The Cosmostatistics Initiative

Reshaping interdisciplinary science development

Artificial Intelligence and Physics
21 March 2019, LAL, Orsay - France

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

The Cosmostatistics Initiative

Reshaping interdisciplinary science development

Artificial Intelligence in Physics
21 March 2019, LAL, Orsay - France

In Astronomy

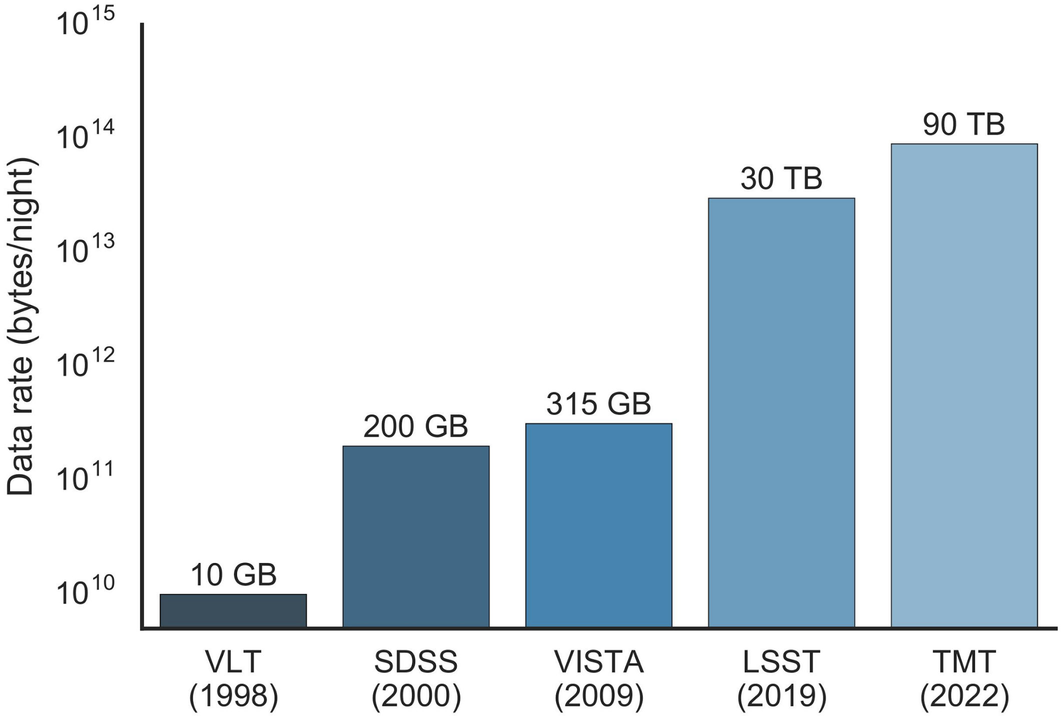
Emille E. O. Ishida

Laboratoire de Physique de Clermont - Université Clermont-Auvergne
Clermont Ferrand, France

Astronomy has been,
traditionally,
an experience of
solitude

The old astronomer, *by Charlie Bowater*

Big data is slowly arriving…



Kremer et al., 2017

… new methods might
take a little longer

The goal of the

Cosmostatistics Initiative

is to speed up this process

while acknowledging

Volatile and competitive job market
Potential contribution of non-astronomers
Diversity of personal and academic background

The COIN Residence Program (CRP)

The COIN Residence Program (CRP)

Step 1 – Choose the people

Who wants to collaborate?

State the rules clearly,

from start

The COIN Residence Program (CRP)

Step 1 - Choose the people

Step 2 - Ask them on which subject they would like to work

The COIN Residence Program (CRP)

Step 1 – Choose the people

Step 2 – Ask them on which subject they would like to work

Step 3 – give them good working conditions

Preparation, **comfort,** motivation

The COIN Residence Program (CRP)

Step 1 – Choose the people

Step 2 – Ask them on wh

would like to work

Step 3 – give them good

conditions

CRP #4, 2017, Clermont Ferrand, France

The COIN Residence Program (CRP)

Step 1 – Choose the people

Step 2 – Ask them on which subject they would like to work

Step 3 – give them go

conditions

CRP #5, 2018, Chania, Greece

The COIN Residence Program (CRP)

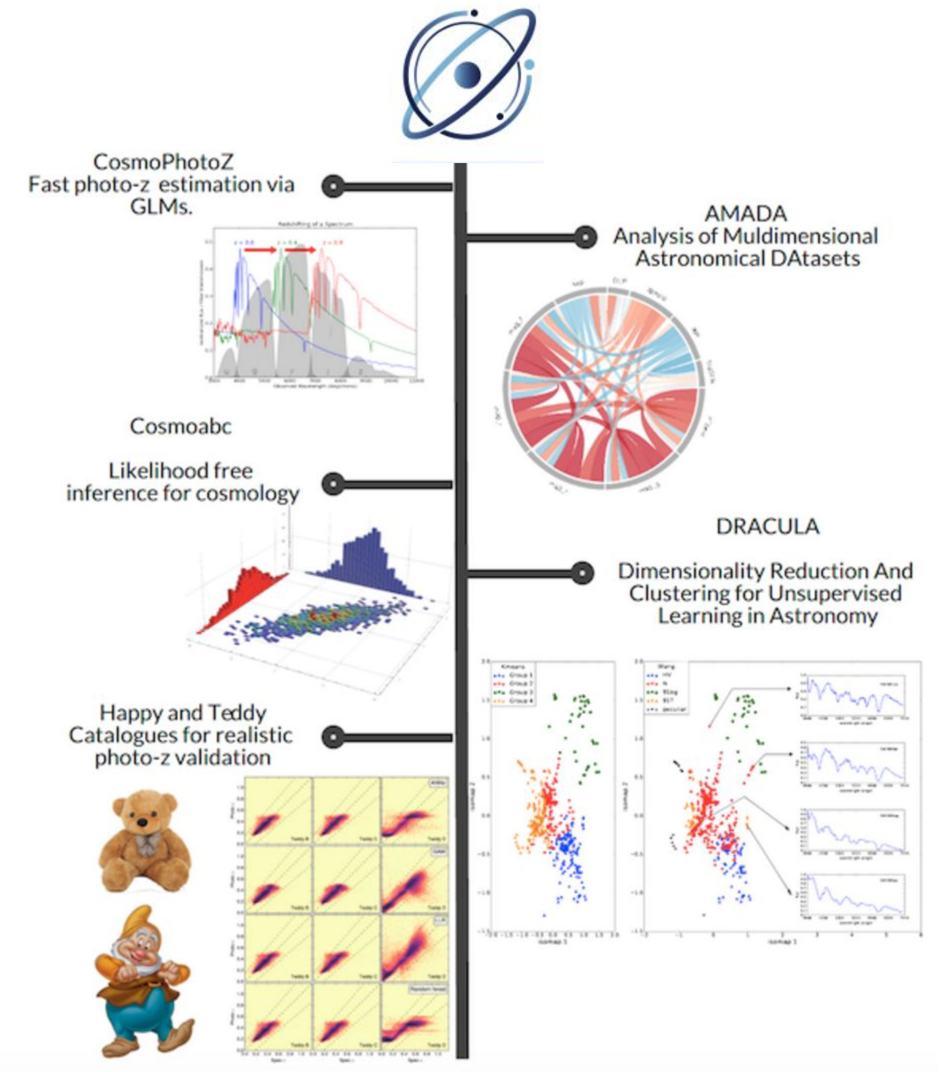
Step 1 – Choose the people

Step 2 – Ask them on which subject they would like to work

Step 3 – give them good working conditions

Step 4 – make sure they do not diverge

In 5 years,
60 researchers
15 countries



CosmoPhotoZ
Fast photo-z estimation via GLMs.

AMADA
Analysis of Muldimensional Astronomical DAtasets

Cosmoabc

Likelihood free inference for cosmology

DRACULA

Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy

Happy and Teddy
Catalogues for realistic photo-z validation

	Paper	Citation
1	GLM I	de Souza *et al.*, 2015
2	GLM II	Elliott *et al.*, 2015
3	GLM III	de Souza *et al.*, 2015
4	AMADA	de Souza & Ciardi, 2015
5	CosmoABC	Ishida *et al.*, 2015
6	DRACULA	Sasdelli *et al.*, 2016
7	AGNlogit	de Souza *et al.*, 2016
8	PhotoZ	Beck *et al.*, 2017
9	AGNgmm	de Souza *et al.*, 2017
10	GalINLA	Gonzalez-Gaitan *et al.*, 2018
11	ActSNclass	Ishida *et al.*, 2018
12	COIN-Gaia	Cantat-Gaudin *et al.*, 2018
13	Hurdle	Hattab *et al.*, 2019
14	SNCosmo	Moews *et al.*, 2018

	Code	Citation
1	CosmoPhotoZ	de Souza *et al.*, 2014,
2	AMADA	de Souza & Ciardi, 2015
3	CosmoABC	Ishida *et al.*, 2015
4	DRACULA	Aguena *et al.*, 2015
5	CoinINLA	Gonzalez-Gaitan *et al.*, 2018

+ 1 galaxy catalog
+ 1 GMM tutorial
+ 2 photoz catalogs
+ 41 open clusters

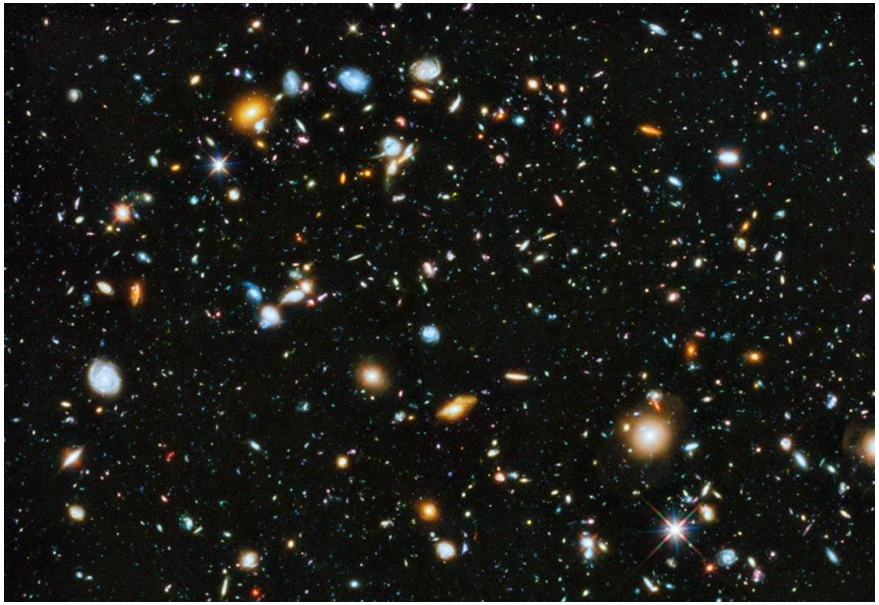
Case study:

Astronomy needs a recommendation system

Types of astronomical data:

Photometry

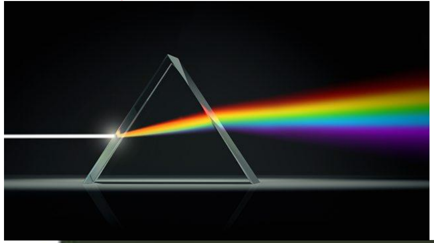
Features
(cheap)



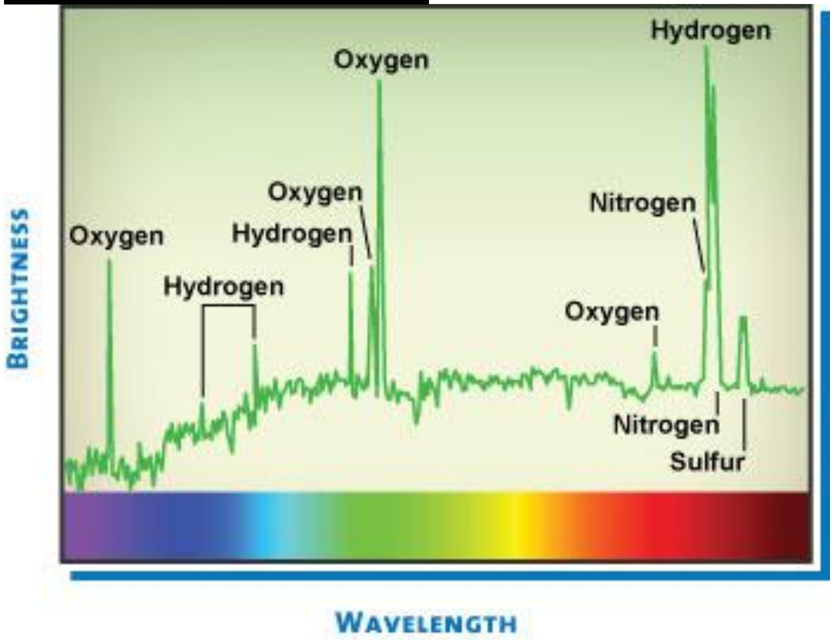
Brightness per object
given wavelength range

Spectroscopy

Used to derive labels (classes)
Very expensive

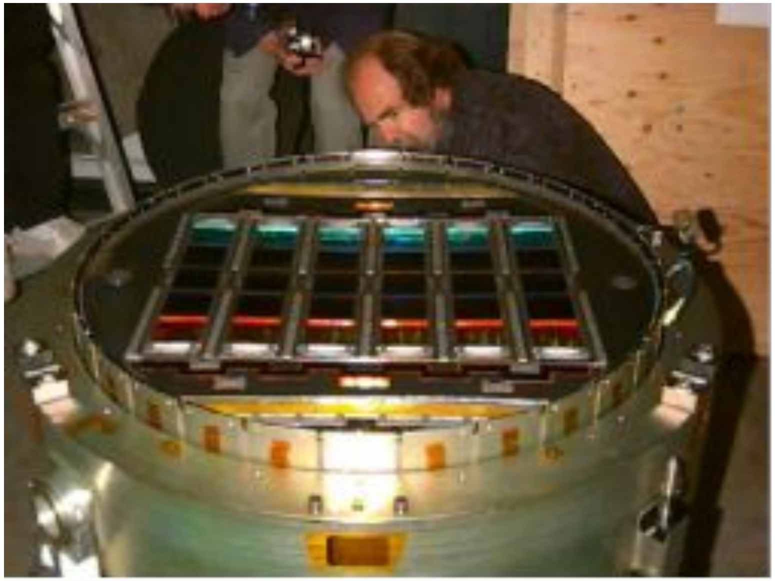


Brightness
given wavelength
per object



Photometry x Spectroscopy

An example from SDSS



Exposure time 2 x 54s

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

Integration time of at least 45 minutes

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

Astronomers would like spectra of everything

Spectroscopy is the key

… which we will not have

DARK ENERGY SURVEY

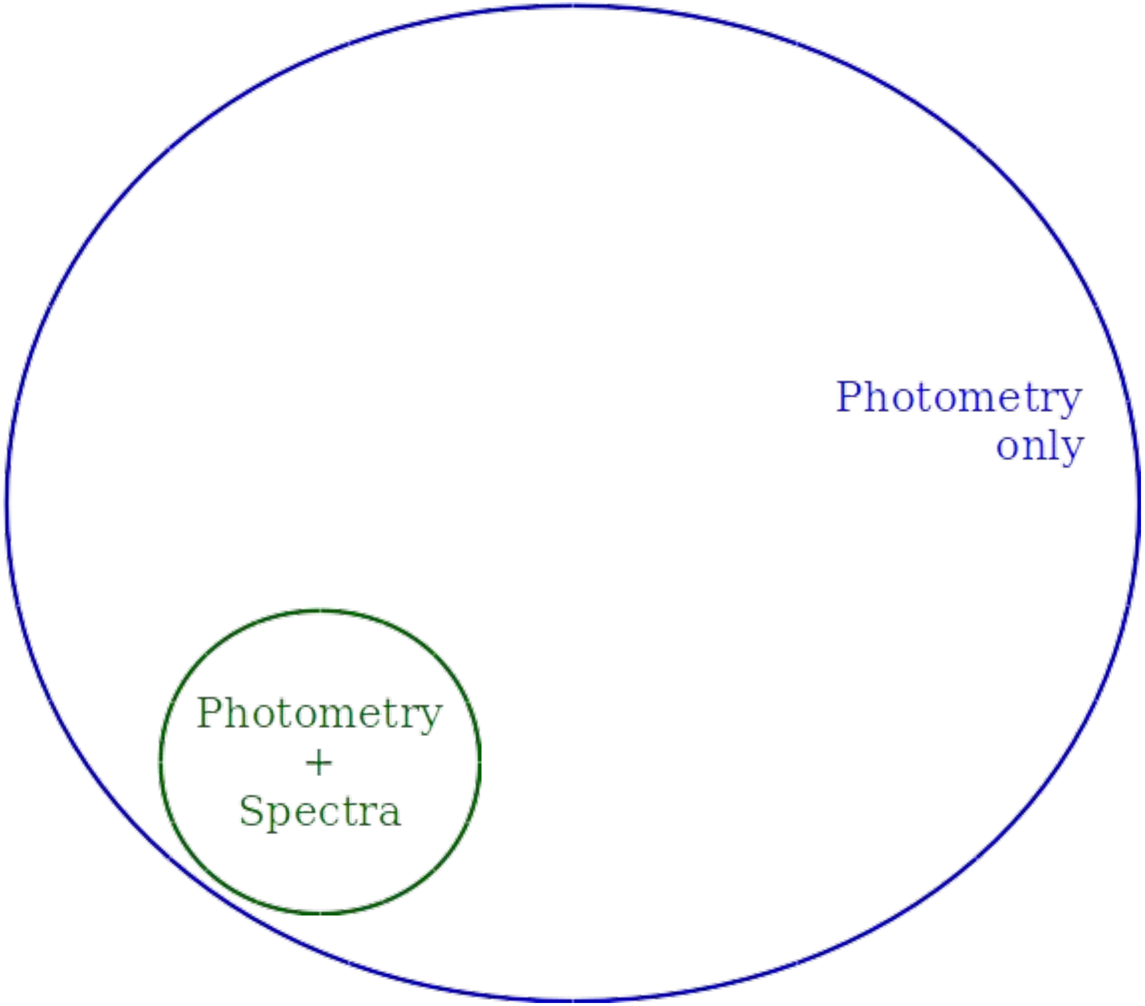
LSST

Large Synoptic Survey Telescope

Big Data in astronomy means more <u>photometry</u>!

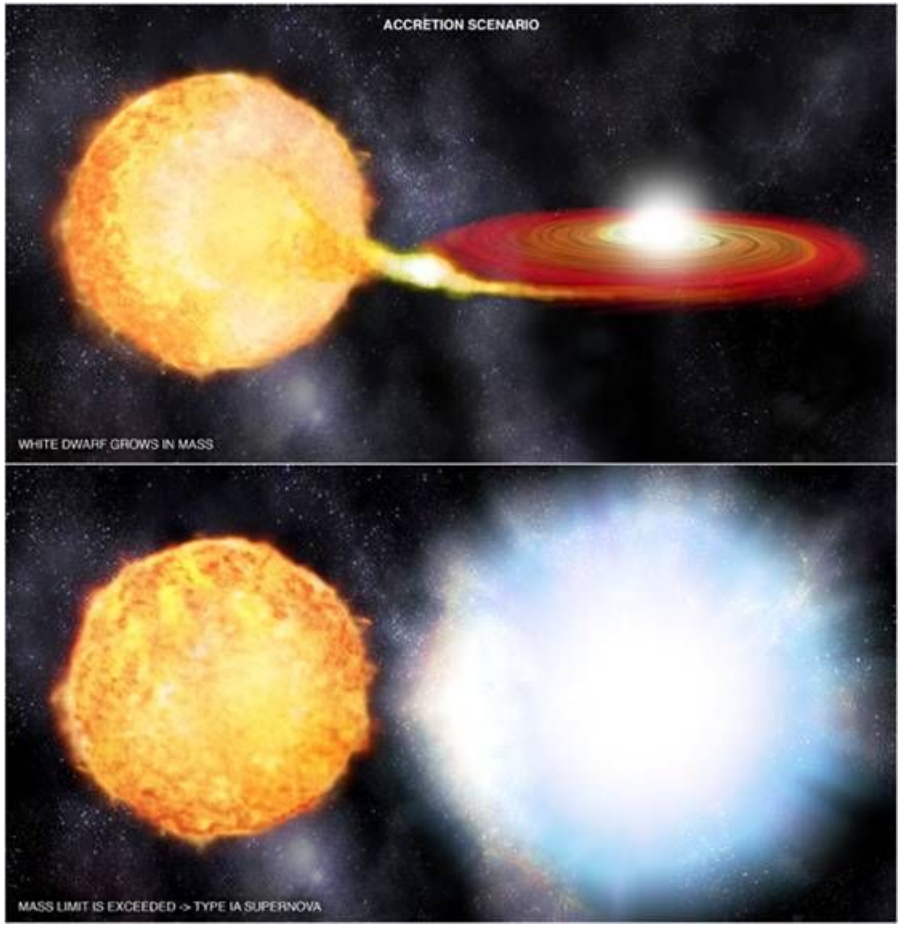
Data situation:

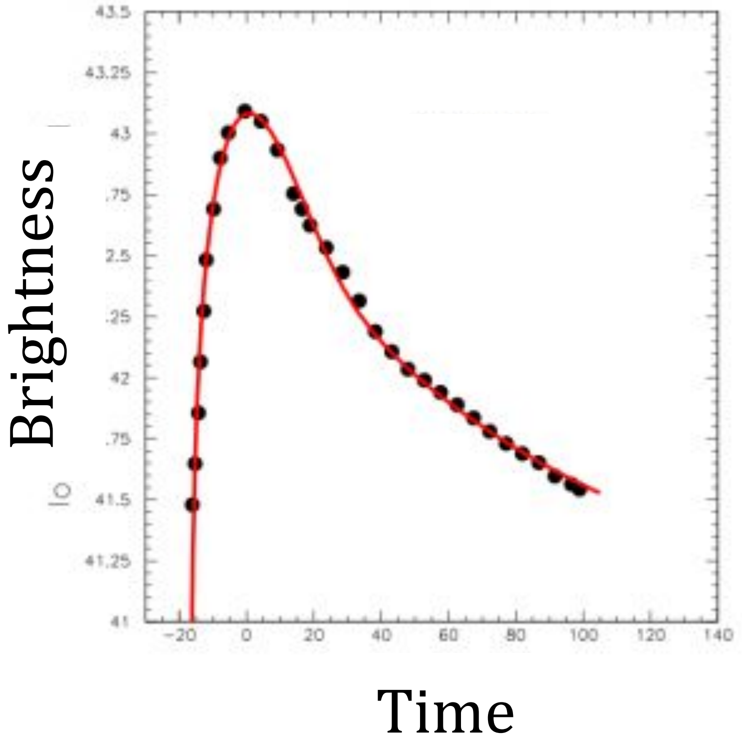
The Problem



Example of things you might want to classify

Supernova Ia


ACCRETION SCENARIO
WHITE DWARF GROWS IN MASS
MASS LIMIT IS EXCEEDED -> TYPE IA SUPERNOVA



https://phys.org/news/2014-08-ia-supernovae-stem-explosion-white.html

Why is this important?

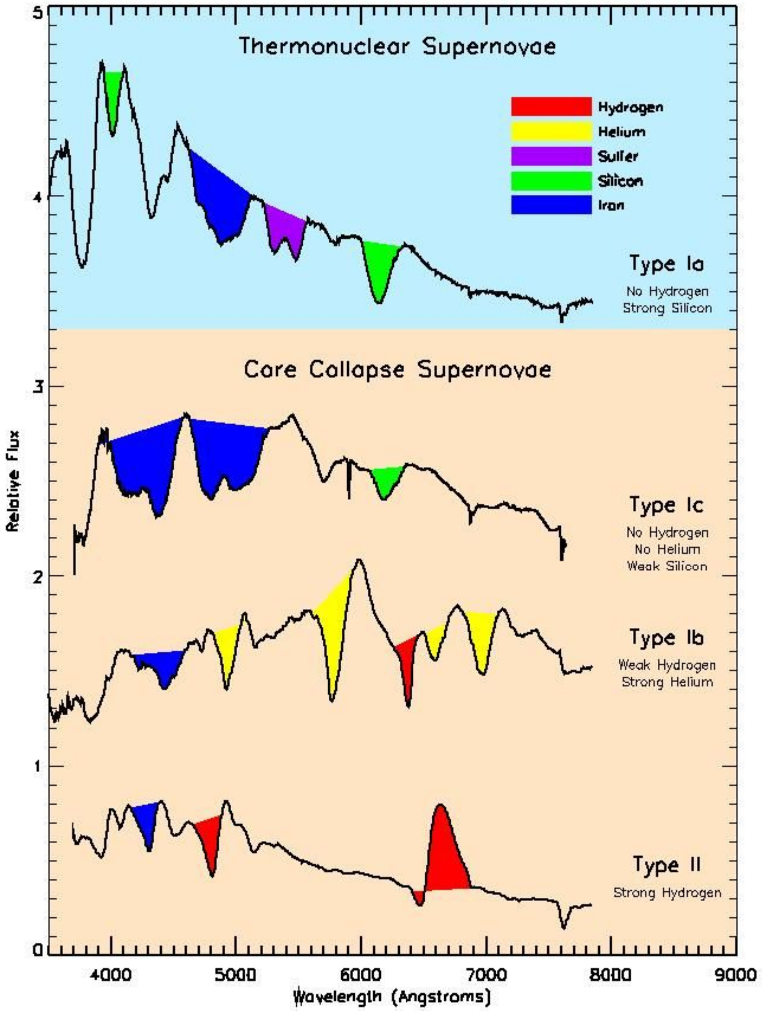
Supernova Ia

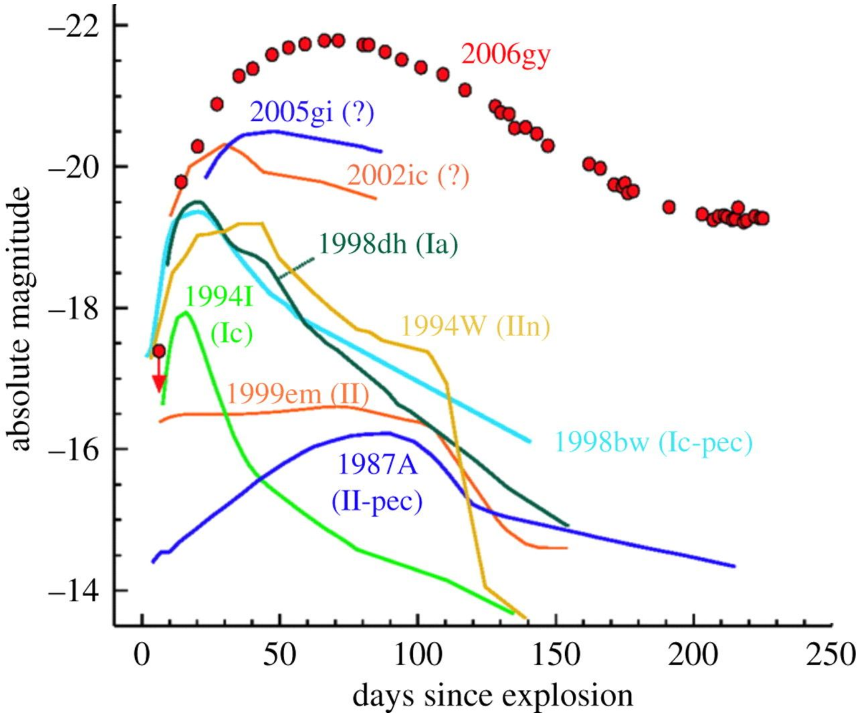
Only Supernova Ia can be used as standard candles

2011

Supernova photometric classification:

It’s complicated!



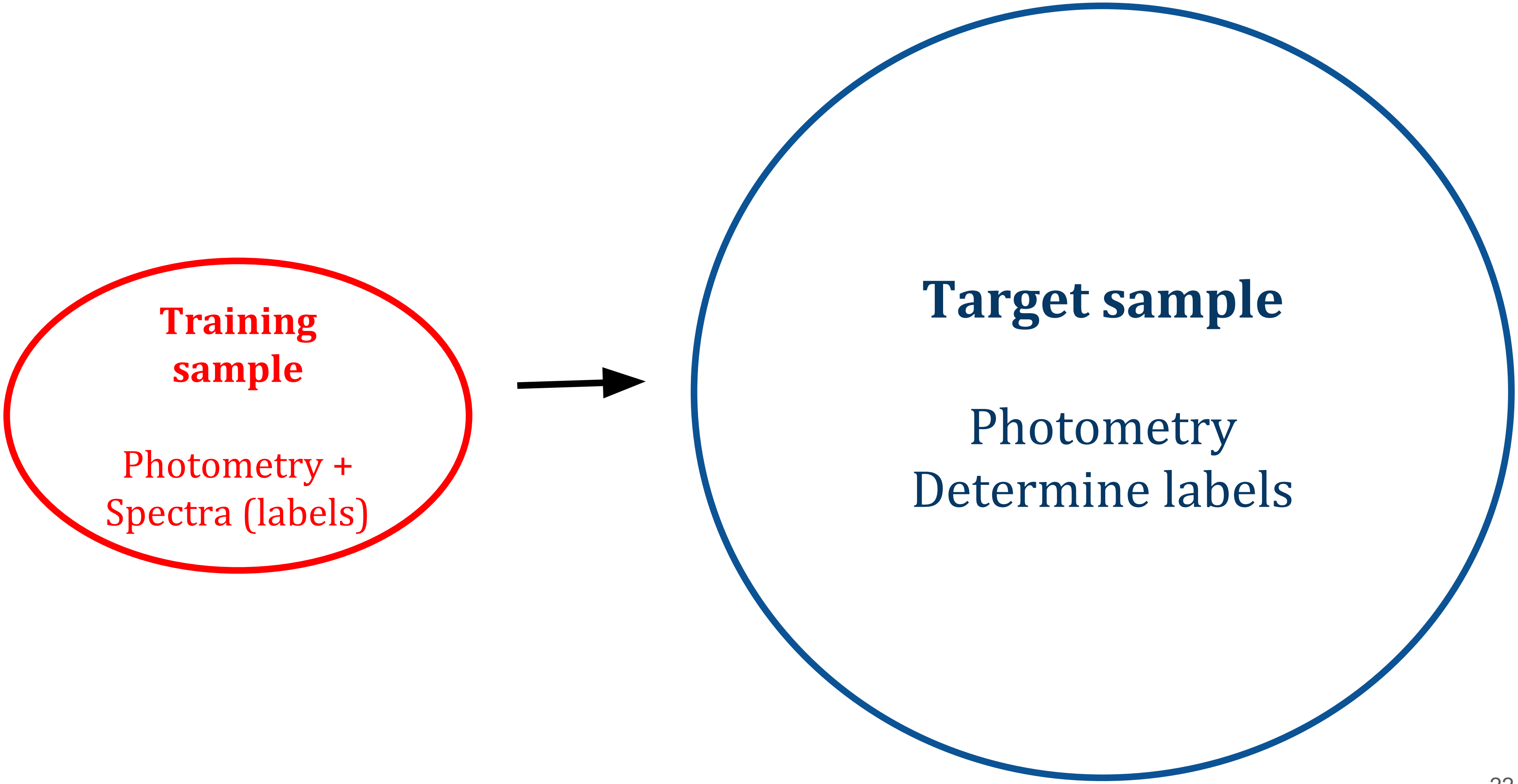


http://rsta.royalsocietypublishing.org/content/370/1960/774

Introduction:

Machine Learning solution

Spectra as labels, photometry as features



Training sample

Photometry +
Spectra (labels)

Target sample

Photometry
Determine labels

Machine Learning solution

Spectra as labels, photometry as features

Training sample

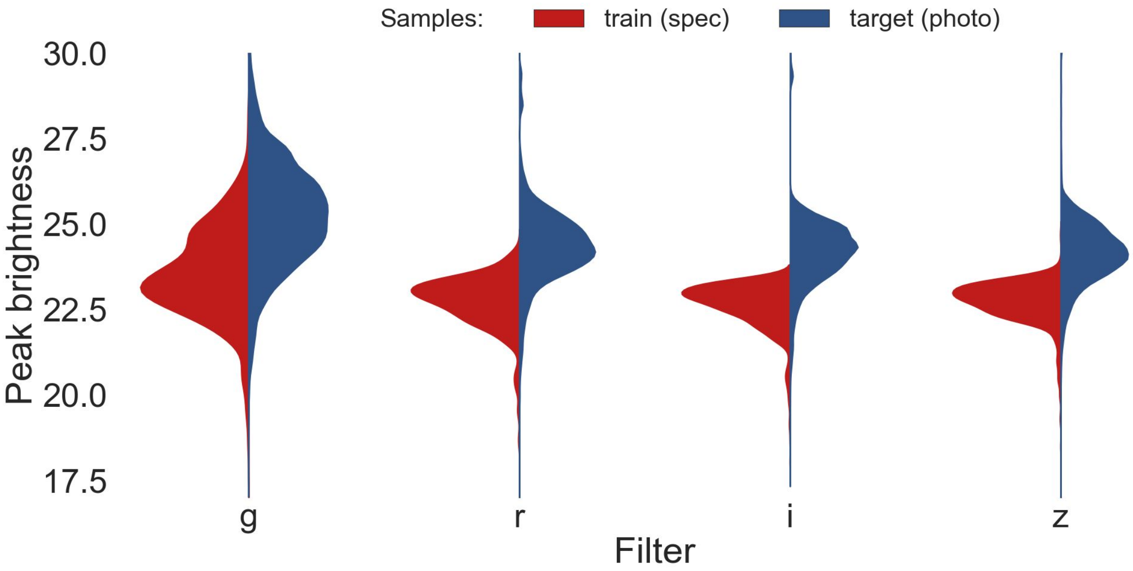
Photometry +
Spectra (labels)

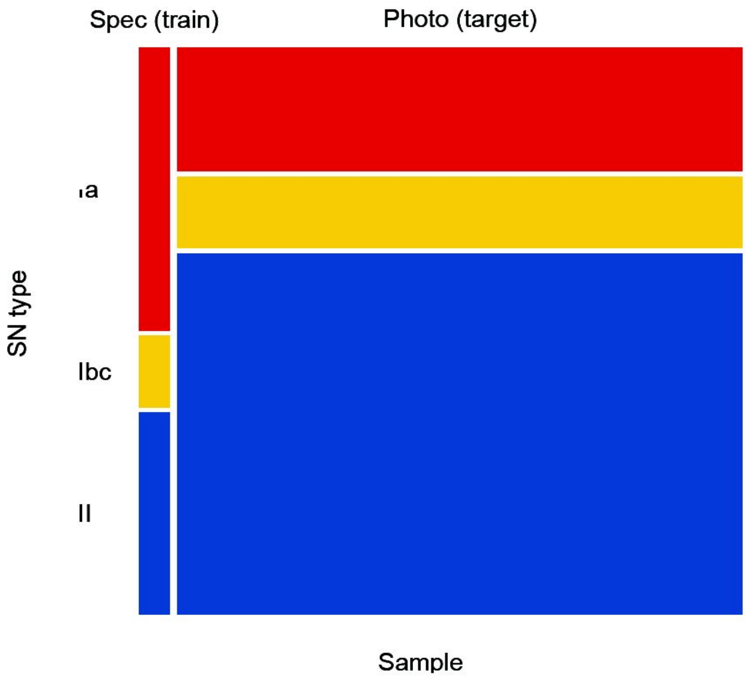
→

Target sample

Photometry
Determine labels

Representativeness

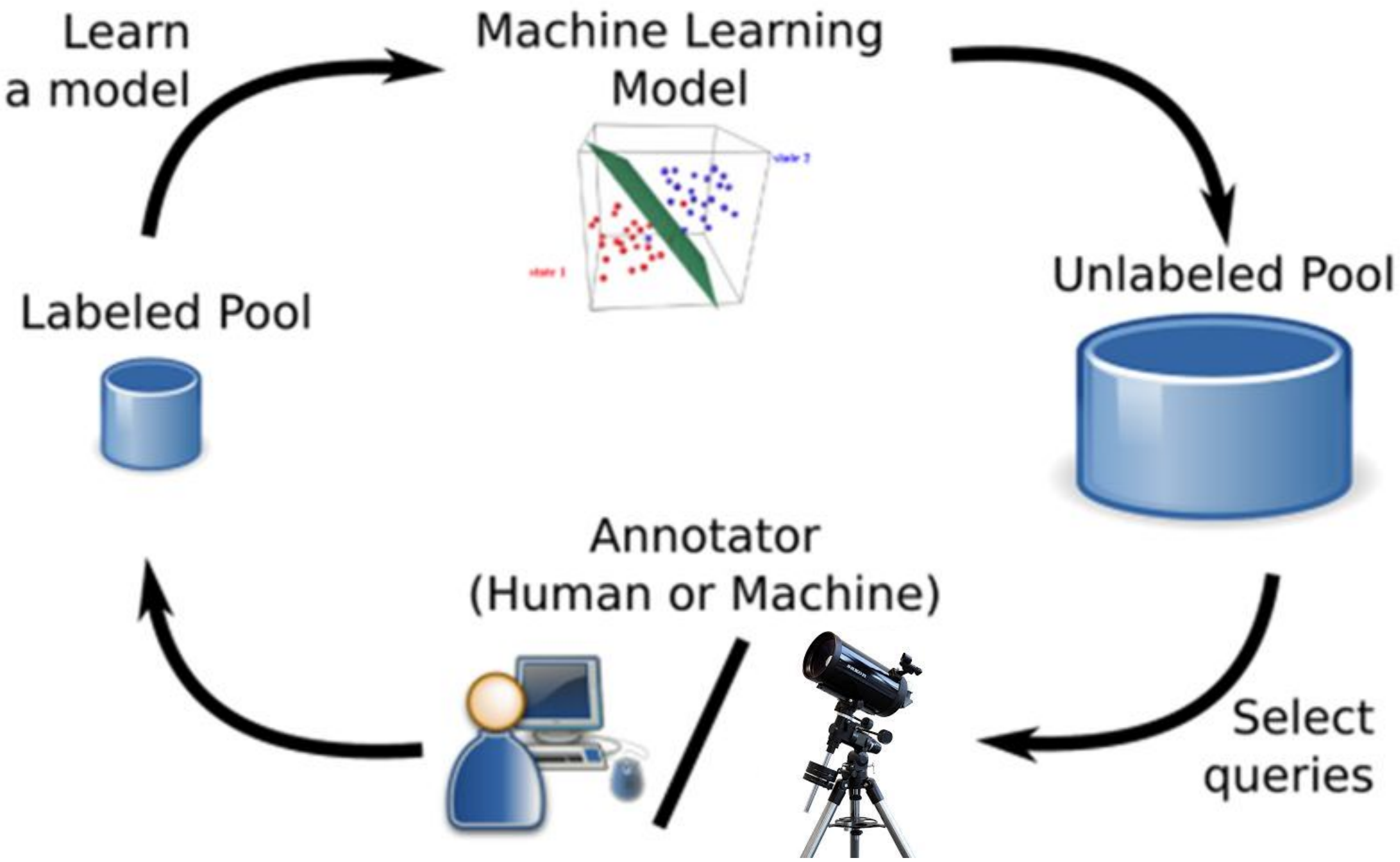




Spectroscopic sample was never meant to be a training set

Active Learning

Optimal classification, minimum training



Learn a model

Machine Learning Model

Unlabeled Pool

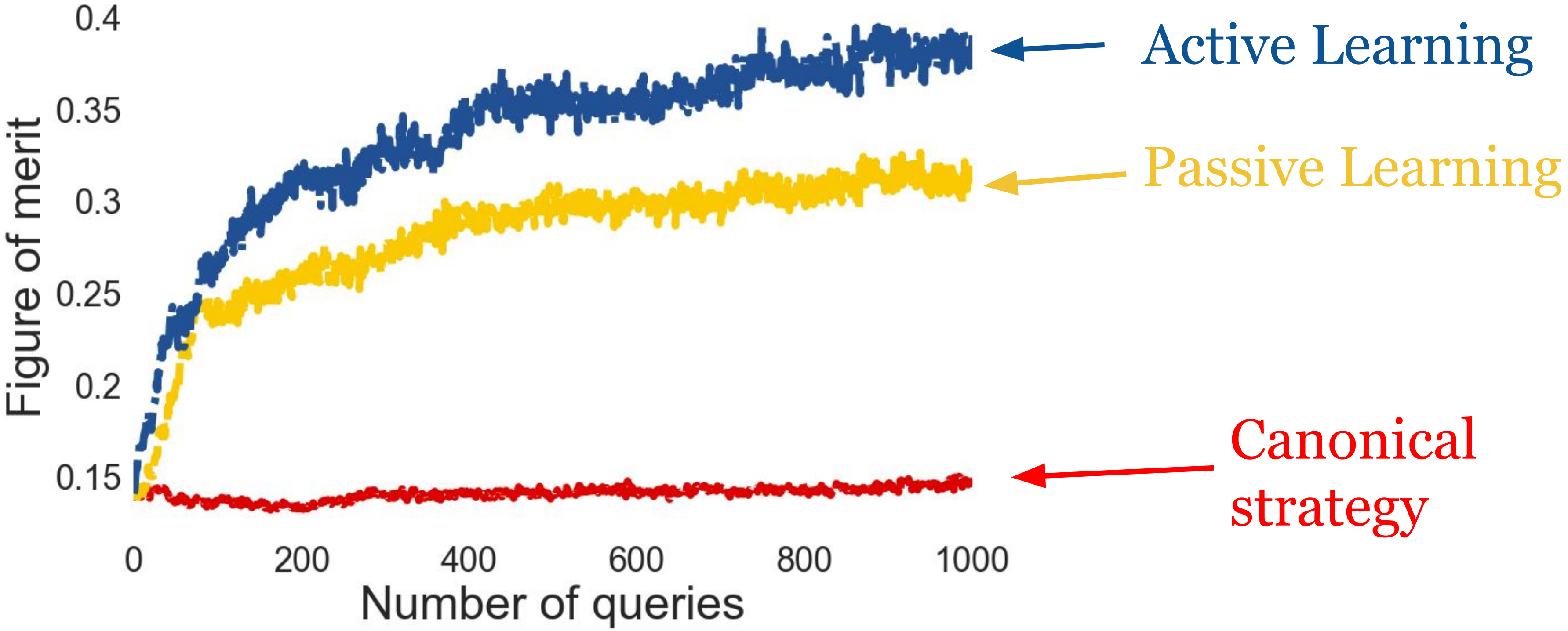
Labeled Pool

Annotator (Human or Machine)

Select queries

AL for SN classification

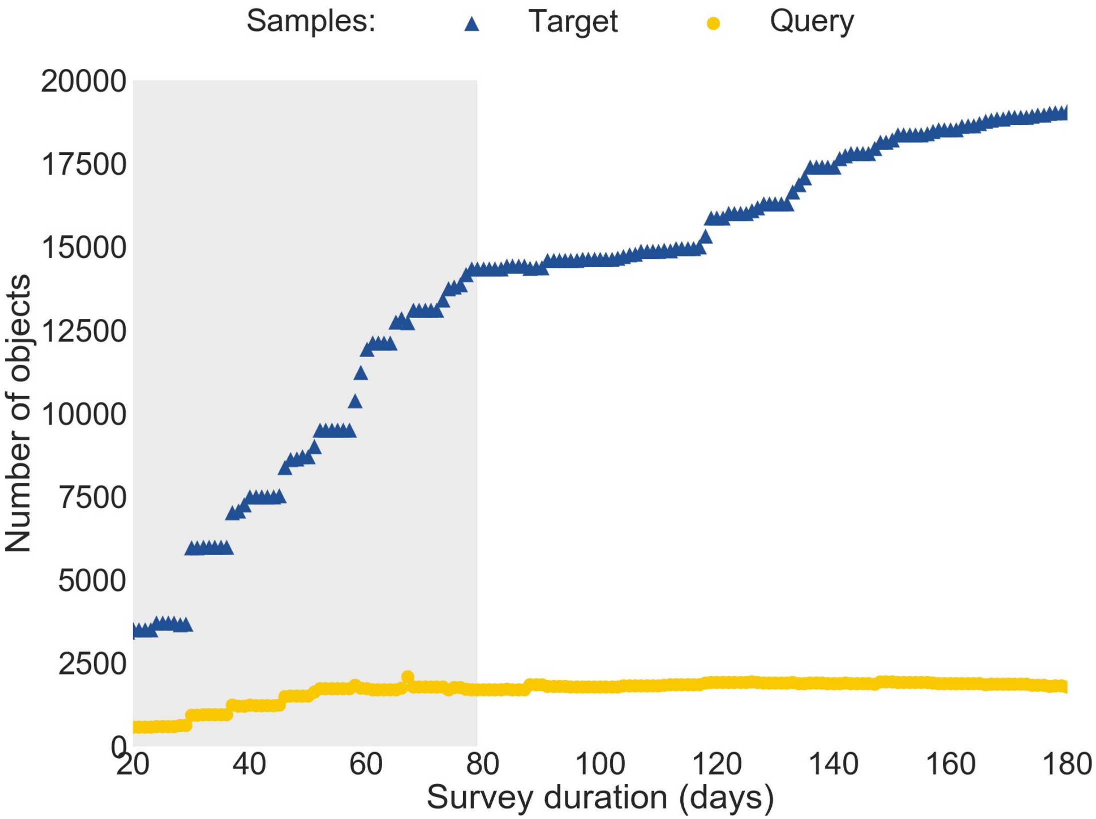
Static results



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

SN are transients

Not everything is available for labelling

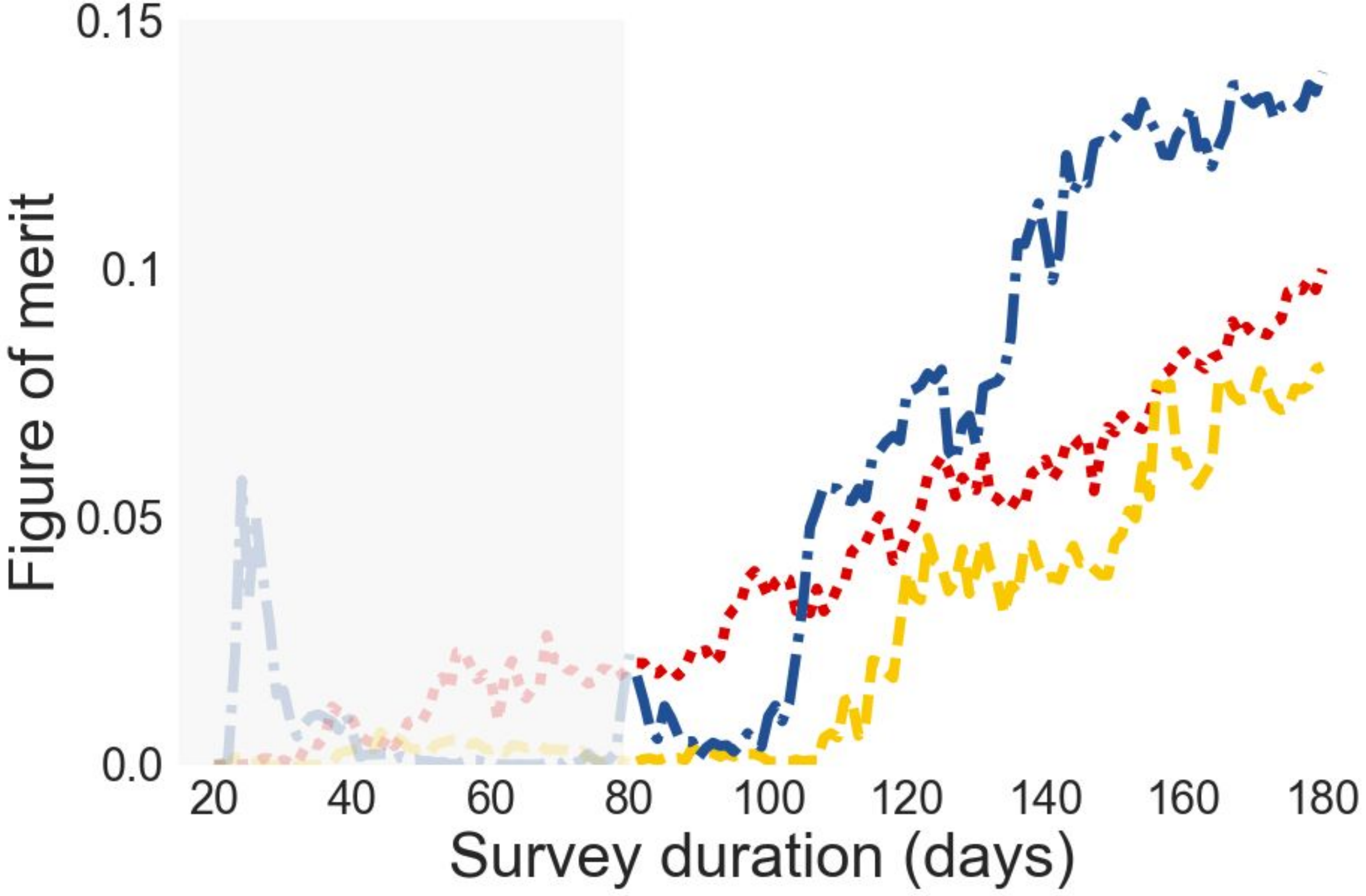


1. Feature extraction done daily **with available observed epochs until then.**

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*

Beginning from scratch



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

Does this solve the problem completely?

No, it is just the best you can do!

Is this the only way of doing it?

Certainly not!

What is next **for this project**?

- ❏ Agreement being drafted with a major telescope to stress test this idea in a more realistic astronomical scenario - world wide coordination with spectroscopic telescopes
- ❏ Adapt this to multi-fiber spectrograph - where should I point the telescope?
- ❏ Issues still to be tackled:
  - ❏ Uncertainties everywhere!
  - ❏ Scalability - LSST will have 2 million alerts/night
  - ❏ Metrics for different science goals
- ❏ Anomaly detection
- ❏ Active Learning for Regression
  - ❏ Representativeness and correlations in uncertainty space

This is a (very unique) group effort!

COIN Residence Program #4

20 - 27 August 2017

Clermont Ferrand, France

Sponsors:

What have we learn from the COIN experience so far?

- The human factor needs to be respected
- True interdisciplinary means freedom and requires trust
- The environment is very important (architecture)
- There is human potential waiting to be used in science in the outside world (for free)
- The most efficient way to work with astronomical data is to have an astronomer friend

Next time, up the mountain!

Application deadline, 10 April

COIN Residence Program #6
Morzine - France, 24 - 31 August 2019

https://cosmostatistics-initiative.org/residence-programs/crp6/

THANK YOU

Extra Slides

The COIN Residence Program (CRP)

Step 1 - Choose the people

Step 2 - Ask them on whi

would like to work

Step 3 - give them good v

conditions

CRP #4, 2017, Clermont Ferrand, France

The COIN Residence Program (CRP)

Step 1 – Choose the people

Step 2 – Ask them on which subject they would like to work

Step 3 – give working conditions

CRP #3, 2016, Budapest, Hungary

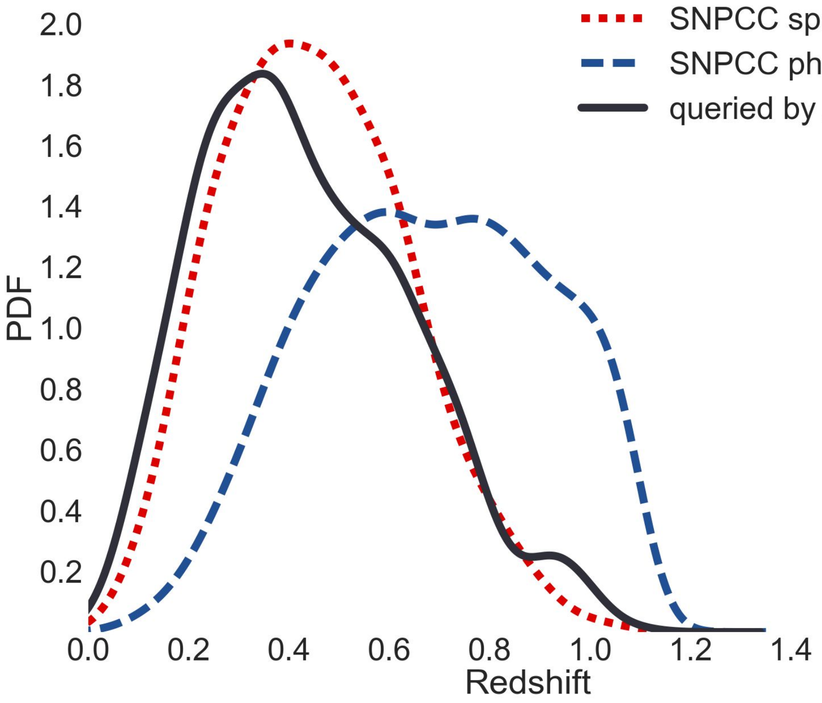
The queried sample

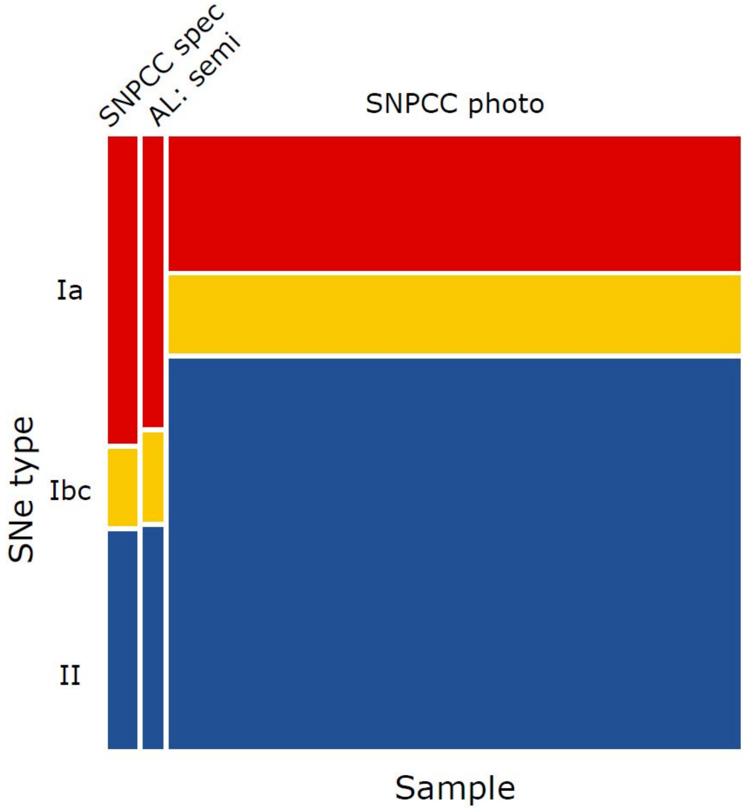
Partial LC, no training, time domain, batch

SNPCC spec:
1103 objects

Queried sample:
800 objects

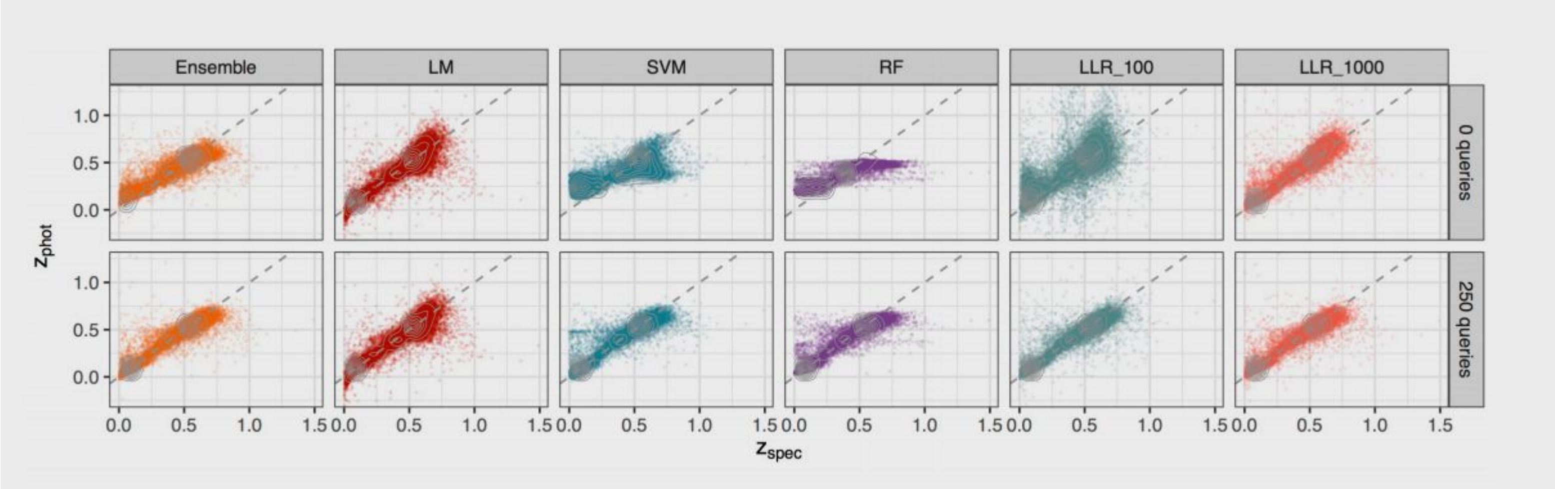
Telescope time:
Queried/spec = 0.999





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

AL for Photo-Z



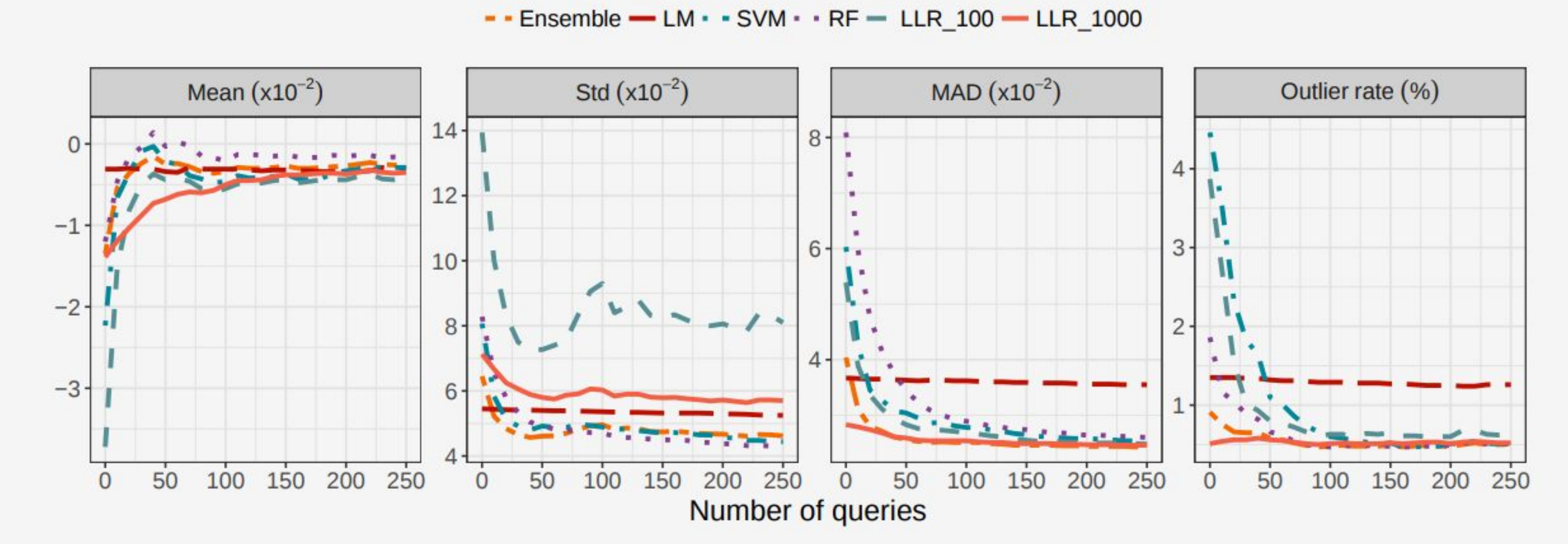
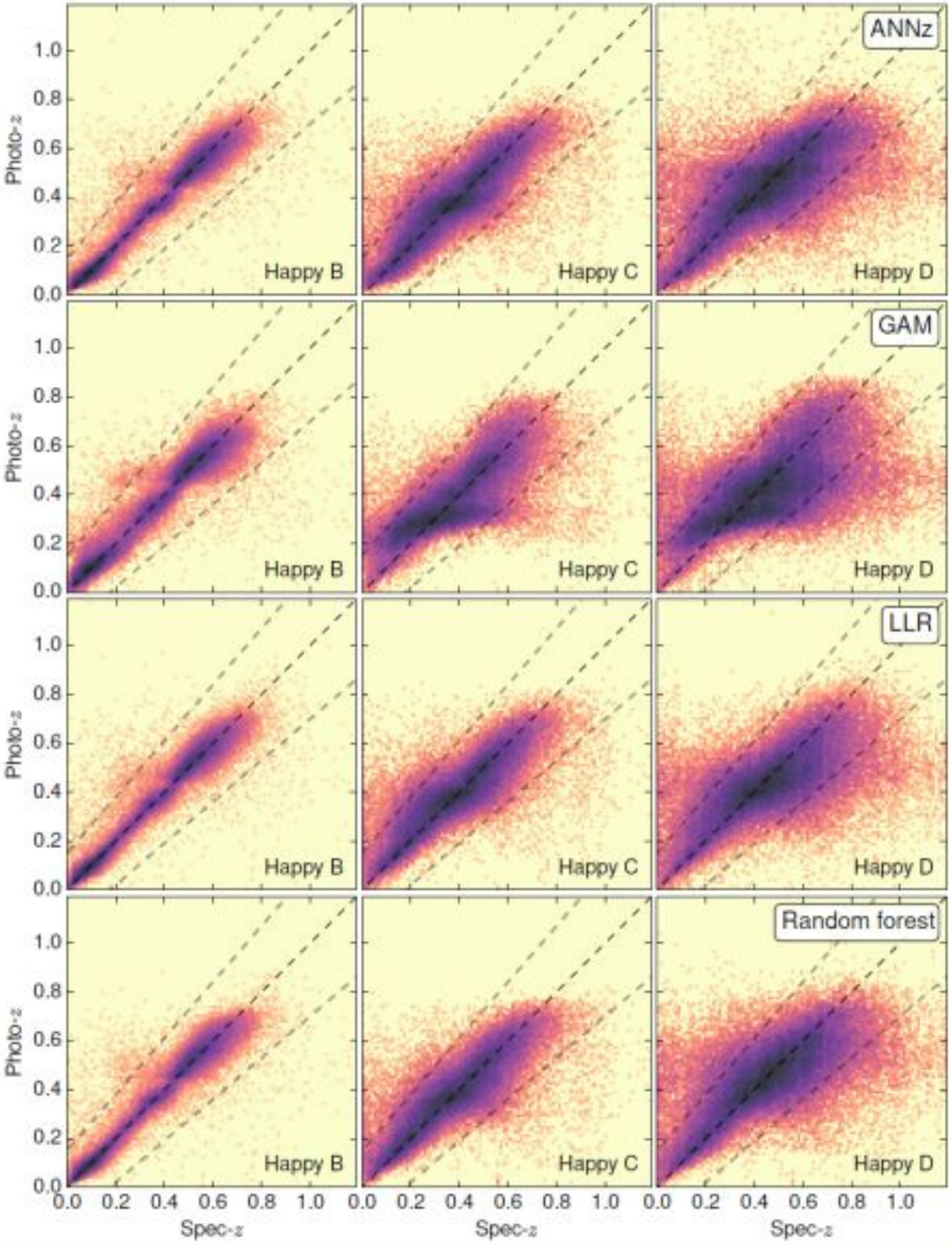


Figure 4. An assessment of the performance of the ensemble model and its constituent models using active learning. Performance diagnostics are shown as a function of the number of queries.

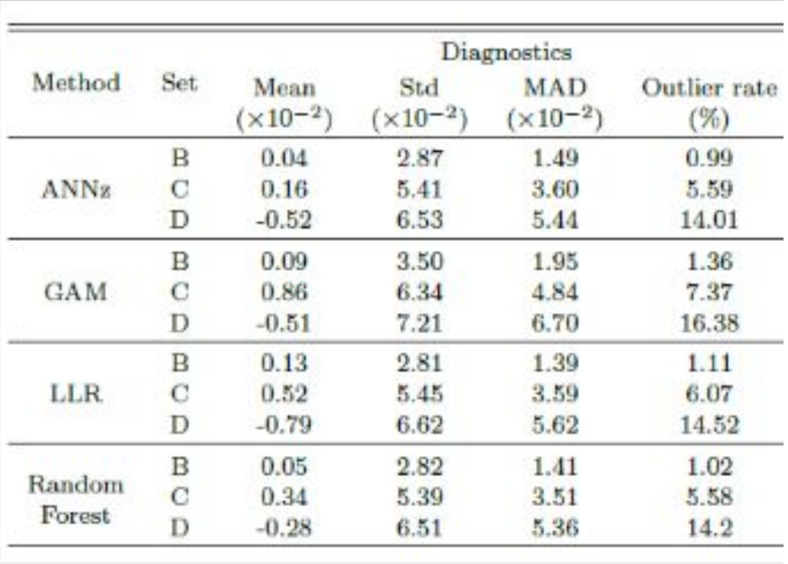
*Vilalta, **Ishida** et al., 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*

Happy catalogue

The effect of coverage + photometric errors



Empirical methods



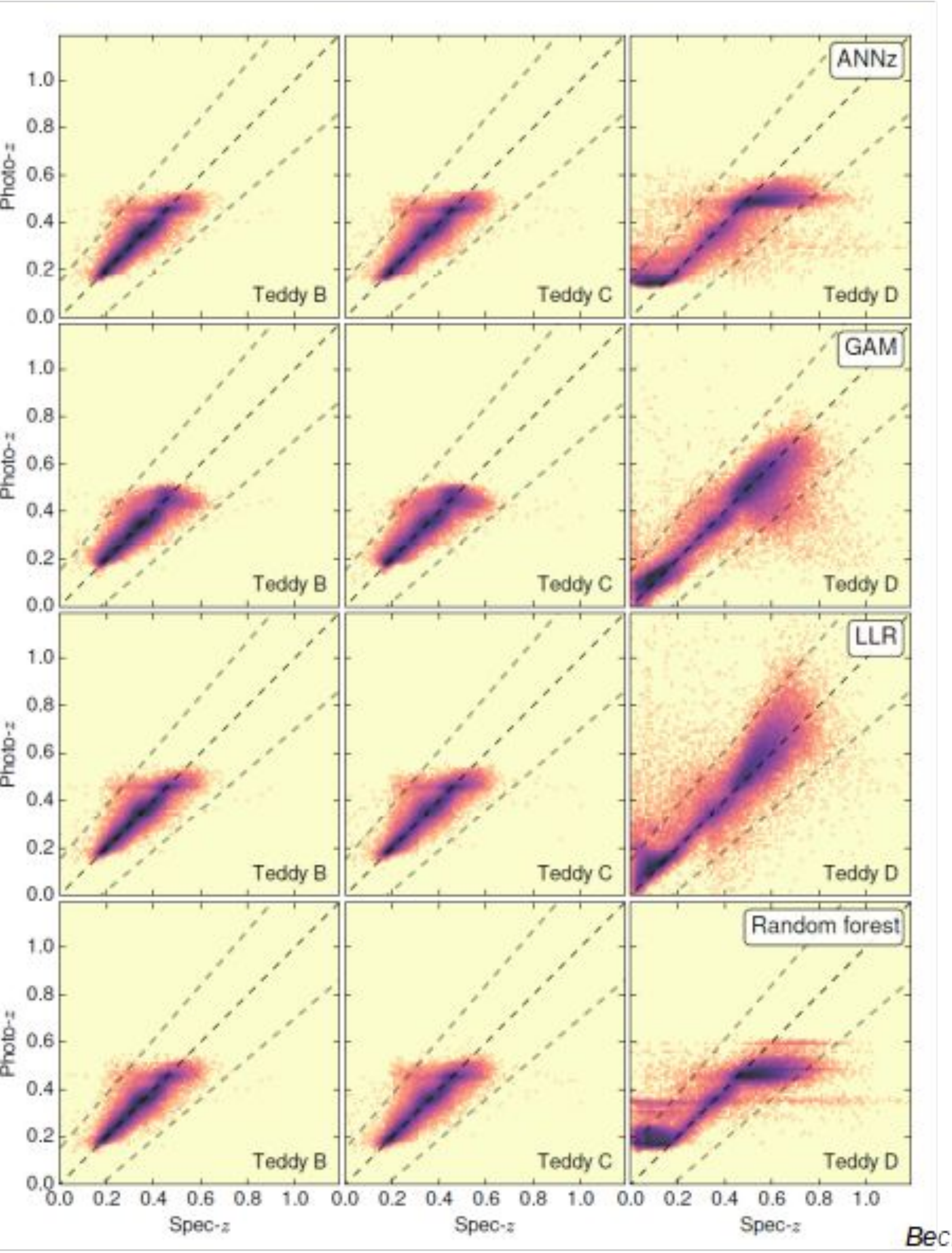
Method	Set	Diagnostics			
		Mean ($\times 10^{-2}$)	Std ($\times 10^{-2}$)	MAD ($\times 10^{-2}$)	Outlier rate (%)
ANNz	B	0.04	2.87	1.49	0.99
	C	0.16	5.41	3.60	5.59
	D	-0.52	6.53	5.44	14.01
GAM	B	0.09	3.50	1.95	1.36
	C	0.86	6.34	4.84	7.37
	D	-0.51	7.21	6.70	16.38
LLR	B	0.13	2.81	1.39	1.11
	C	0.52	5.45	3.59	6.07
	D	-0.79	6.62	5.62	14.52
Random Forest	B	0.05	2.82	1.41	1.02
	C	0.34	5.39	3.51	5.58
	D	-0.28	6.51	5.36	14.2

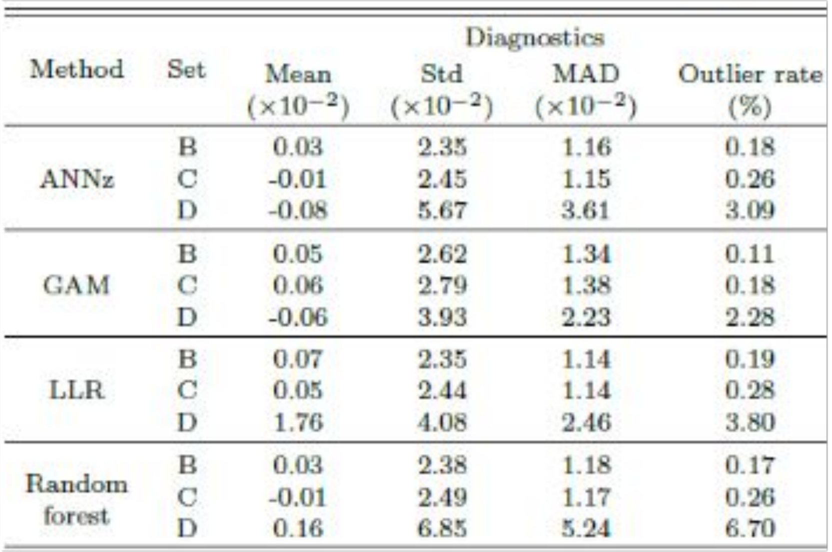
Beck et al., astro-ph:1701.08748, MNRAS in press

Teddy catalogue

The effect of color coverage

Empirical methods





Method	Set	Diagnostics Mean ($\times 10^{-2}$)	 Std ($\times 10^{-2}$)	 MAD ($\times 10^{-2}$)	 Outlier rate (%)
ANNz	B	0.03	2.35	1.16	0.18
	C	-0.01	2.45	1.15	0.26
	D	-0.08	5.67	3.61	3.09
GAM	B	0.05	2.62	1.34	0.11
	C	0.06	2.79	1.38	0.18
	D	-0.06	3.93	2.23	2.28
LLR	B	0.07	2.35	1.14	0.19
	C	0.05	2.44	1.14	0.28
	D	1.76	4.08	2.46	3.80
Random forest	B	0.03	2.38	1.18	0.17
	C	-0.01	2.49	1.17	0.26
	D	0.16	6.85	5.24	6.70

Beck et al., astro-ph:1701.08748, MNRAS in press

Urgency:

The data Paradigm

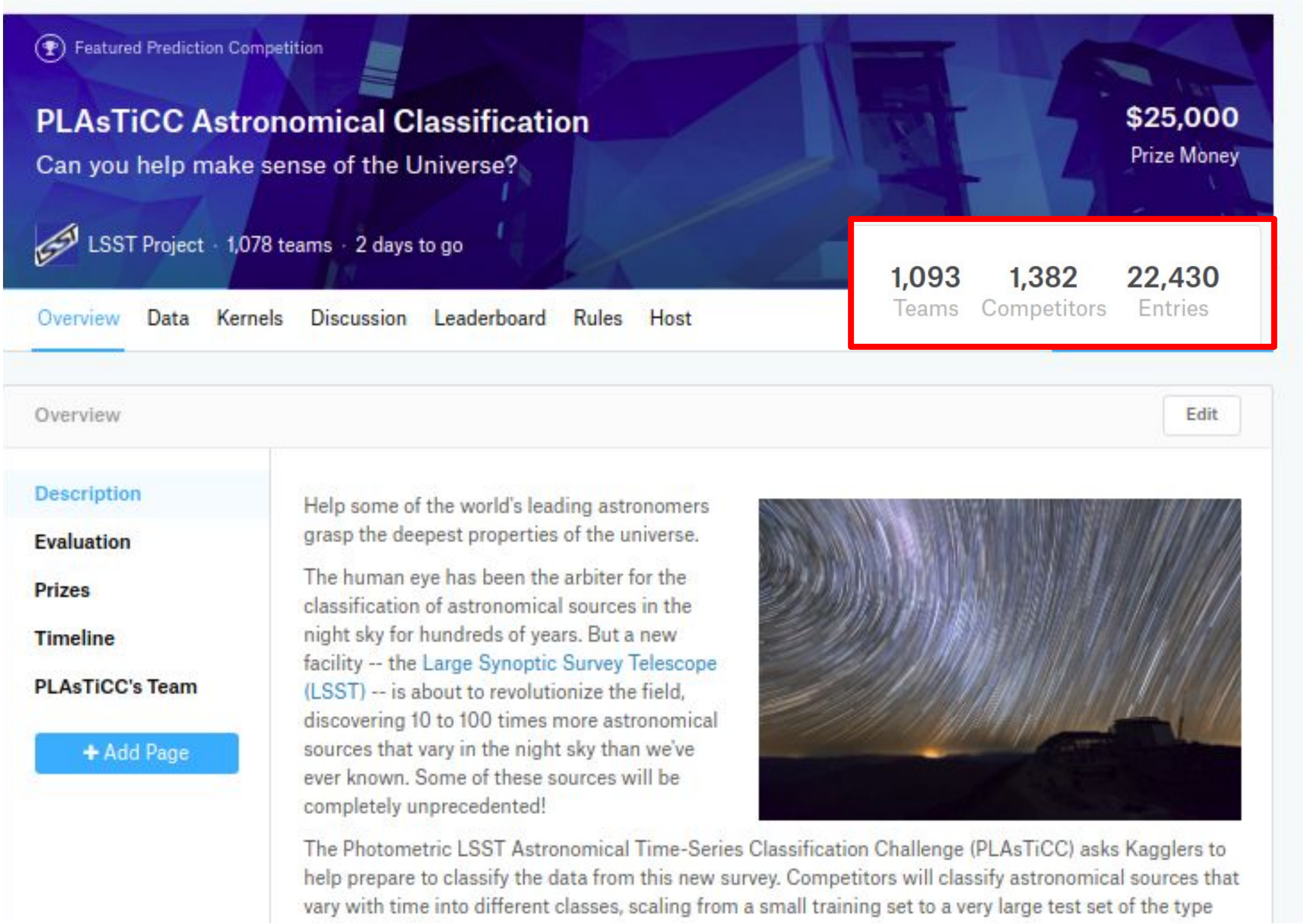
year	Number of supernova
1998	42
2014	740
2025	> 10 000

2 million alerts/day
15 TB/day

40 nights of LSST
↓
entire Google database

A. Connelly, TED2014

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



PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

LSST Project · 1,078 teams · 2 days to go

$25,000
Prize Money

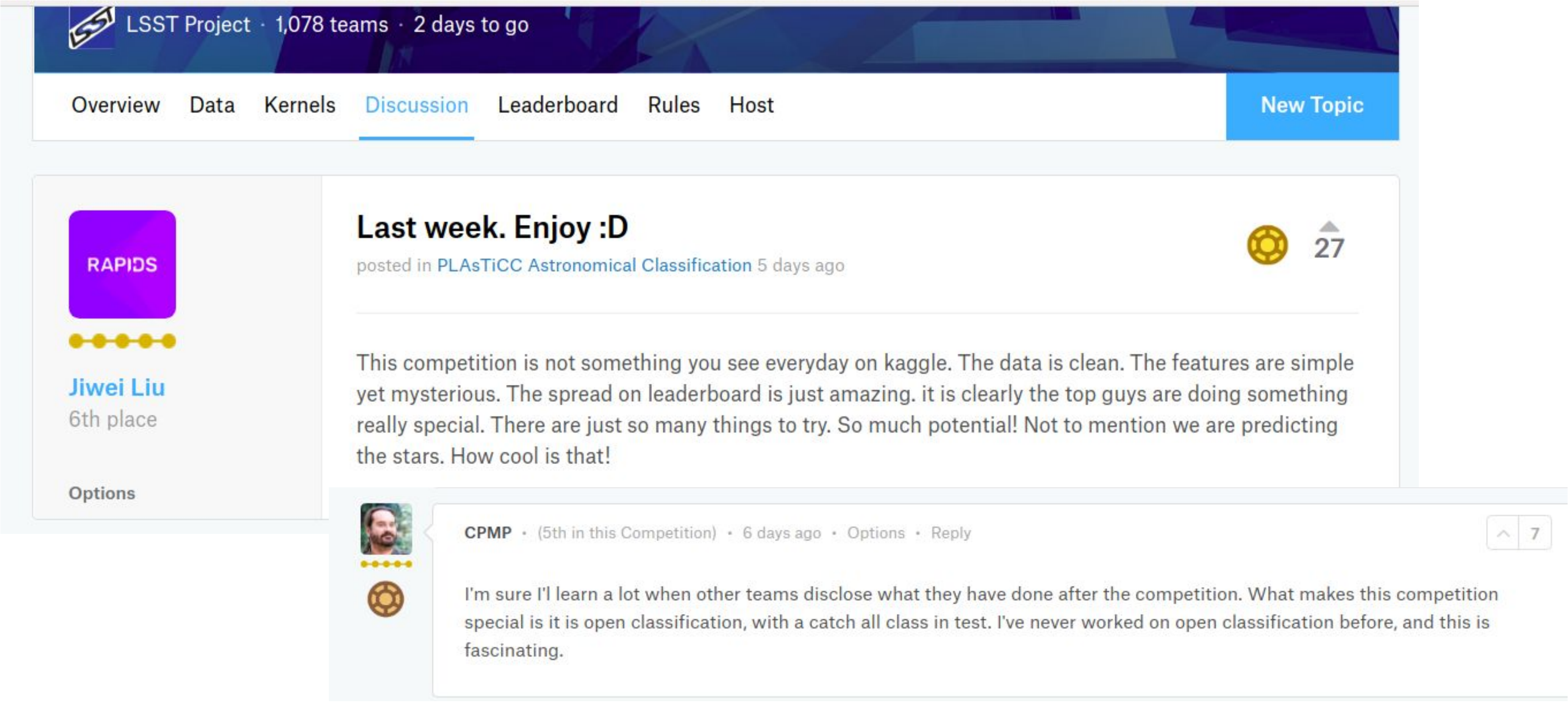
1,093	1,382	22,430
Teams	Competitors	Entries

Help some of the world's leading astronomers grasp the deepest properties of the universe.

The human eye has been the arbiter for the classification of astronomical sources in the night sky for hundreds of years. But a new facility -- the Large Synoptic Survey Telescope (LSST) -- is about to revolutionize the field, discovering 10 to 100 times more astronomical sources that vary in the night sky than we've ever known. Some of these sources will be completely unprecedented!

The Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) asks Kagglers to help prepare to classify the data from this new survey. Competitors will classify astronomical sources that vary with time into different classes, scaling from a small training set to a very large test set of the type

The repercussion



Last week. Enjoy :D

posted in PLAsTiCC Astronomical Classification 5 days ago

Jiwei Liu
6th place

This competition is not something you see everyday on kaggle. The data is clean. The features are simple yet mysterious. The spread on leaderboard is just amazing. it is clearly the top guys are doing something really special. There are just so many things to try. So much potential! Not to mention we are predicting the stars. How cool is that!

CPMP · (5th in this Competition) · 6 days ago · Options · Reply

I'm sure I'l learn a lot when other teams disclose what they have done after the competition. What makes this competition special is it is open classification, with a catch all class in test. I've never worked on open classification before, and this is fascinating.

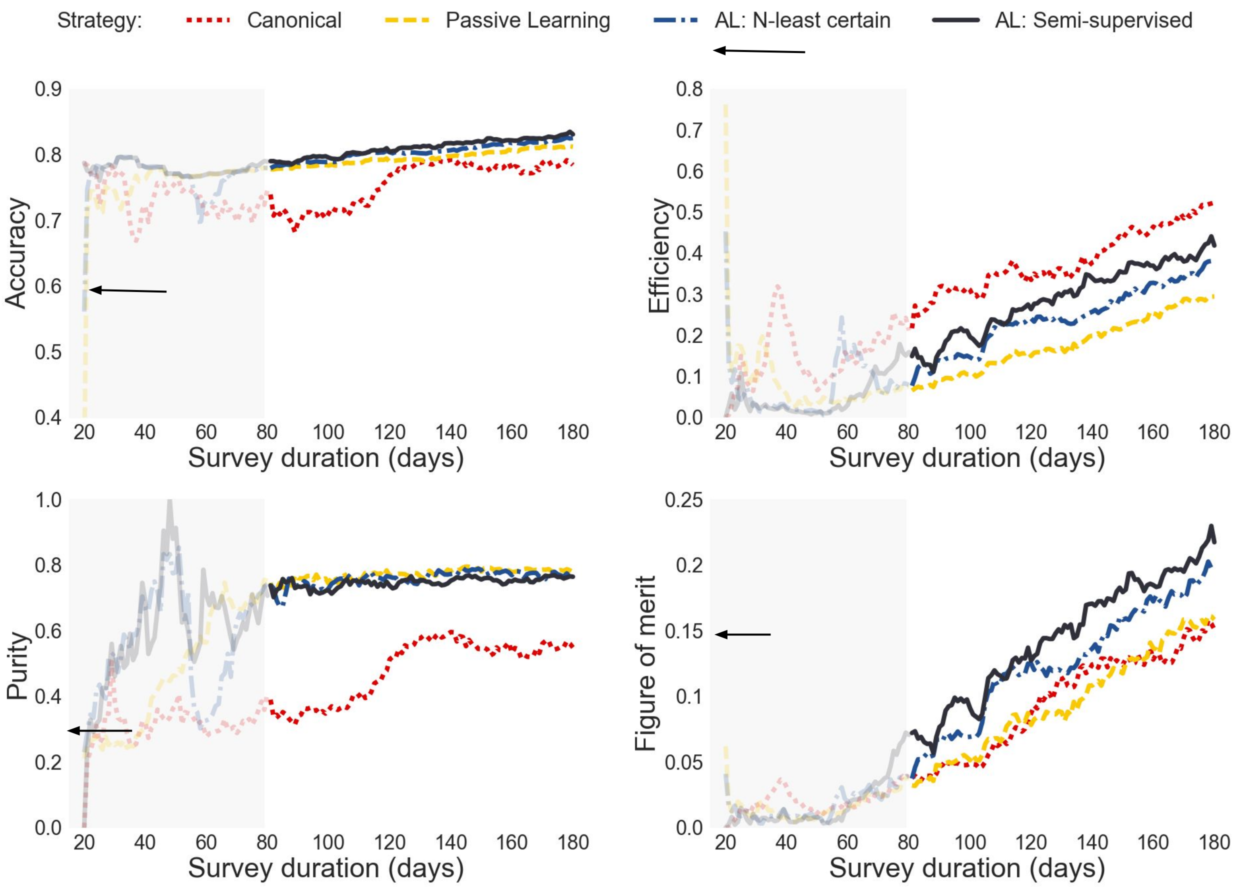
Our data is extremely complex …

...and this is an opportunity...

https://www.kaggle.com/c/PLAsTiCC-2018/discussion/74292

Batch Mode

Partial LC, no initial training, time domain



The arrow shows traditional Full light-curve results with full SNPCC spec

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