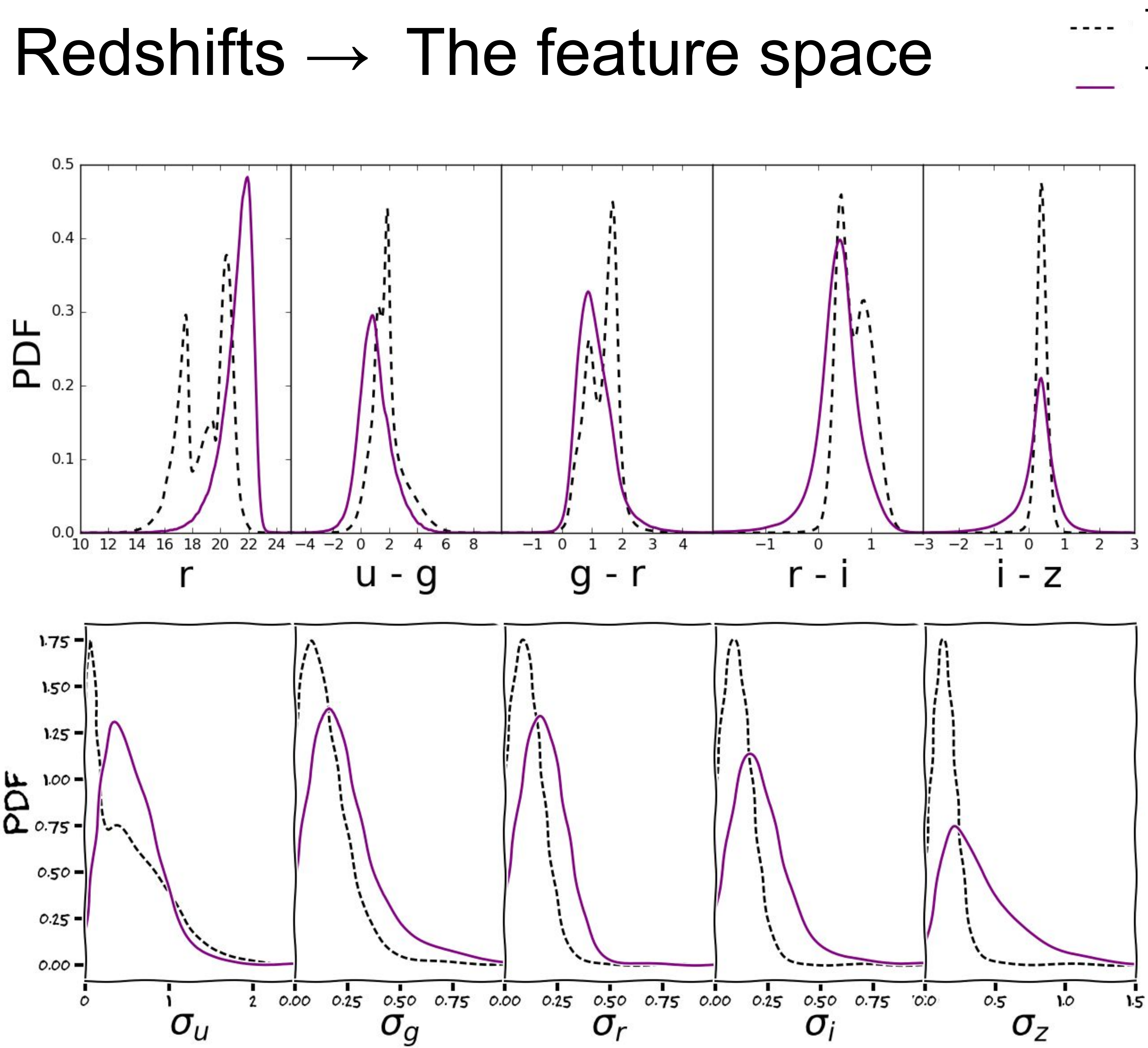
Each sample has
~ 75000 lines
5 features + errors

Happy

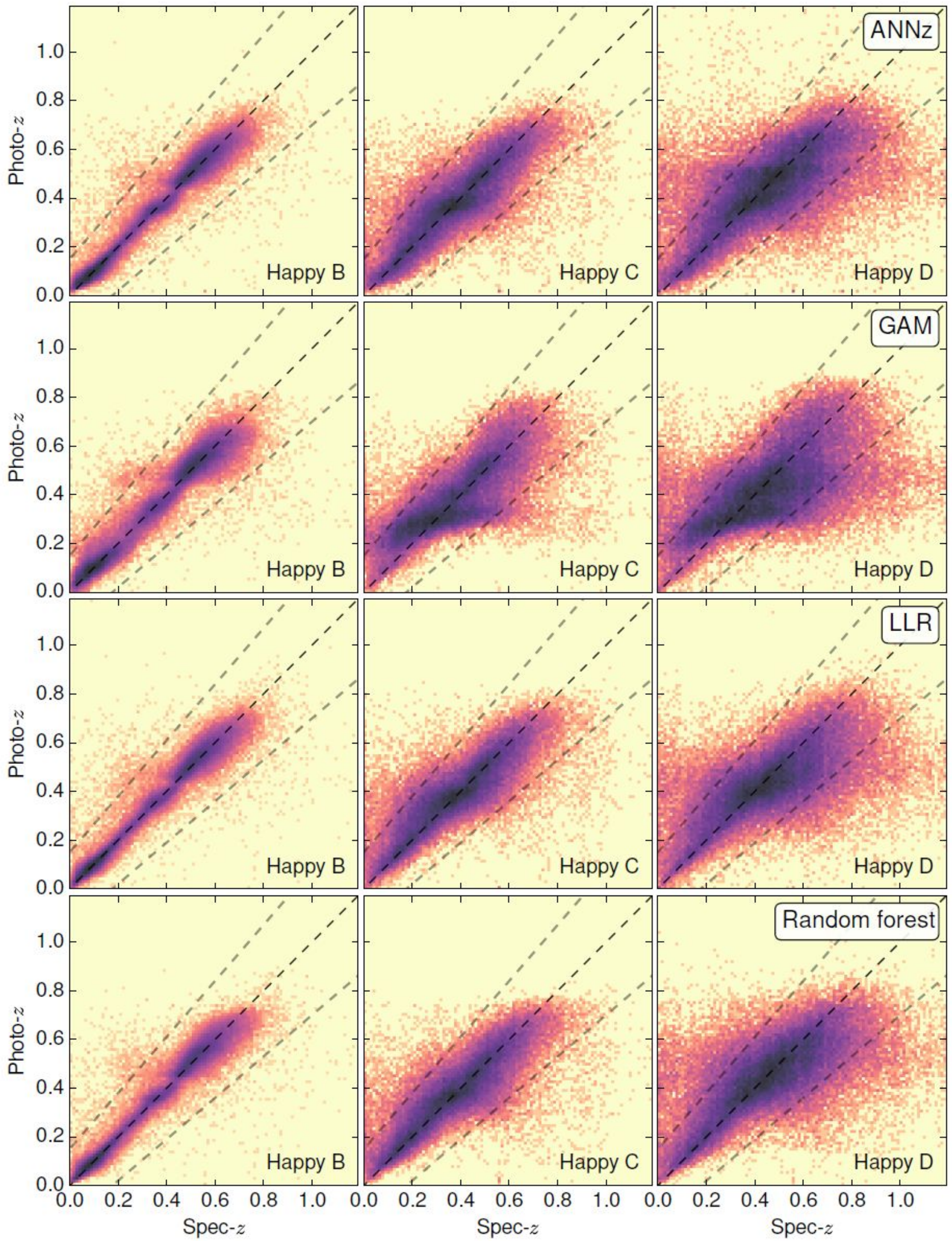
Sample A → training

Sample B → target sample, ideal case
Sample C → target sample, data quality similar to A
Sample D → target sample, realistic case
Larger errors when compared to A

Redshifts → The feature space



Beck et al.,
astro-ph:1701.08748,
MNRAS 2017



Happy catalogue

The effect of coverage + photometric errors

Beck et al., astro-ph:1701.08748, MNRAS 2017