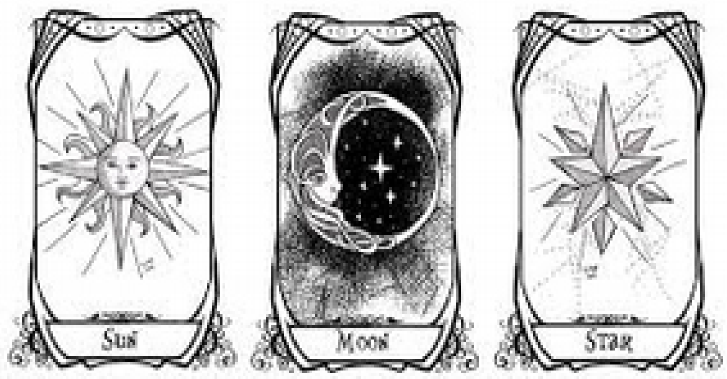
Astronomy in the 21st century:
Drowning in data, Starving for knowledge

Emille E. O. Ishida

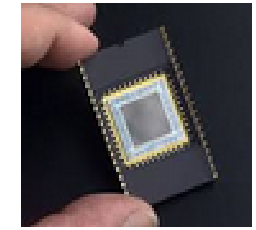
Laboratoire de Physique Corpusculaire - Université Blaise Pascal
Clermont Ferrand, France

The Big Picture

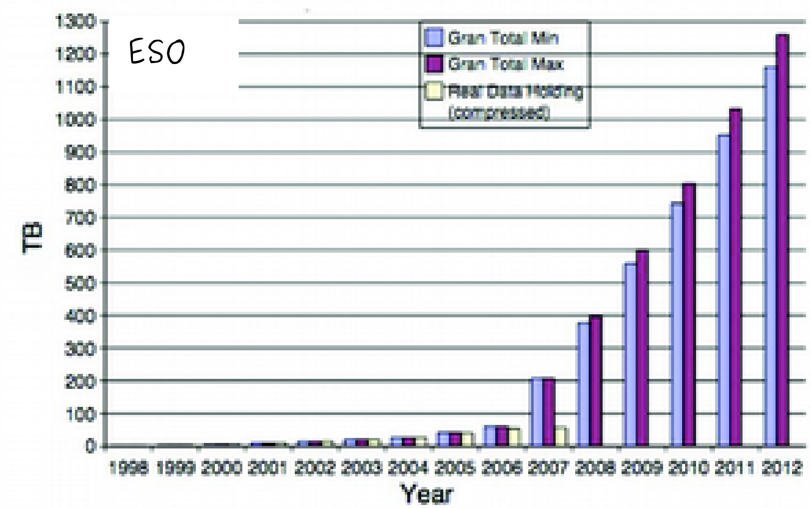
Astronomy began with 3 elements



In 1969, the CCD



Data



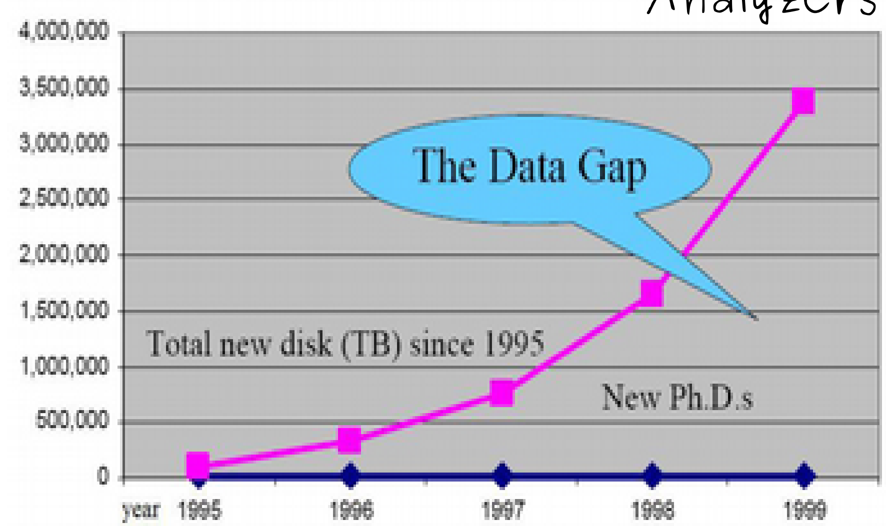
In the 17th century...

1 telescope

1 observer

1 object

Analyzers



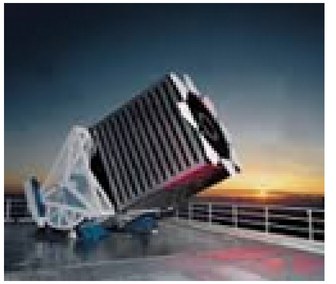
Grossman (2001)

An illustrative example:

SDSS – Sloan Digital Sky Survey

1992 – 2000: planning
2001 – 2009: observing

2.5m mirror
New Mexico, USA



1992

2.5 Terapixels of images

10 TB of raw data

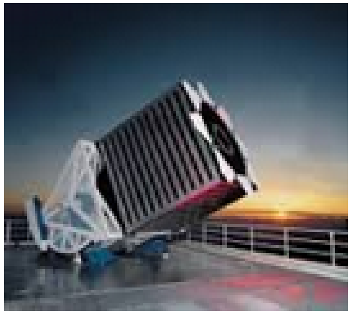
0.5 TB catalogs

An illustrative example:

SDSS – Sloan Digital Sky Survey

1992 – 2000: planning
2001 – 2009: observing

2.5m mirror
New Mexico, USA



1992

2.5 Terapixels of images

10 TB of raw data

0.5 TB catalogs

How to deliver 0.5 TB of useful data to all users?

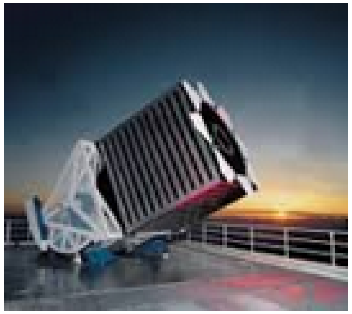
A. Szalay, https://www.youtube.com/watch?v=FlcdG4hUn1Q

An illustrative example:

SDSS – Sloan Digital Sky Survey

1992 – 2000: planning
2001 – 2009: observing

2.5m mirror
New Mexico, USA

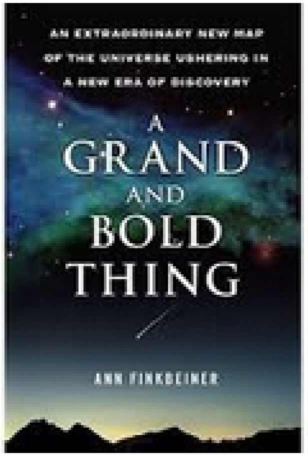


1992

2.5 Terapixels of images
10 TB of raw data
0.5 TB catalogs

2009

5 Tpx of images
120TB processed data
35TB catalogs



by Ann K. Finkbeiner, 2012

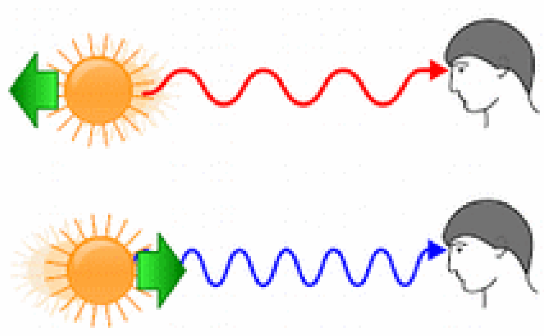
A. Szalay, https://www.youtube.com/watch?v=FlcdG4hUn1Q

Case study:

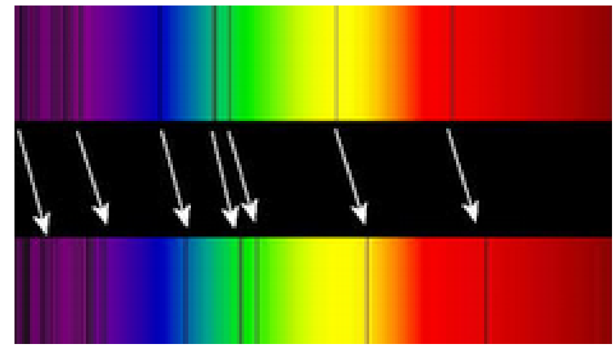
Photometric Redshifts

Redshifts (z) ↔ Distances

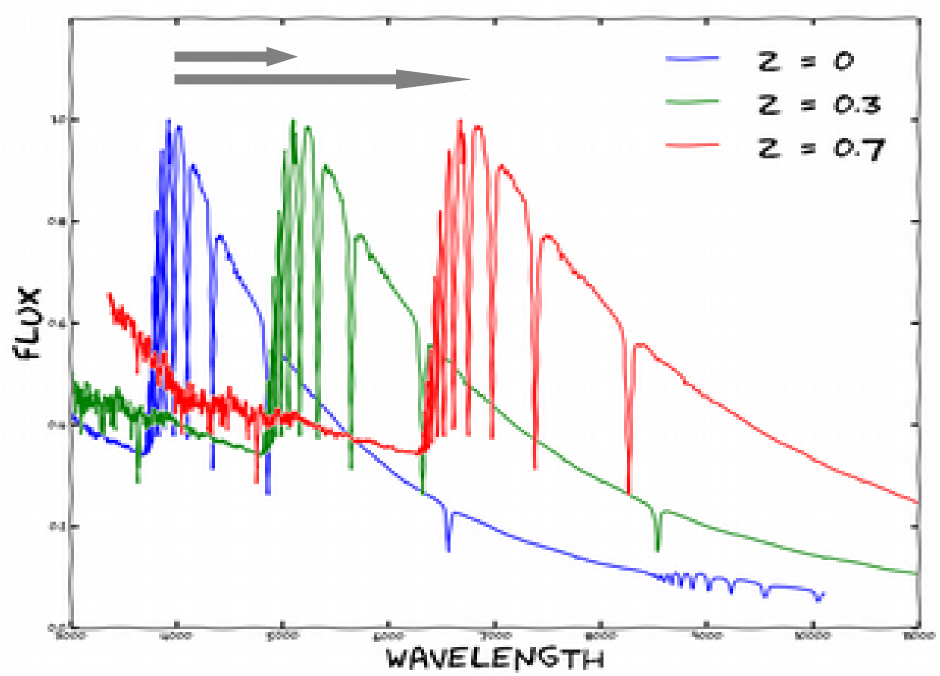
Idea



Observation



or



Velocities

+

Cosmological model

=

distances

+

finite light velocity

History

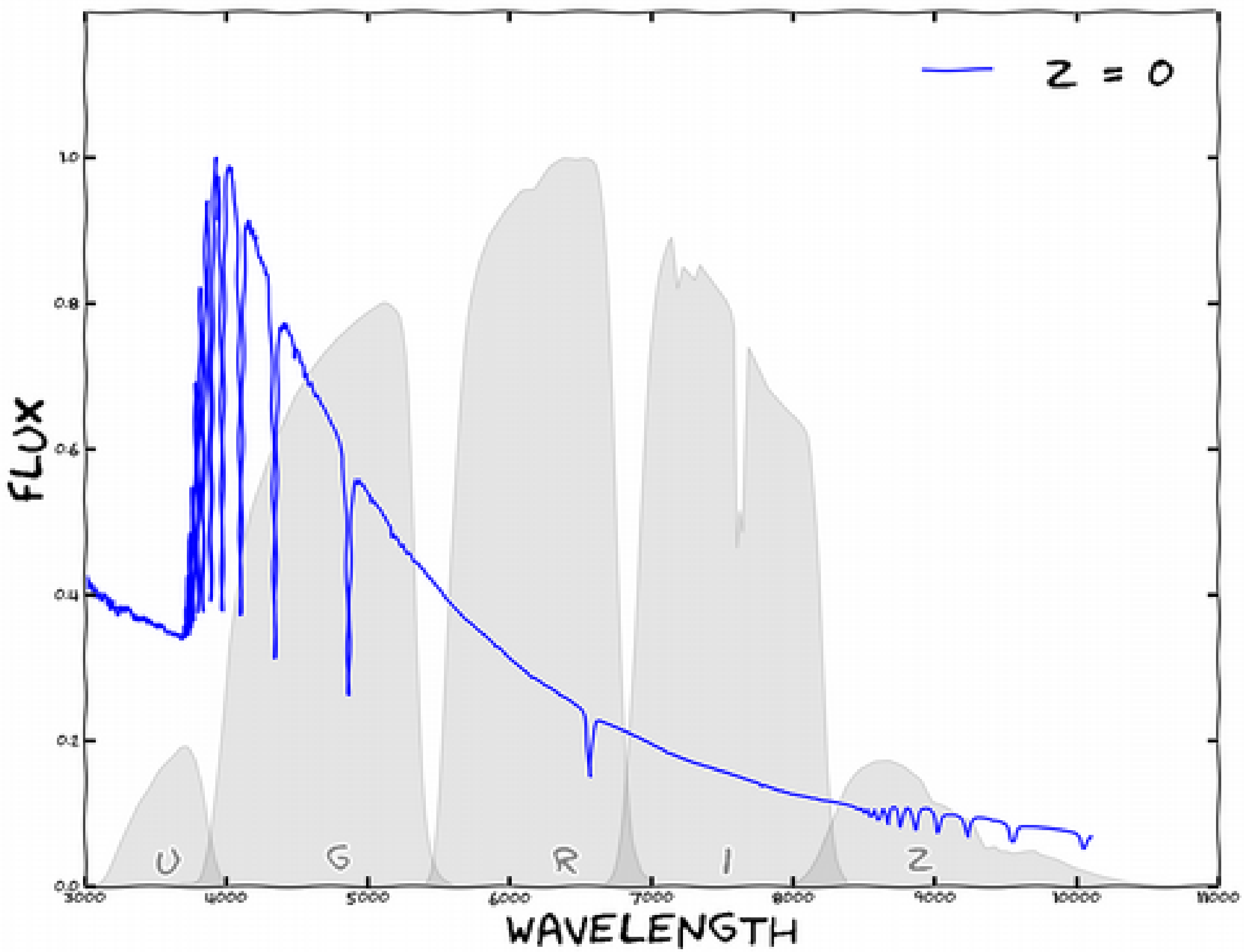
Redshifts are important!

Photometric Redshifts

Spectra are expensive!

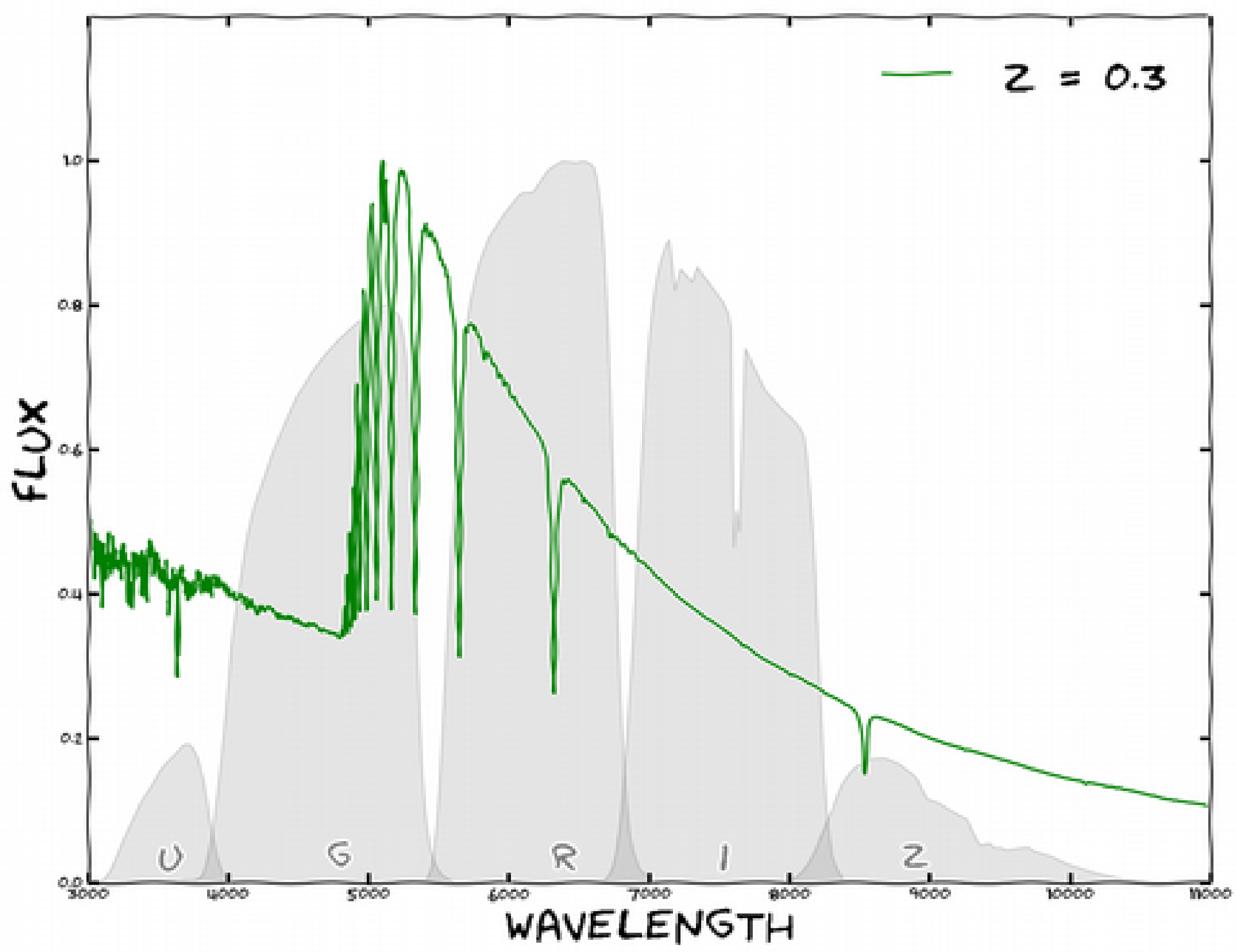
Photometric Redshifts

Alternative measurement
300 dim → 5 dim



Photometric Redshifts

Alternative measurement
300 dim →5 dim



Photometric Redshifts

Alternative measurement
300 dim → 5 dim

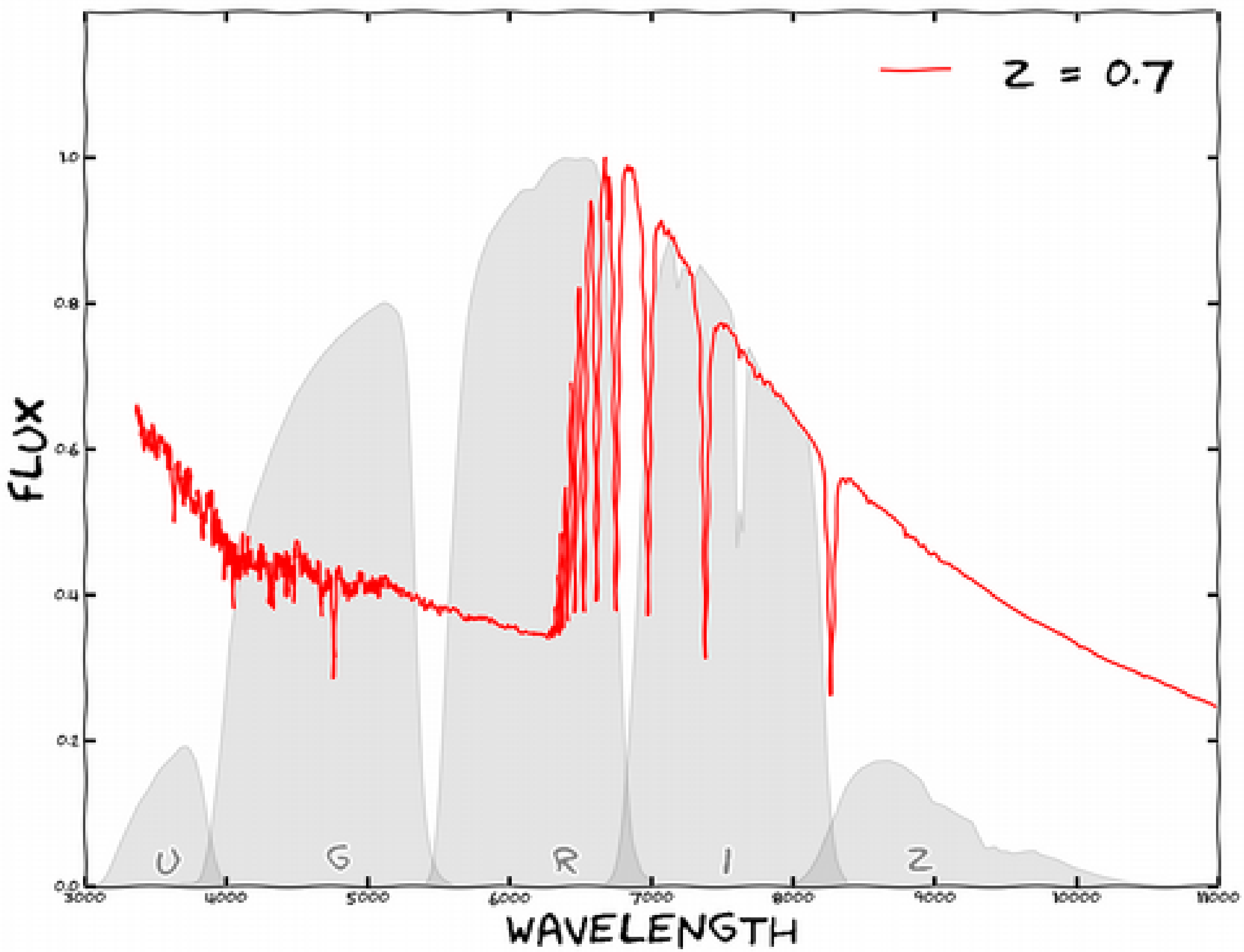


Photo-z: a regression problem

u-g	g-r	r-i	i-z	redshift
2.07	1.39	0.48	0.27	0.31
1.54	1.58	0.54	0.42	0.34
1.03	1.76	0.67	0.37	0.41
2.17	1.30	0.43	0.30	0.19
...	...	...	...	...
1.36	1.72	0.52	0.36	0.32

Photo-z: a regression problem

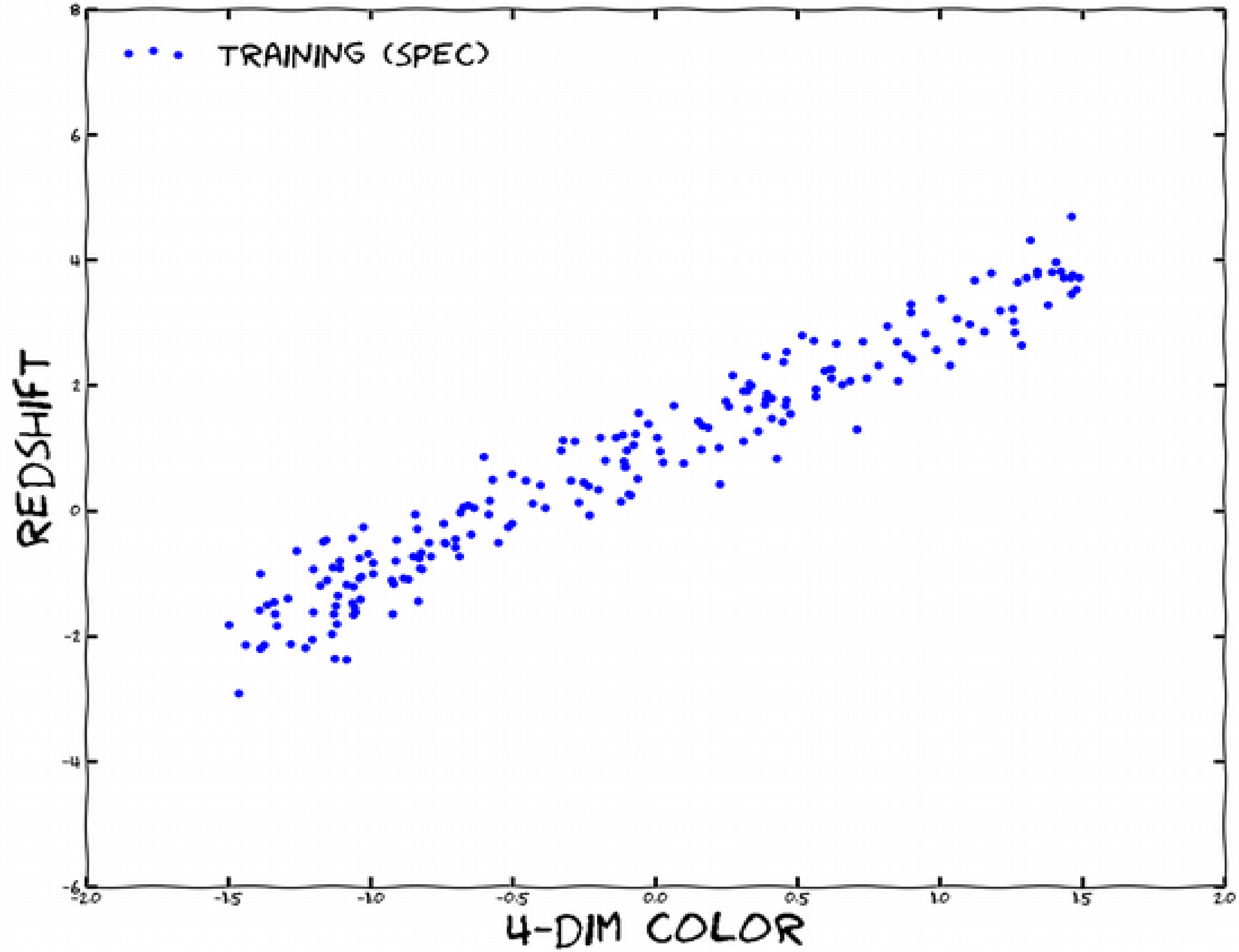


Photo-z: a regression problem

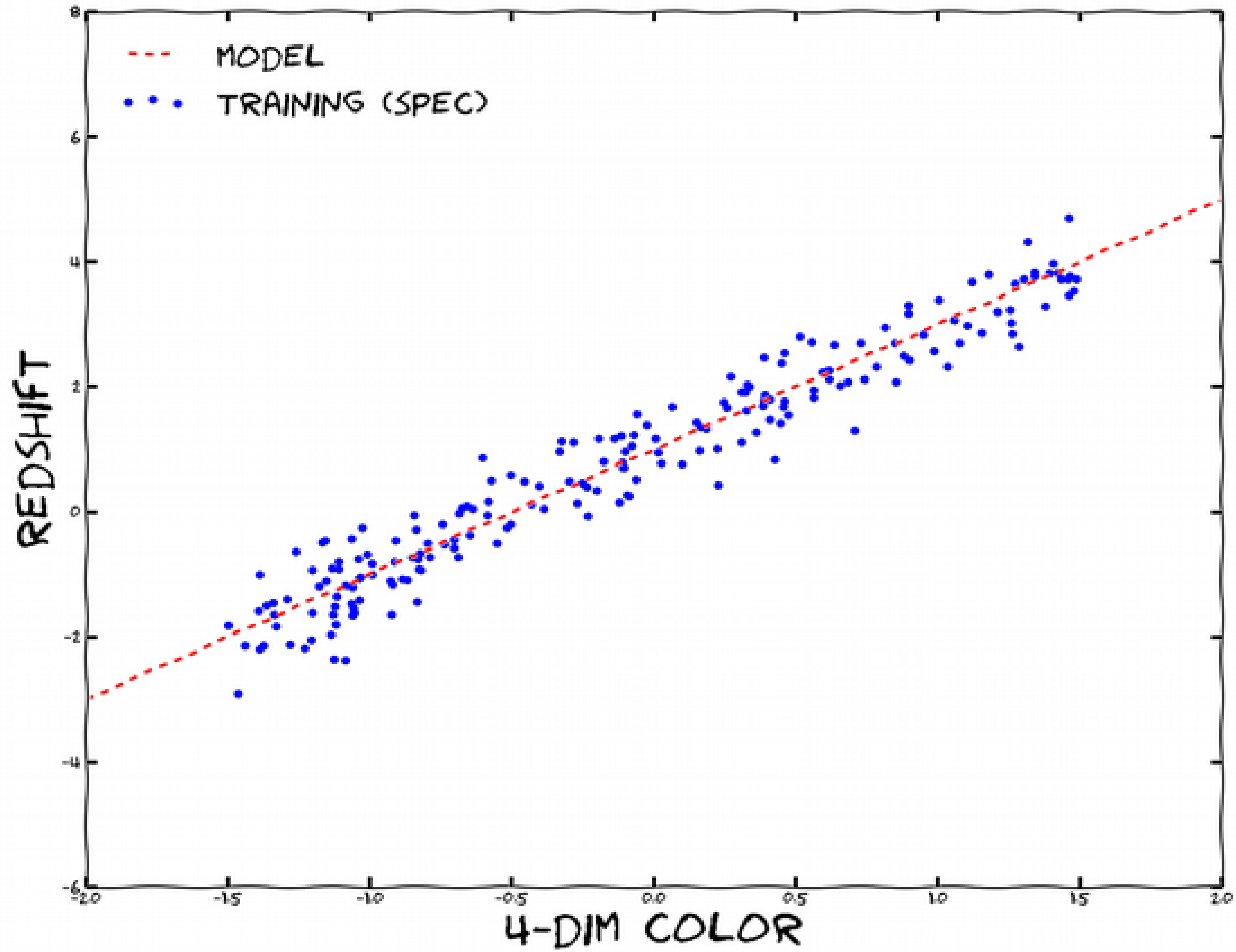


Photo-z: a regression problem

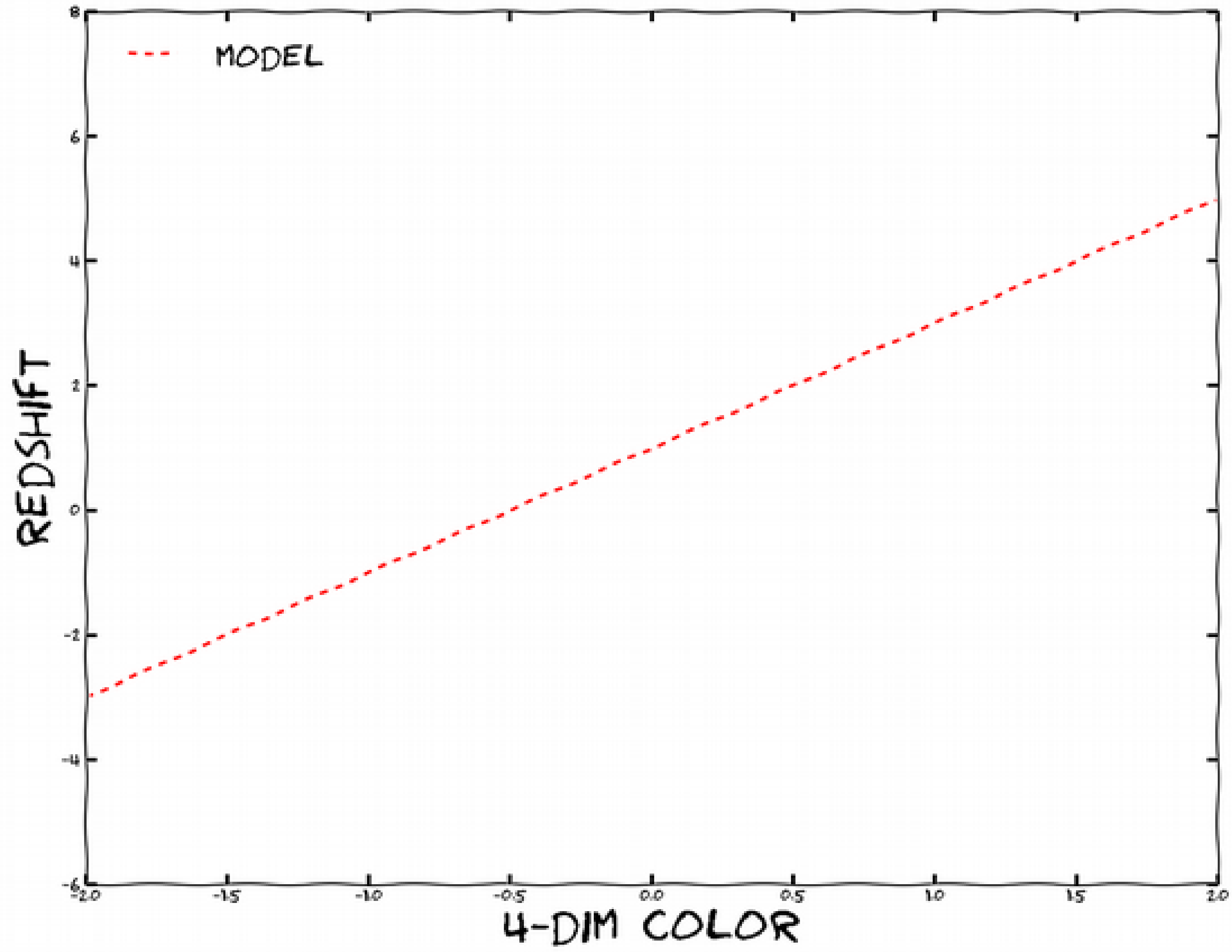


Photo-z: a regression problem

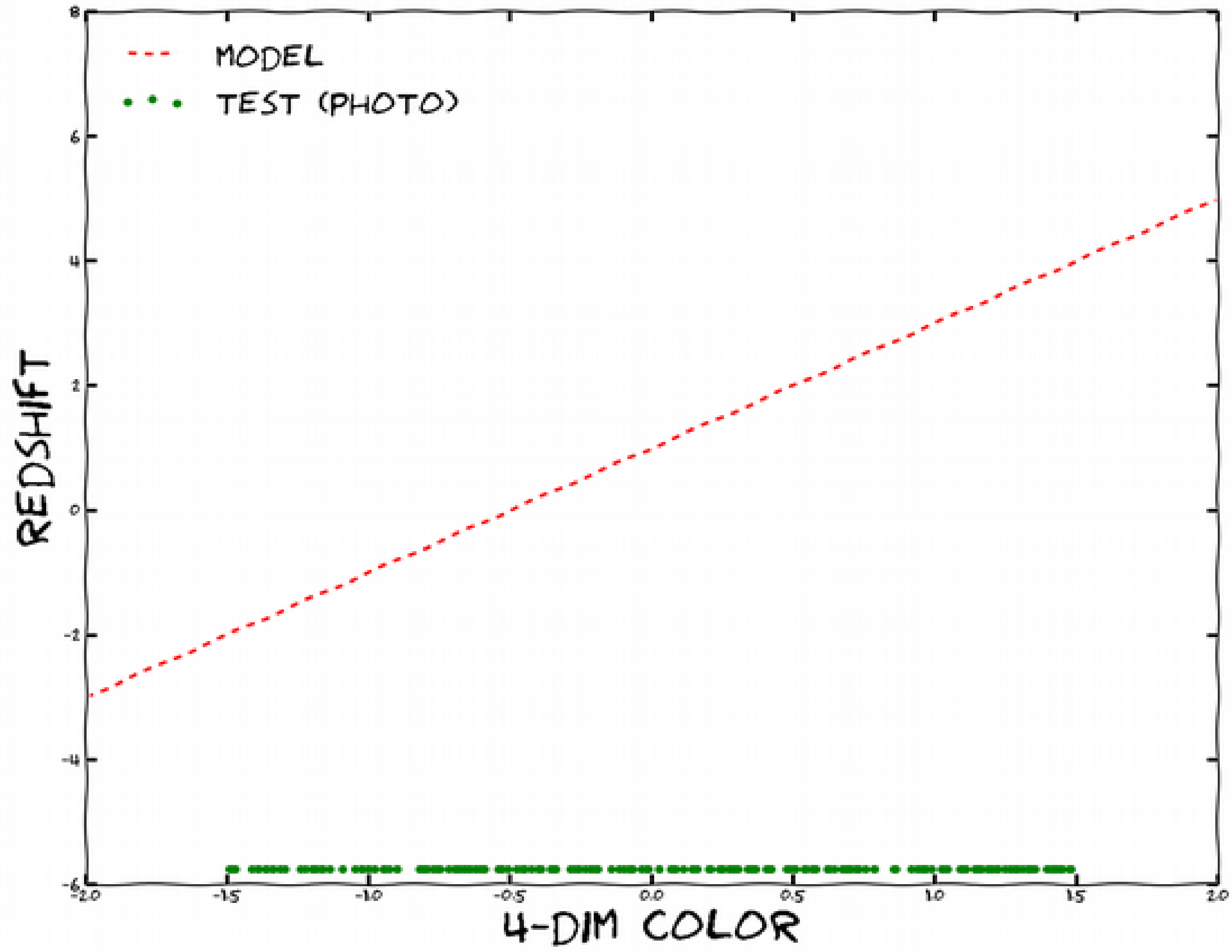


Photo-z: a regression problem

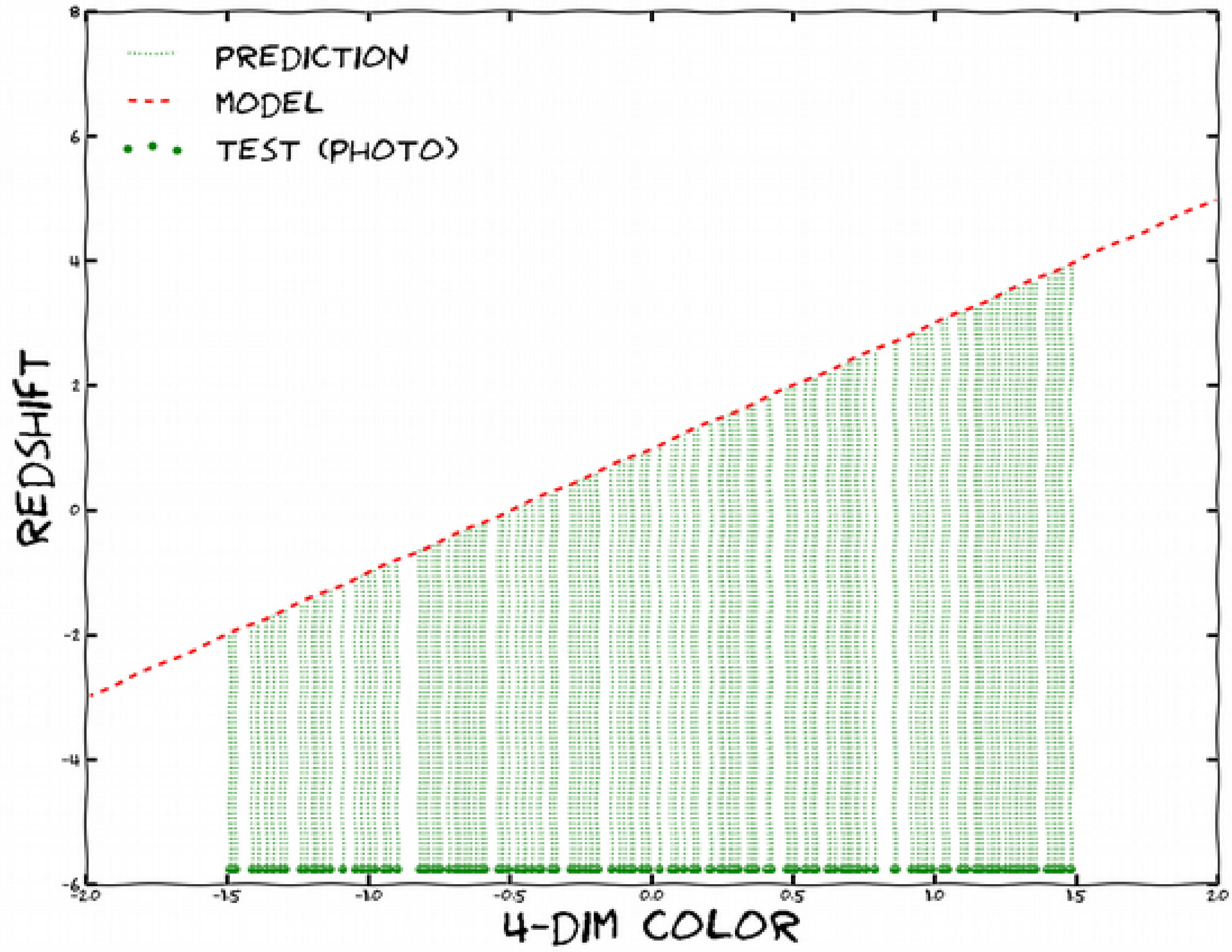
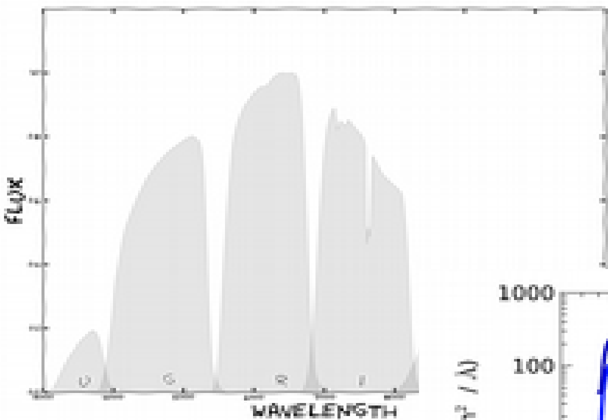
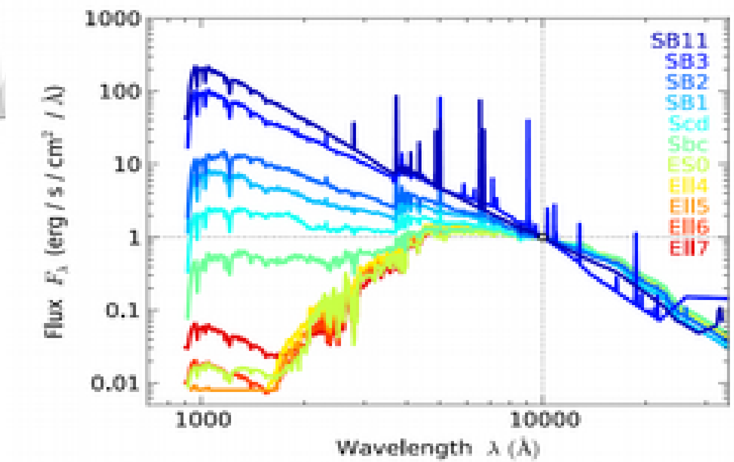


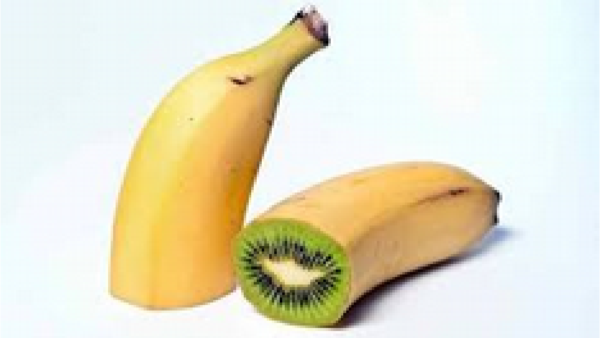
Photo-z: methods

Template fitting





Hybrid



Machine Learning

Photo-z: methods

Machine
Learning

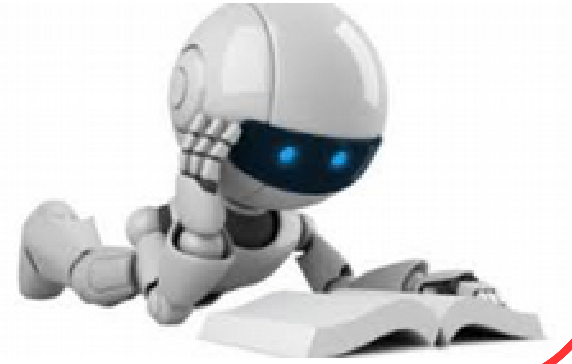
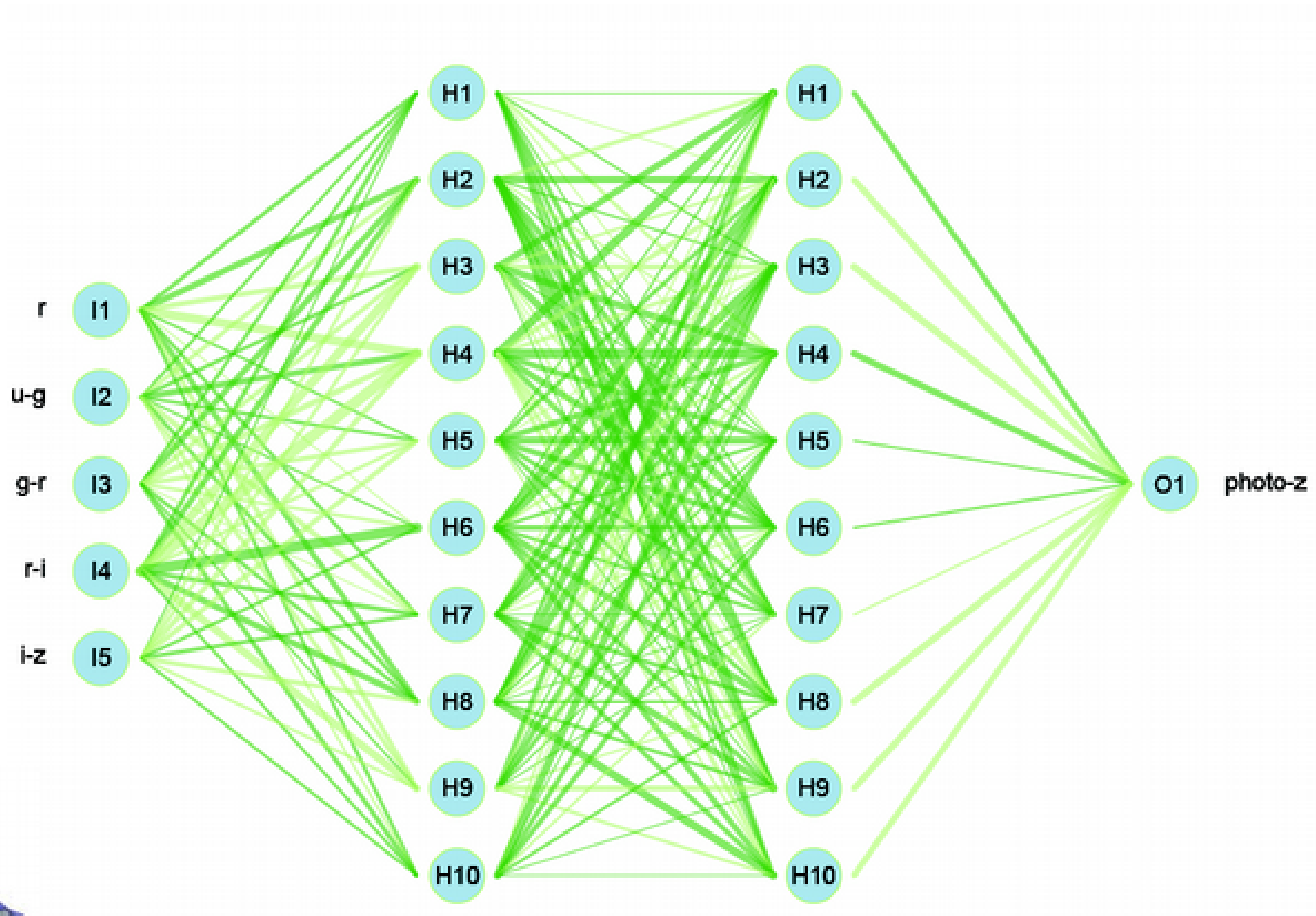


Photo-z: Artificial Neural Networks

Used in astronomy since 1990's



Plot by Rafael S. de Souza

Photo-z: Artificial Neural Networks

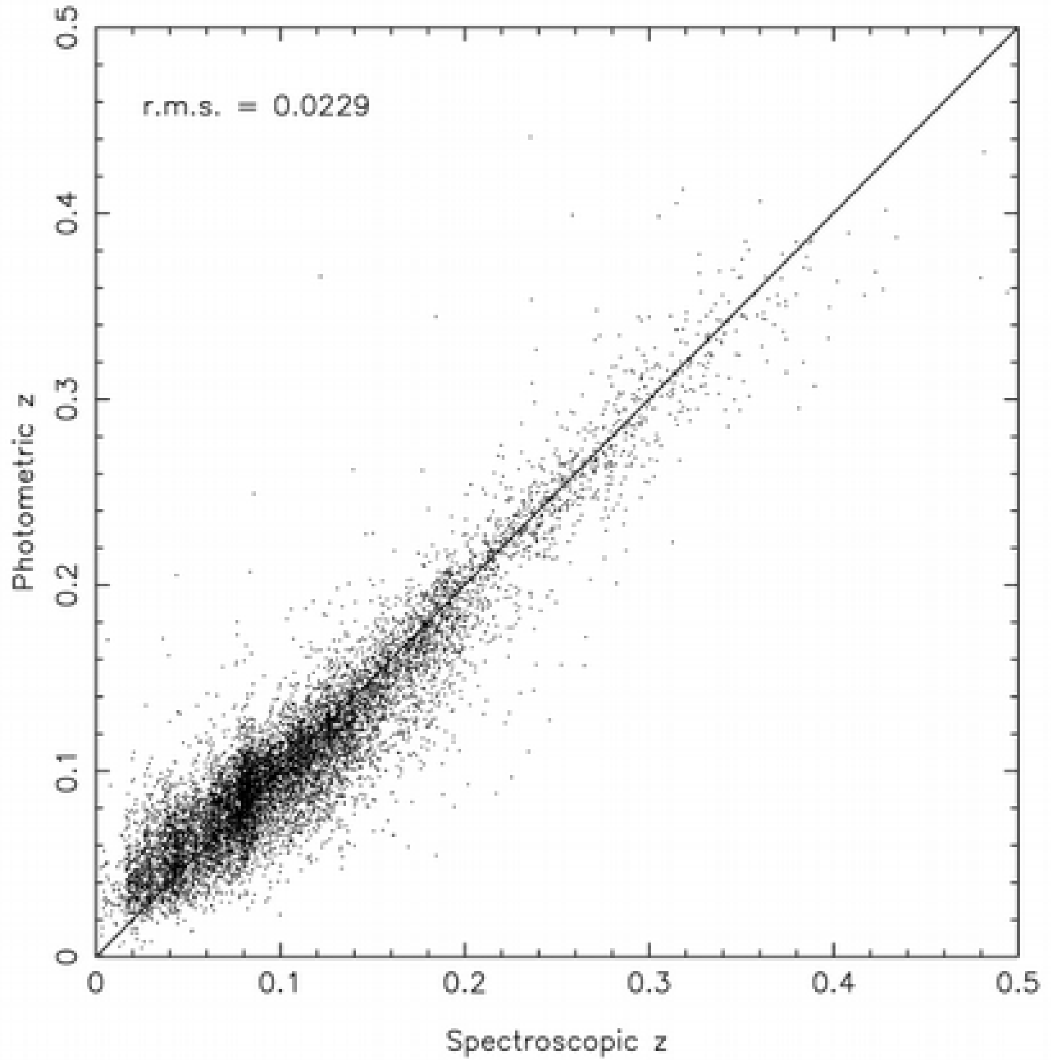


FIG. 2.— Spectroscopic *vs.* photometric redshifts for ANNz applied to 10,000 galaxies randomly selected from the SDSS EDR.

Collister & Lahav, 2003

Photo-z: Nearest Neighbors

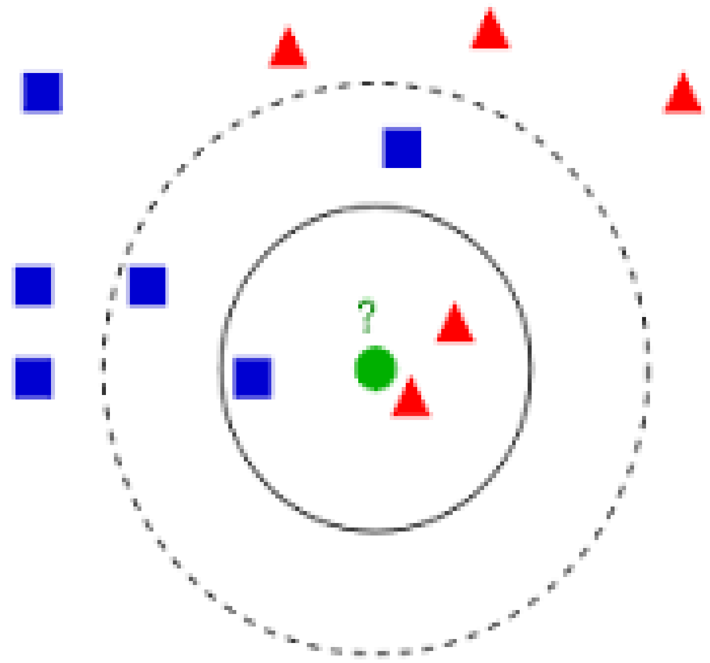
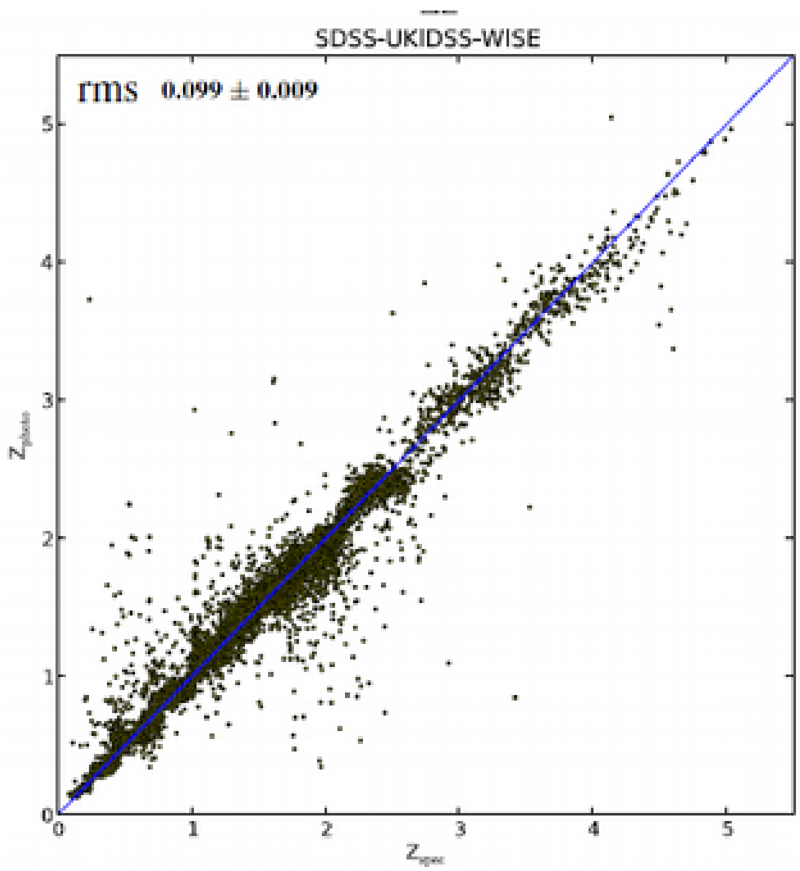
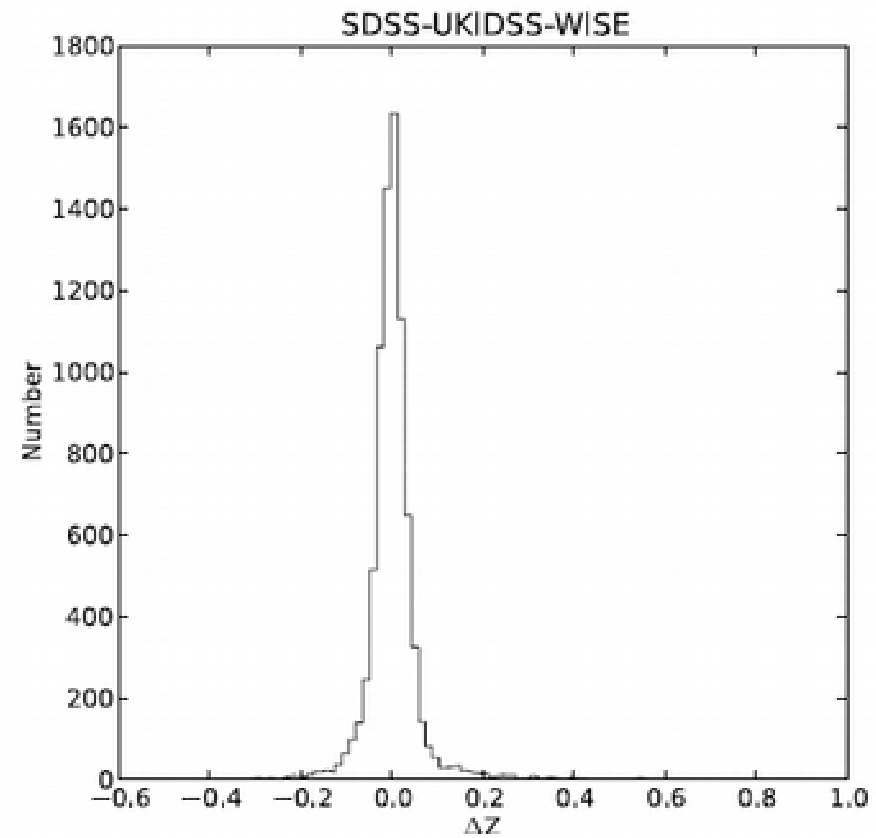


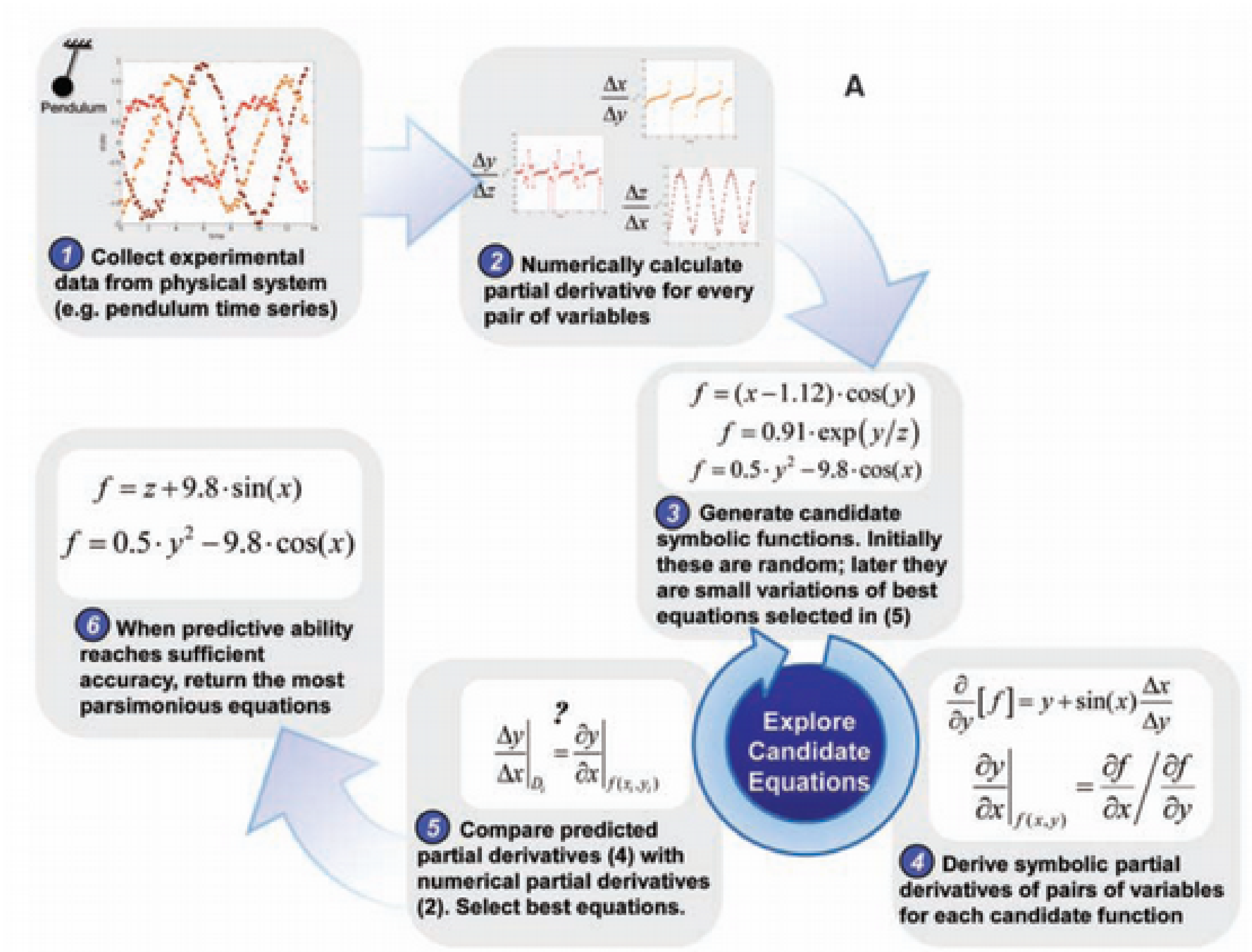
Photo-z: Nearest Neighbors





Zhang et al., 2013

Photo-z: Symbolic Regression



Schimdt & Lipson, Science, 324 (2009)

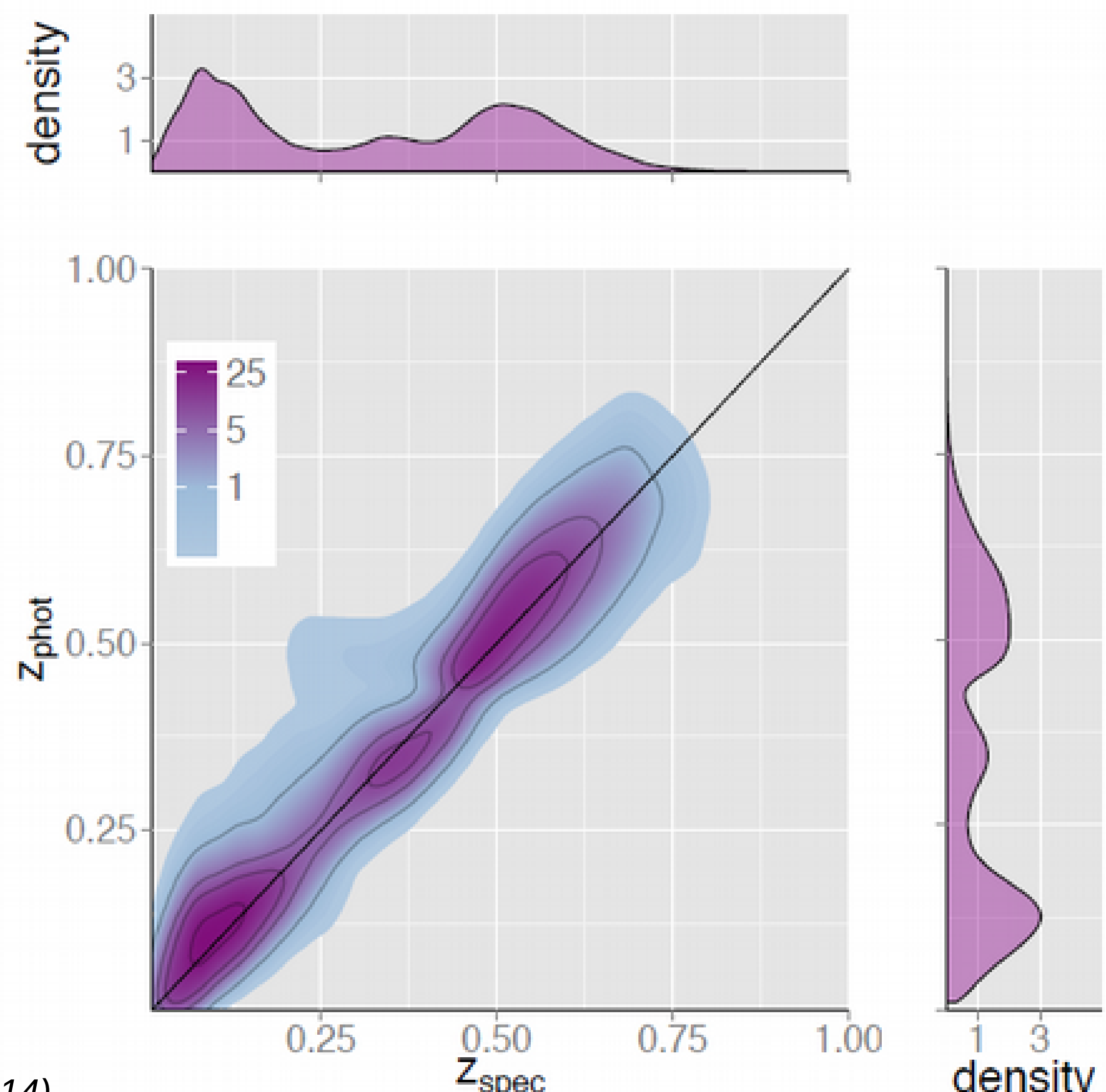
Photo-z: Symbolic Regression

Final expression:

$$z_{\mathrm{phot}} = \frac{0.4436r - 8.261}{24.4 + (g-r)^2(g-i)^2(r-i)^2 - g + 0.5152(r-i)}.$$

Parametric model

$$\sigma_{(z_{\mathrm{phot}} - z_{\mathrm{spec}})/(1+z_{\mathrm{spec}})} \approx 0.0449$$

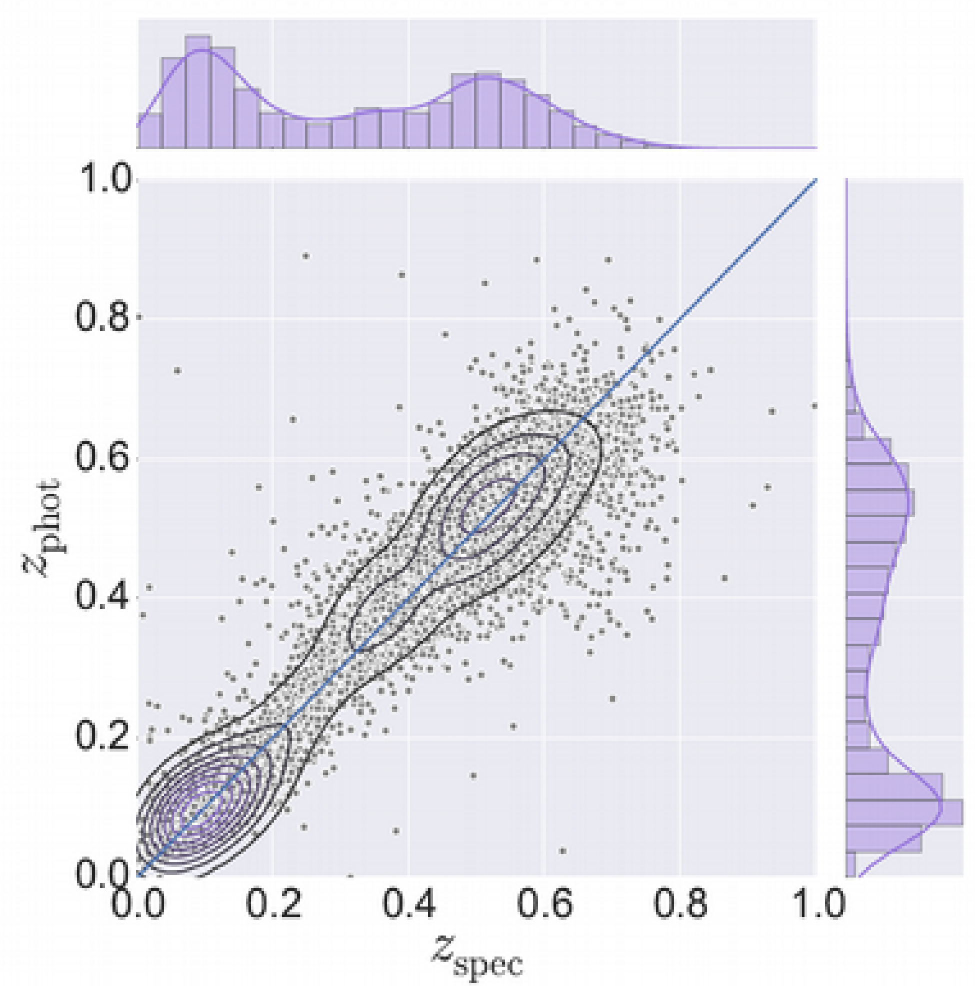


Pre-COIN paper:
*Krone-Martins, **Ishida** & de Souza, MNRASL 443 (2014)*

Photo-z: Generalized Linear Models

$$rms(\Delta z) \sim 0.034$$

Statistical model

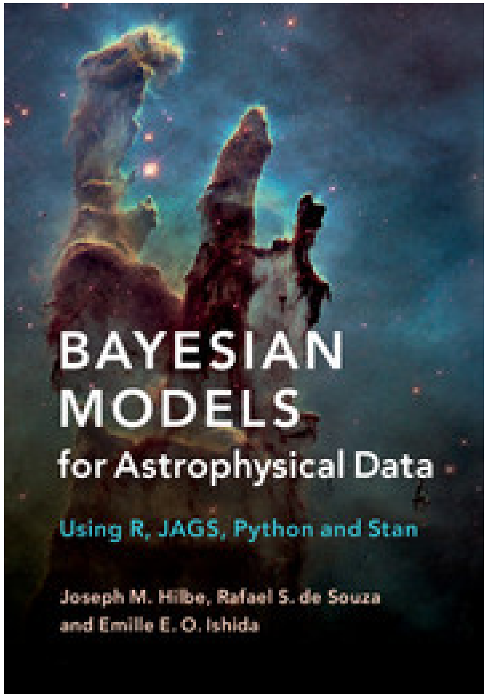


From COIN Residence Program #1:
Elliot et al. (incl. ***Ishida****), Astronomy & Computing, 10 (2015)*

Photo-z: Generalized Linear Models

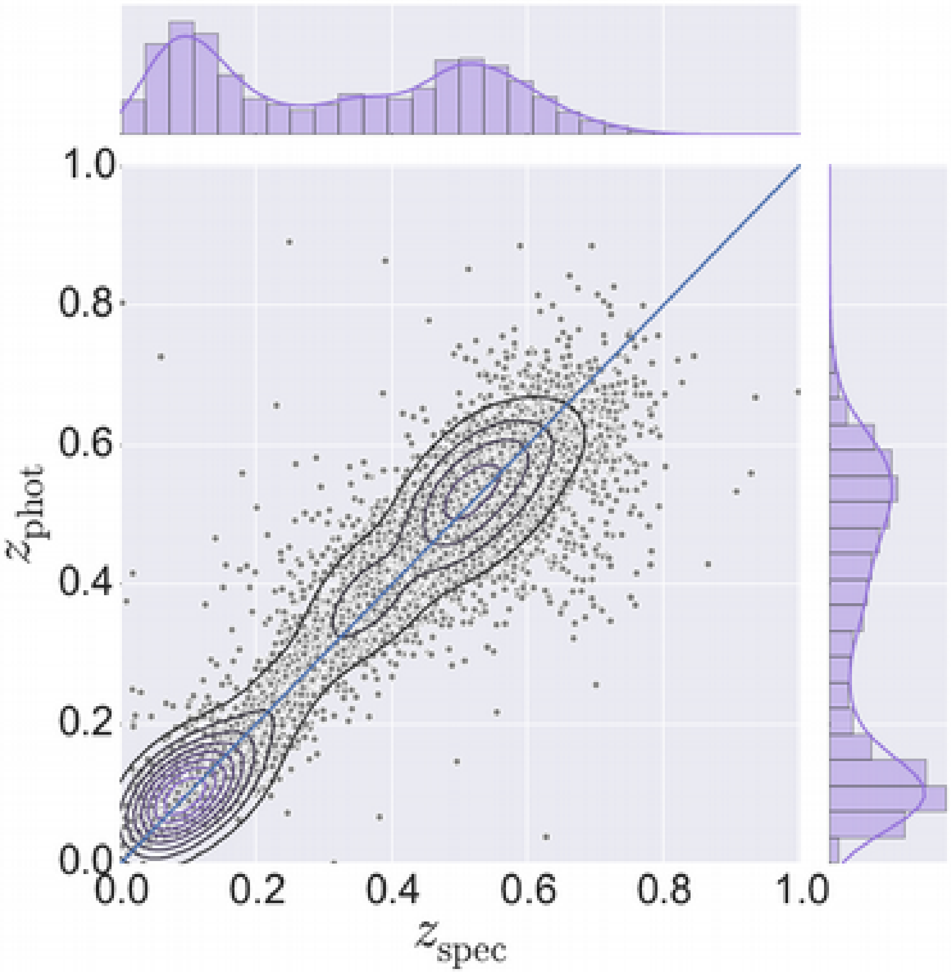
More on GLMs
(Bayesian approach):

https://github.com/RafaelSdeSouza/ADA8



Cambridge
University Press
May/2017

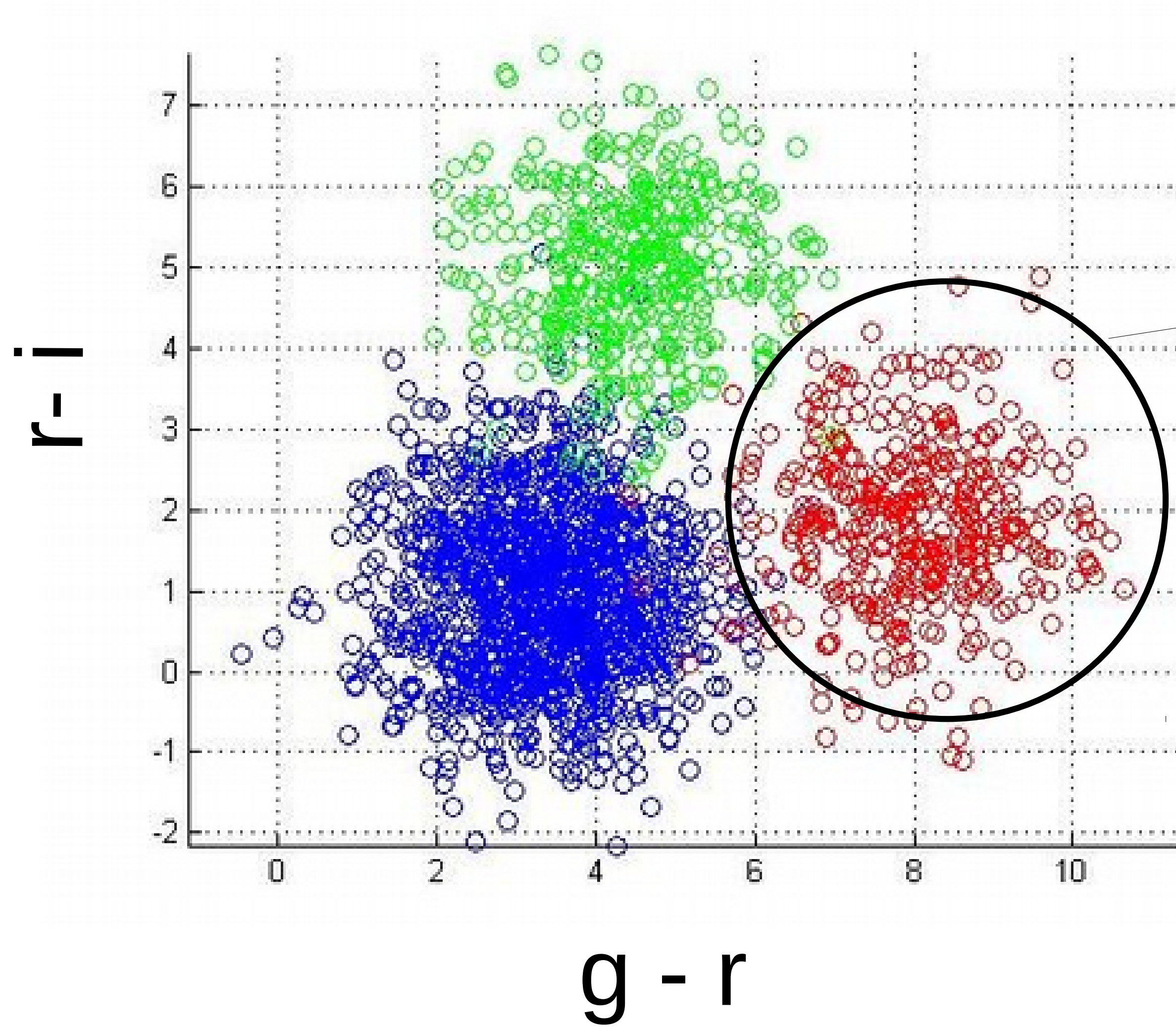
$rms(\Delta z) \sim 0.034$



From COIN Residence Program #1:
Elliot et al. (incl. ***Ishida****), Astronomy & Computing, 10 (2015)*

Photo-z: Local Linear Regression

Nearest neighbors



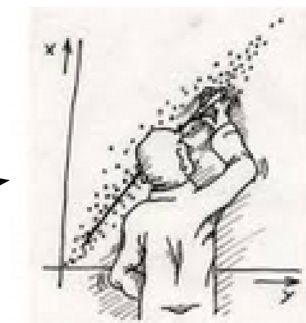
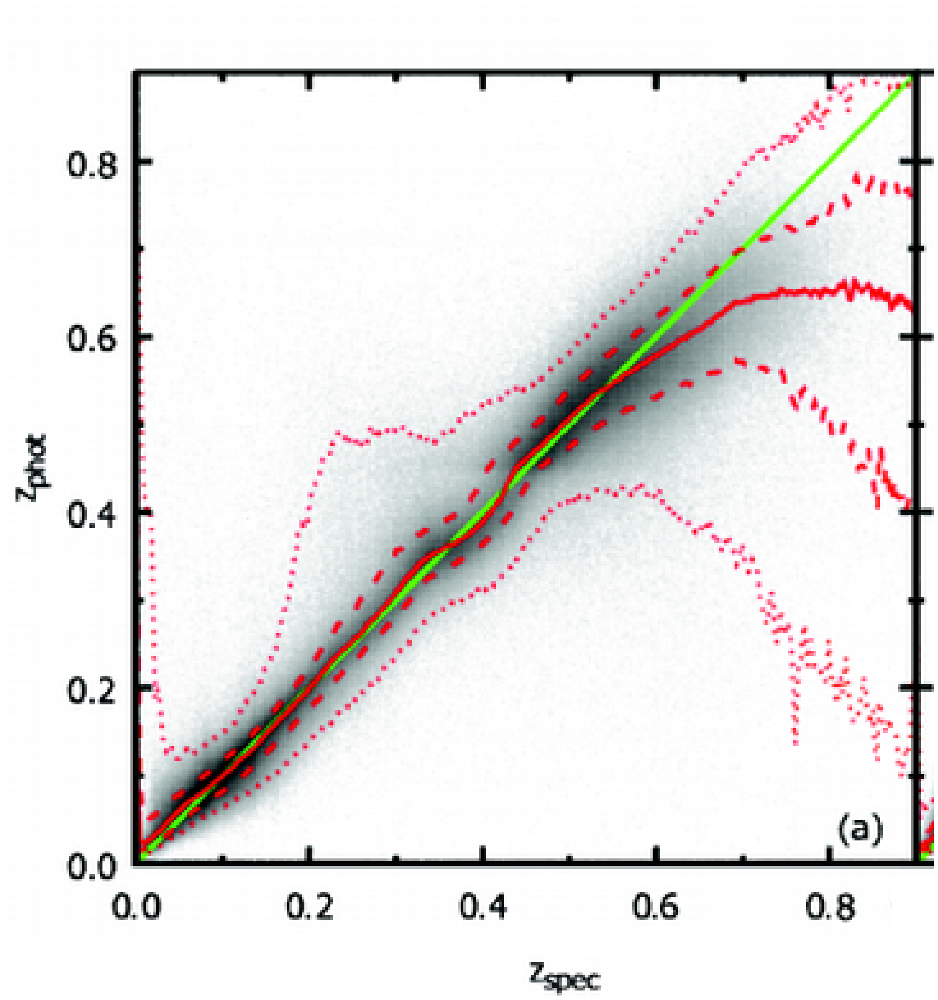


Photo-z: Local Linear Regression

Official SDSS DR12 Photoz method

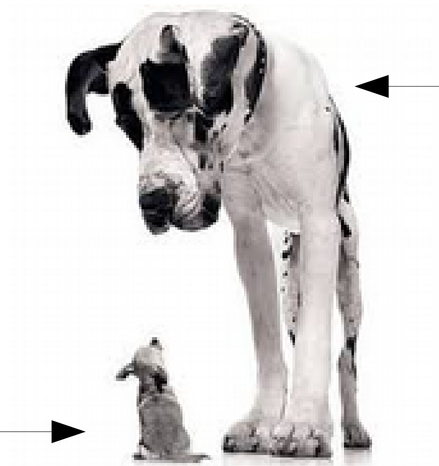


$\sigma\left(\Delta z_{\mathrm{norm}}\right) = 0.0205$

Beck et al., MNRAS 460 (2016)

Summary of results:

Challenges



test

training

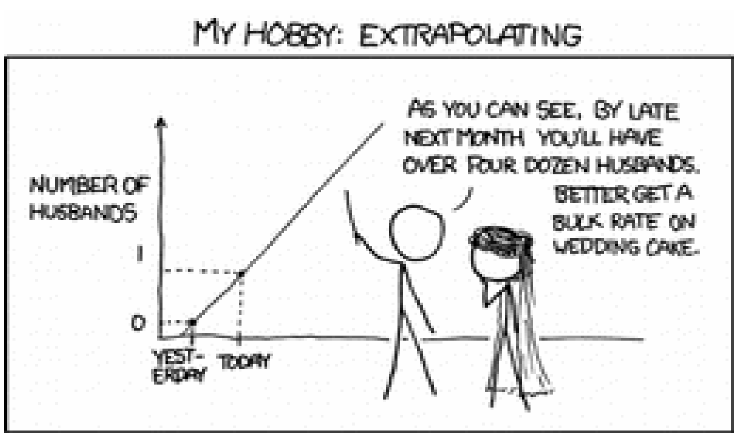
Measurement errors

Impossible to get a representative training sample

Training will always be:

- **Brighter**
- **Closer**
- **Higher data quality**
- **Diverse population**

Supervised methods cannot extrapolate



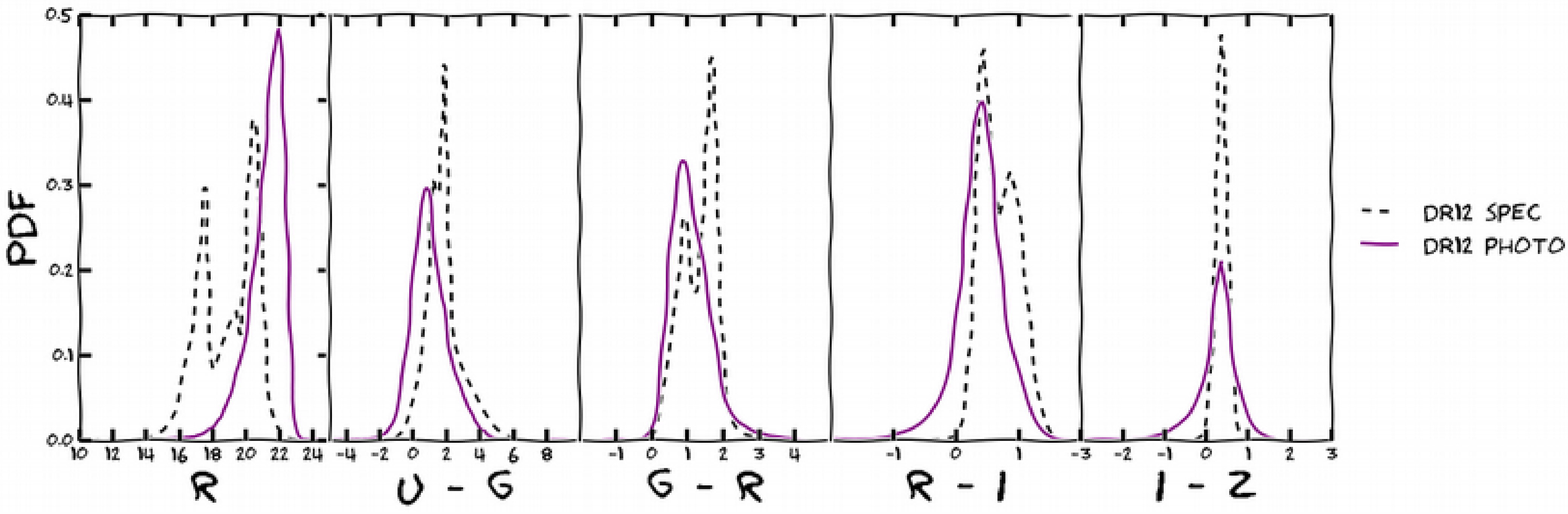
Challenges

Impossible to get a representative training sample

Training will always be:

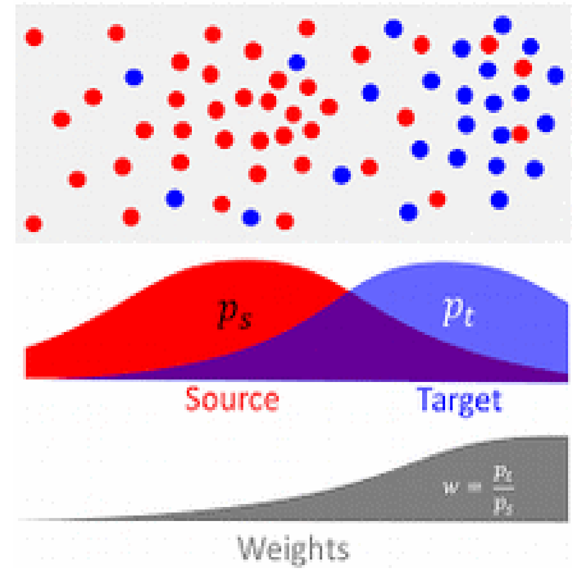
- **Brighter**
- **Closer**
- **Higher data quality**
- **Diverse population**

The quest for representativeness

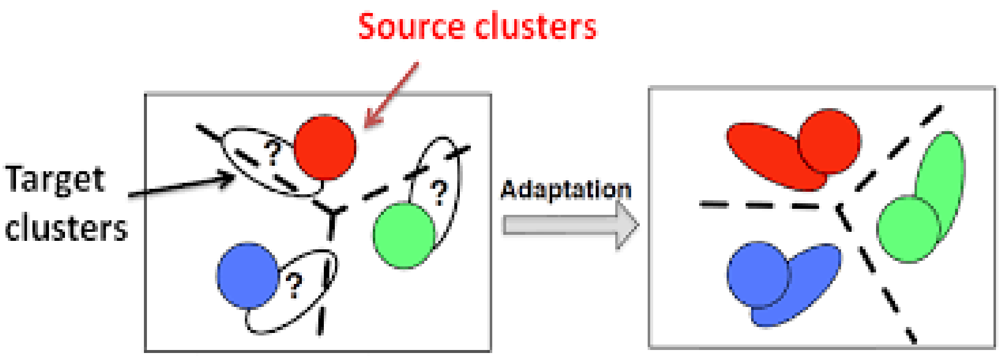


Domain Adaptation

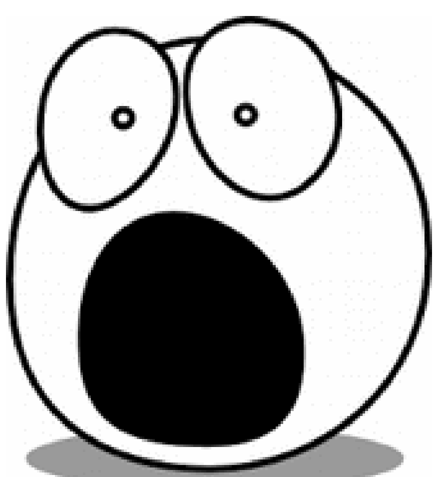
Give weights



Define boundaries



Feature space transformation

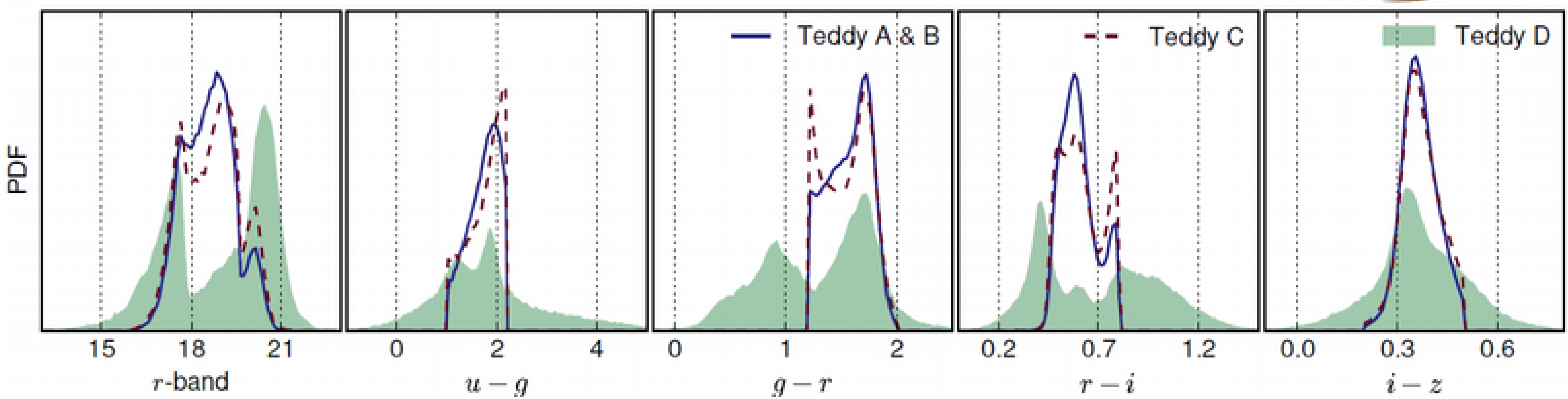


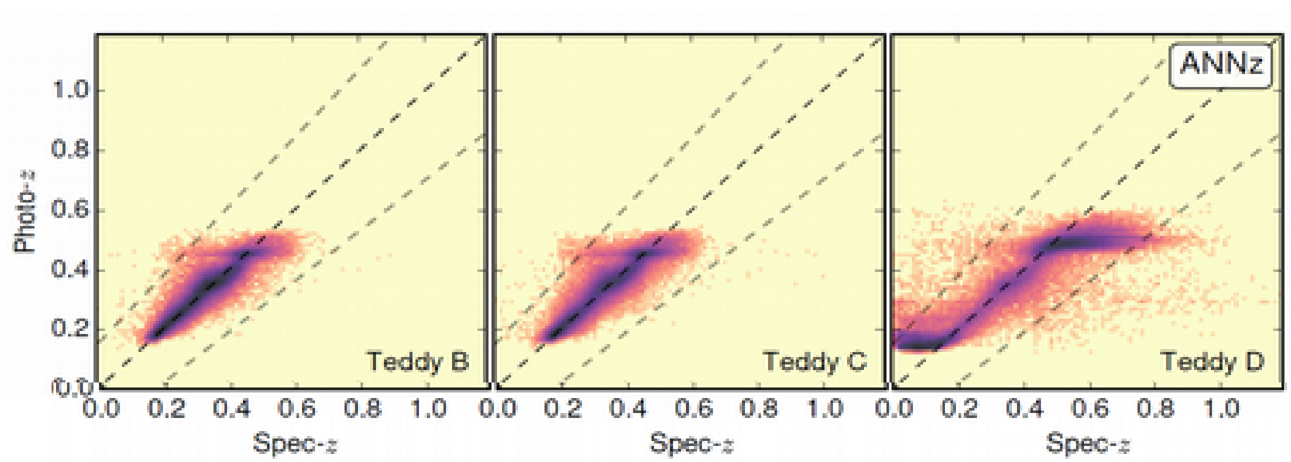
CRP #3 – Budapest, 2016

Teddy catalogue

Probing the effect of 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)



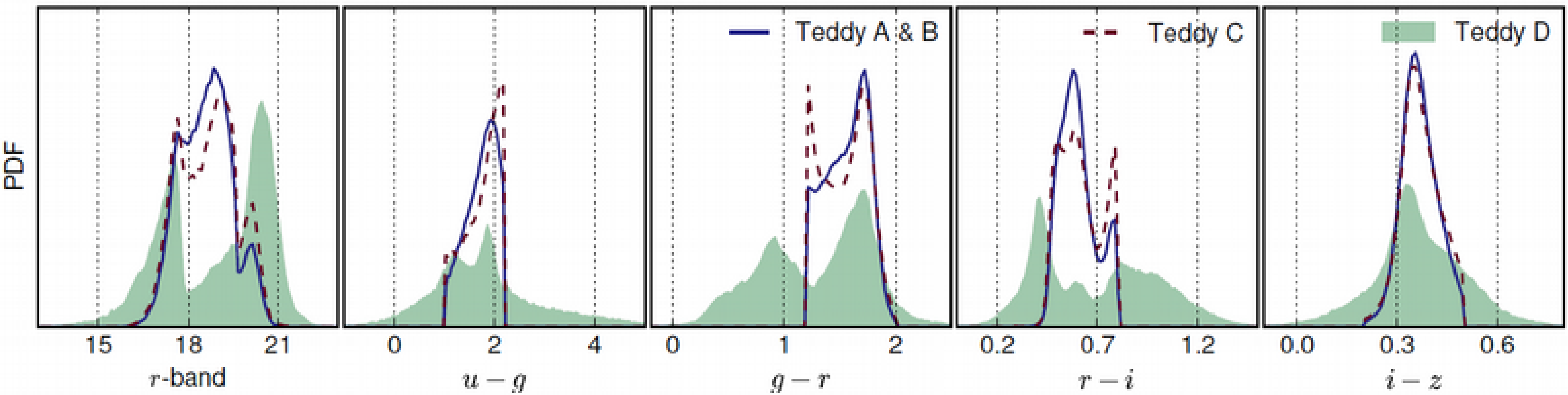


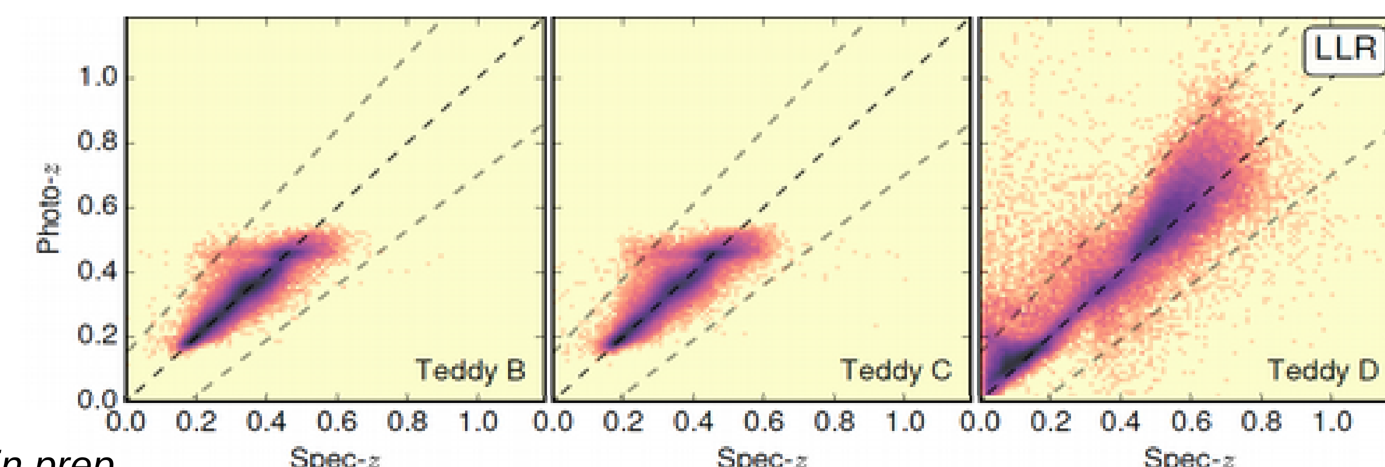
From COIN Residence Progrm #3 – in prep

Teddy catalogue

Probing the effect of 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)

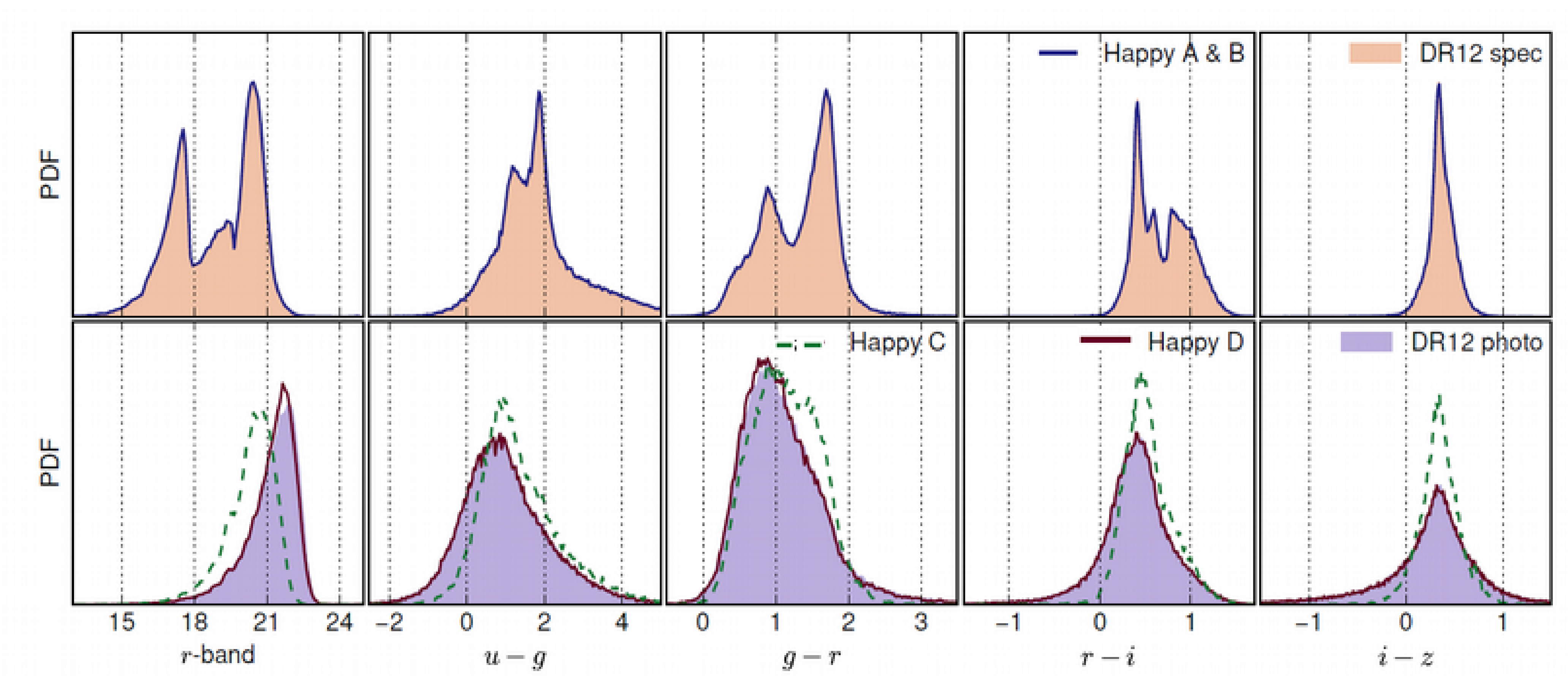




From COIN Residence Progrm #3 – in prep

Happy catalogue

The effect of coverage + photometric errors

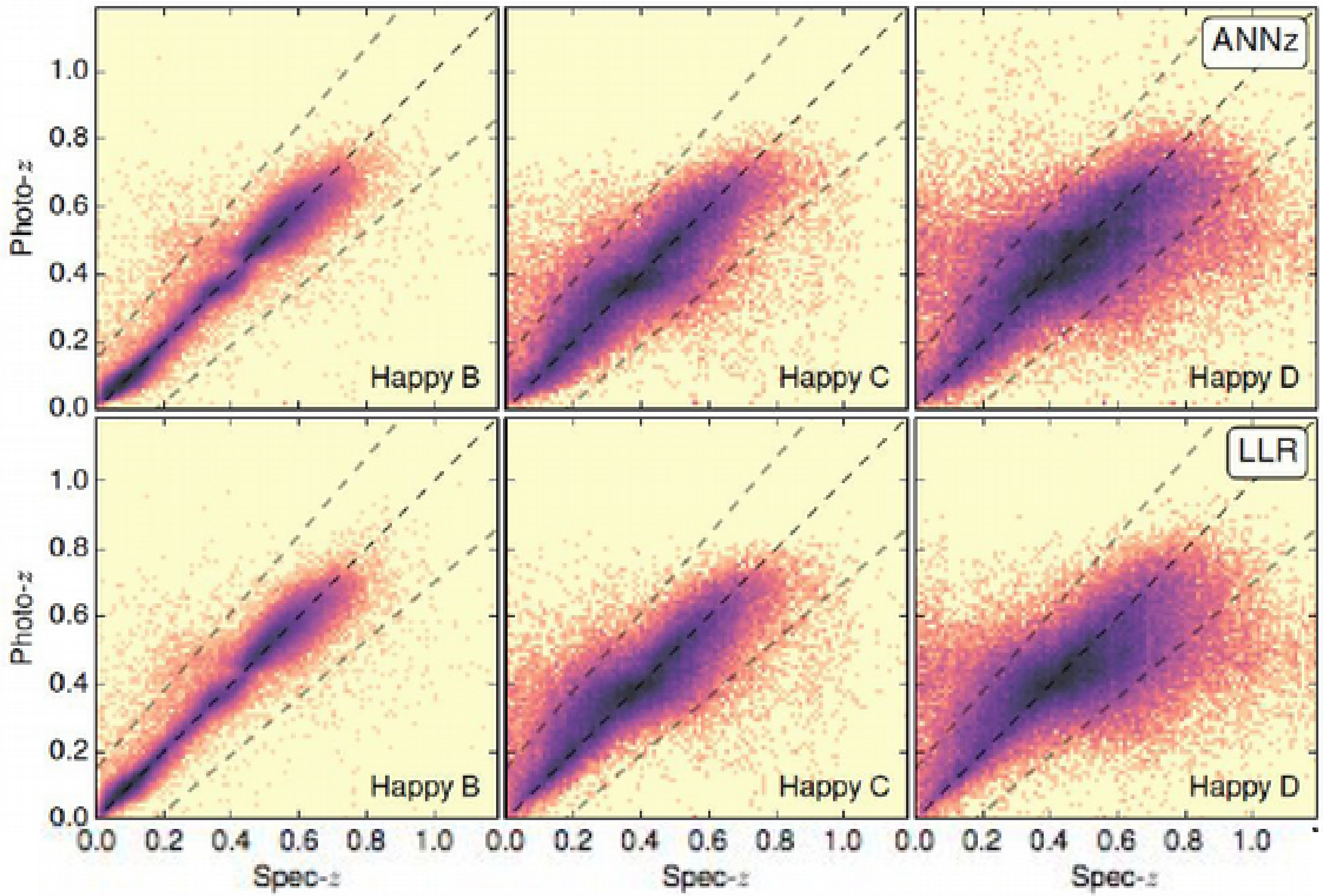


From COIN Residence Program #3 – in prep

Happy catalogue

The effect of coverage + photometric errors

representative   same data quality   realistic



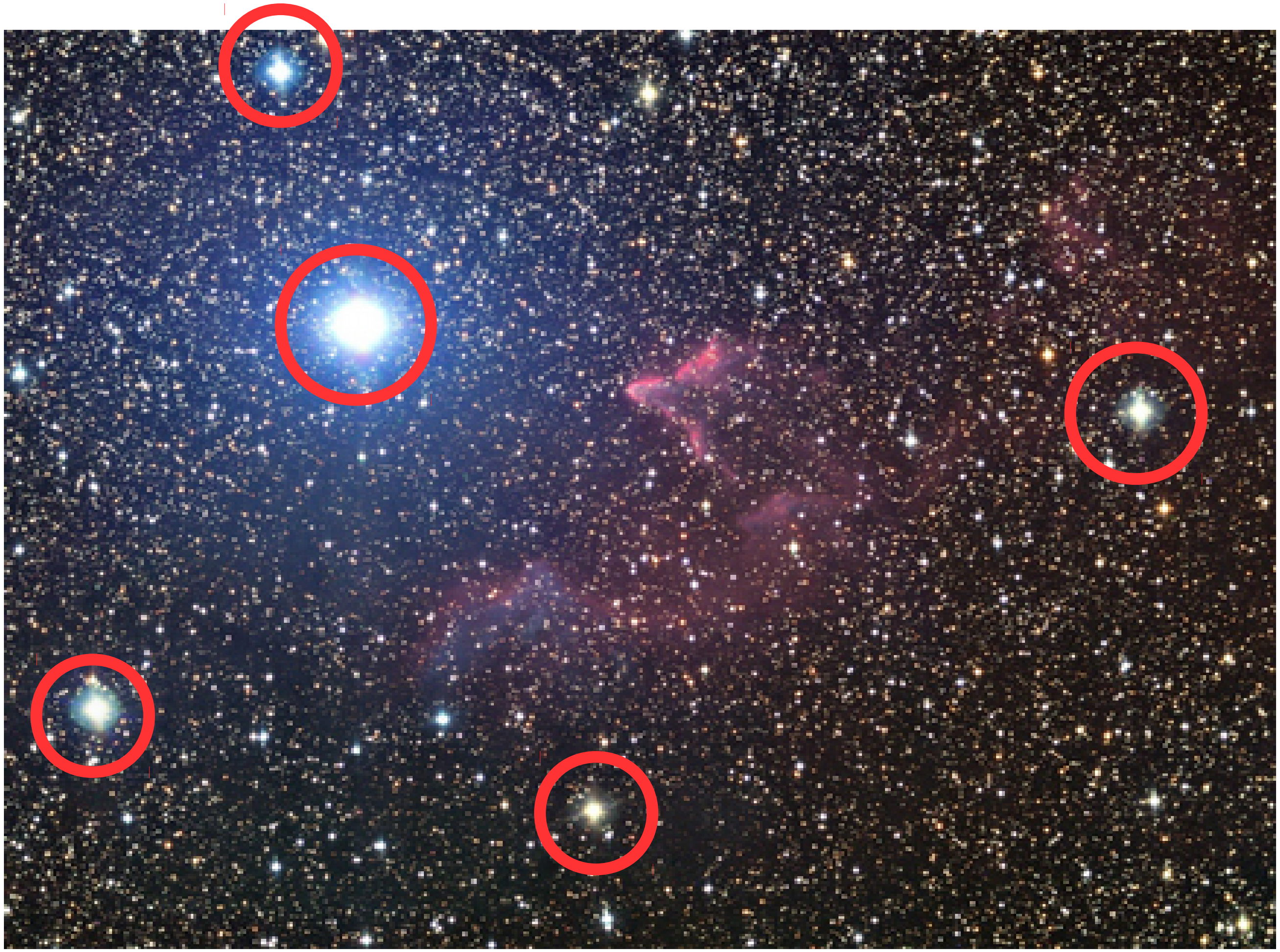
From COIN Residence Program #3 – in prep

The Big Picture

(data perspective)

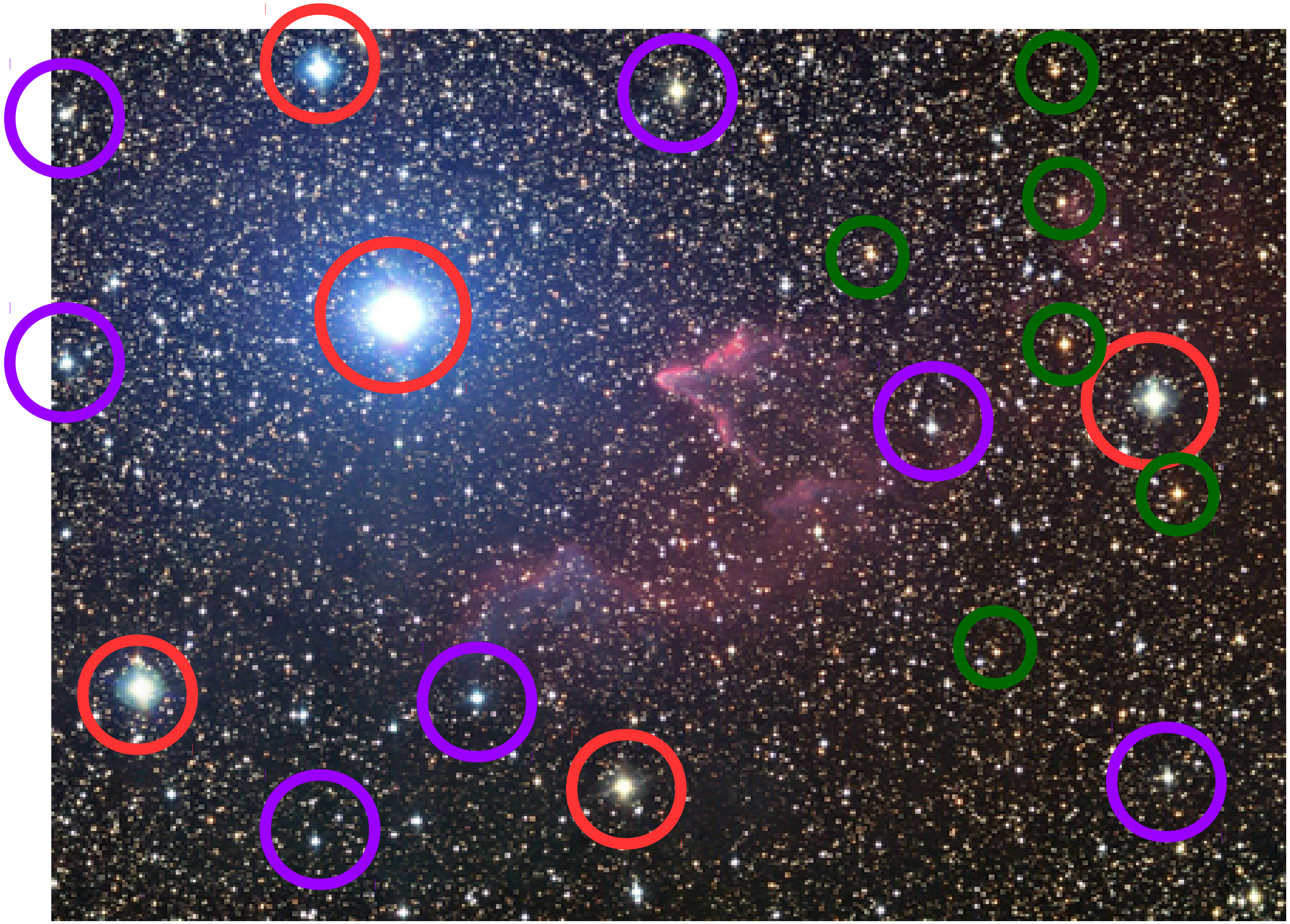
How are spectroscopic sets constructed?

Take spectra for learning and determine everything else



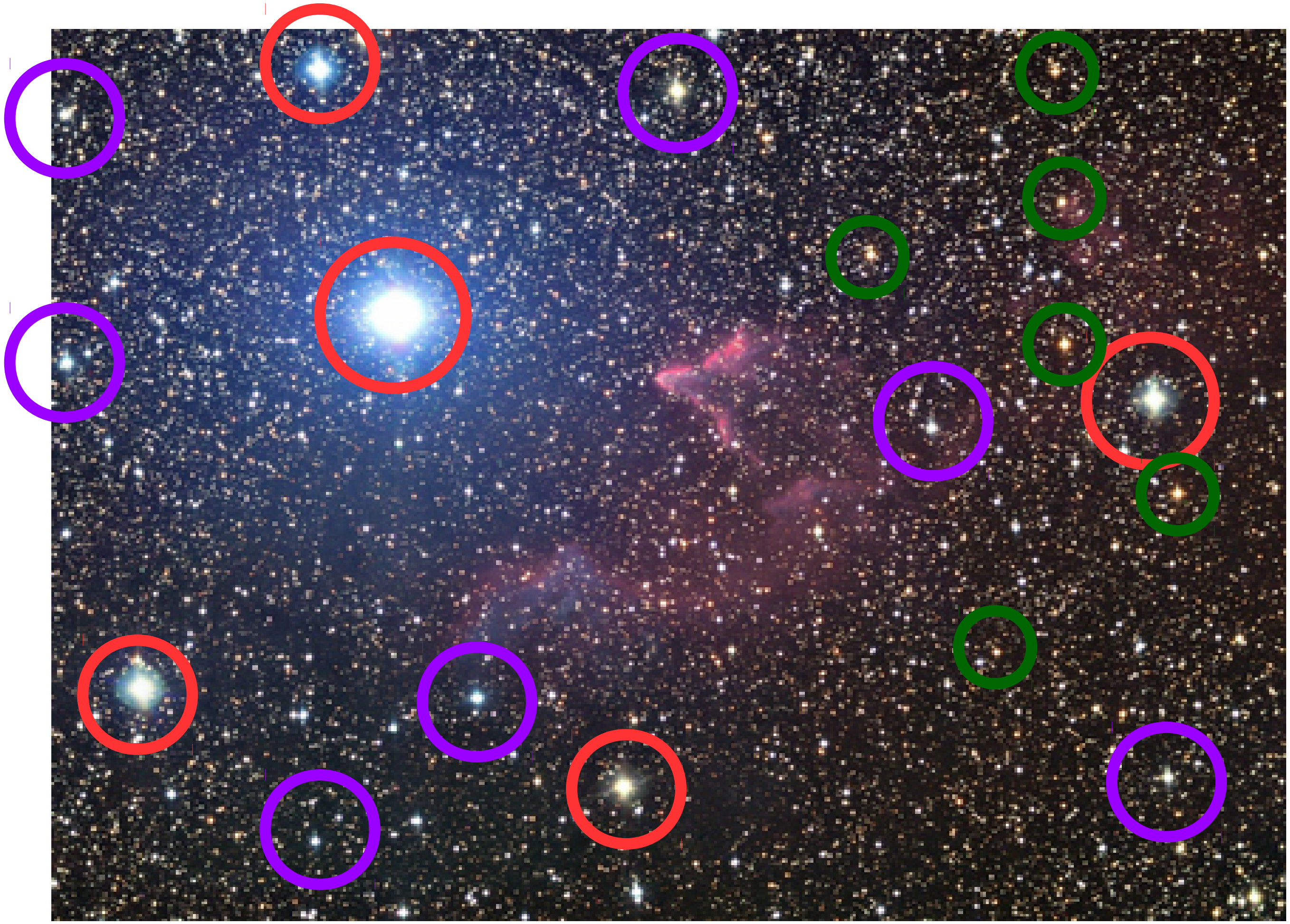
Alternative approach

Landmark selection + Active Learning



Alternative approach

Landmark selection + Active Learning



Take home message

Astronomy has evolved …

…there is still a long way to go

Astronomers won't do it alone

The REAL goal is HUMAN learning


THANK YOU!