COIN

Cosmostatistics Initiative

https://github.com/COINtoolbox

The Cosmostatistics Initiative

MAESTRO, June/2017

Emille E. O. Ishida

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

Long term:

Contribute to the establishment of Astrostatistics and Astroinformatics as full fledged scientific disciplines

Long term:

Contribute to the establishment of Astrostatistics and Astroinformatics as full fledged scientific disciplines **ASAP!**

Long term:

Contribute to the establishment of Astrostatistics and Astroinformatics as full fledged scientific disciplines **ASAP!**

Short term:

Make astronomers, statisticians, computer scientists and data experts understand each other …

WHILE doing science!

COIN's activities *cannot* be merely pedagogical

Long term:

Contribute to the establishment of Astrostatistics and Astroinformatics as full fledged scientific disciplines **ASAP!**

Short term:

Make astronomers, statisticians, computer scientists and data experts understand each other …

WHILE doing science!

Directive:

Significantly contribute to the CV of our members

Long term:

Contribute to the establishment of Astrostatistics and Astroinformatics as full fledged scientific disciplines **ASAP!**

Short term:

Make astronomers, statisticians, computer scientists and data experts understand each other …

WHILE doing science!

Try to remember:
they might work as robots, but they are not!

The COIN Residence Program - CRP

Annual meetings

Conference

Workshop

Hackathon

The COIN Residence Program - CRP

Annual meetings

~~Conference~~

Workshop

Hackathon

The COIN Residence Program - CRP

Annual meetings

~~Conference~~

~~Workshop~~

Hackathon

The COIN Residence Program - CRP

Annual meetings

~~Conference~~

~~Workshop~~

~~Hackathon~~

The COIN Residence Program - CRP

A non-profit start-up?

Annual meetings

John Johnson/HBO

https://www.theroadtosiliconvalley.com/moving/comparing-sydney-silicon-valley/

The COIN Residence Program - CRP

A non-profit start-up?

Annual meetings

John Johnson/HBO

CRP #2, UK, 2015

CRP #3, Budapest, 2016

CRP #2, UK, 2015

https://www.theroadtosiliconvalley.com/moving/comparing-sydney-silicon-valley/

Choosing the participants

The COIN Residence Program

The COIN Residence Program

What we can guarantee up front

Lots of coffee

1 house

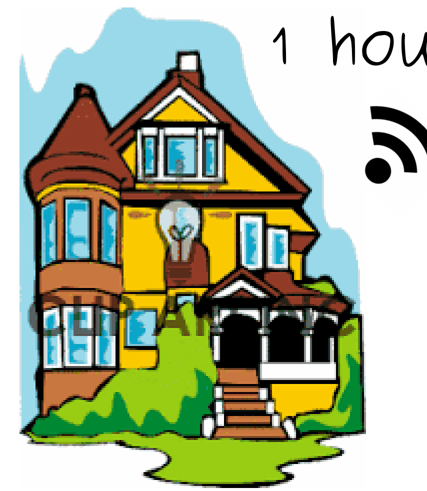
paper

The COIN Residence Program

What we can guarantee up front

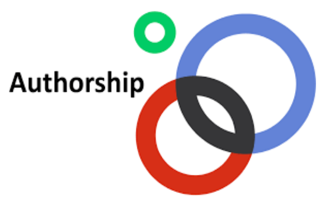
Lots of coffee

1 house



paper

What we can NOT guarantee up front



Talks

The COIN Residence Program

What we can guarantee up front

Lots of coffee

1 house

paper

What we can NOT guarantee up front

Authorship

Talks

What we require from participants

Choosing the projects

Questions posed by the organizers

1. What do you know how to do?
2. What do you like to do?
3. What would you like to learn?

Participants can propose projects.
Everyone votes to which project will be selected

Questions posed by the organizers

1. What do you know how to do?
2. What do you like to do?
3. What would you like to learn?

Participants can propose projects.
Everyone votes to which project will be selected

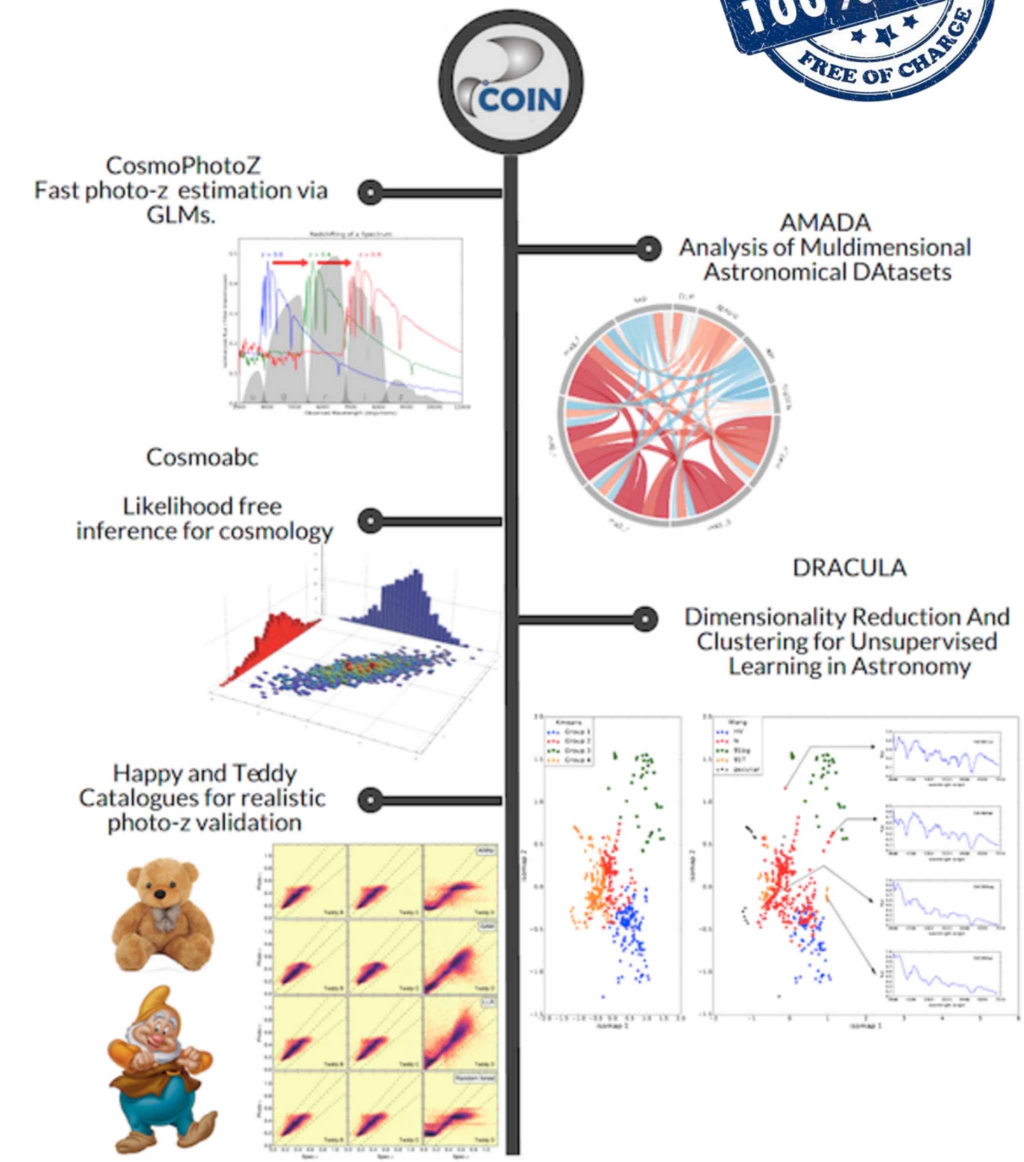
From CRP #2, UK - 2015

1	2	3
Proj. Name: HBM SNe host galaxies Method: Hierarchical Bayesian Models Object: SNe/GRB Data: TBD Manager : TBD Scientific Question: **Can HBM improve our understanding of the bias between type Ia SN hubble residuals and their host galaxy properties?** Participants: Heather, Rafael, Emille, Joseph, Mohammad, Paniez, Miguel, Madhura	Project type: application Method: Spatial Statistics - INLA Object: TBD Data: Spatio-Temporal evolution of chemical abundances in primordial galaxies Scientific Question: **Explore Spatial Statistics (spatstat, INLA) capabilities for astronomical research** Participants: Marina, Mariana, Sandro, Alberto, Rafael, Ewan, Joseph, Miguel, Mohammad, Eric?	Proj. Name: Galaxy AGN connection Method: Bayesian Logistic Regression Objects: Galaxy bars, rings, AGNs ... Data: TBD Manager: TBD Scientific Question: **How do the properties of the galaxy influence its probability to host an AGN?** Participants: Marina, Rafael, Alberto, Joseph, Mohammad, Arlindo, Malu, Paula

4	5	6
Proj. Name: Review on Variable selection Method: different variable selection algorithms Object: Data: Manager: Scientific Question: **What is the state of art methodology to subset the best predictors for multivariate regression in astronomical datasets?** Participants: Alberto, Rafael, Zoe, Luke, Paniez, Miguel, Arlindo, Bruce, Alan, Yabebal, Malu	Proj. Name: MAchine learning in SNe spectra Method: Dim. Reduction + unsupervised learning Object: Sne Data: SN spectral series Manager: Michele Scientific Question: **Use unsupervised/ semi-supervised learning to identify subtypes of SNe** Participants: Michele, Emille, Rafael, Ricardo, Paolo, Fabian, Arlindo, Paniez	

Does it work?

COIN products

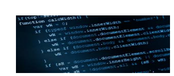


60 researchers from **15** countries

Scientific outcomes

In 3 years

	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



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

+ 1 galaxy catalog
+ 1 GMM tutorial
+ 2 photoz catalogs

COIN products are open source!

Infographic by Rafael S. de Souza

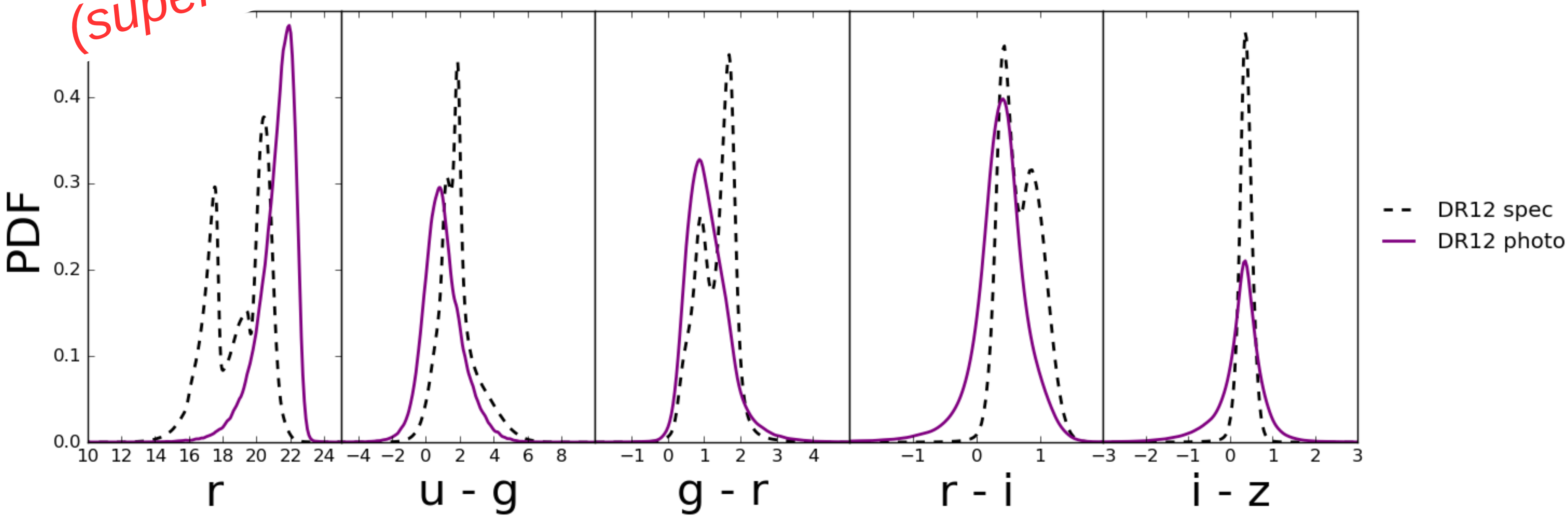
Example from CRP #4, Budapest 2016

The problem with text-book ML:

Representativeness

For photo-z
(supervised regression)

Features (columns)
distributions

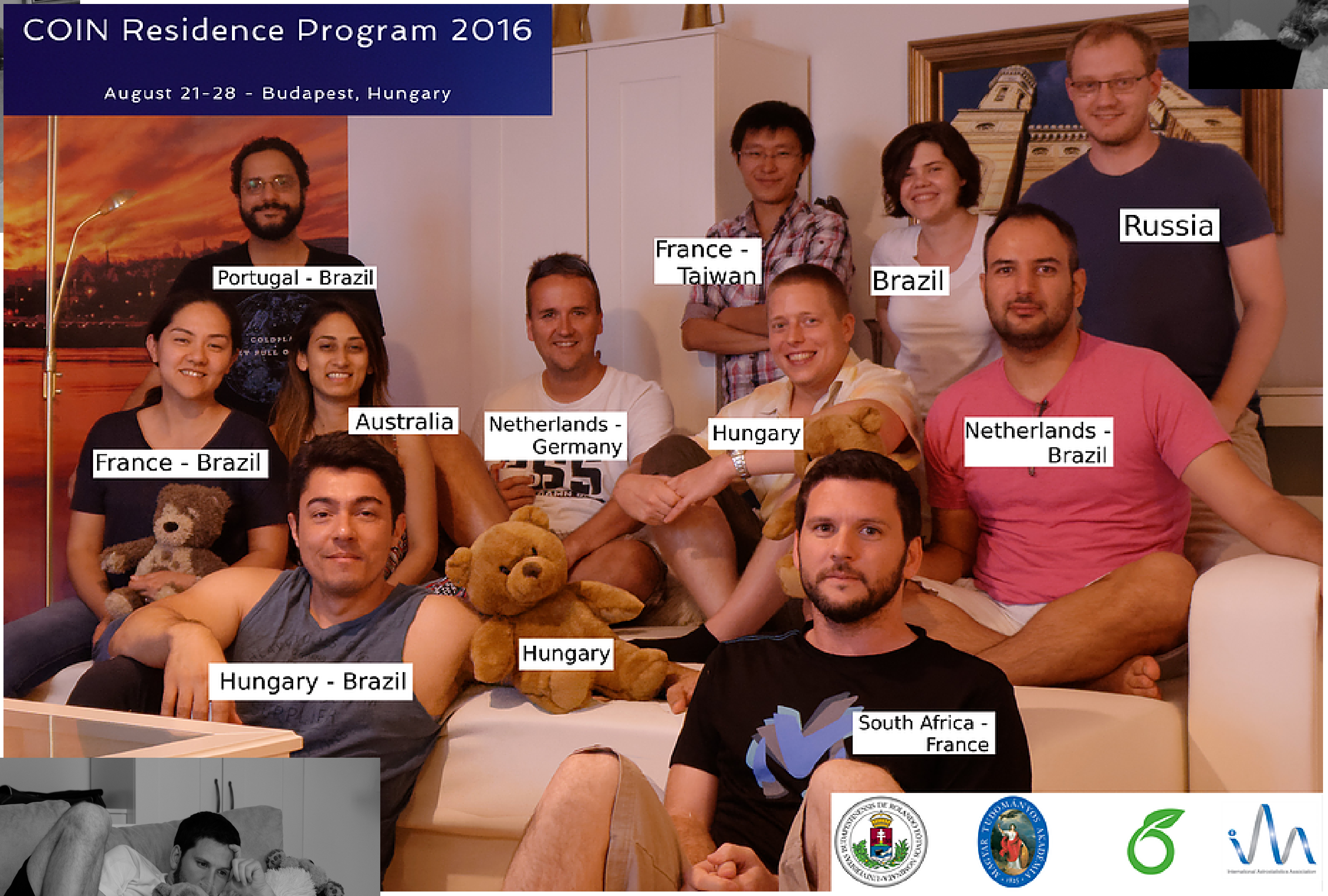


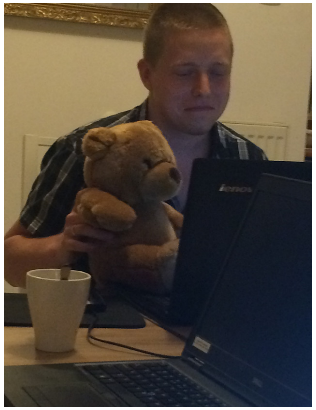
Spec → training/validation
Photo → test

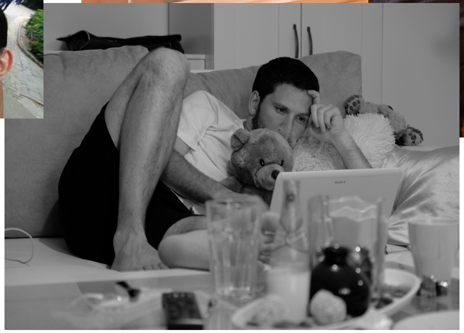
What if you wish to try a method which address the realistic situation?

COIN Residence Program #3

http://iaacoin.wixsite.com/crp2016





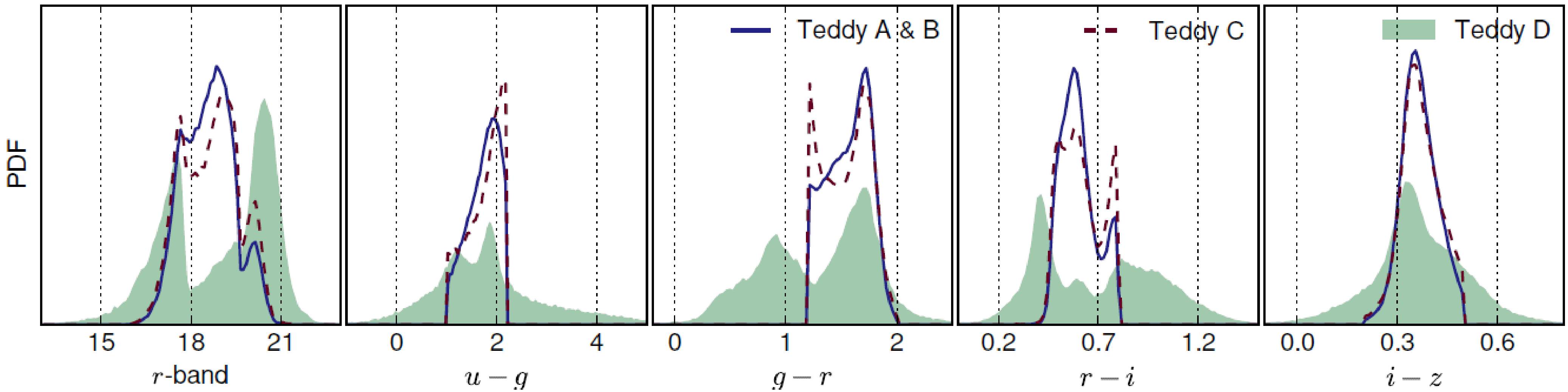


COIN
Cosmostatistics Initiative

Teddy catalogue

The effect of color coverage

A /B follow SDSS spec distribution
B is completely representative of A
C has the same coverage but slightly different shape
D has a wider domain in r-mag and color (no coverage)



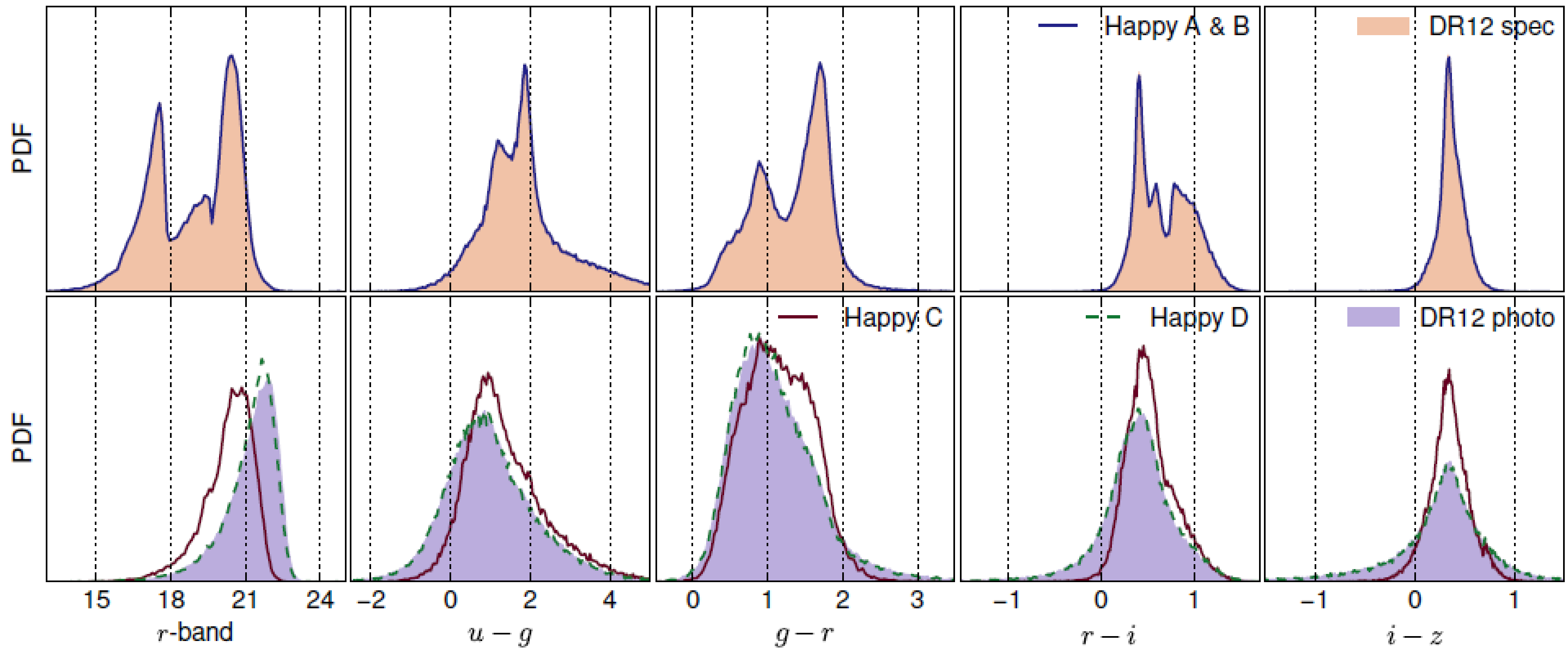
Beck, Lin, ***Ishida*** *et al., astro-ph:1701.08748, MNRAS (2017), 468, issue 4, pp. 4323-4339 – from CRP #3*

Happy catalogue

The effect of coverage + photometric errors

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

- A /B follow SDSS spec distribution
- B is completely representative of A
- C was constructed performing a nearest neighbor between the SDSS-DR12 photo sample and the extended spec sample but with a cut on photometric errors
- D is the same of C but without the photometric error cut.
- Consequently, D follows exactly the SDSS-DR12 photo sample distribution

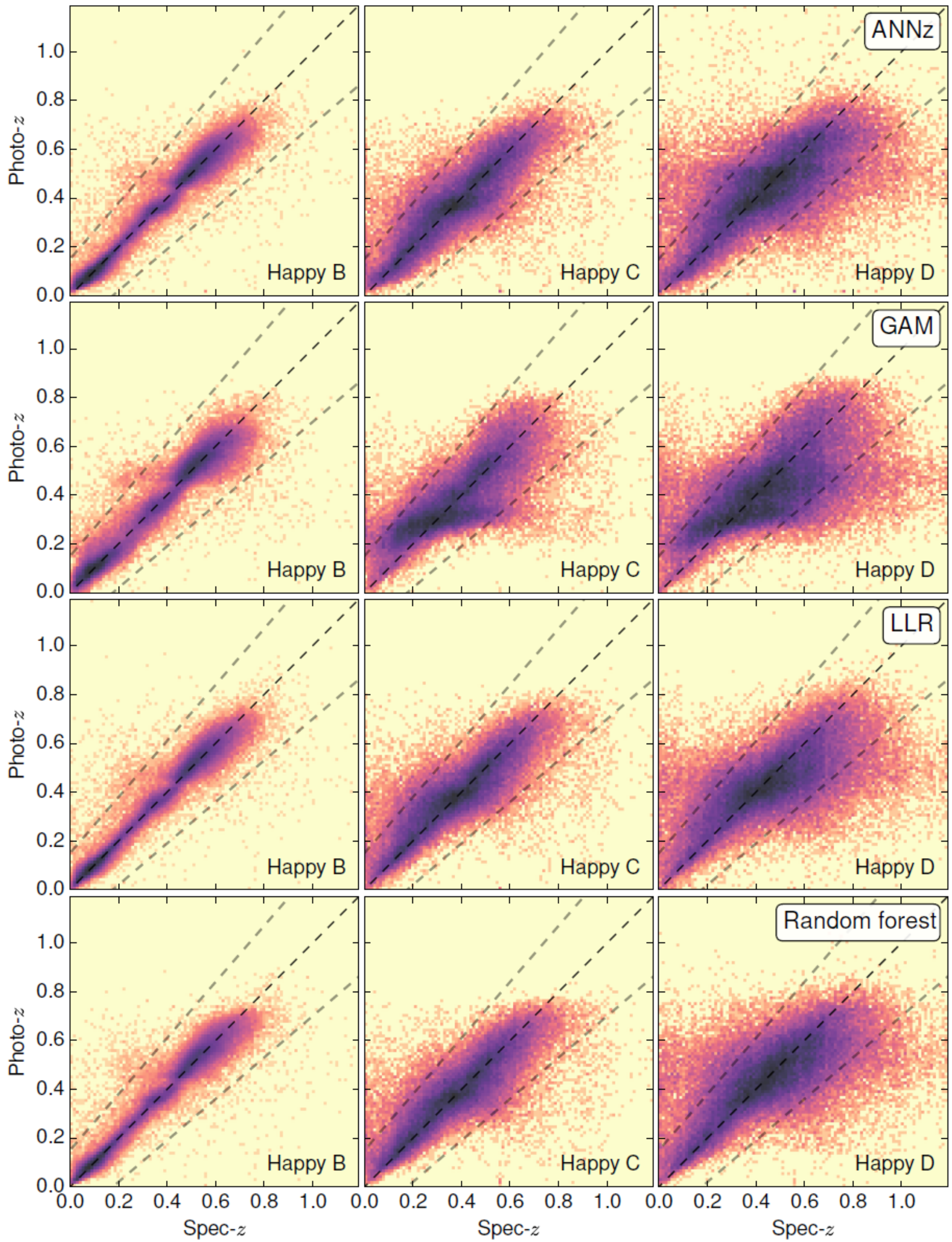


Happy

*Beck, Lin, **Ishida** et al., astro-ph:1701.08748, MNRAS (2017), 468, issue 4, pp. 4323-4339 – from CRP #3*

Happy catalogue

The effect of coverage + photometric errors



Empirical methods

Method	Set	Diagnostics Mean ($\times10^{-2}$)	Std ($\times10^{-2}$)	MAD ($\times10^{-2}$)	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

https://github.com/COINtoolbox/photoz_catalogues

README.md	Readme minor	5 months ago

README.md

Teddy and *Happy* photo-z catalogues

This repository contains the photometric redshift catalogues presented in Beck et al., 2017 - *On the realistic validation of photometric redshifts, or why Teddy will never be Happy*.

This is one of the products of the third edition of the COIN Residence Program, which took place in August/2016 in Budapest (Hungary).

Any questions/suggestions should be sent to iaa.coin@gmail.com .

A general overview of both catalogues is given bellow. Check the individual folders for detailed information on the files presented here.

Teddy

This catalogue was designed to isolate the effect of limited spectroscopic sample coverage in colour/magnitude space.

It is constructed from the SDSS DR12 spectroscopic sample and is maintained by Chieh-An Lin (CEA, France)

Happy

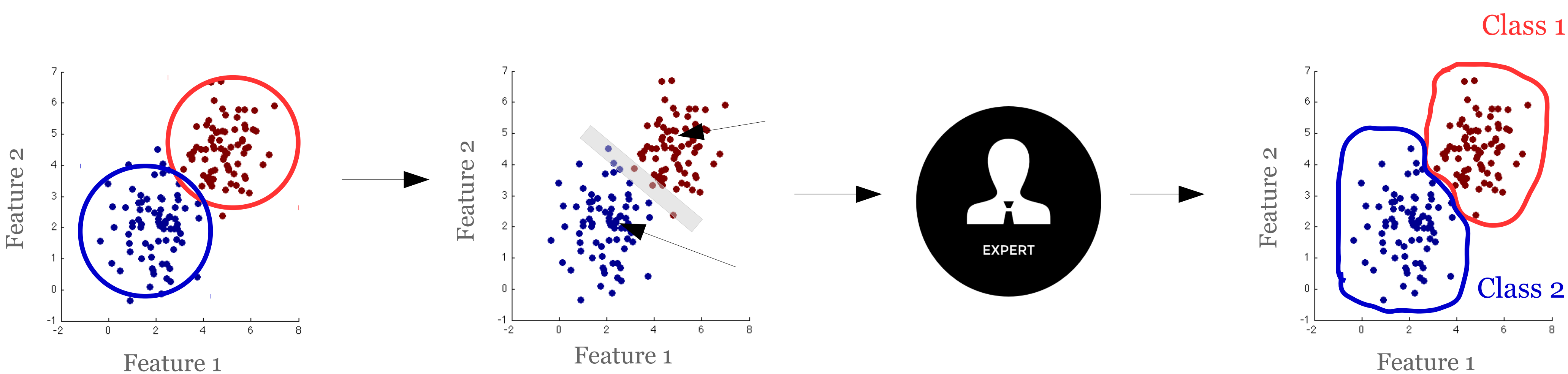
This catalogue was designed to reproduce the effect of distinct photometric error distributions and their convolution with colour/magnitude space coverage between the spectroscopic and photometric samples.

All photometry was taken from SDSS DR12, and spectroscopy was gathered from a set of different sources (see Beck et al., 2017 for further details).

Happy is maintained by Robert Beck (ELTE, Hungary).

Potential solution: Active Learning

SAMSI & COIN, in prep



Step 1

Standard Classification

Step 2

Identify problematic as well as representative elements

Step 3

Make a query

Step 4

Classify

The future of COIN

https://iaacoin.wixsite.com/crp2017

Find people ...

Find people ...

Solve bureaucracy

Find people ...

Solve bureaucracy

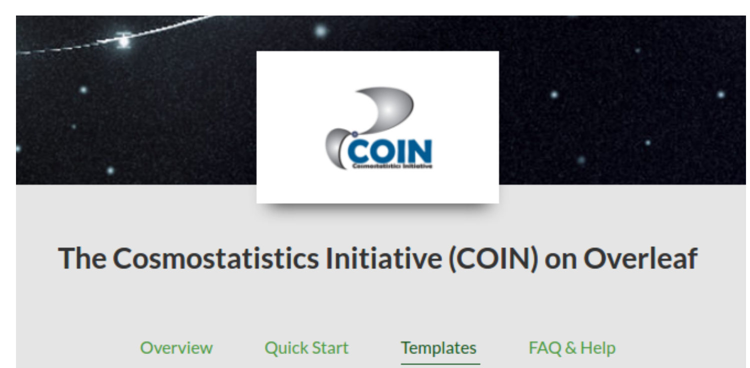
Find people ...

Solve bureaucracy

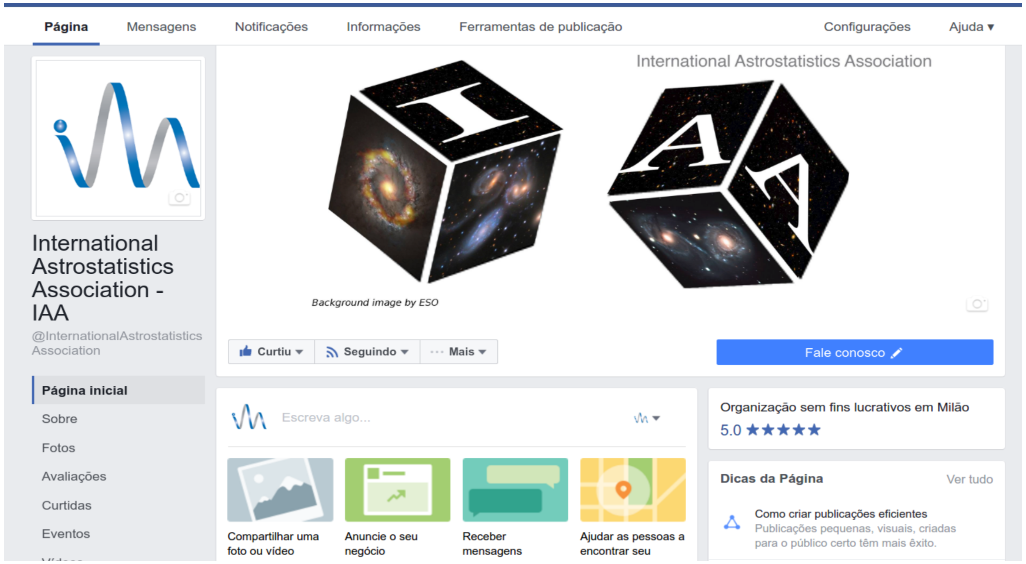
Find people ...

Solve bureaucracy

Let it go ...



IAA facebook page



COIN on twitter

Registrations will open July 7th!

Do we need to re-think the academic model?


Thank you

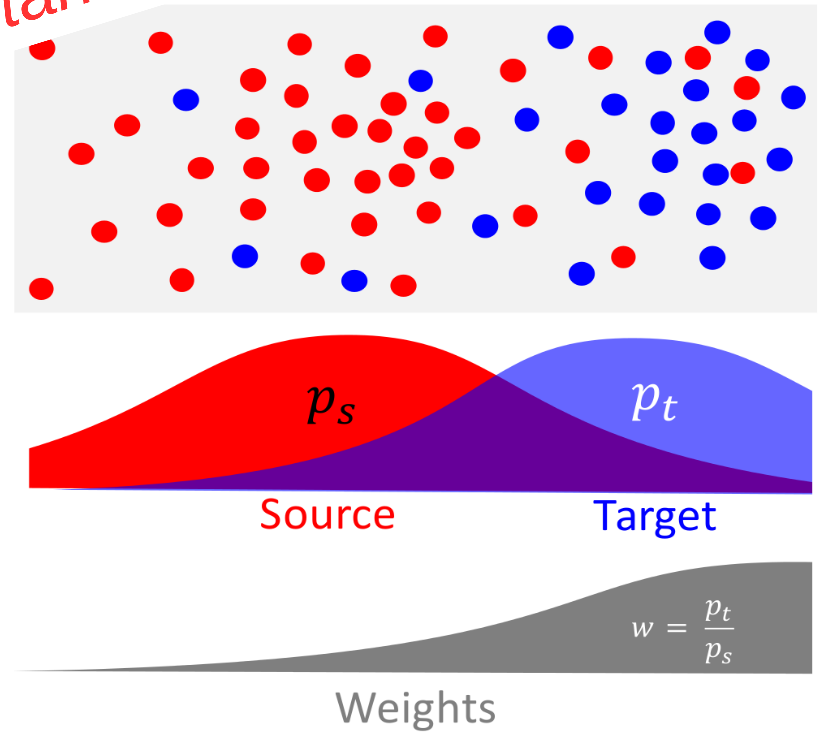
COIN

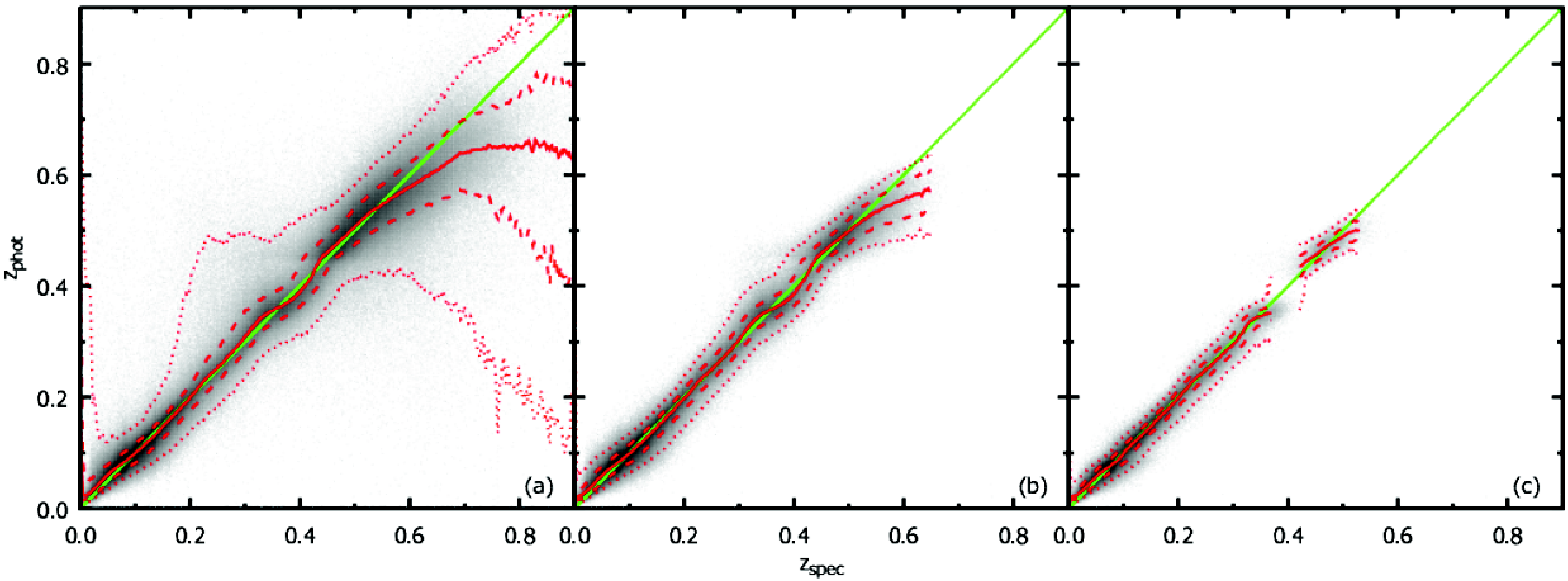
Cosmostatistics Initiative

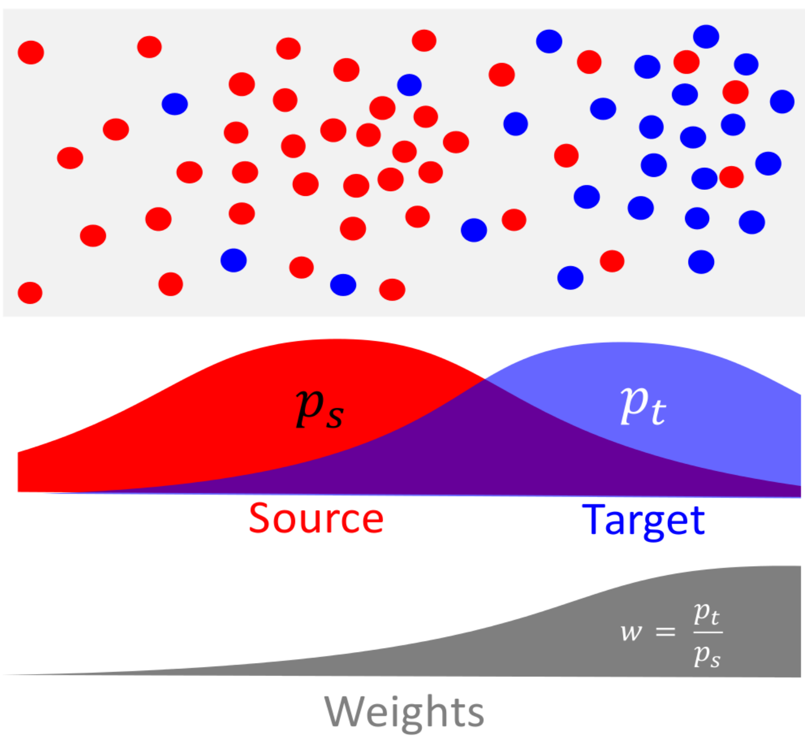
https://github.com/COINtoolbox

Attempted solutions: Domain Adaptation

instance weighting







Photometric redshifts for the SDSS Data Release 12

Róbert Beck, László Dobos, Tamás Budavári, Alexander S. Szalay, István Csabai

Mon Not R Astron Soc (2016) 460 (2): 1371-1381.

Attempted solution: Domain Adaptation

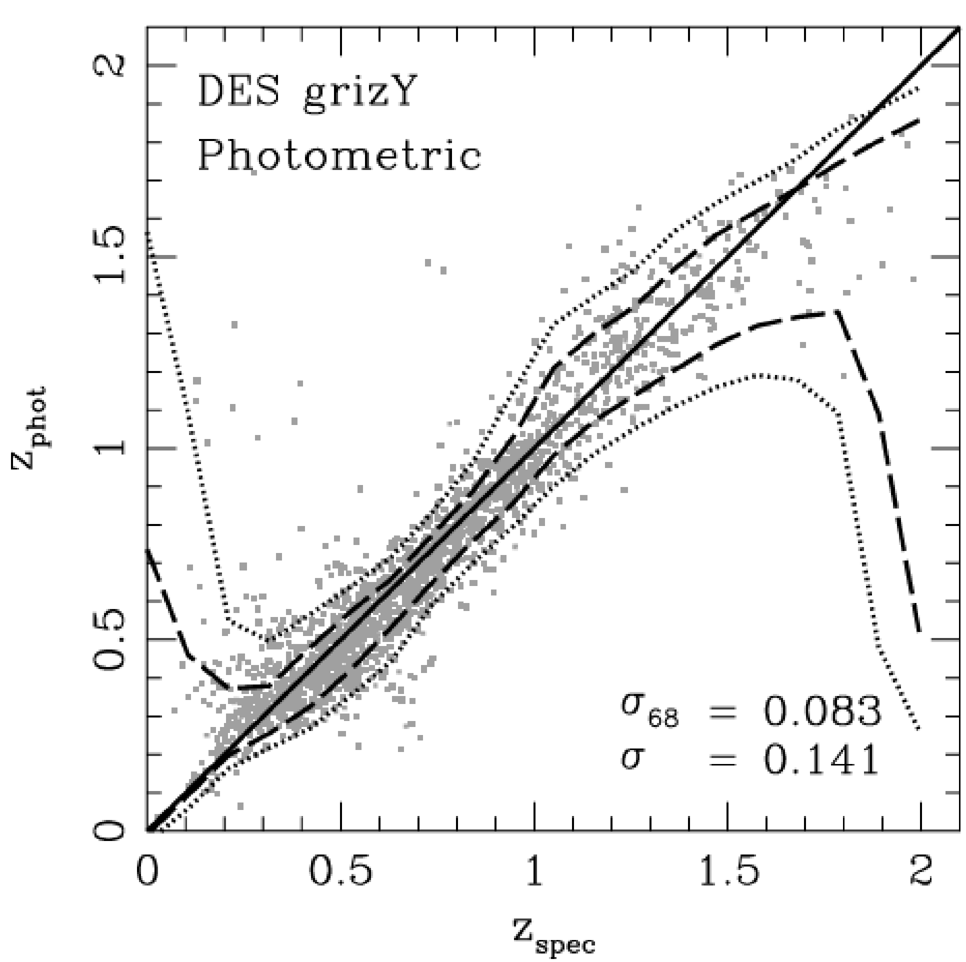
supervised regression

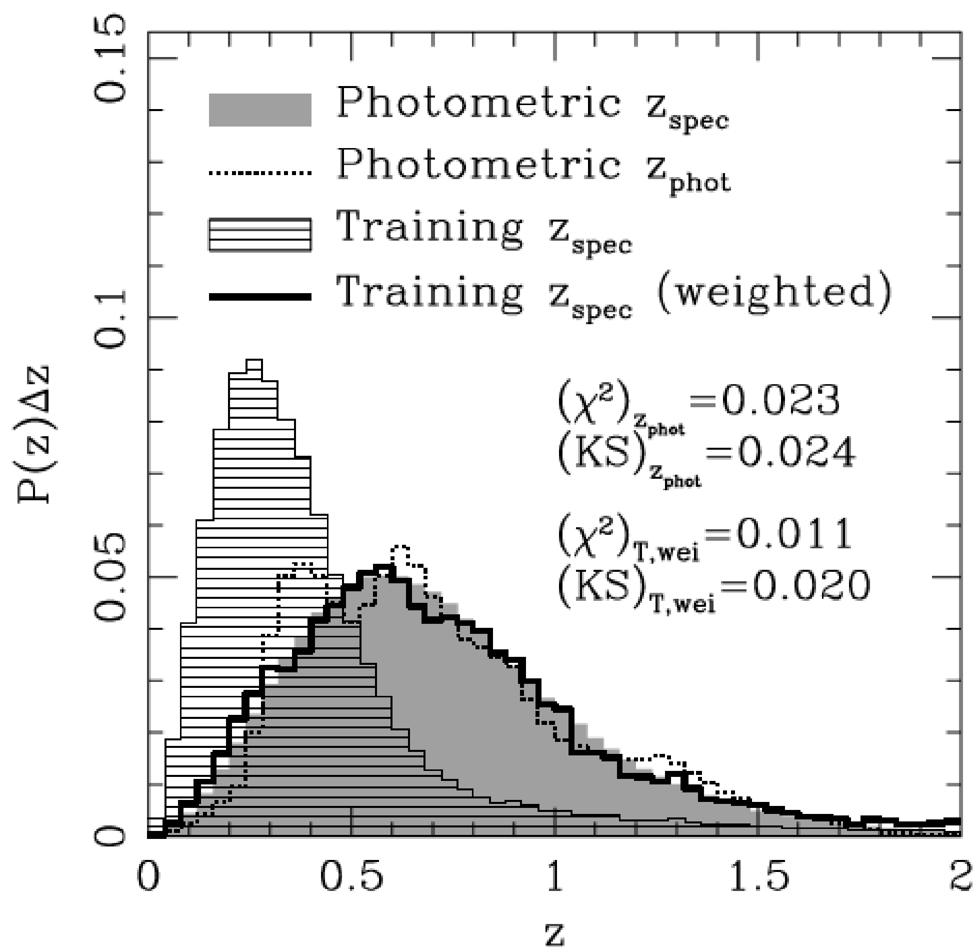
Estimating the redshift distribution of photometric galaxy samples

Marcos Lima,[1,2★] Carlos E. Cunha,[2,3] Hiroaki Oyaizu,[2,3] Joshua Frieman,[2,3,4] Huan Lin[4] and Erin S. Sheldon[5]

Mon. Not. R. Astron. Soc. **390,** 118–130 (2008)

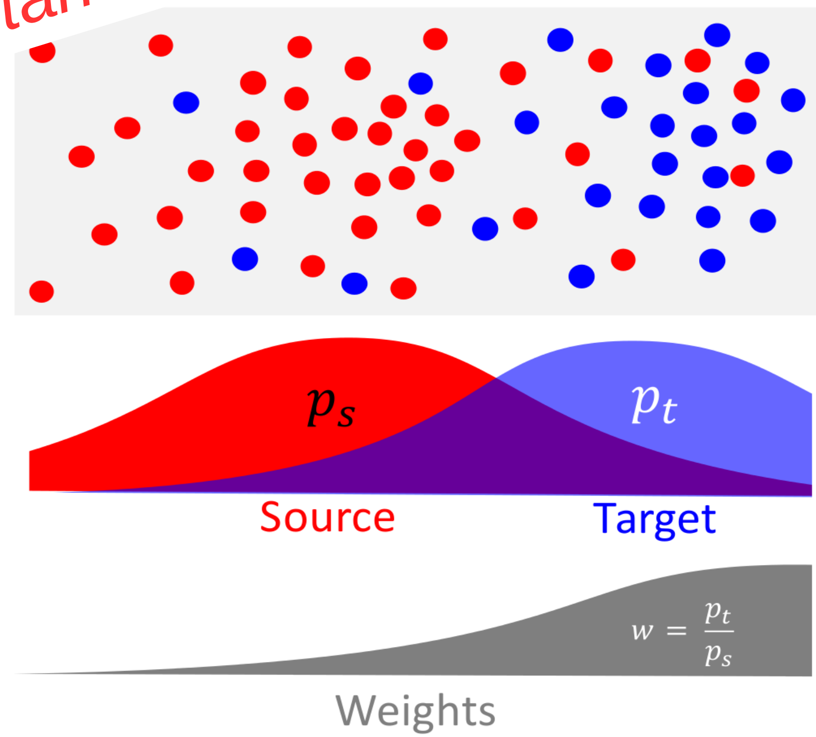
instance weighting





Attempted solutions: Domain Adaptation

instance weighting



supervised regression

Estimating the redshift distribution of photometric galaxy samples

Marcos Lima,[1,2★] Carlos E. Cunha,[2,3] Hiroaki Oyaizu,[2,3] Joshua Frieman,[2,3,4] Huan Lin[4] and Erin S. Sheldon[5]

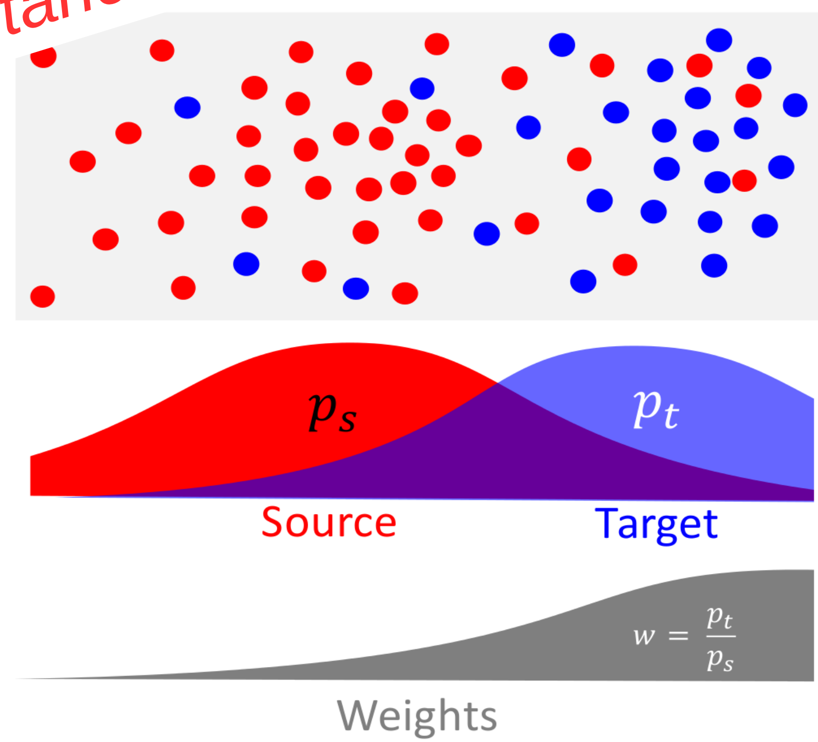
Mon. Not. R. Astron. Soc. **390,** 118–130 (2008)

Works well when data is not sparse, and there is coverage!

Attempted solutions: Domain Adaptation

supervised regression

instance weighting



Estimating the redshift distribution of photometric galaxy samples

Marcos Lima,[1,2★] Carlos E. Cunha,[2,3] Hiroaki Oyaizu,[2,3] Joshua Frieman,[2,3,4] Huan Lin[4] and Erin S. Sheldon[5]

Mon. Not. R. Astron. Soc. **390,** 118–130 (2008)

Works well when data is not sparse, and there is coverage!

Photometric redshifts for the SDSS Data Release 12

Róbert Beck, László Dobos, Tamás Budavári, Alexander S. Szalay, István Csabai

Mon Not R Astron Soc (2016) 460 (2): 1371-1381.

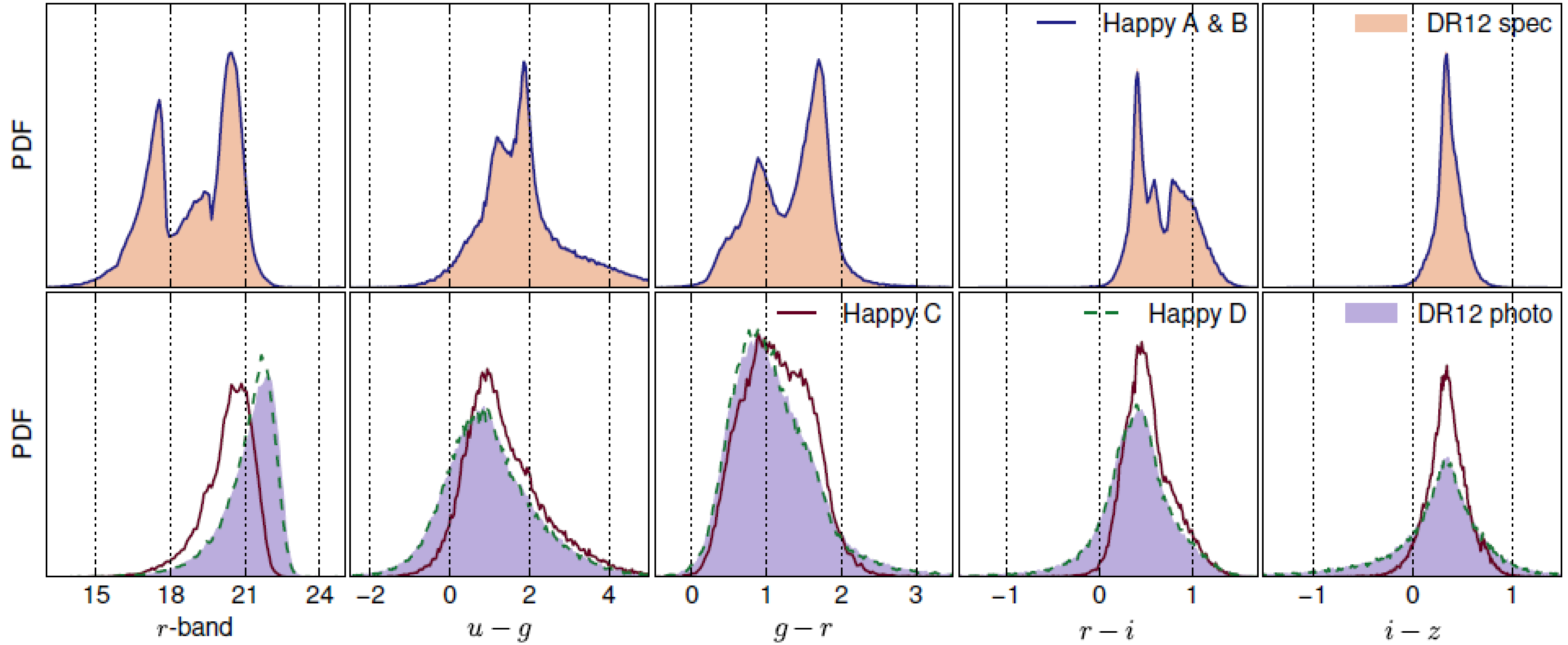
If there is no coverage, identify problematic areas and discard!

Happy catalogue

The effect of coverage + photometric errors

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

- A /B follow SDSS spec distribution
- B is completely representative of A
- C was constructed performing a nearest neighbor between the SDSS-DR12 photo sample and the extended spec sample but with a cut on photometric errors
- D is the same of C but without the photometric error cut.
- Consequently, D follows exactly the SDSS-DR12 photo sample distribution



*Beck, Lin, **Ishida** et al., astro-ph:1701.08748, MNRAS (2017), 468, issue 4, pp. 4323-4339 – from CRP #3*

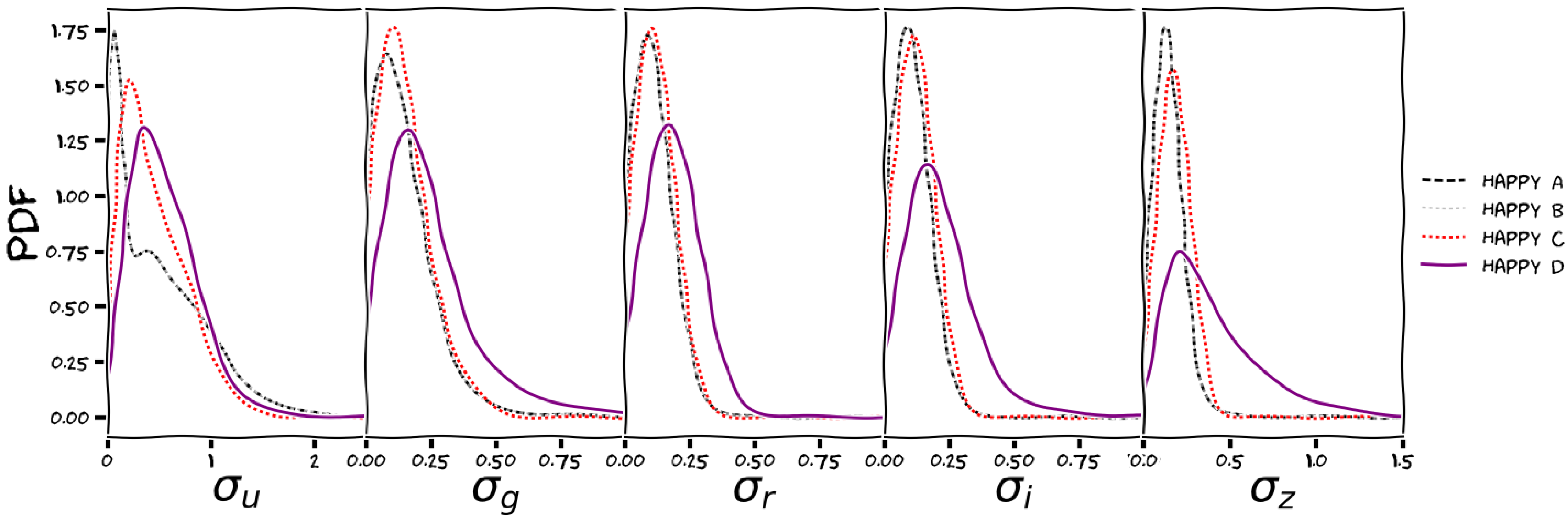
Happy catalogue

The effect of coverage + photometric errors

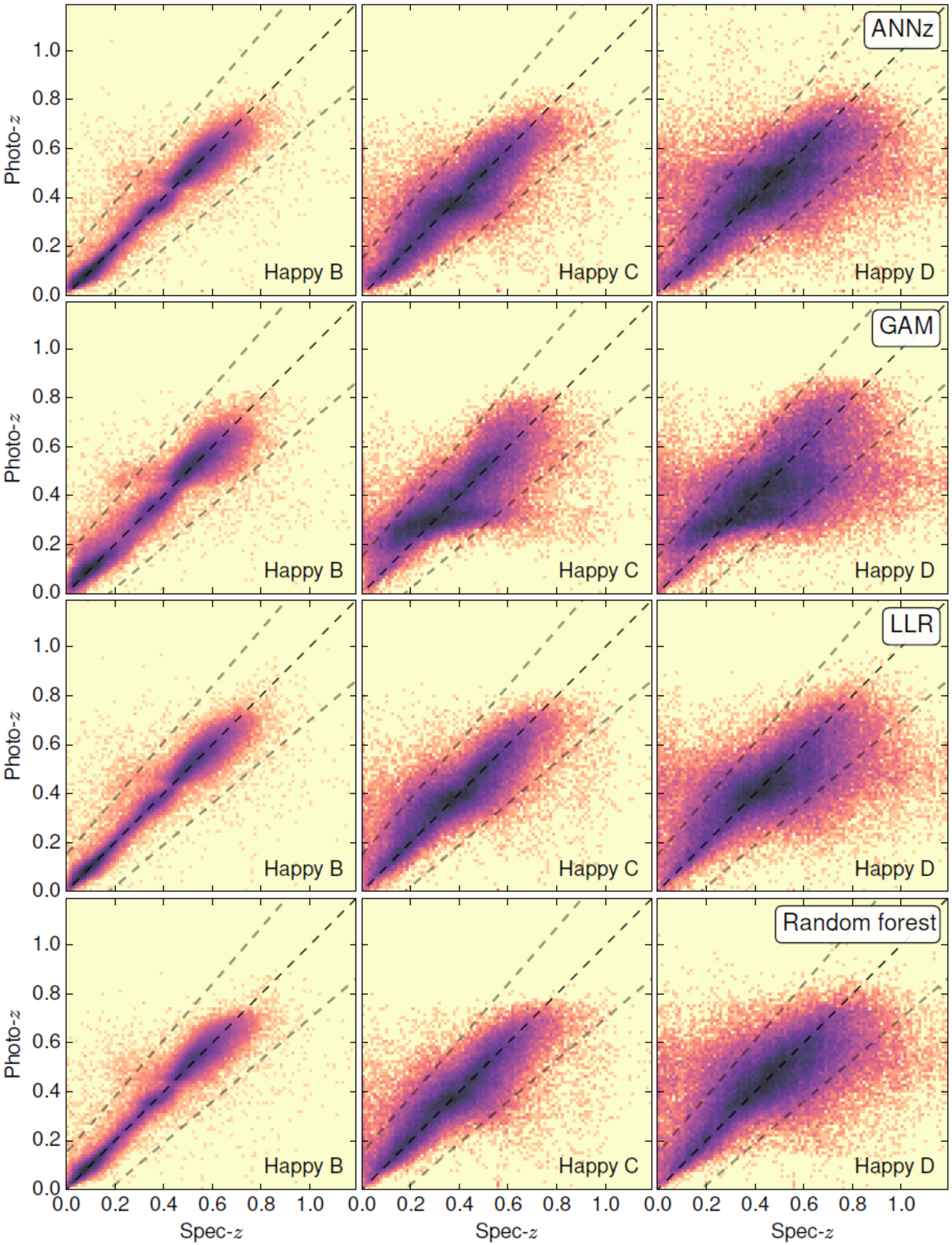
Photometry from SDSS
Spec-z from many different surveys leads to larger photometric errors and consequently wide domain in r-band and color

- A /B follow SDSS spec distribution
- B is completely representative of A
- C was constructed performing a nearest neighbor between the SDSS-DR12 photo sample and the extended spec sample but with a cut on photometric errors
- D is the same of C but without the photometric error cut.
- Consequently, D follows exactly the SDSS-DR12 photo sample distribution

Error distributions correlate with features



Beck, Lin, ***Ishida*** *et al., astro-ph:1701.08748, MNRAS (2017), 468, issue 4, pp. 4323-4339, from CRP #3*



Happy catalogue

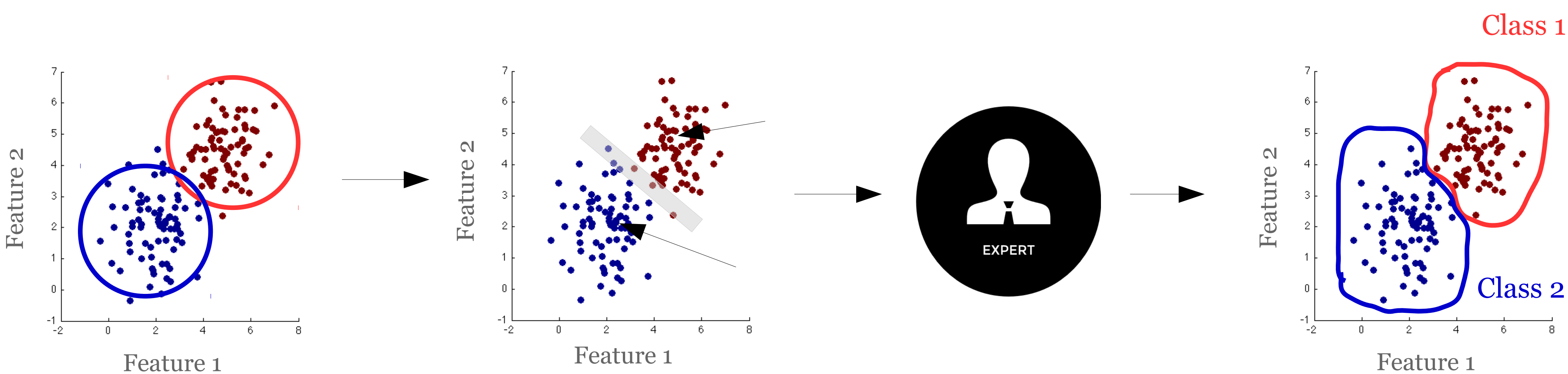
The effect of coverage + photometric errors

Method	Set	Diagnostics Mean ($\times10^{-2}$)	Std ($\times10^{-2}$)	MAD ($\times10^{-2}$)	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, Lin, **Ishida** et al., astro-ph:1701.08748, MNRAS (2017), 468, issue 4, pp. 4323-4339 – from CRP #3*

Potential solution: Active Learning

SAMSI & COIN, in prep



Step 1

Standard Classification

Step 2

Identify problematic as well as representative elements

Step 3

Make a query

Step 4

Classify

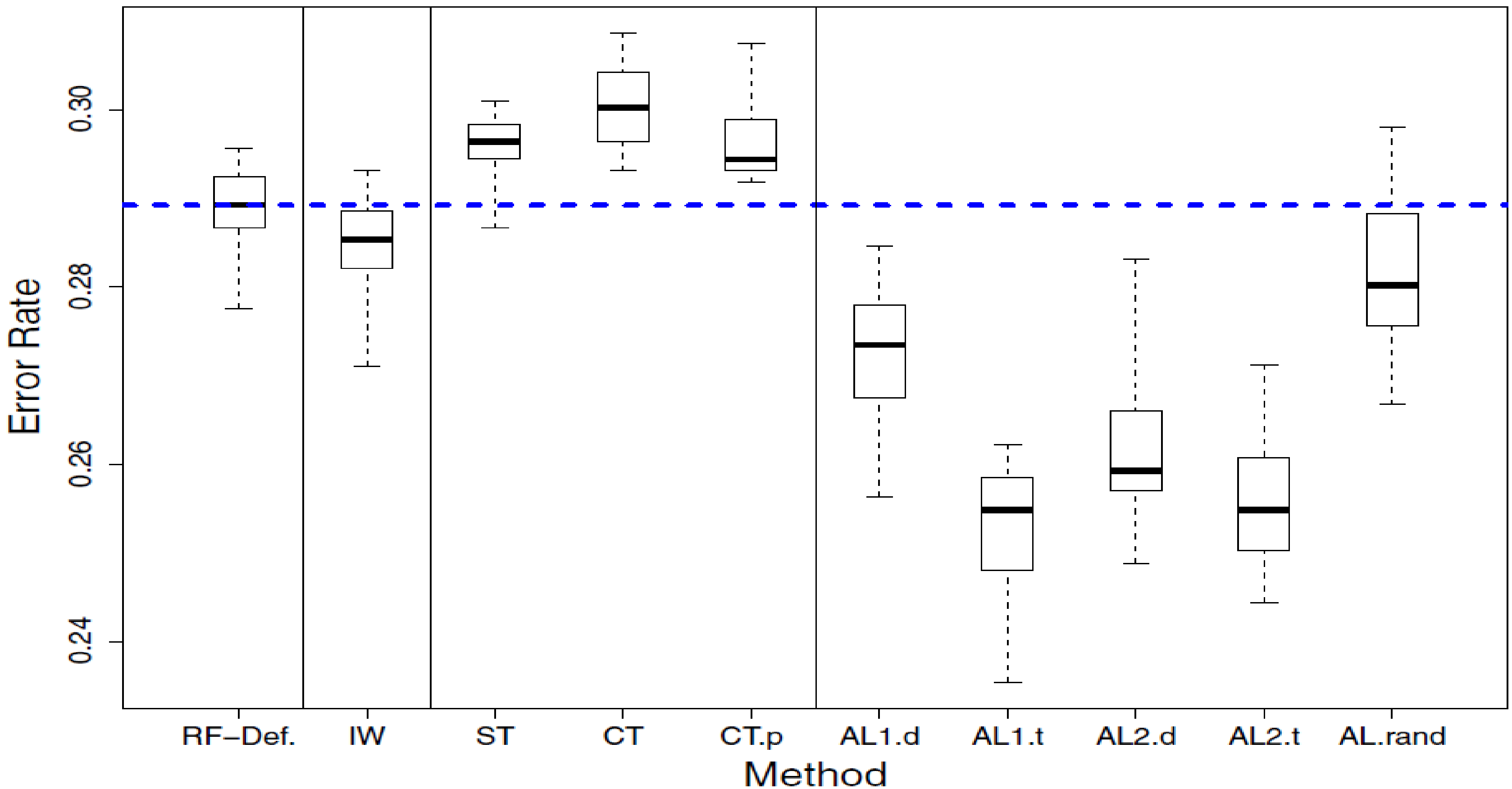
Background: **Active Learning in Astronomy**

ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION

Joseph W. Richards[1,2], Dan L. Starr[1], Henrik Brink[3], Adam A. Miller[1], Joshua S. Bloom[1], Nathaniel R. Butler[1], J. Berian James[1,3], James P. Long[2], and John Rice[2]

supervised classification

The Astrophysical Journal, 744:192 (19pp), 2012 January 10



Background: Active Learning in Astronomy

Automated Supernova Ia Classification Using Adaptive Learning Techniques

Kinjal Dhar Gupta*, Renuka Pampana*, Ricardo Vilalta*, Emille E. O. Ishida†, Rafael S. de Souza‡

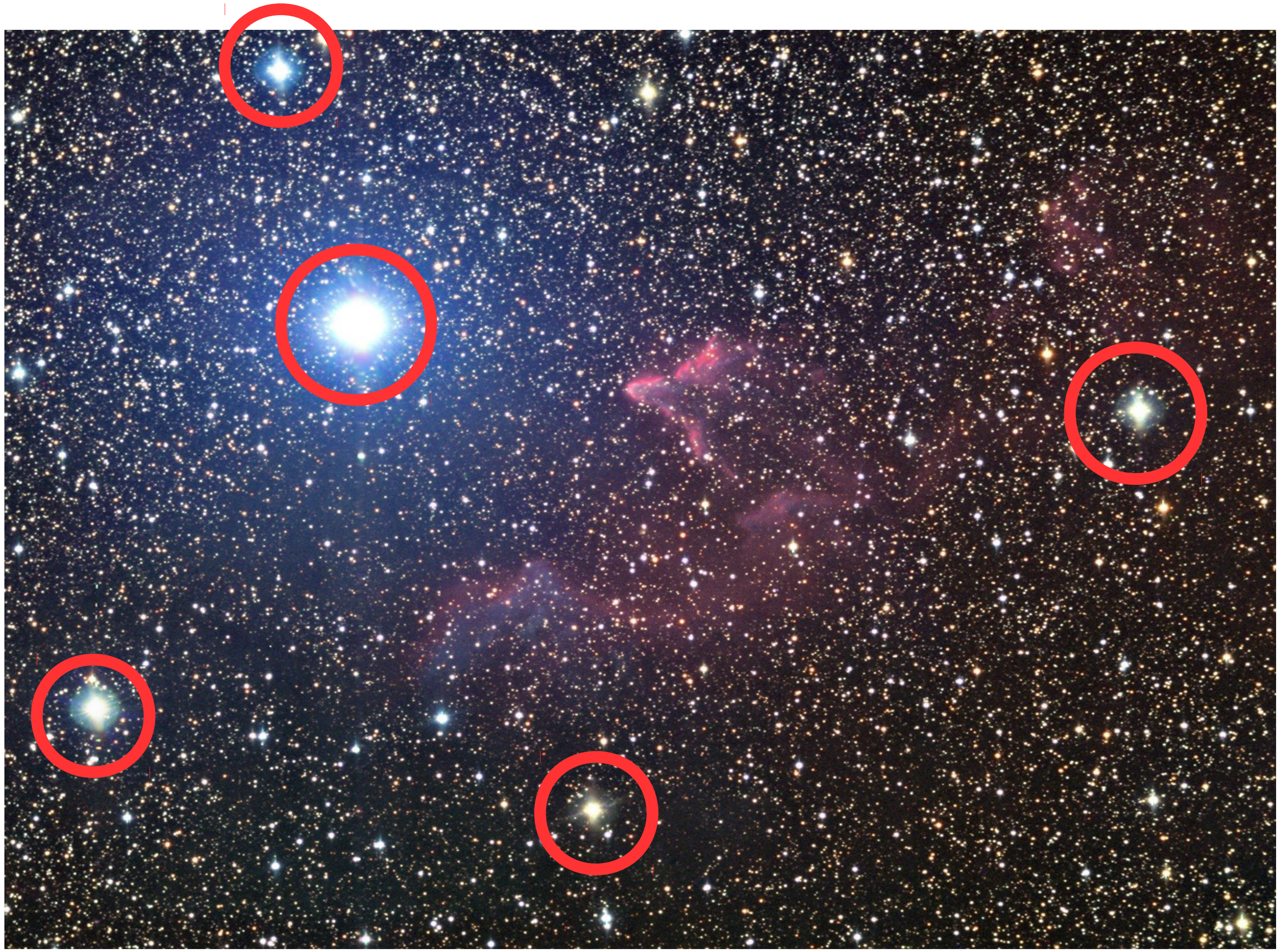
supervised classification

*Dhar Gupta et al. (**incl. Ishida**), 2016 IEEE Symposium in Computational Intelligence, Athens*

How are spectroscopic sets constructed?

How are spectroscopic sets constructed?

Take spectra for learning and determine everything else

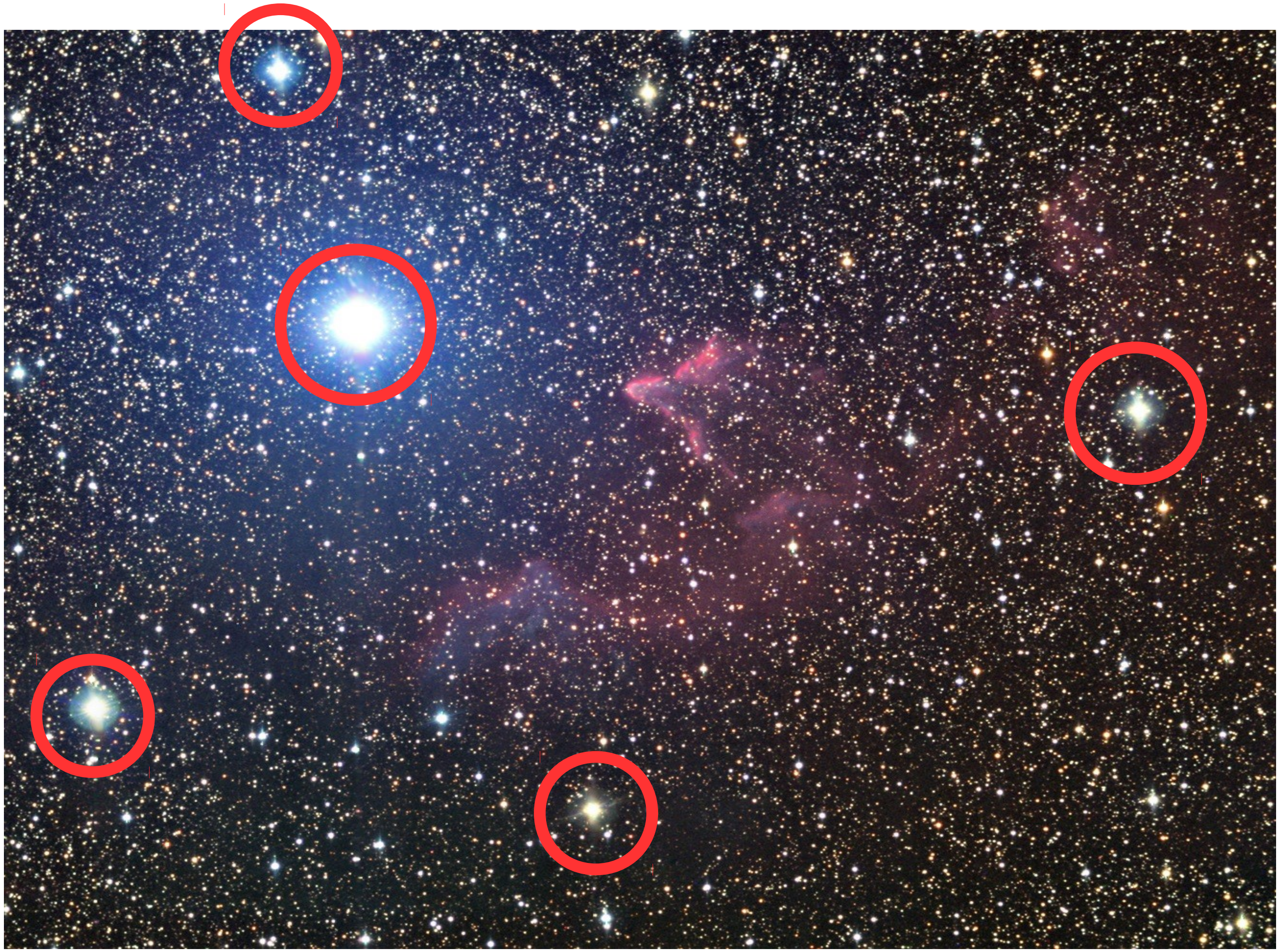


Alternative approach

Landmark selection

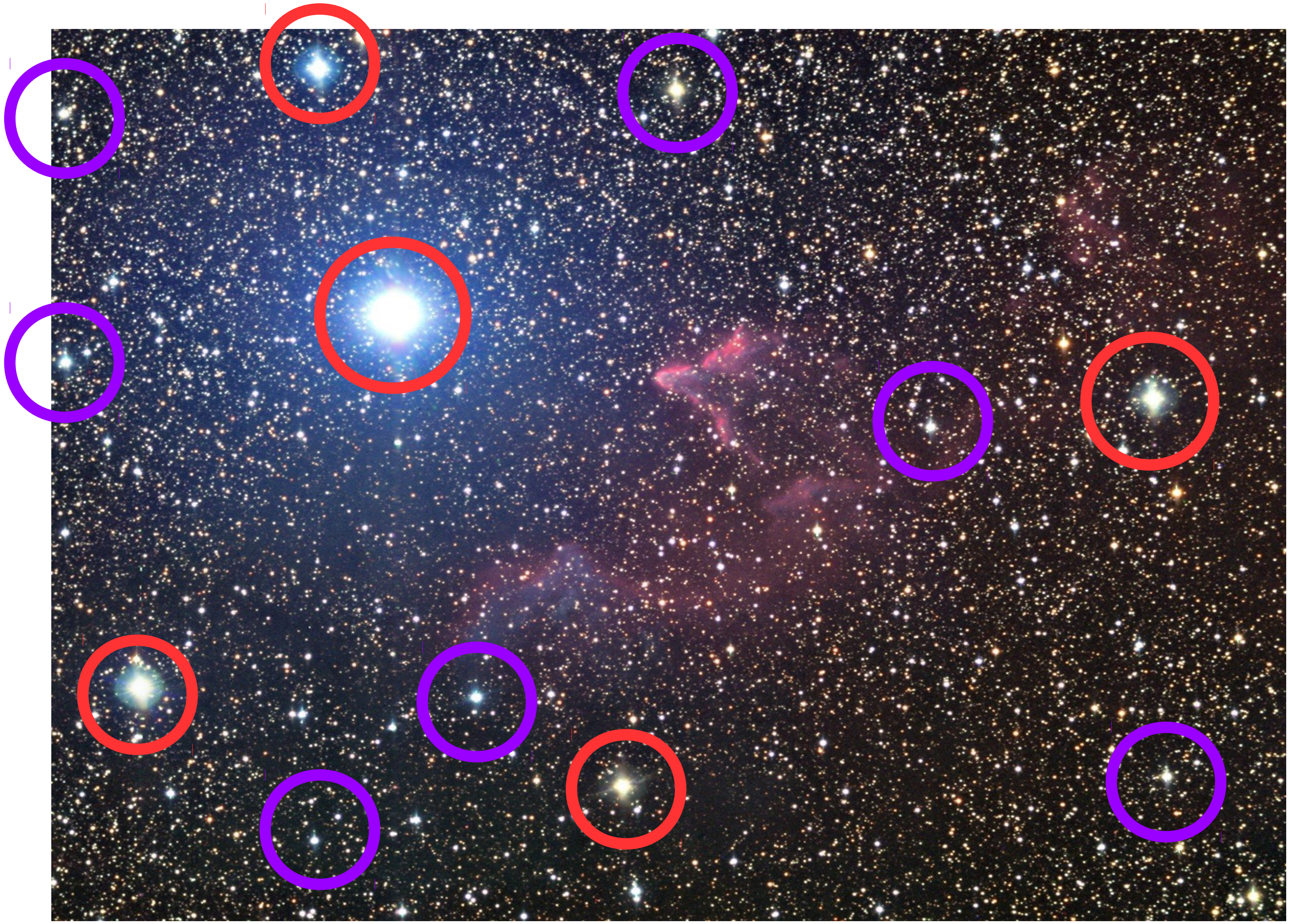
Alternative approach

Landmark selection



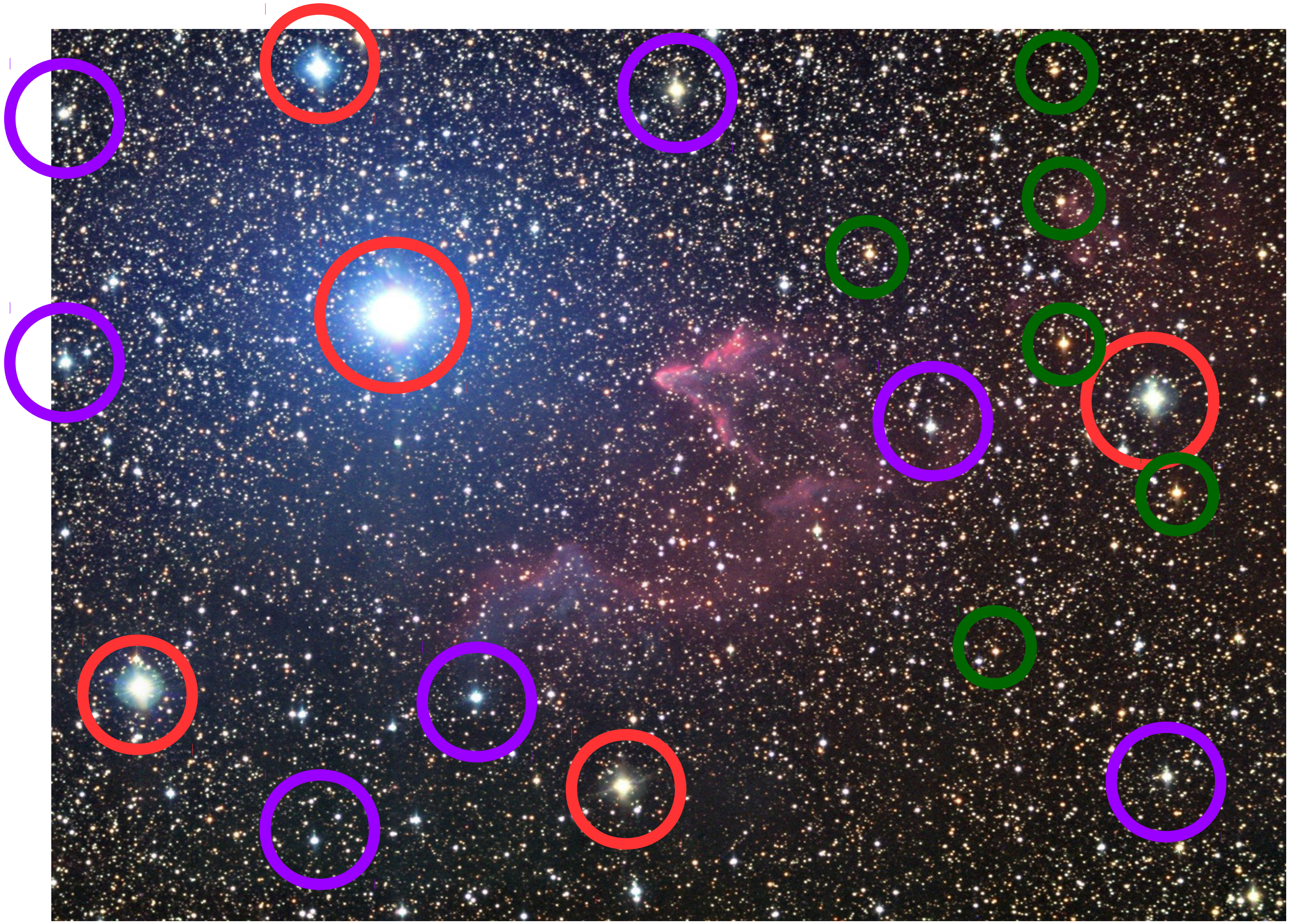
Alternative approach

Landmark selection



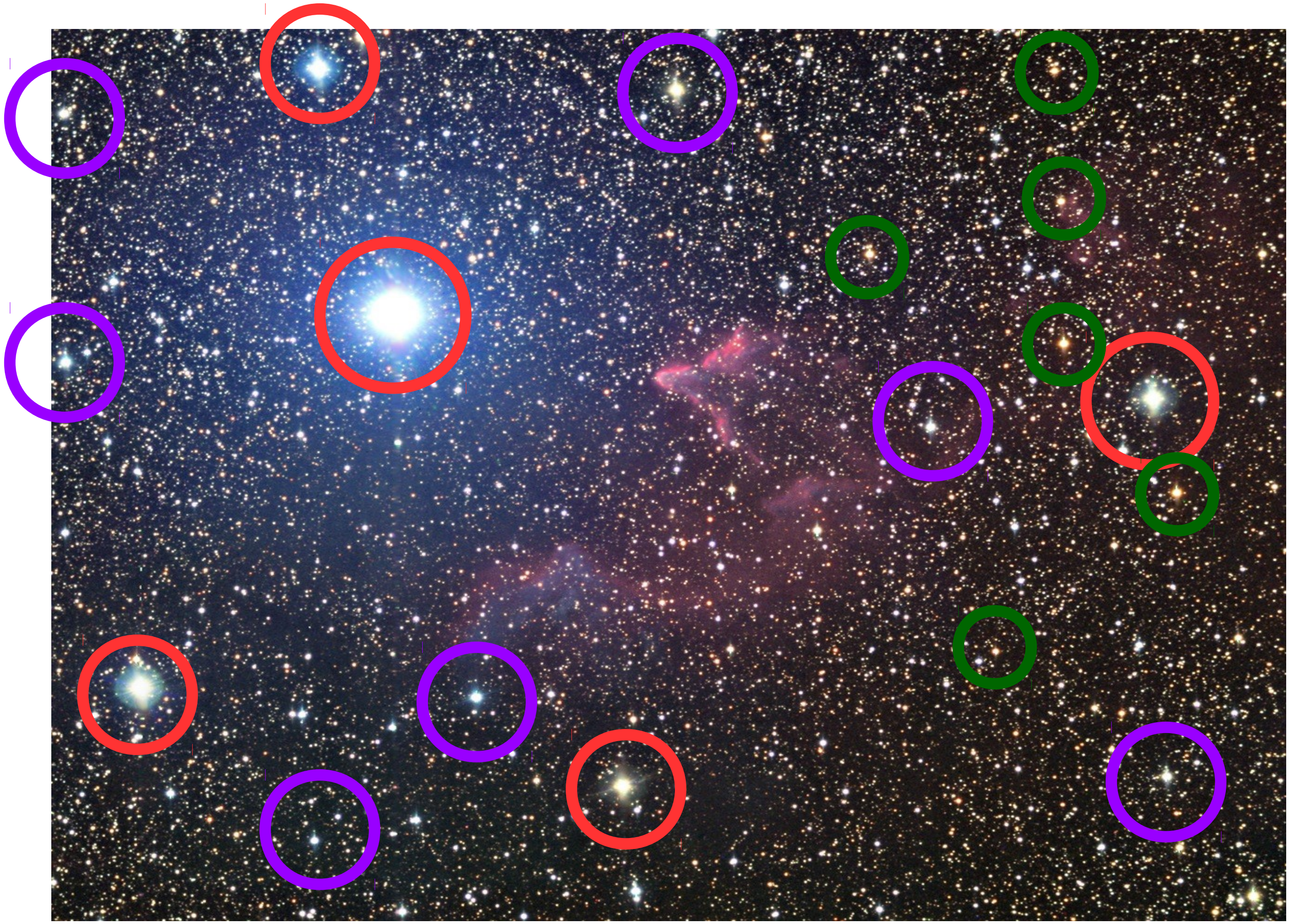
Alternative approach

Landmark selection



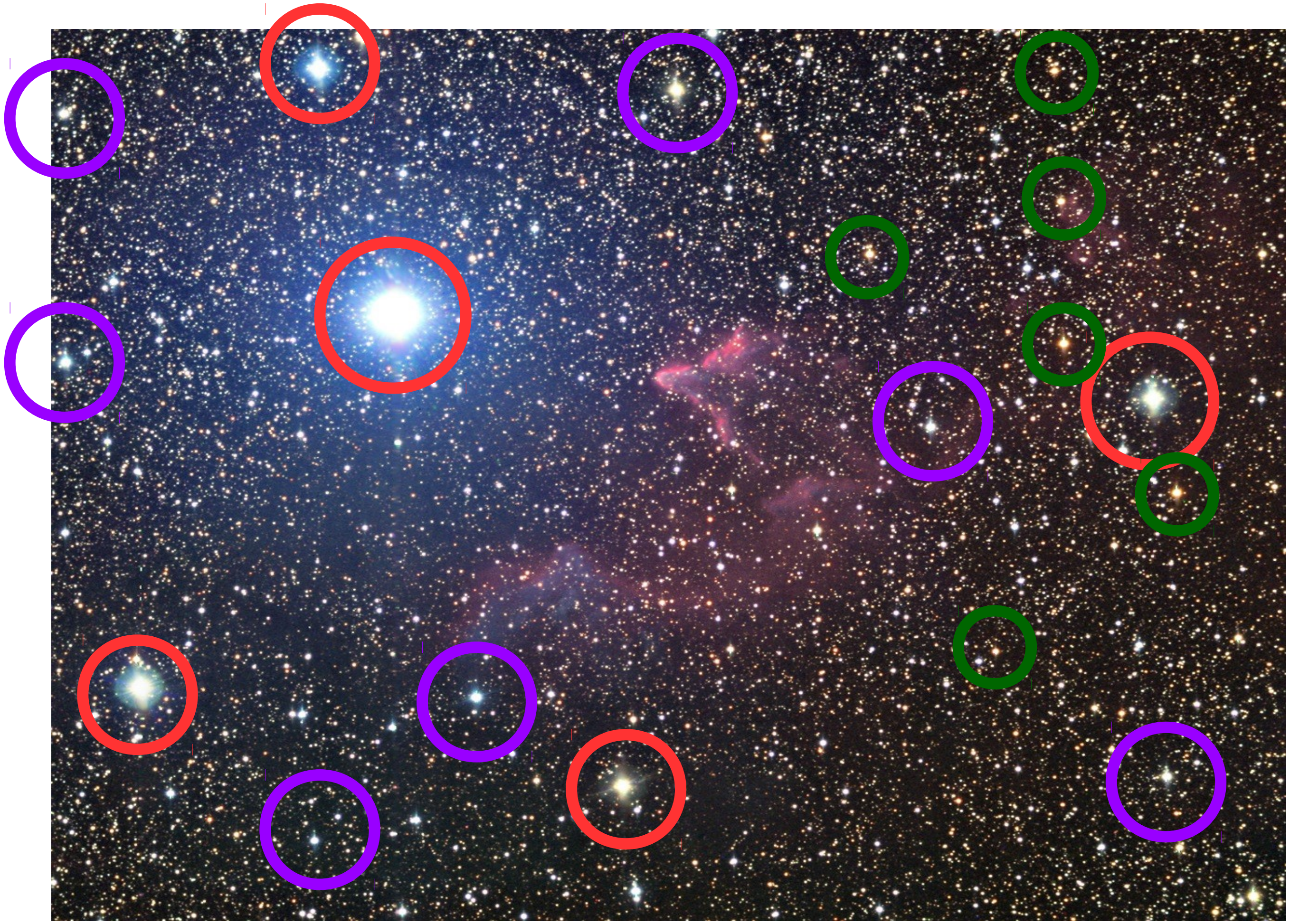
Alternative approach

Landmark selection + Active Learning



Alternative approach

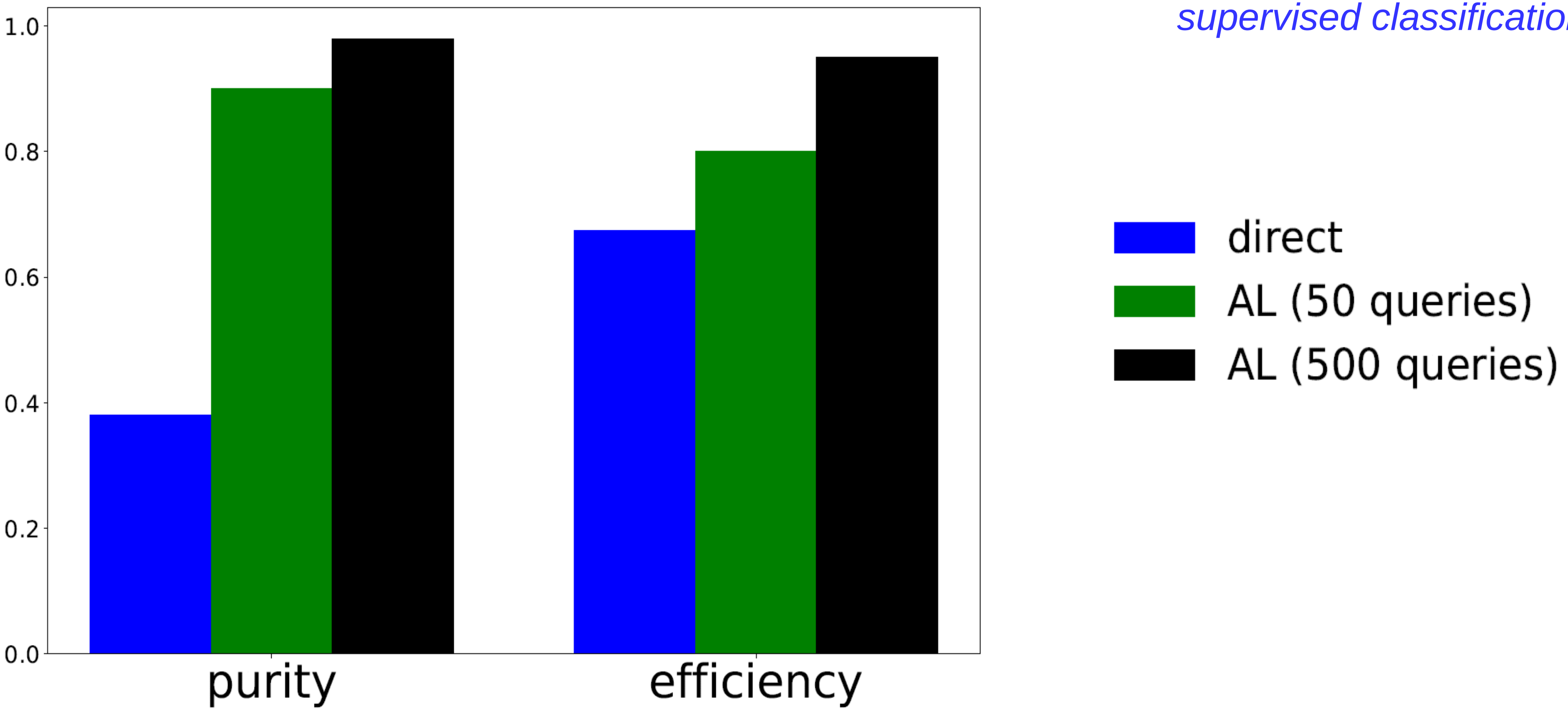
Landmark selection + Active Learning



Background: Active Learning in Astronomy

Automated Supernova Ia Classification Using Adaptive Learning Techniques

Kinjal Dhar Gupta*, Renuka Pampana*, Ricardo Vilalta*, Emille E. O. Ishida†, Rafael S. de Souza‡



supervised classification

*Dhar Gupta et al. (**incl. Ishida**), 2016 IEEE Symposium in Computational Intelligence, Athens*

Active Learning for supervised regression?

Active Learning for supervised regression?

Main tasks:

Regression:

- photometric redshift
- stellar parameters determination

...

Clustering:

- SN Ia spectra characterization
- galaxy spectral classification

...

Classification:

- detection
- star/galaxy separation
- galaxy morphology
- variable stars
- supernova

...

Anomaly/Novelty detection:

- unforeseen new objects
- detection error analysis
- identification of predicted objects

...

What about the future?

urgent: Build a support community

urgent: Build a support community

urgent: Build a support community

Solve bureaucracy

urgent: # Build a support community

Solve bureaucracy

The REAL goal is HUMAN learning