Optimizing spectroscopic follow-up for supernova classification with active learning

Massive stars and supernovae - 5-9 Nov. 2018 – Bariloche, Argentina

Santiago González-Gaitán

CENTRA, Instituto Superior Técnico
Universidade de Lisboa, Portugal

On behalf of the COIN Collaboration:
E. Ishida, R. Beck, R.S. de Souza, A. Krone-Martins, J.W. Barrett, N. Kennamer, R. Vilalta, J.M. Burgess, B. Quint, A.Z. Vitorelli, A. Mahabal, E. Gangler

centra

COIN

Cosmostatistics Initiative

TÉCNICO LISBOA

Optimizing spectroscopic follow-up for supernova classification with active learning
(Ishida et al. 2018)

https://github.com/COINtoolbox/ActSNClass

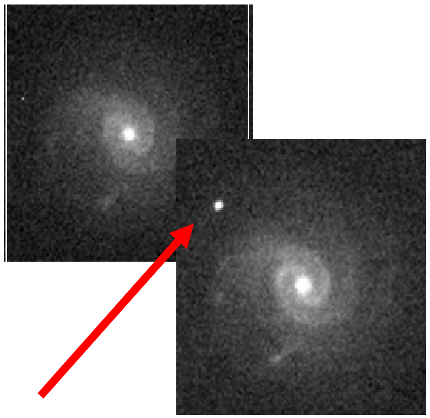
Spatial field reconstruction with INLA: Application to IFU galaxy data
(González-Gaitán et al. 2018)

https://github.com/COINtoolbox/Galaxies_INLA

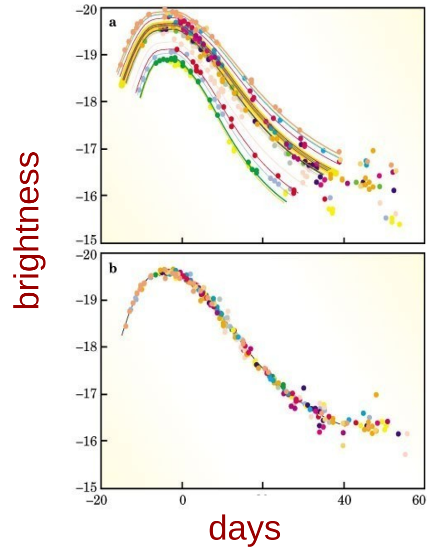
COIN
Cosmostatistics Initiative

Supernova Cosmology

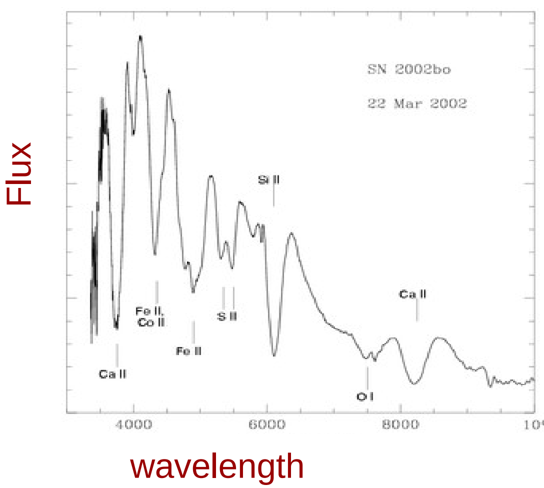
1. detection



2. photometry

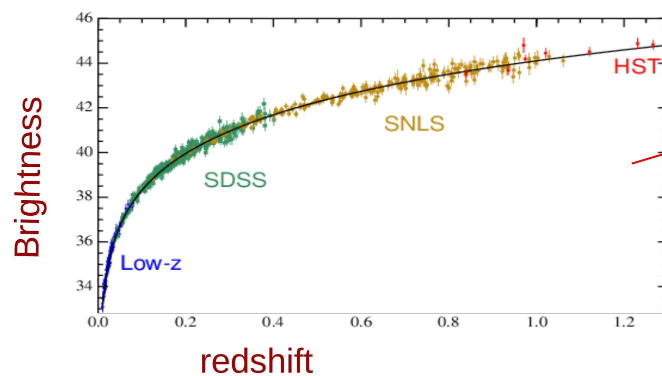


3. spectroscopy



Distance (redshift) + classification

4. standardization + cosmological fit

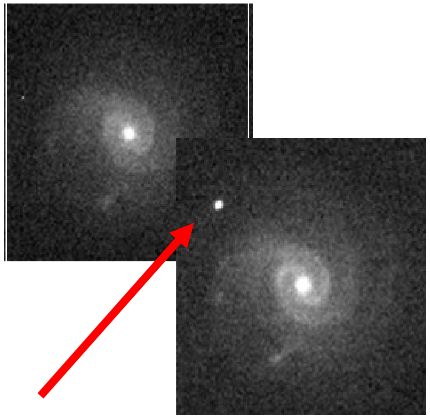


year	Number of supernova
1998	42
2014	740

Credit: E. Ishida

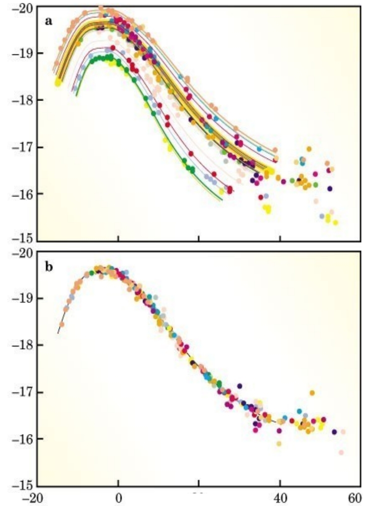
Supernova Cosmology

1. detection



2. photometry

brightness

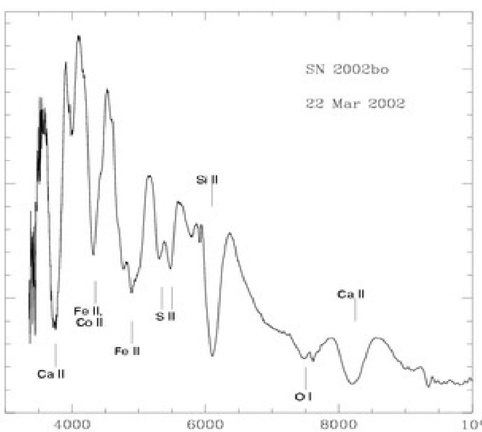


days

SURVEY	Number of supernovae (spec)	Number of supernovae (phot)
SDSS	375	750
SNLS	290	690

3. spectroscopy

Flux



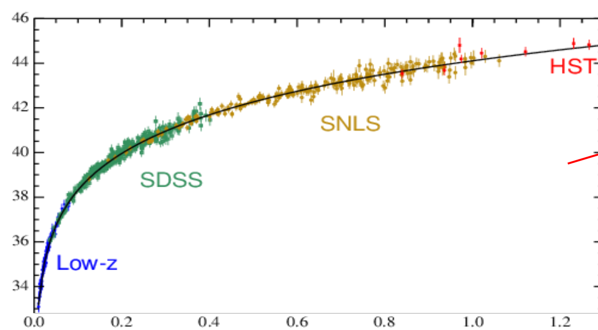
wavelength

Distance (redshift)

+ classification

4. standardization + cosmological fit

Brightness

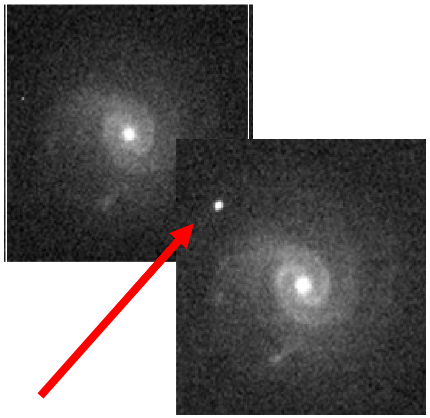


redshift

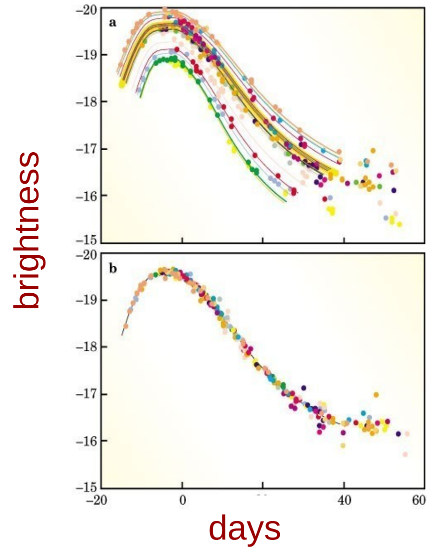
year	Number of supernova
1998	42
2014	740

Supernova Cosmology

1. detection

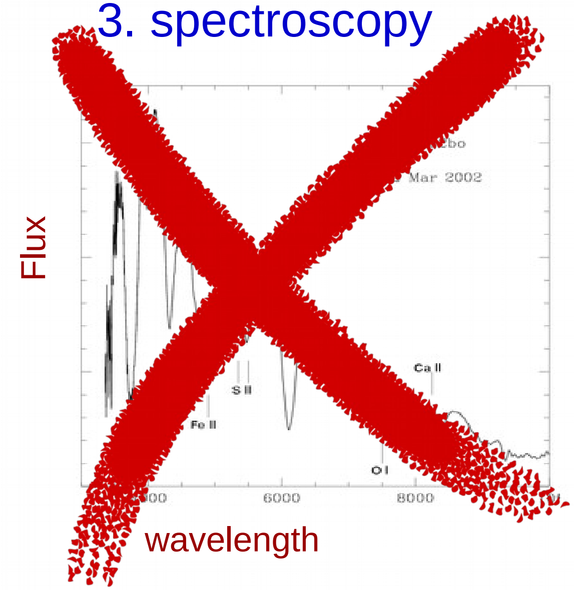


2. photometry



SURVEY	Number of supernovae (spec)	Number of supernovae (phot)
SDSS	375	750
SNLS	290	690

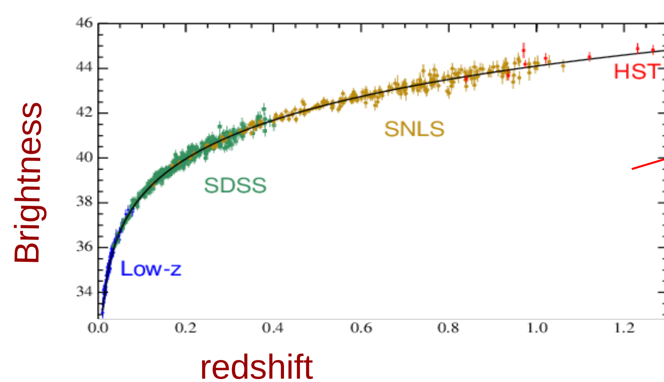
3. spectroscopy



Distance (redshift)

+ classification

4. standardization + cosmological fit



year	Number of supernova
1998	42
2014	740

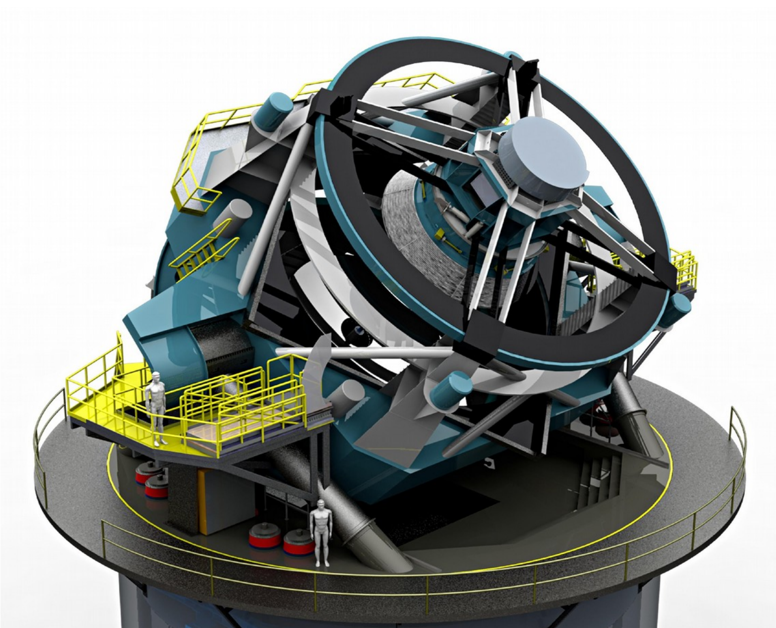
Big Data (in astronomy) → Large Scale Sky Surveys

year	Number of supernova
1998	42
2014	740
2025	> 10 000

2 million alerts/day
15 TB/day

40 nights of LSST

entire Google database



https://www.lsst.org/

LSST DESC Dark Energy Science Collaboration

LSST Large Synoptic Survey Telescope

Credit: E. Ishida

Photometric classification

"Brute-force" approaches:

e.g. color-color, color-mag cuts, template fits and cuts

(Poznanski+02, Johnson & Crotts06, Sullivan+06)

Machine learning: supervised learning

e.g. decision trees, random forest, neural networks

(Richards+12, Ishida & Souza13, Karpenka+13, Lochner+13M Möller+16, Dai+18)

Supernova Photometric Challenge (SNPCC), Kessler et al. 2010

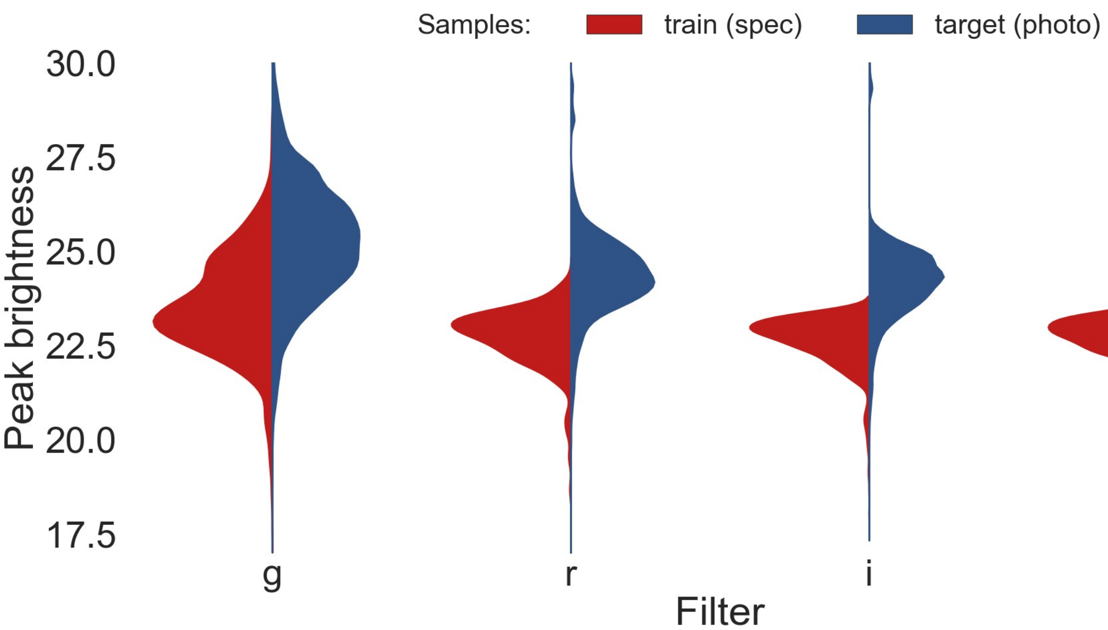
Number of objects: 20000 (2000 training) Number of classes: 3 (Ia, II, Ibc)

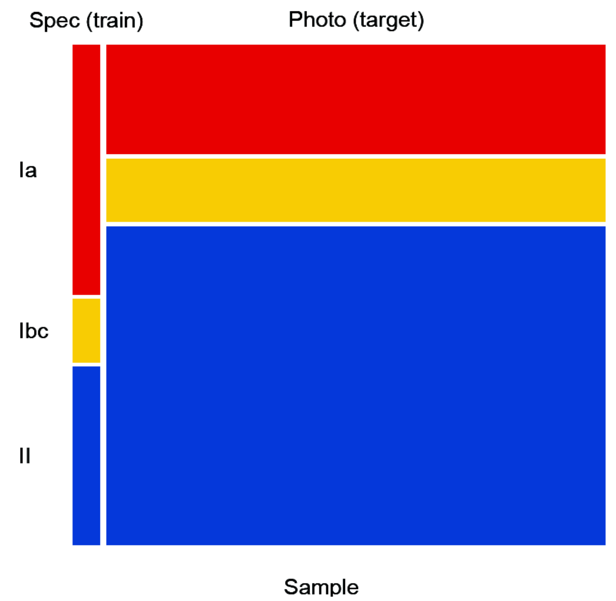
Efficiency $\frac{N_{\rm Ia}^{\rm true}}{N_{\rm Ia}^{\rm TOT}}$

Purity $\frac{N_{\rm Ia}^{\rm true}}{N_{\rm Ia}^{\rm true} + W_{\rm Ia}^{\rm false} N_{\rm Ia}^{\rm false}}$

Figure of Merit (FoM): $\frac{N_{\rm Ia}^{\rm true}}{N_{\rm Ia}^{\rm TOT}} \times \frac{N_{\rm Ia}^{\rm true}}{N_{\rm Ia}^{\rm true} + W_{\rm Ia}^{\rm false} N_{\rm Ia}^{\rm false}} = \epsilon_{\rm Ia} + PP_{\rm Ia}.$

Representativeness





From COIN Residence Program #4, ***Ishida*** *et al., 2018 – arXiv:astro-ph/1804.03765*

The Data: post-SNPCC simulations – *Kessler et al., 2010*

How to construct training samples which optimize photometric classification results?

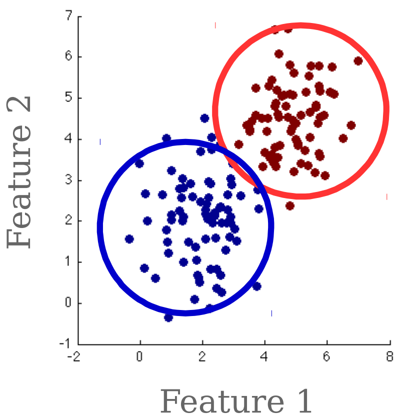
Given known observational constraints...

Credit: E. Ishida

Active Learning

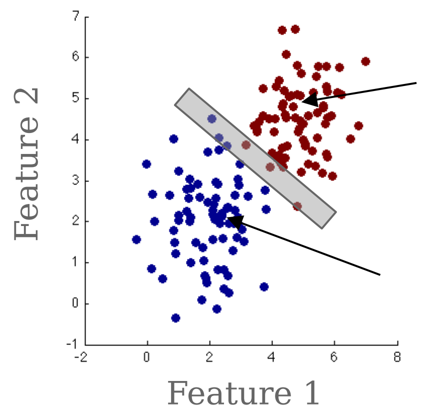
or Optimal Experimental Design

“Can machines learn with fewer labeled training instances if they are allowed to ask questions?”



Step 1

Standard Classification

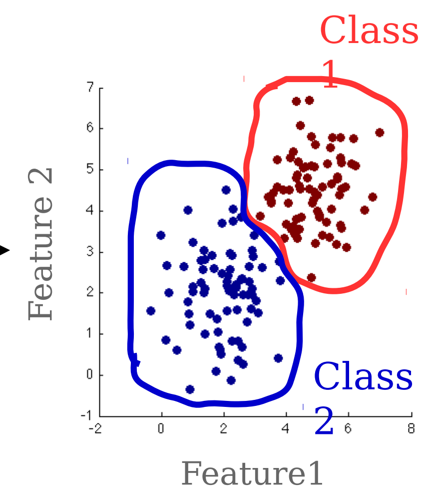


Step 2

Identify problematic as well as representative elements

Step 3

Make a query



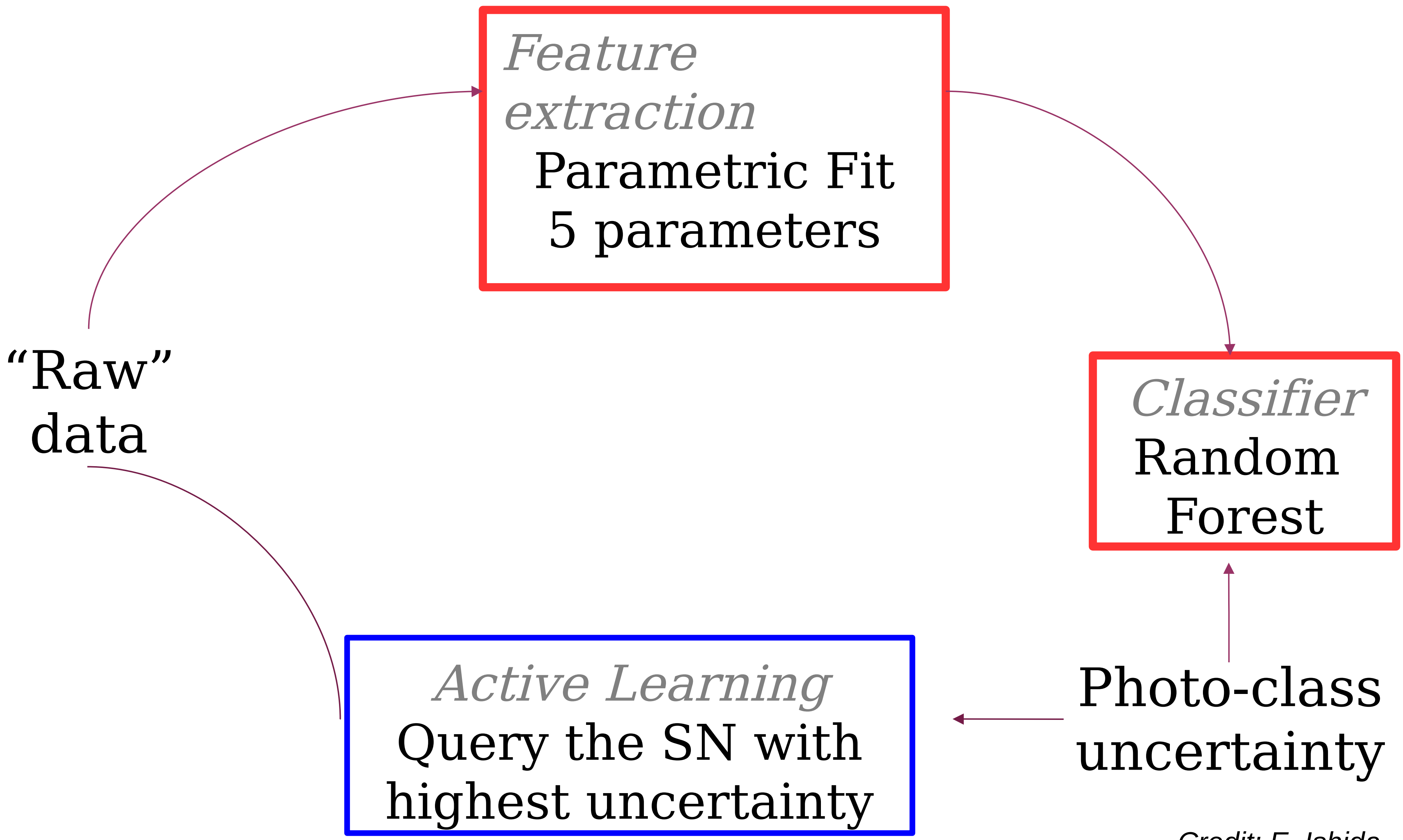
Step 4

Classify

Credit: E. Ishida

AL for Supernova classification

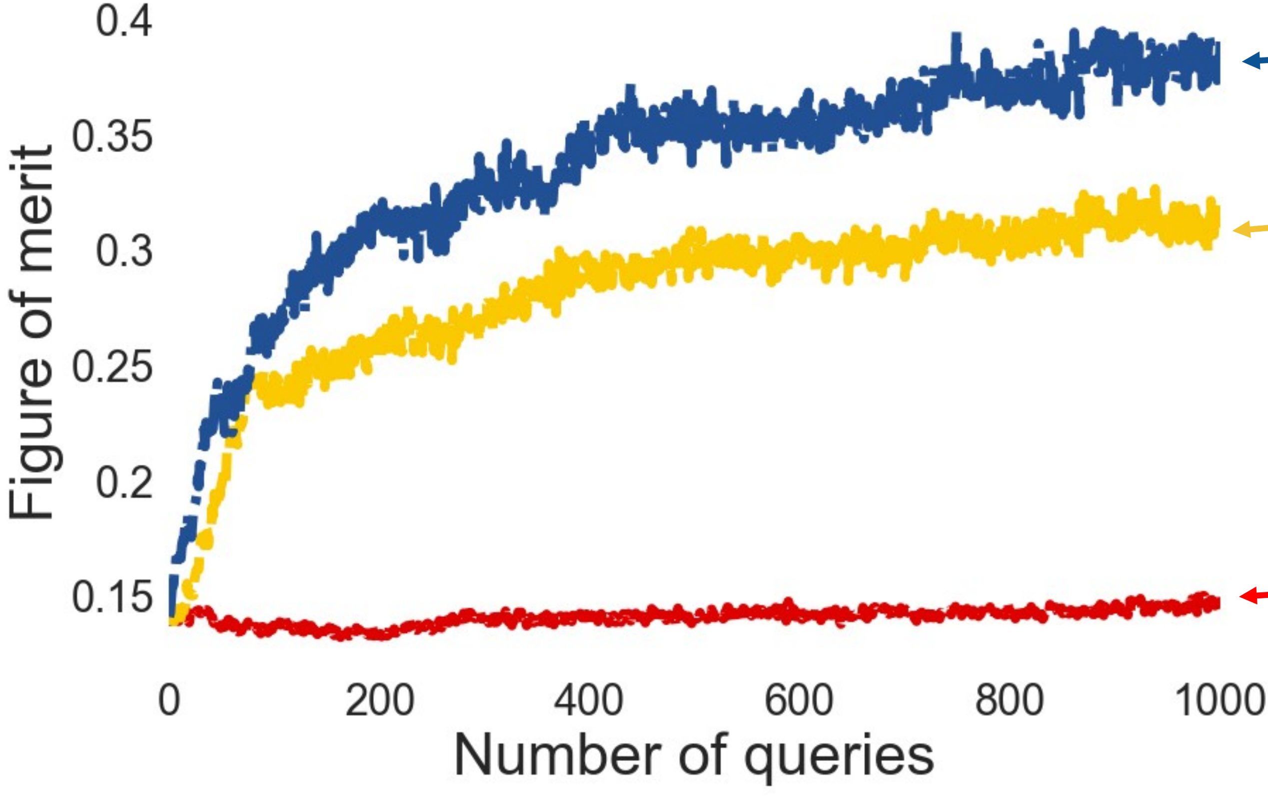
Our strategy



Credit: E. Ishida

AL for SN classification

Static results
(full survey)



Active Learning
(most informative SN)

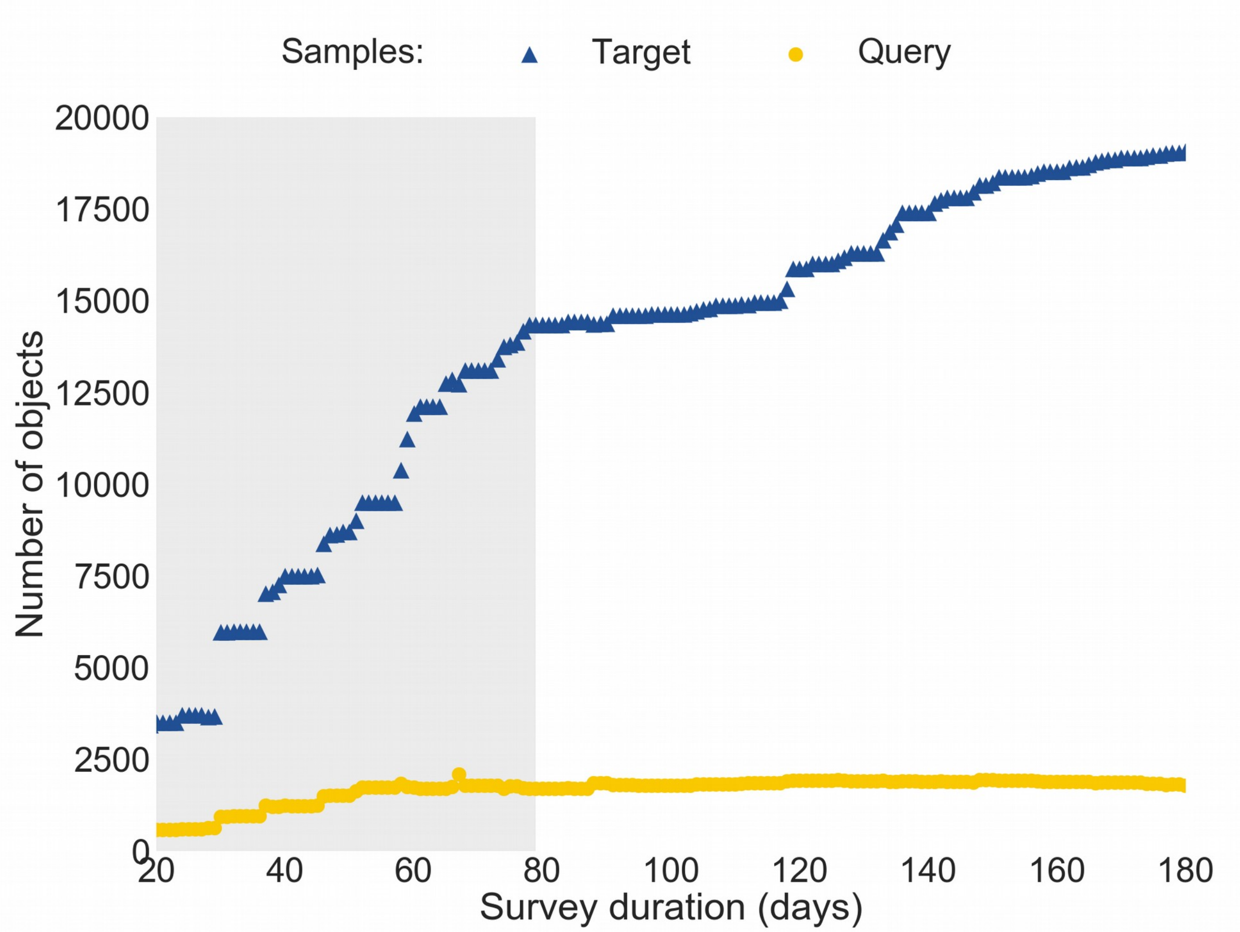
Passive Learning
(random selection from phot pool)

Canonical strategy
(selection from phot sample similar to spec)

Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Time Domain

Survey evolution



1. Feature extraction done daily **with available observed epochs until then** (partial LC fits).

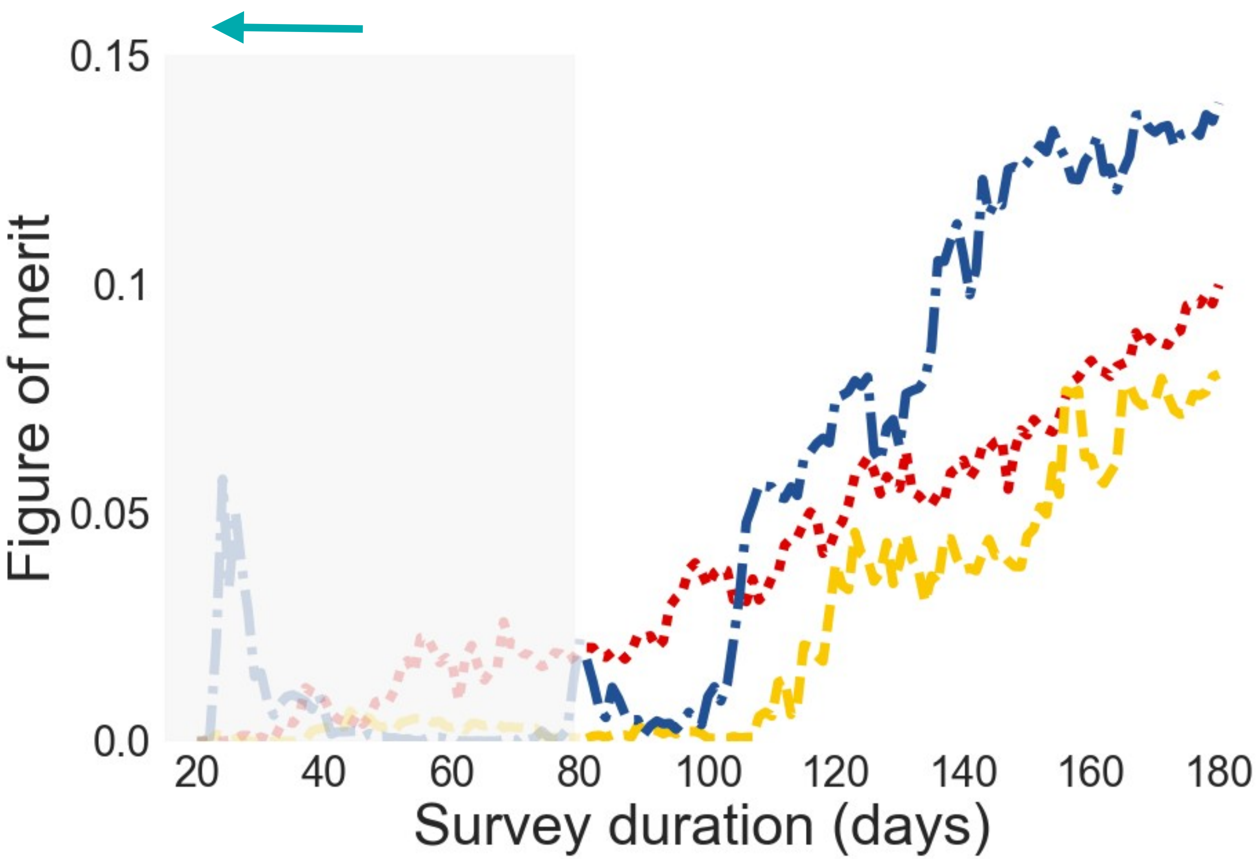
2. Query sample is also re-defined daily: objects with **r-mag < 24**

3. **No** need for an **initial training** sample

Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Partial LC, no training

Time domain



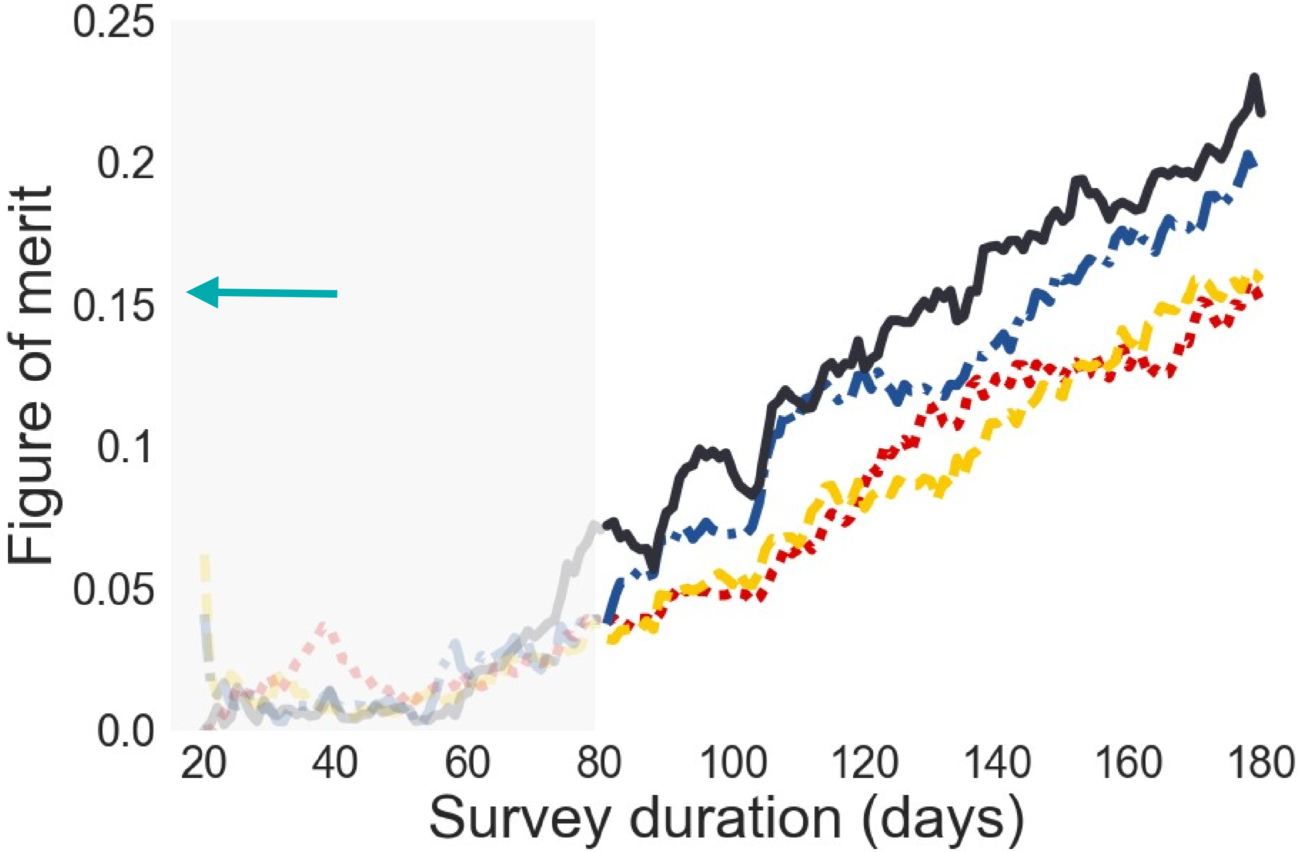
The arrow shows <u>traditional</u> full light-curve results with full SNPCC spec

Similar FoM with 168 spec (15% of full spec sample)

Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Batch Mode

Partial LC, no initial training, time domain



Batch mode: instead of 1 classification per night, set of N SNe queried

Two types:

- N-least certain
- Semi-supervised uncertainty sampling

45% better FoM (70% of full spec sample)

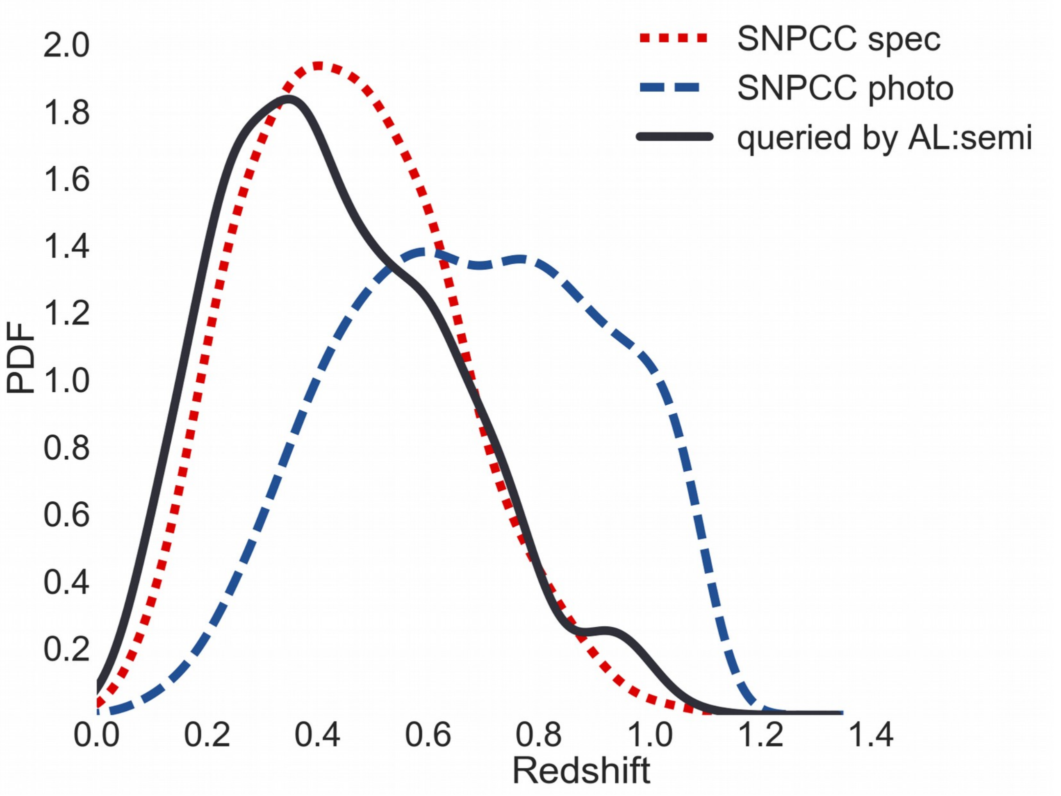
Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

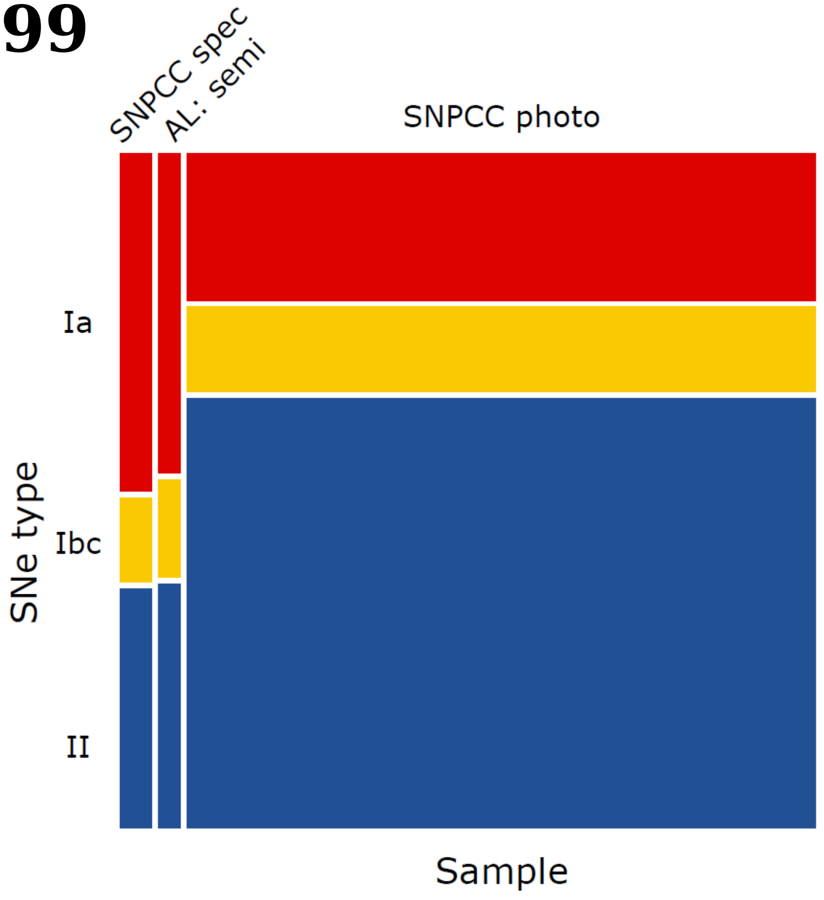
The queried sample

Partial LC, no training, time domain, batch

SNPCC spec: Queried sample:
Telescope time:
1103 objects 800 objects

Telescope time: Queried/spec = 0.999





Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Summary

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

Active Learning <u>designed</u> for astronomical data

This is a group effort!

The Cosmostatistics Initiatve (COIN) was born in

Cosmo21 - Lisbon, 2014!

PLAsTiCC

Photometric LSST Astronomical Time-series Classification Challenge

A data challenge aimed to prepare a **larger** community for the LSST data paradigm

- PI: Renee Hlozek, simulations: Rick Kessler, deployment: Emille Ishida
- SNANA simulations → Light curves in observer-frame (no images!)
- 3 years worth of LSST data, ~ 100 MB
- ~ 10^7 objects
- Around 20 transient models (galactic and extra-galactic, periodic and non-periodic)

RELEASED!

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

- Please respect model-information policy: ``don't ask, don't tell"

- Not all models will be present in the training sample
- Supervised classification + novelty detection
- Deployment: kaggle + RAMP

Slide by Emille Ishida

Thank you!

COIN

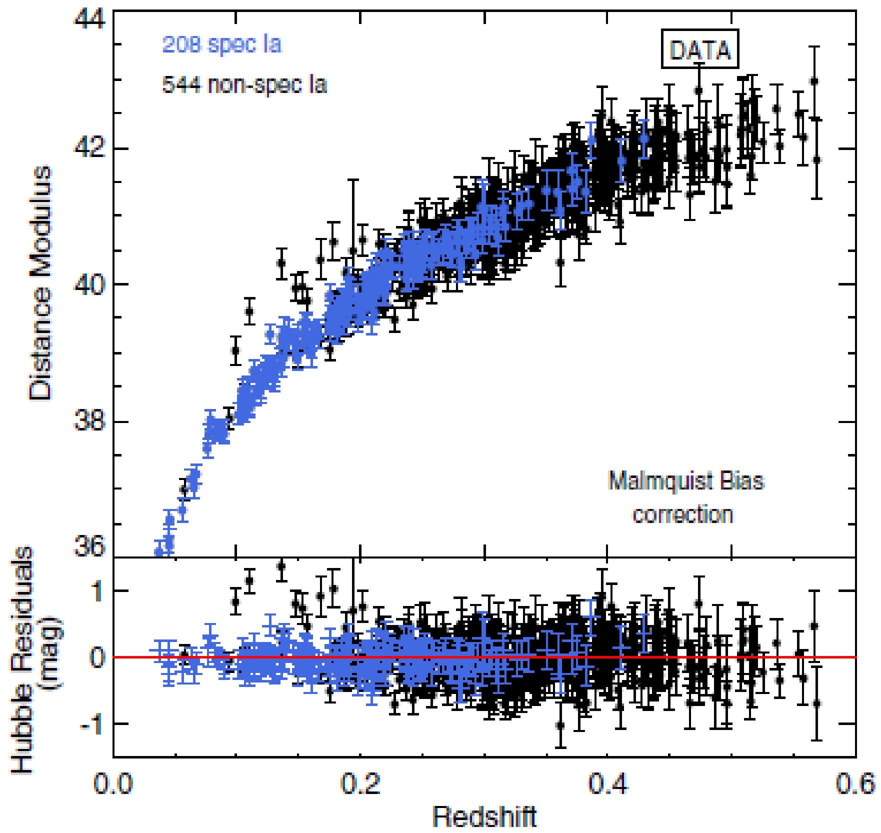
Cosmostatistics Initiative

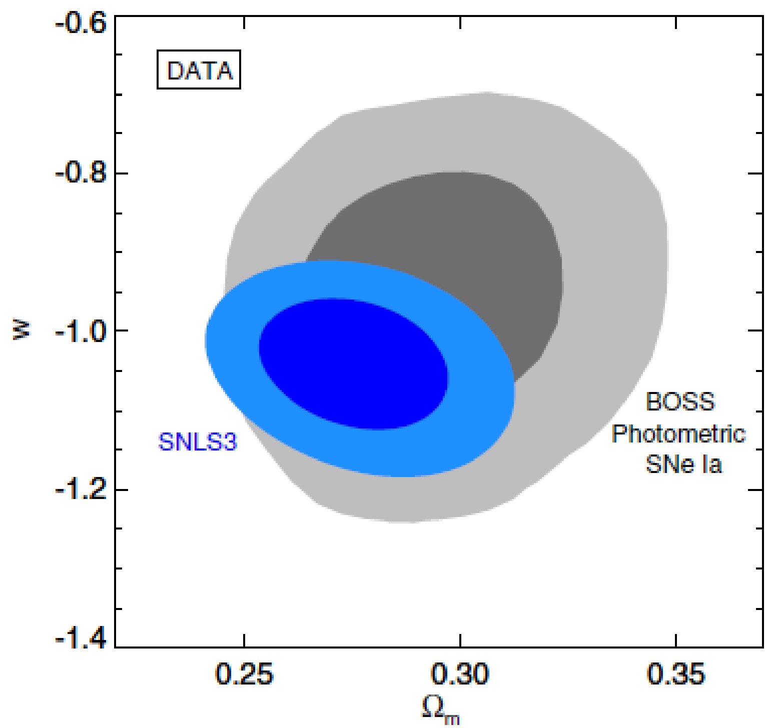
http://cointoolbox.github.io/

Supernova Cosmology

Instead of SN spectroscopy: **host galaxy spectroscopy**

SDSS: Campbell+13

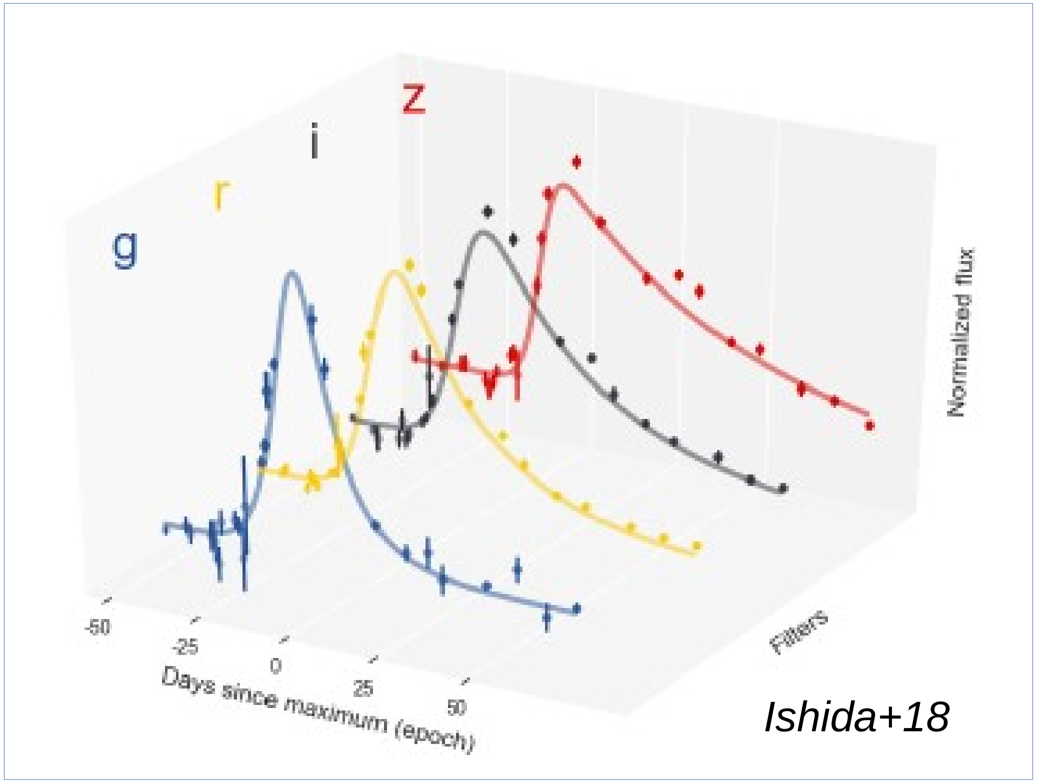




4% contamination-> no effect on cosmology

Currently in DES: Generating HD with ~2500 photometric SNe Ia

Feature extraction: parametric fits



Ishida+18

Generic parametric function (Bazin+09) with 5 parameters for each filter: *A.B,t0,tf,tr*

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

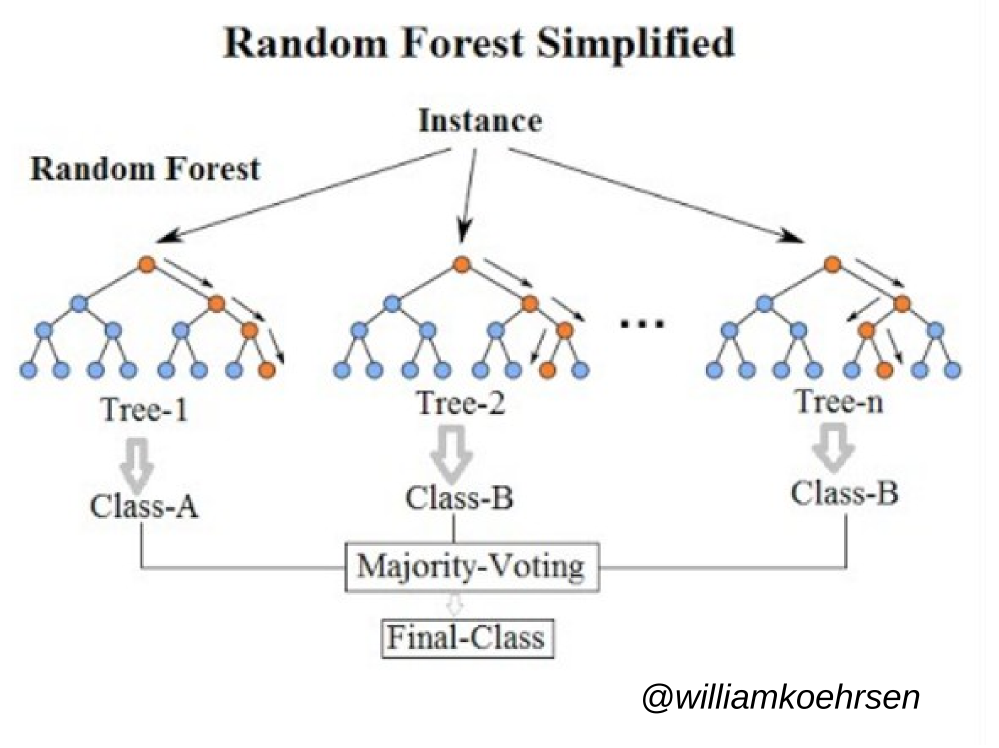
Advantages: easy and fast, homogeneous
Disadvantages: may introduce biases, fit depens on data

WARNING: There are certainly better choices of feature extraction!!
(e.g. Lochner+16, Naul+18)

Classifier: Random Forest

Random forest is a machine learning algorithm made of averages of multiple decision trees trained over different sub-sets
Decision tree is a series of questions on features to give a probable class

Two classes: Ia vs non-Ia

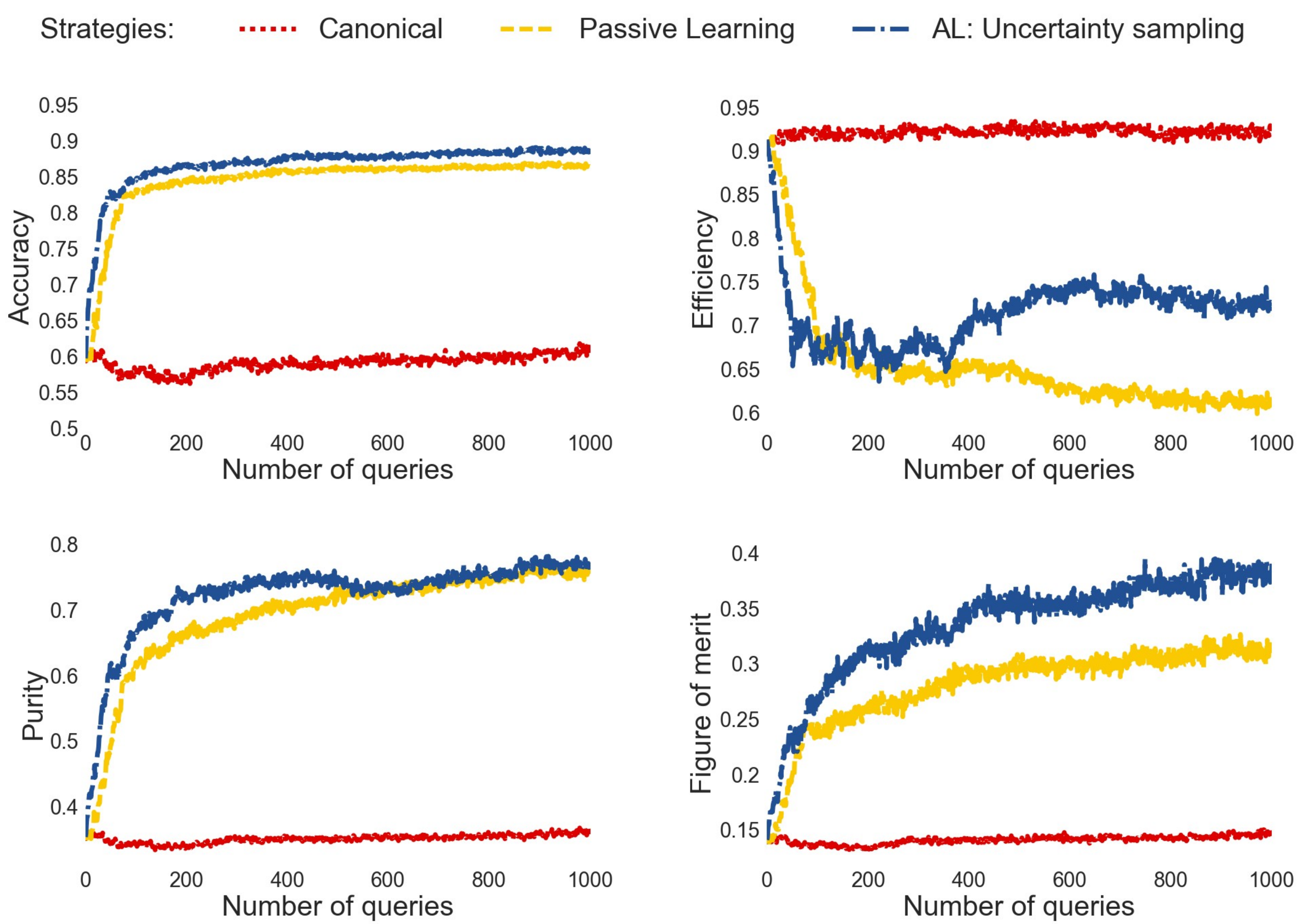


Use of *scikit-learn* with 1000 trees with P(Ia) is % of trees voting for Ia

WARNING: There may be better choices of classifiers!!
(e.g. Lochner+16)

AL for SN classification

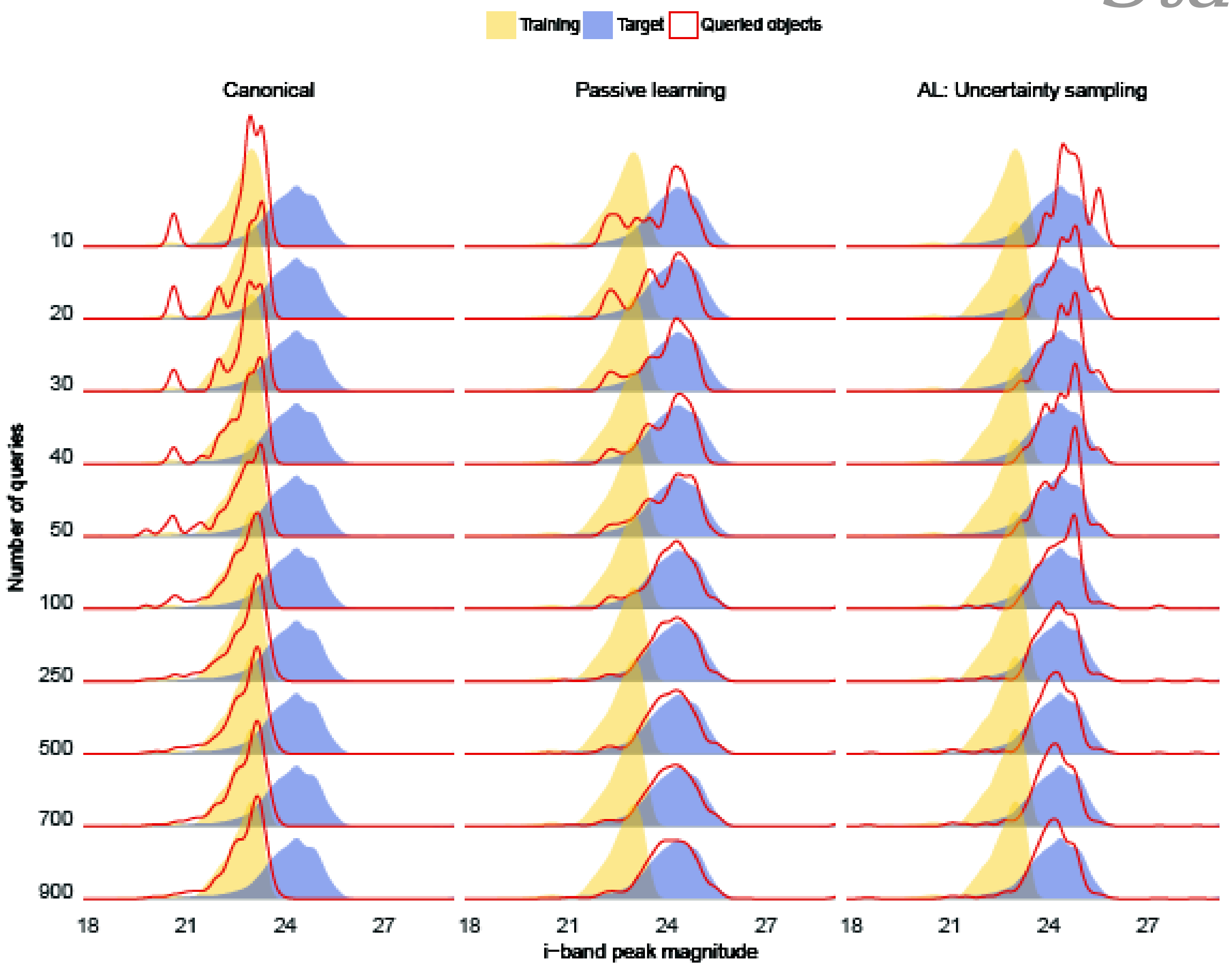
Static results
(full survey)



Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

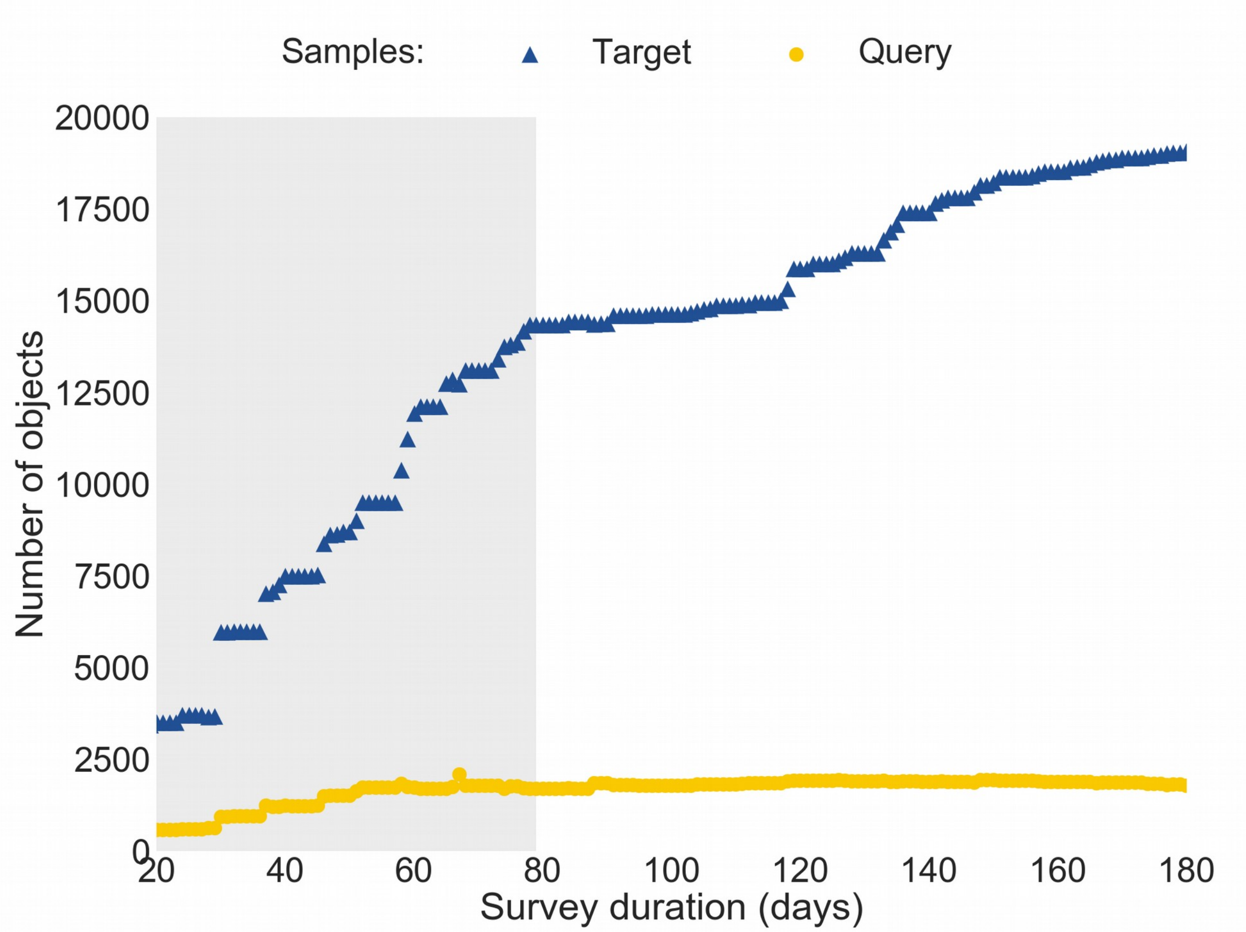
AL for SN classification

Static results
(full survey)



Time Domain

Survey evolution

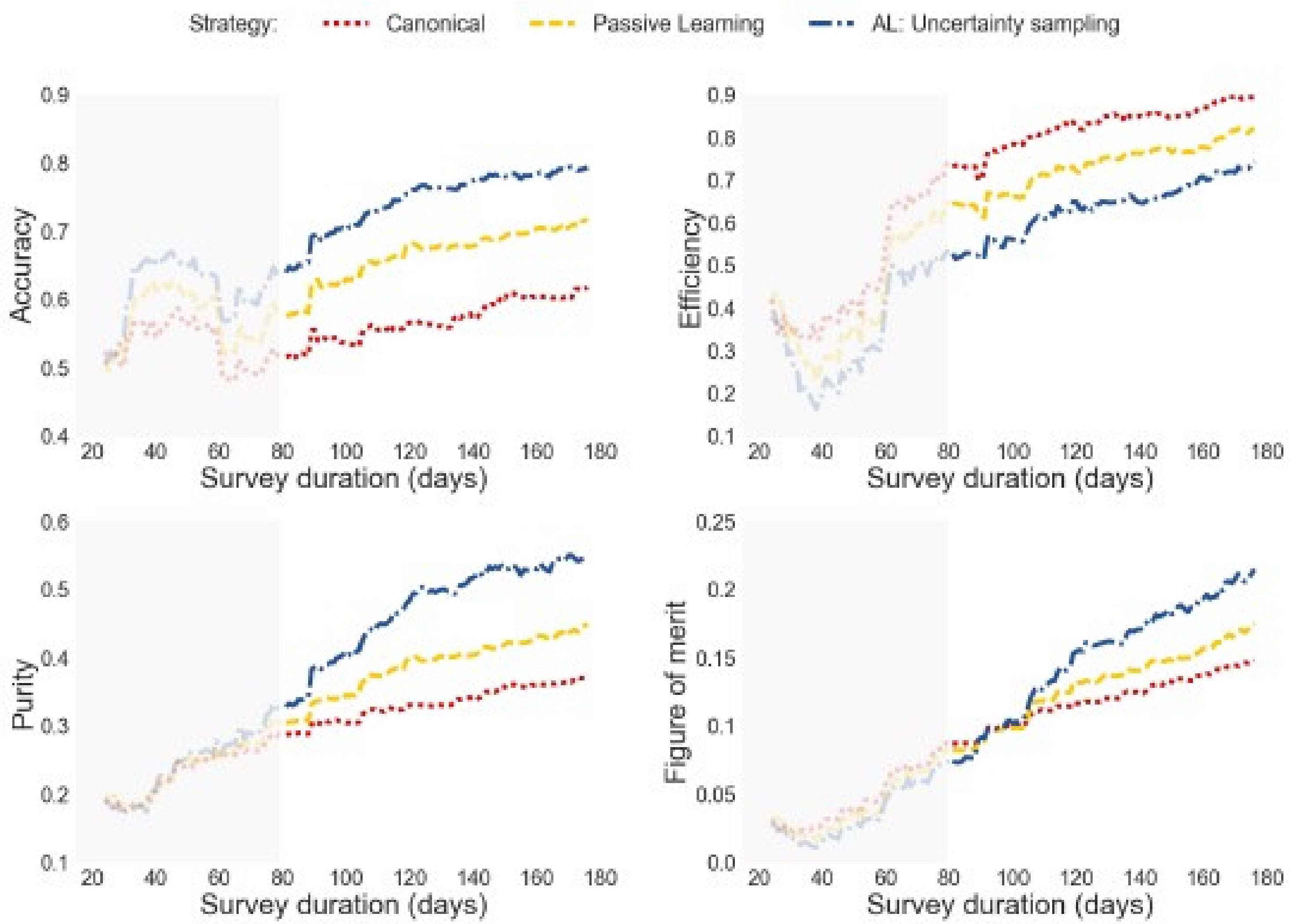


- Need at least 5 days per SN/filter and r-mag< 24 for query: *wait 20 days of survey*
- *Query sample:* objects visible at the time (new faint SNe are in target sample, may move to query and then fade towards target)
- *Build-up phase*: < 80d

Ishida *et al., 2018 - arXiv:astro-ph/1804.03765 - from* ***CRP #4***

Partial LC, initial training

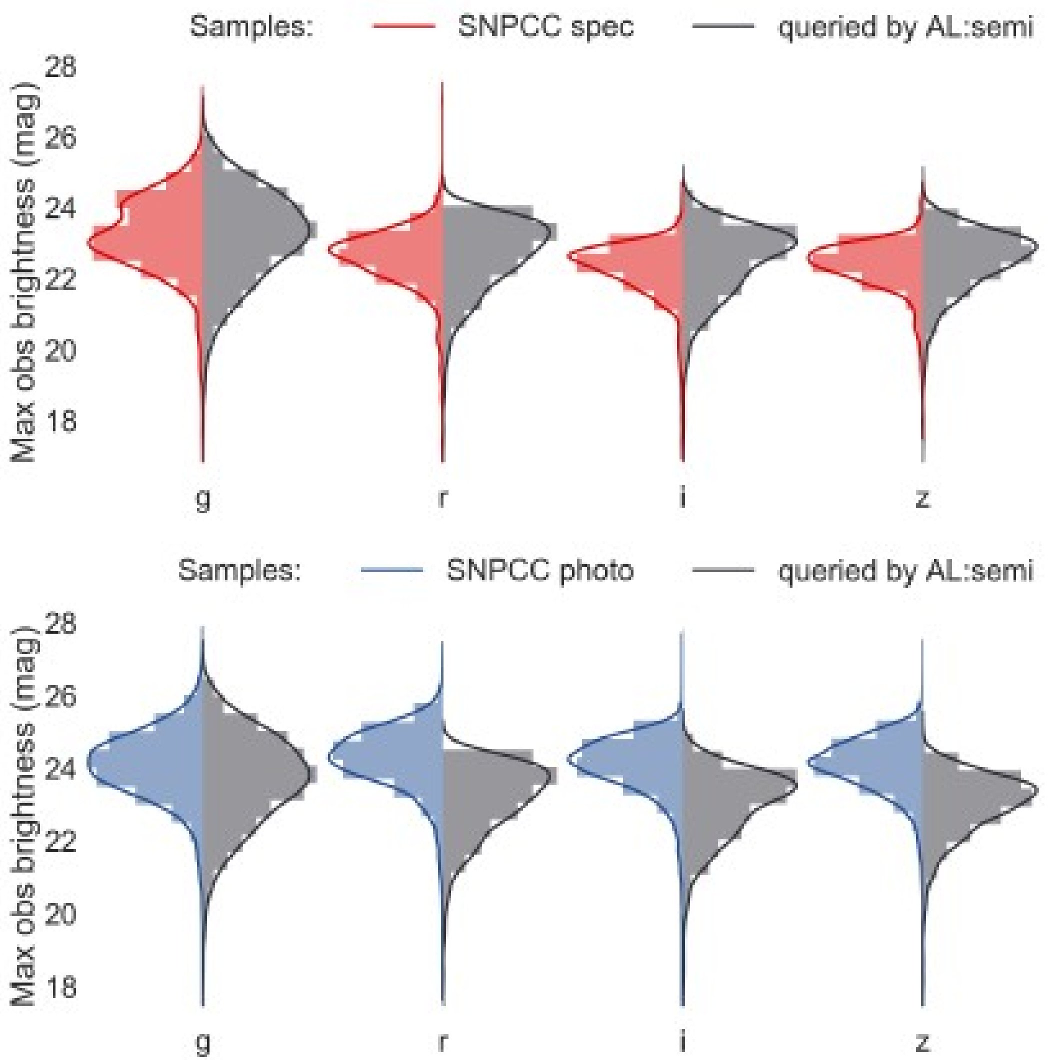
Time domain



Ishida *et al., 2018 - arXiv:astro-ph/1804.03765 - from* ***CRP #4***

The queried sample

Partial LC, no training, time domain, batch



Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Telescope time

- Telescope time can be added as a cost function () instead of a constraint (r-mag<24)

- Integration time estimation required to achieve a given SNR=10 considering magnitude and noise (sky and readout noise) – Bolte 2015

- For training, spectrum considered at max, for queried objects at the time of query.

- Ratio of SNPCC spec sample to objects of semi-supervised AL:

  queried/spec = 0.9992 (2.9s)

- With overheads, it would be significantly less because 26% less objects