

Machine Learning: Challenges and Opportunities in LSST

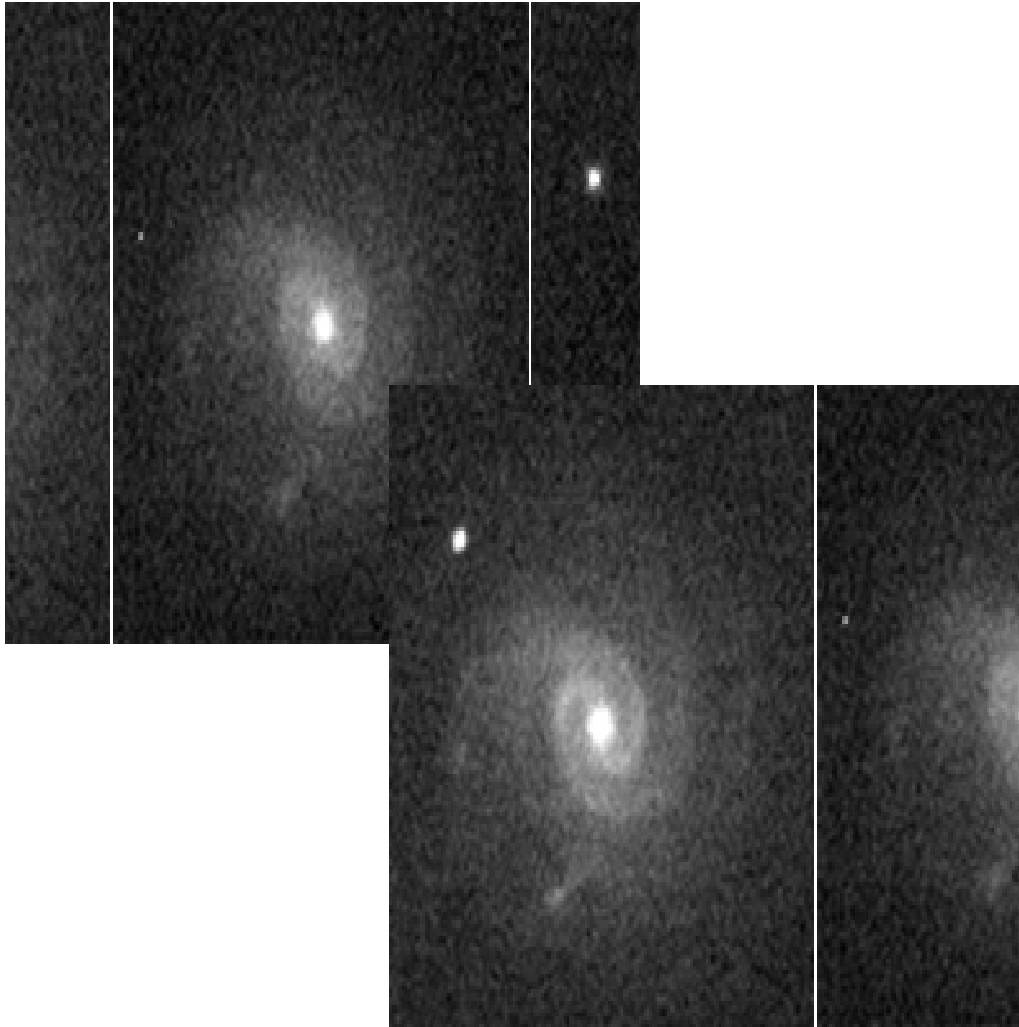
CAS at CC-IN2P3, September/2017

Emille E. O. Ishida

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

Types of data to be delivered by LSST

Images



Types of data to be delivered by LSST

1. Images



2. Catalogs

A collage of overlapping spreadsheets or catalogs, illustrating the types of data delivered by LSST. The spreadsheets contain columns of numerical data, likely representing astronomical parameters such as magnitude, position, and color. The text is repeated across multiple overlapping sheets, creating a sense of depth and volume of data.

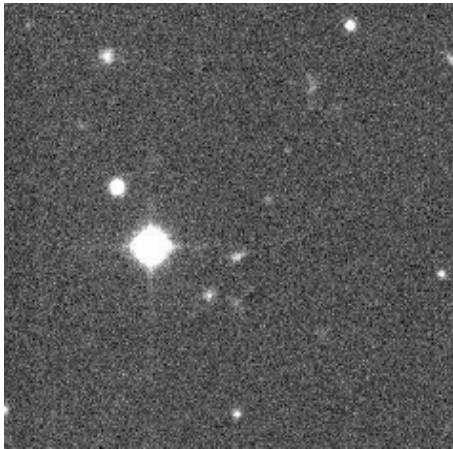
Either way... learn by example!



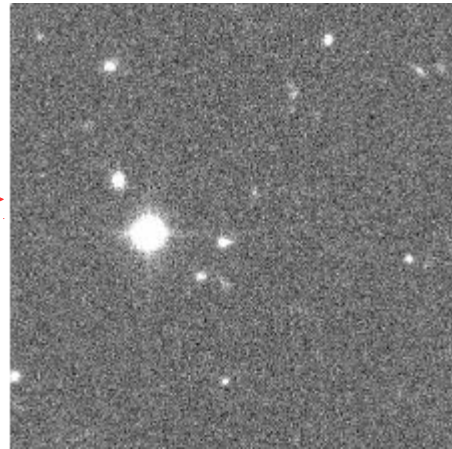
Machines will learn a lot!

Image data: Identification of Transients

Science Image



Background



Subtracted

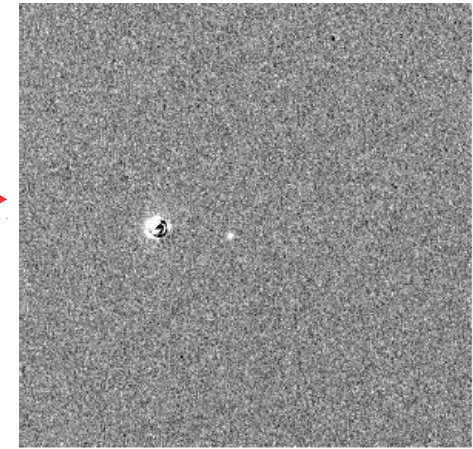


Image data: Identification of Transients

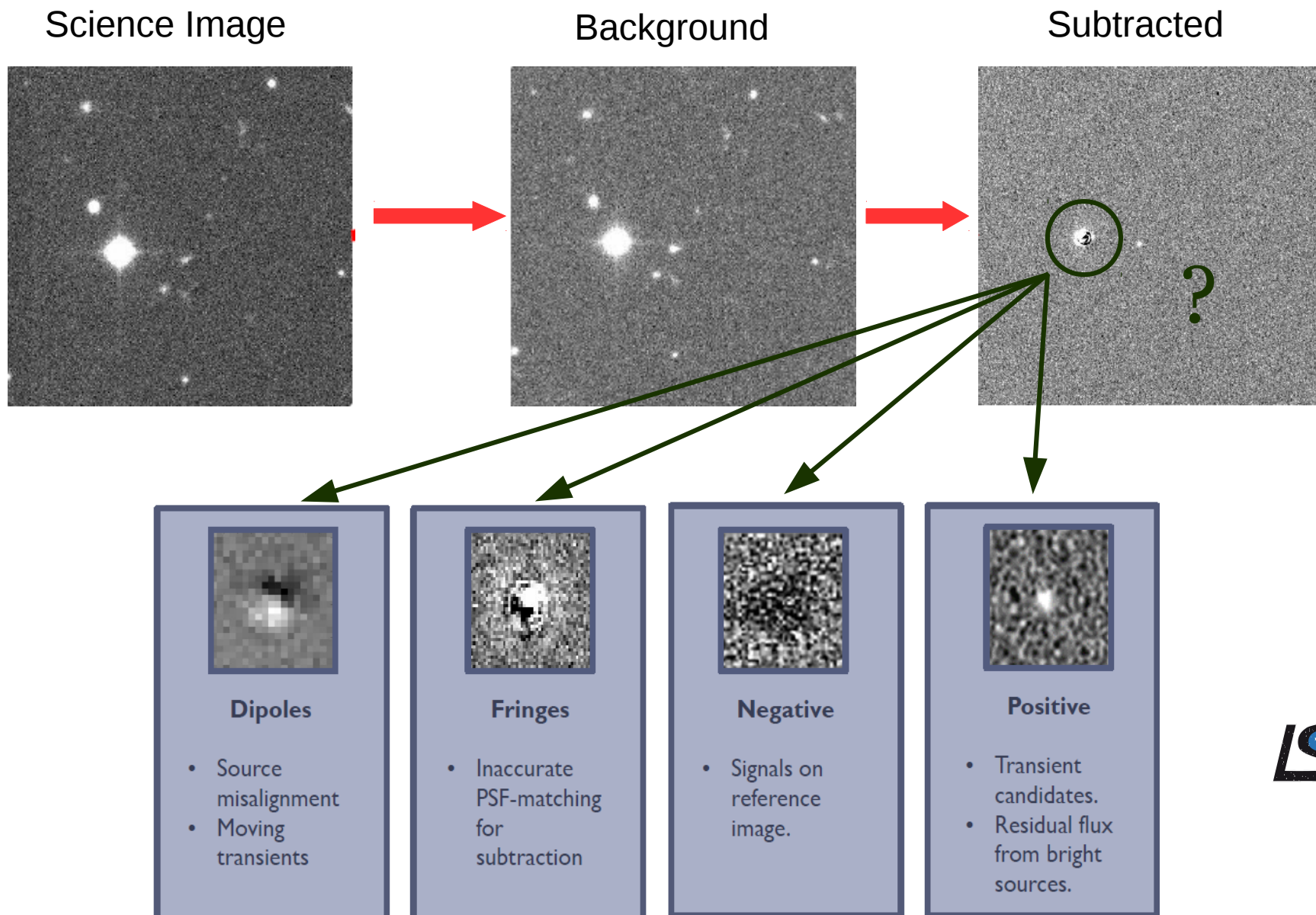
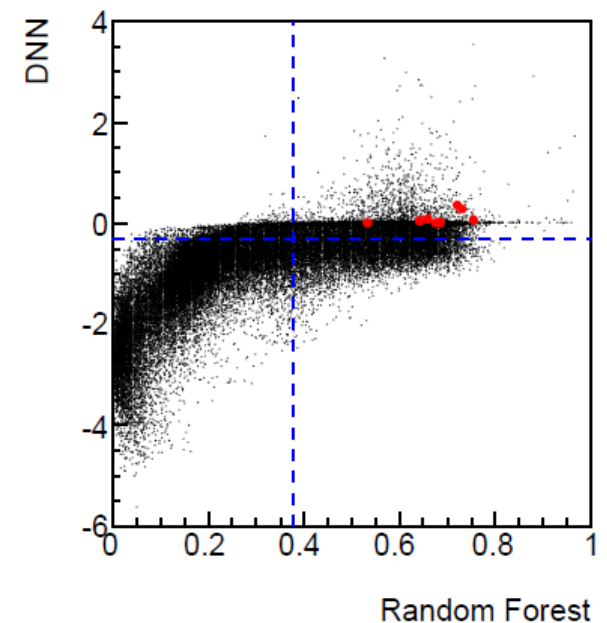
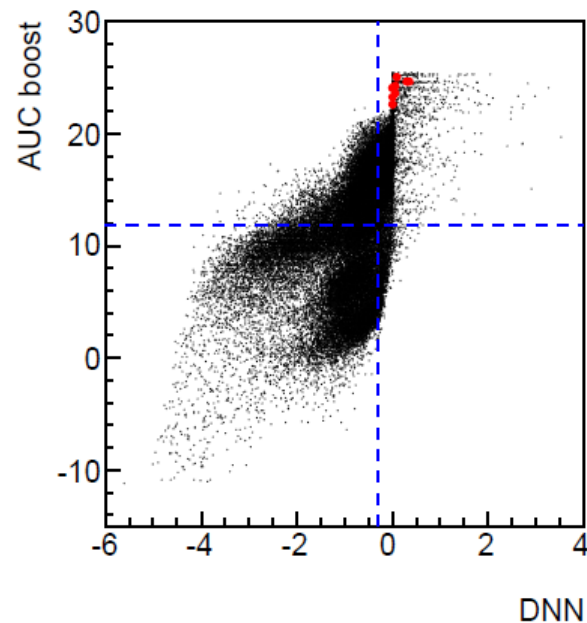
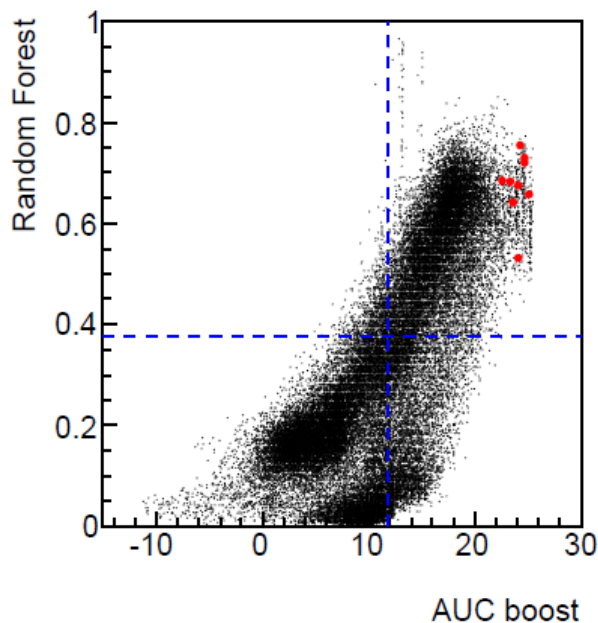


Image data: Identification of Transients

Example:

Committee of ML algorithms applied in the **identification of optical transients**

Application of Random Forest, Boosting and Deep Neural Networks

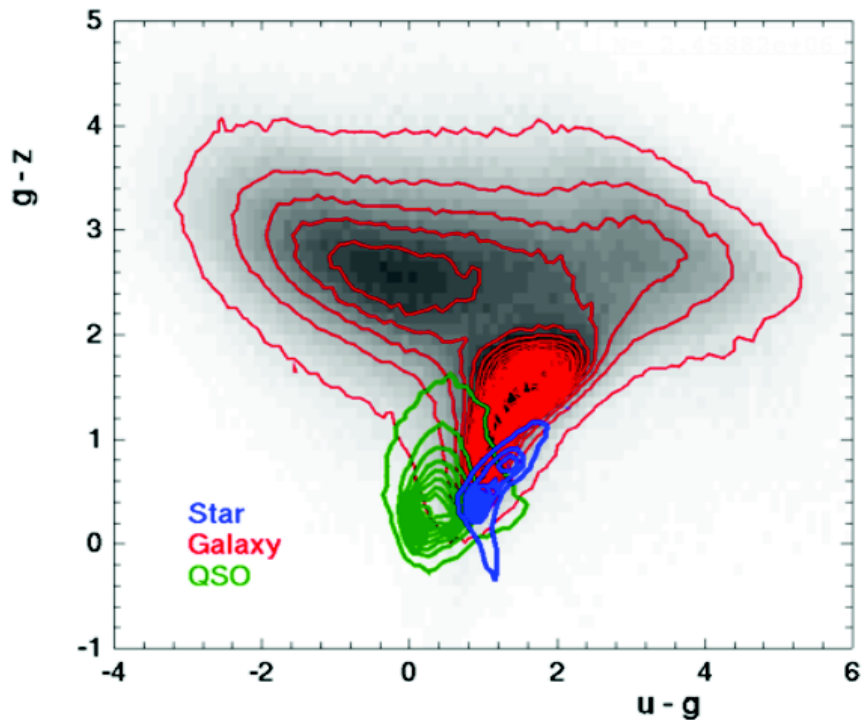


Catalog data: **Star/Galaxy separation**

Catalog data: **Star/Galaxy separation**



Colour-colour diagram

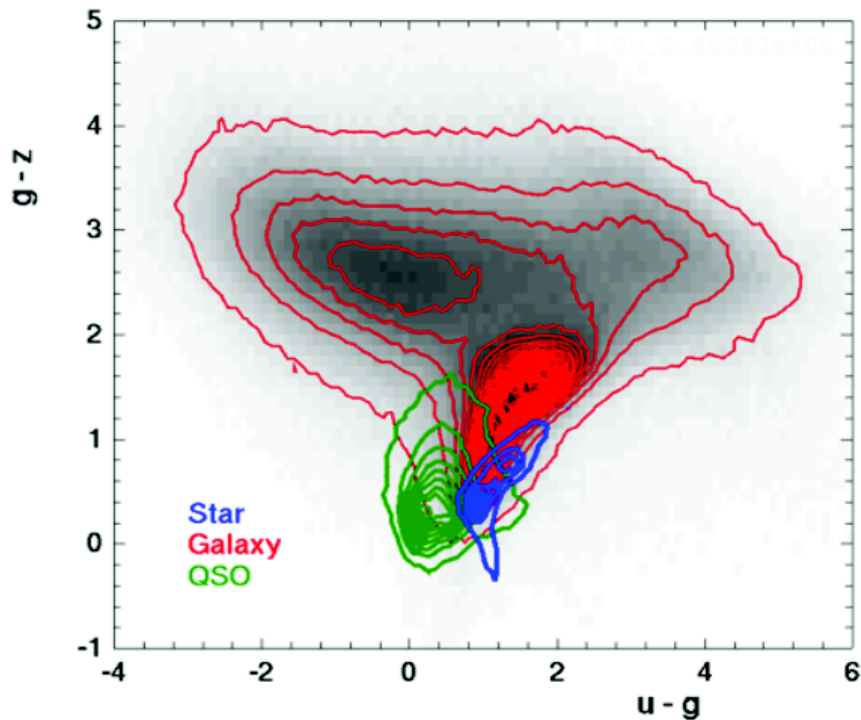


galaxies are redder
in average

Catalog data: Star/Galaxy separation

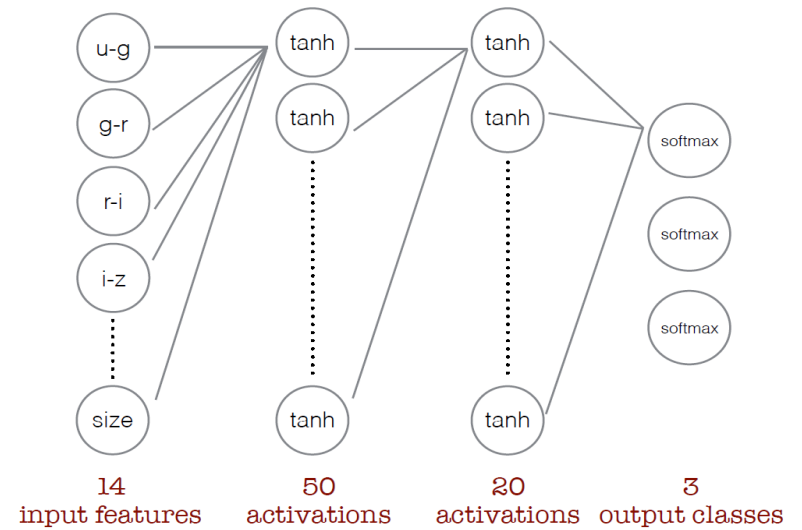


Colour-colour diagram



galaxies are redder
in average

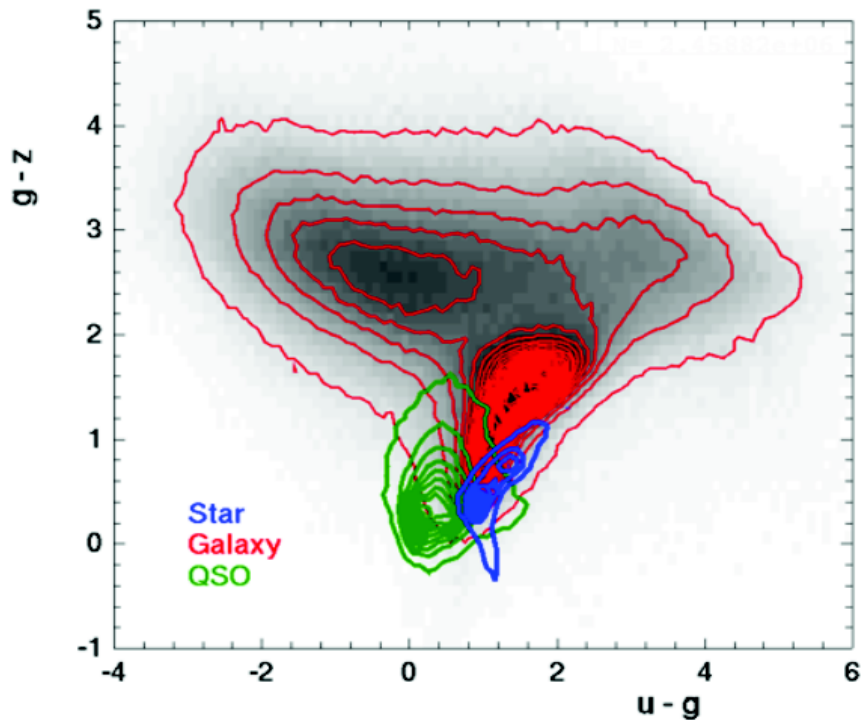
The perceptron



Catalog data: Star/Galaxy separation

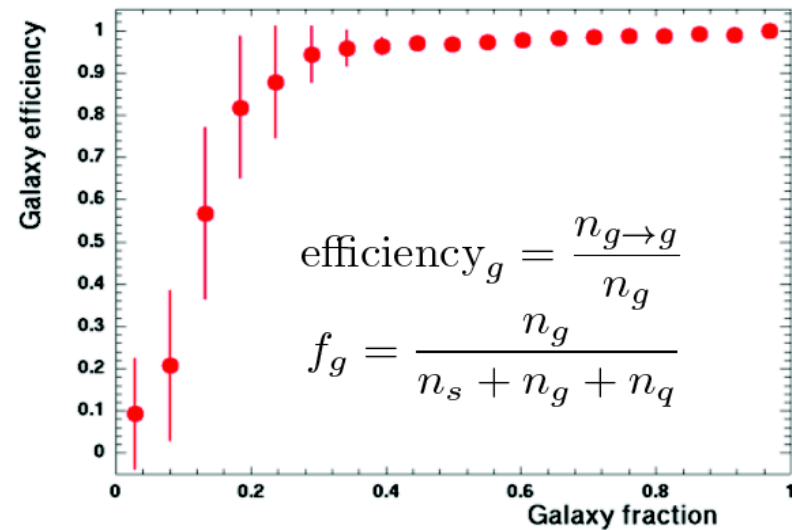
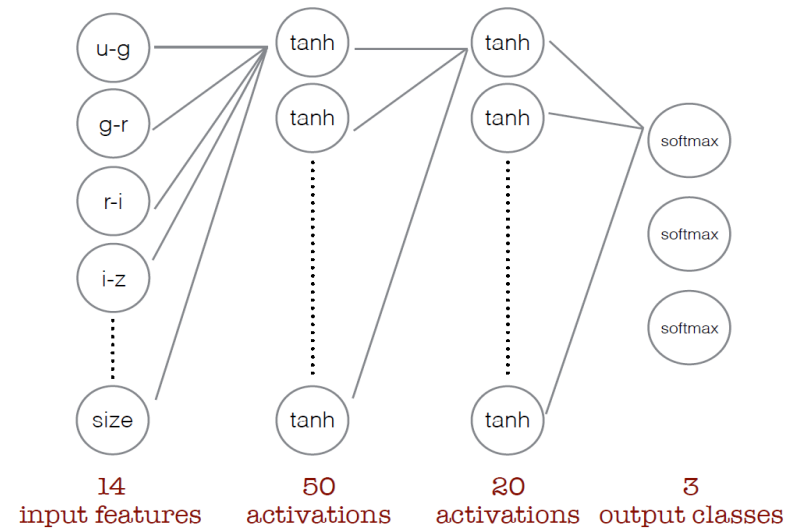


Colour-colour diagram



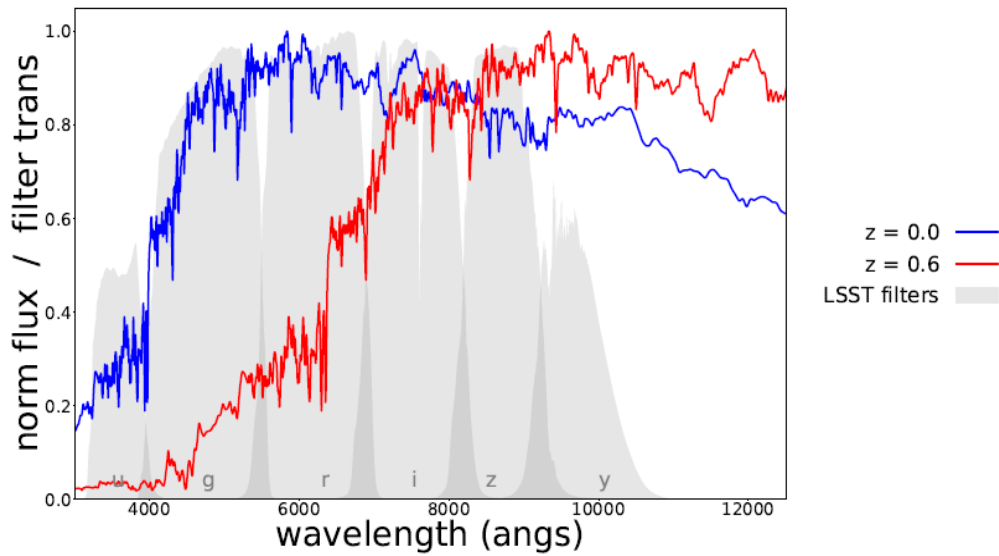
galaxies are redder
in average

The perceptron



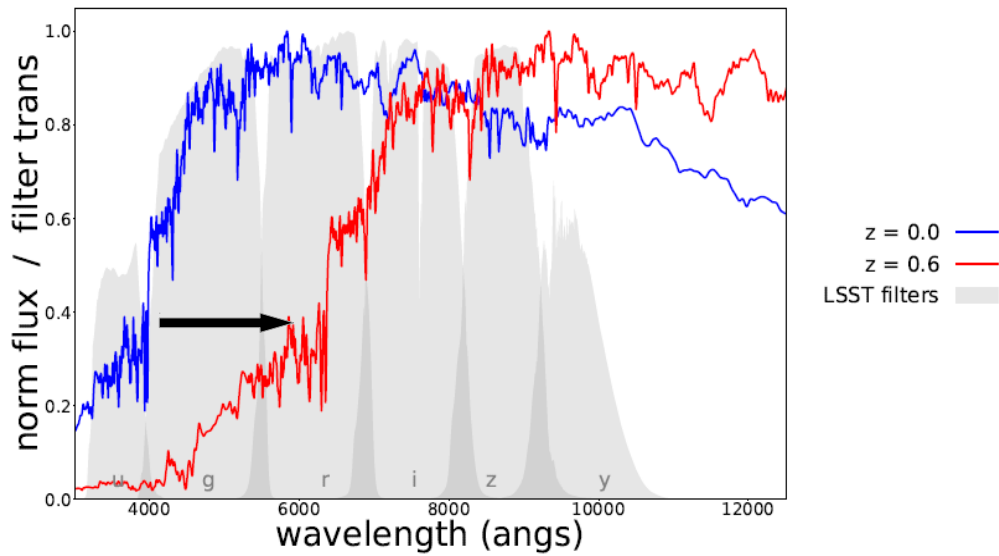
Catalog data: Determination of distances

spectroscopy



Catalog data: Determination of distances

spectroscopy

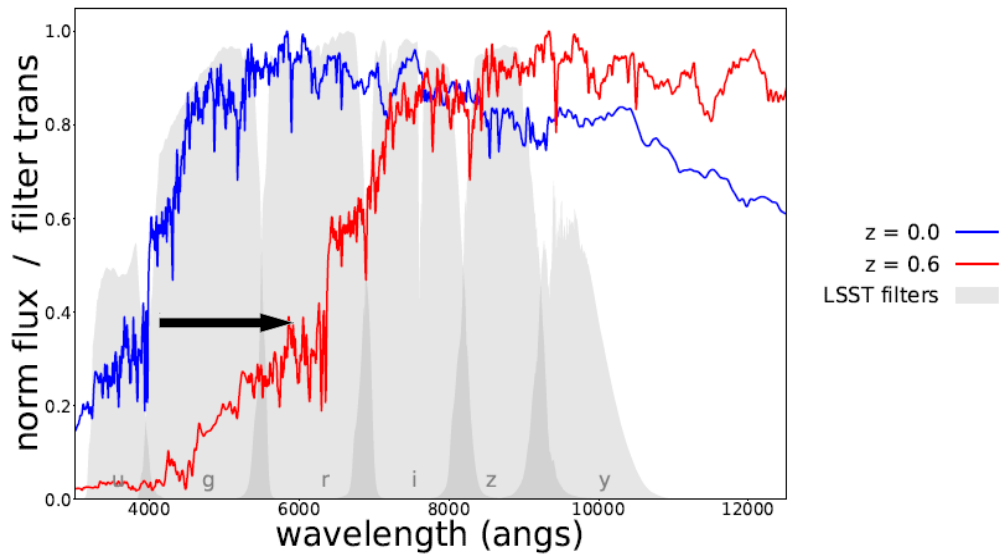


Catalog data: Determination of distances

Spectroscopy

High resolution

Expensive

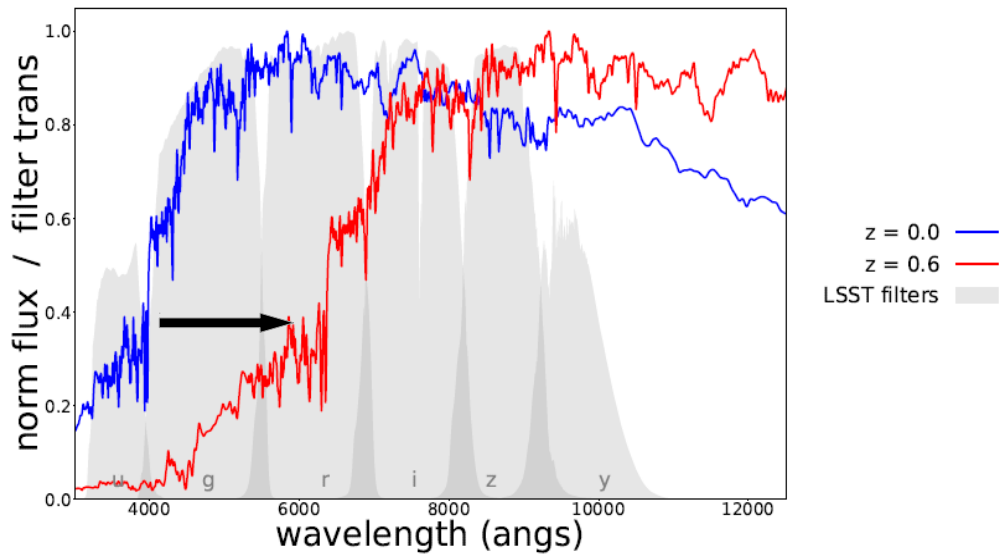


Catalog data: Determination of distances

spectroscopy

High resolution

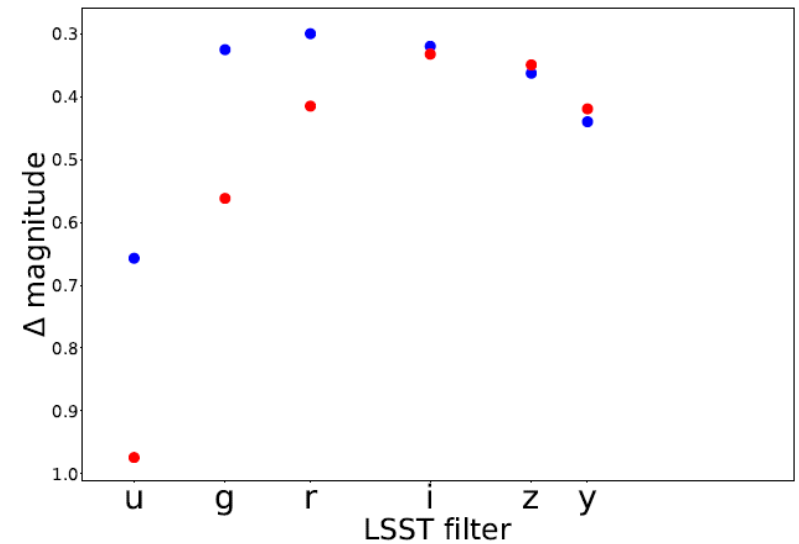
Expensive



photometry

Cheap

Low resolution

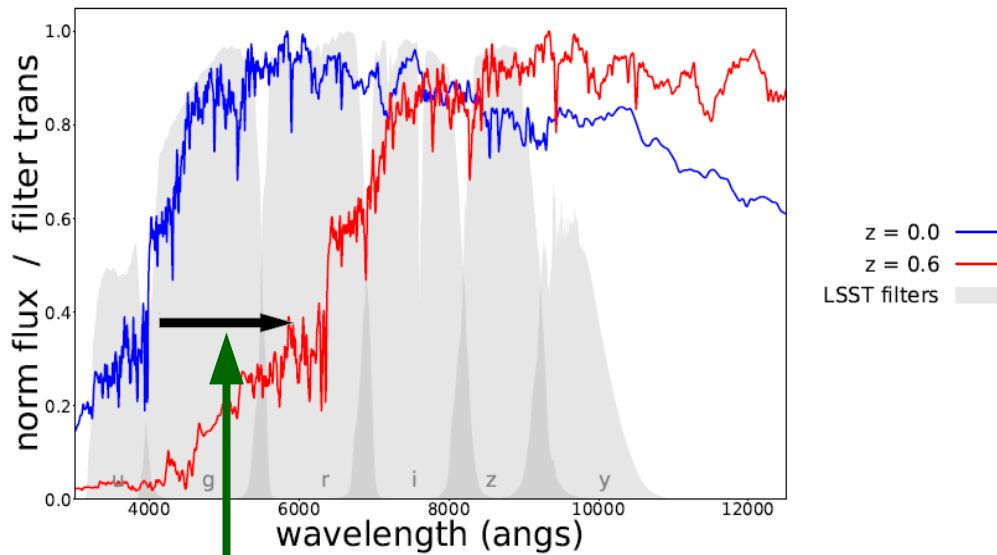


Catalog data: Determination of distances

spectroscopy

High resolution

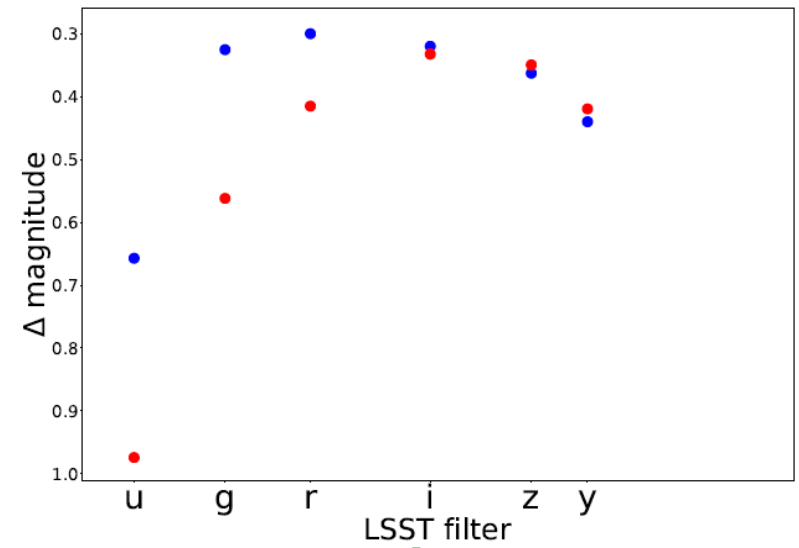
Expensive



photometry

Cheap

Low resolution



Catalog data: Determination of distances

Example:
Artificial Neural Networks

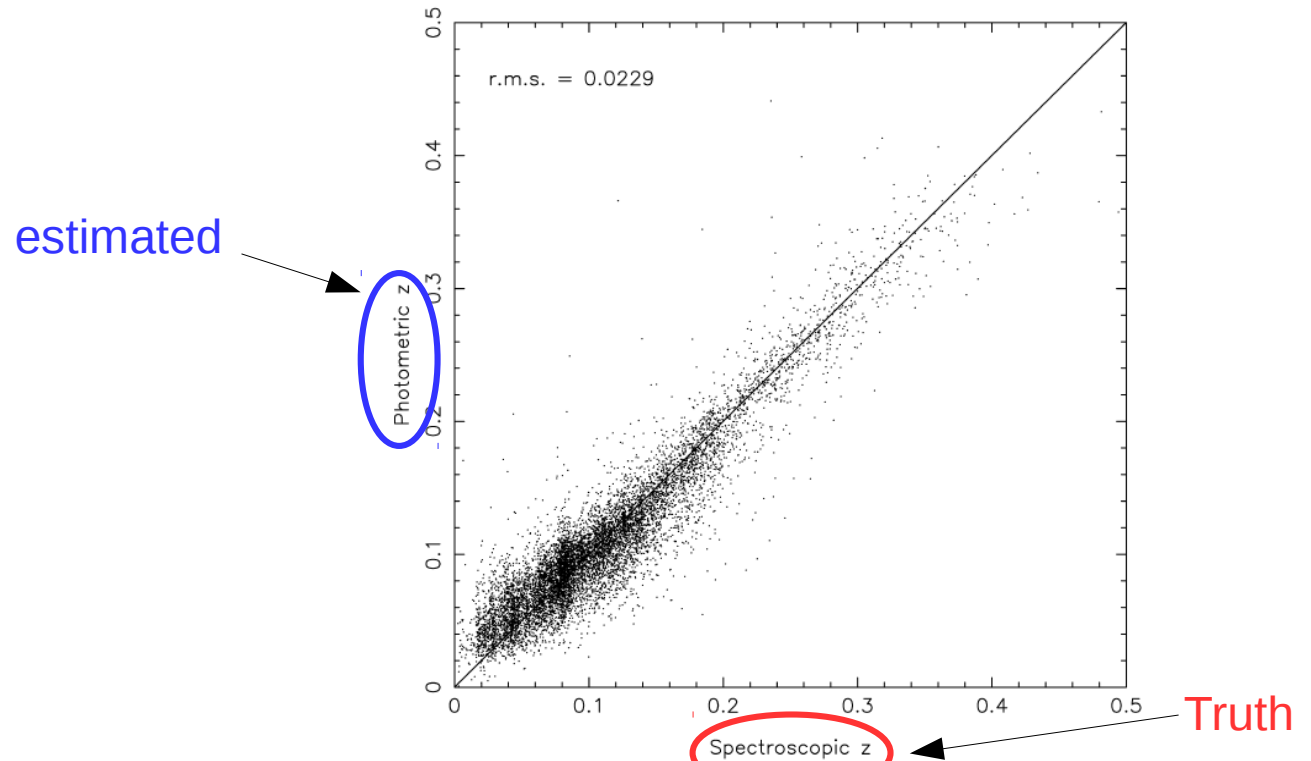
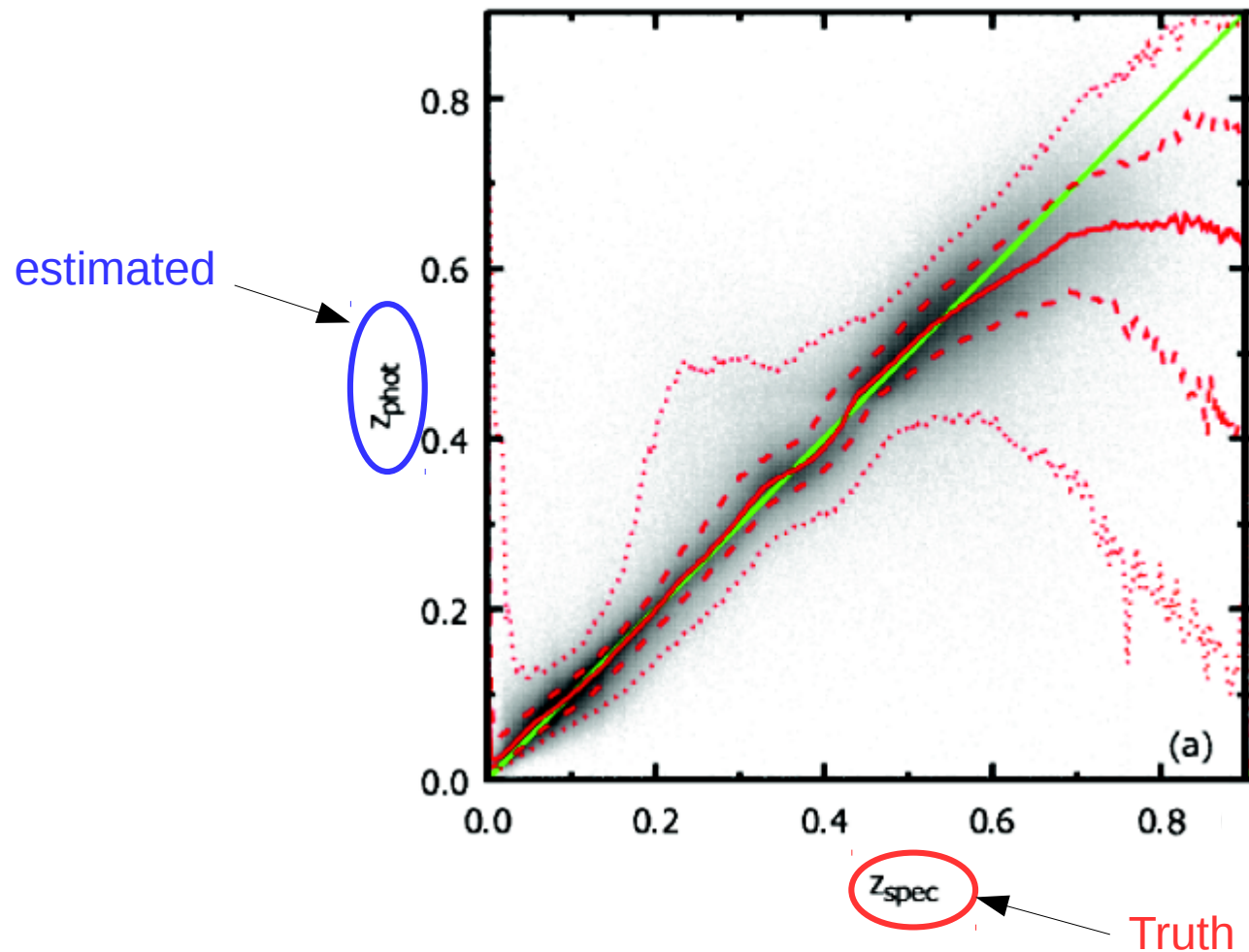


FIG. 2.— Spectroscopic vs. photometric redshifts for ANNz applied to 10,000 galaxies randomly selected from the SDSS EDR.

Catalog data: Determination of distances

Example:
Local Linear Regression

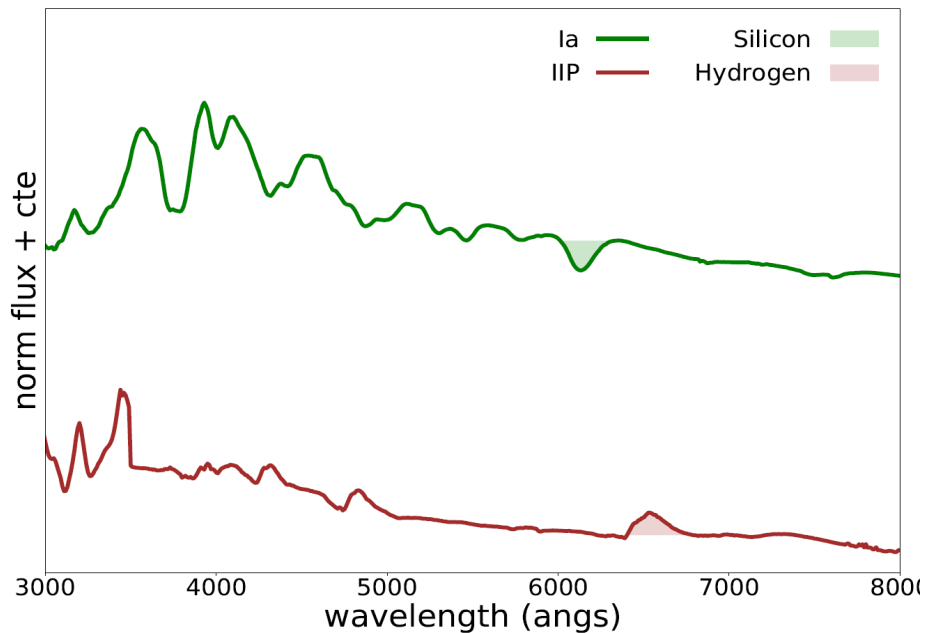


Catalog data: Supernova Classification

spectroscopy

High resolution

Expensive

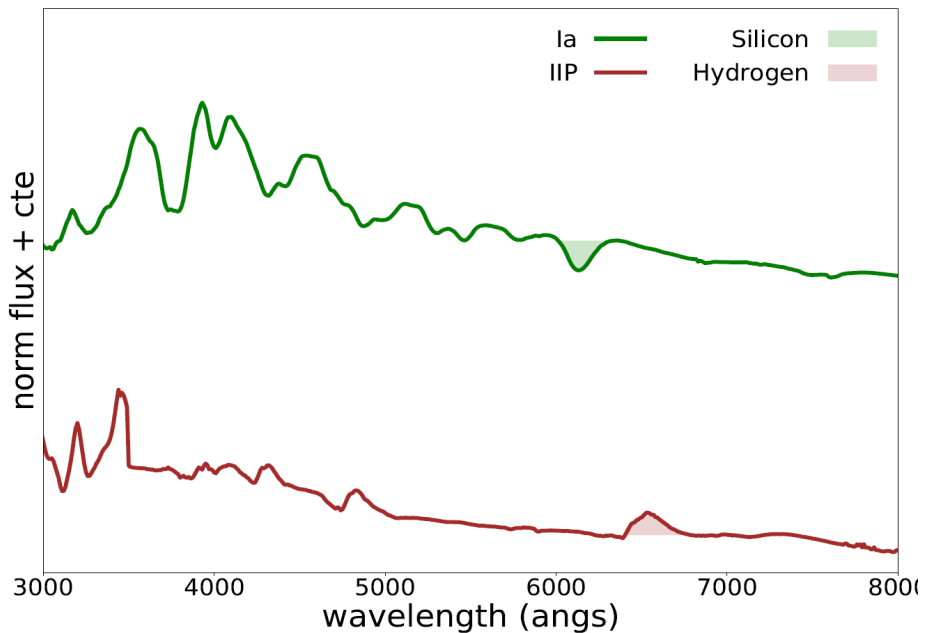


Catalog data: Supernova Classification

spectroscopy

High resolution

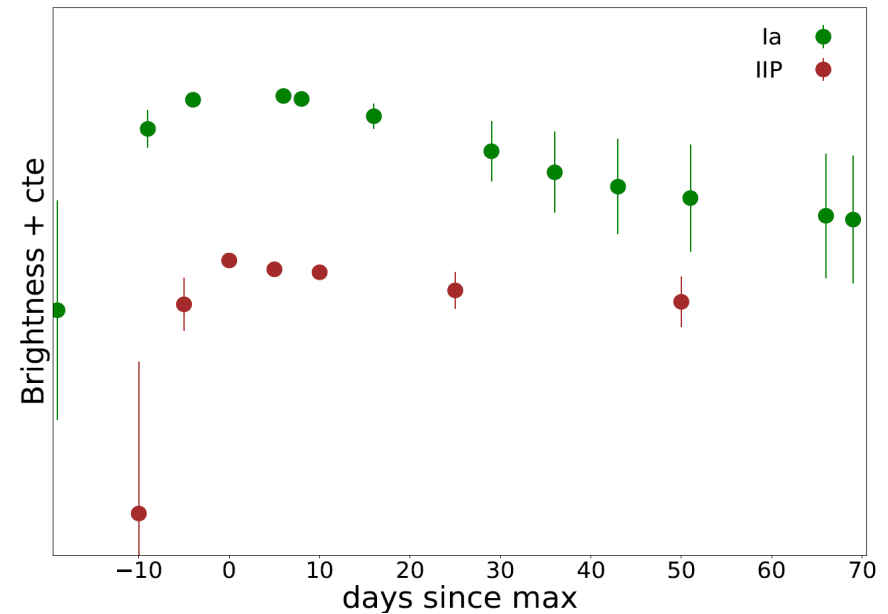
Expensive



photometry

Cheap

Low resolution

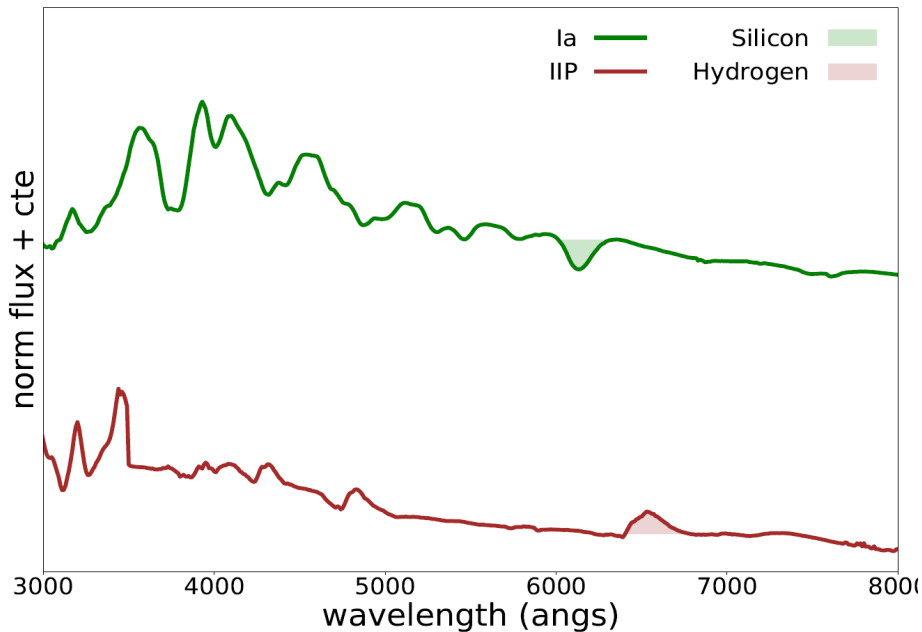


Catalog data: Supernova Classification

spectroscopy

High resolution

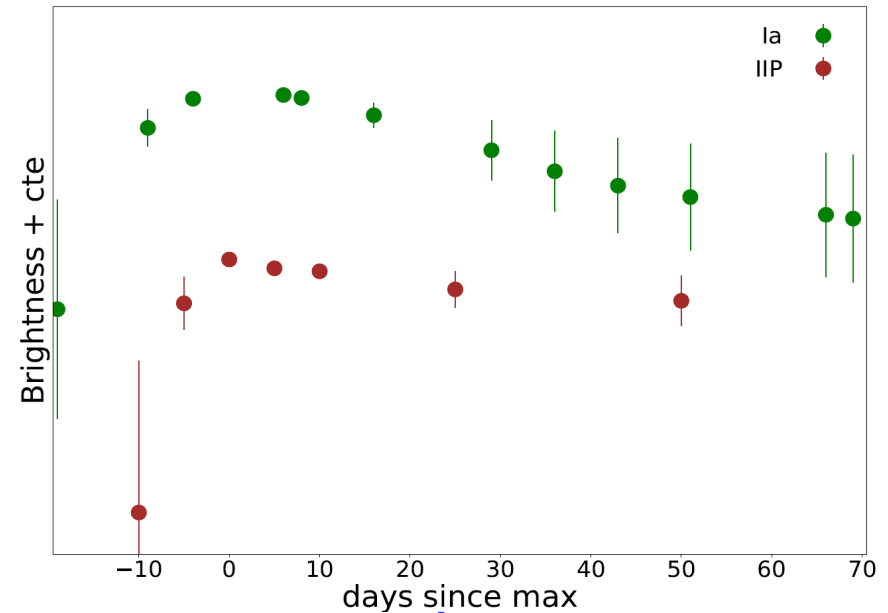
Expensive



photometry

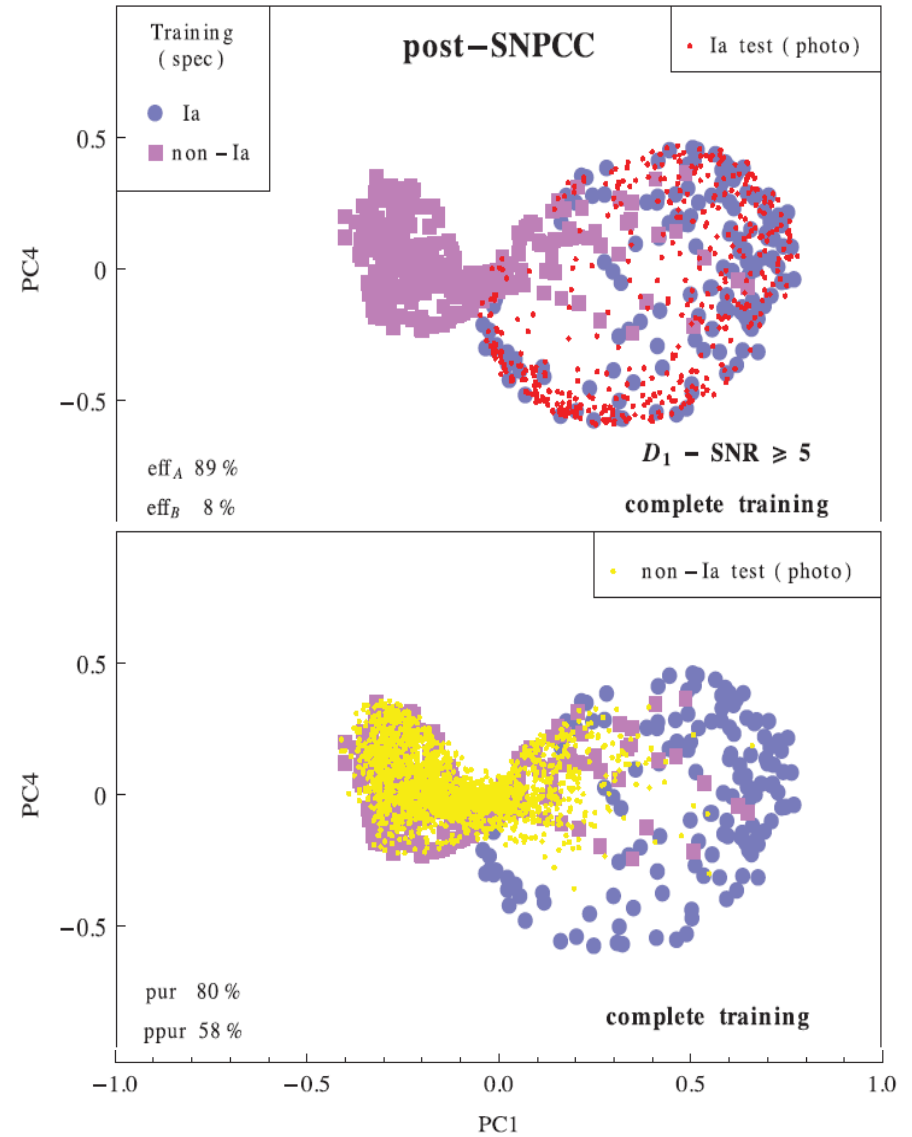
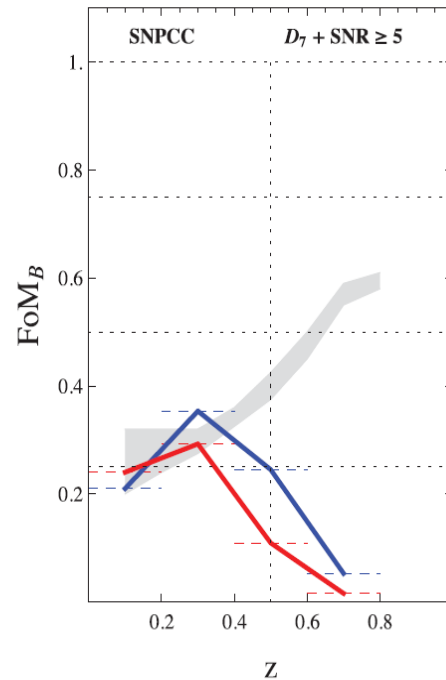
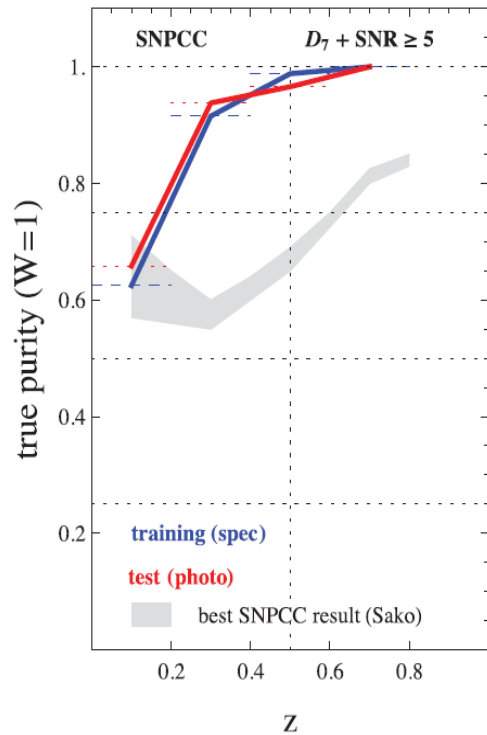
Cheap

Low resolution



Catalog data: Supernova Classification

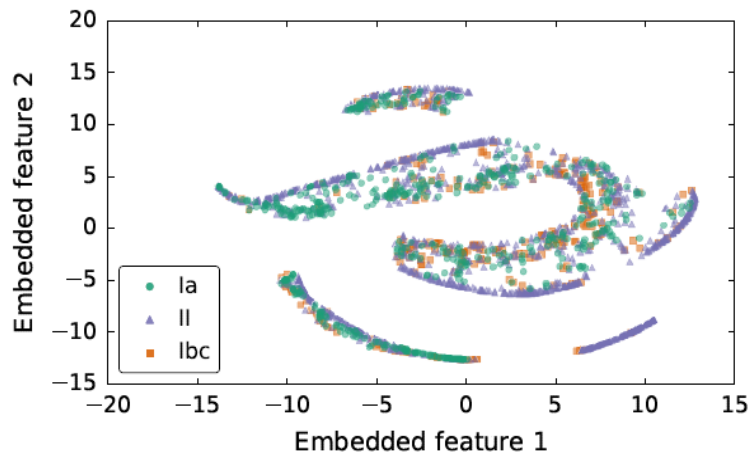
Example:
Kernel PCA + Nearest Neighbor



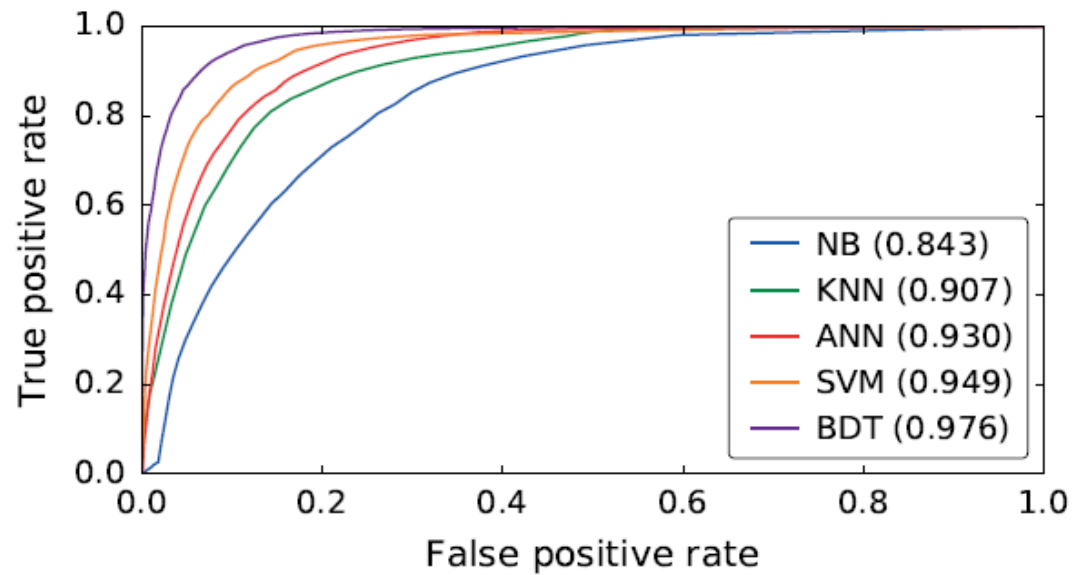
Catalog data: Supernova Classification

Example:

Wavelet decomposition + Boosted Decision Trees



(a) SALT2 without redshift (5 features)



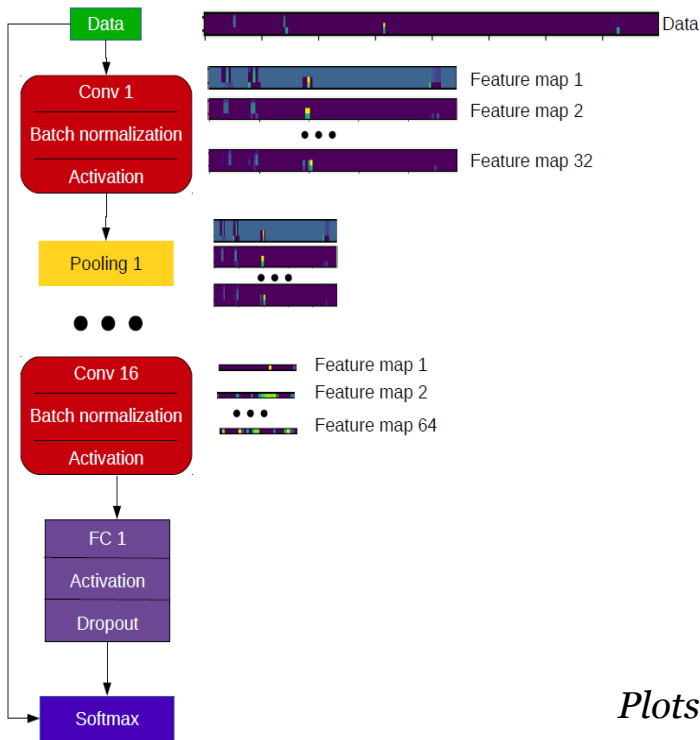
(a) SALT2 model, no redshift

Catalog data: Supernova Classification

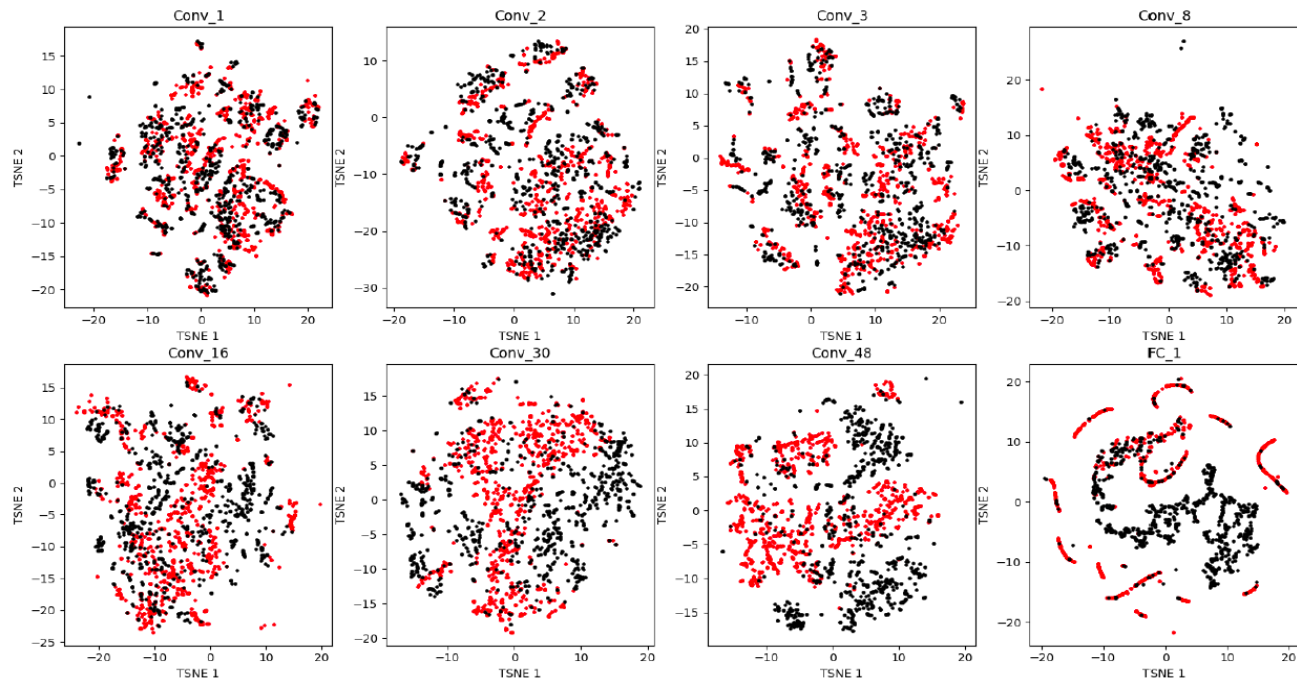
Example:
Convolution Neural Network



Network architecture

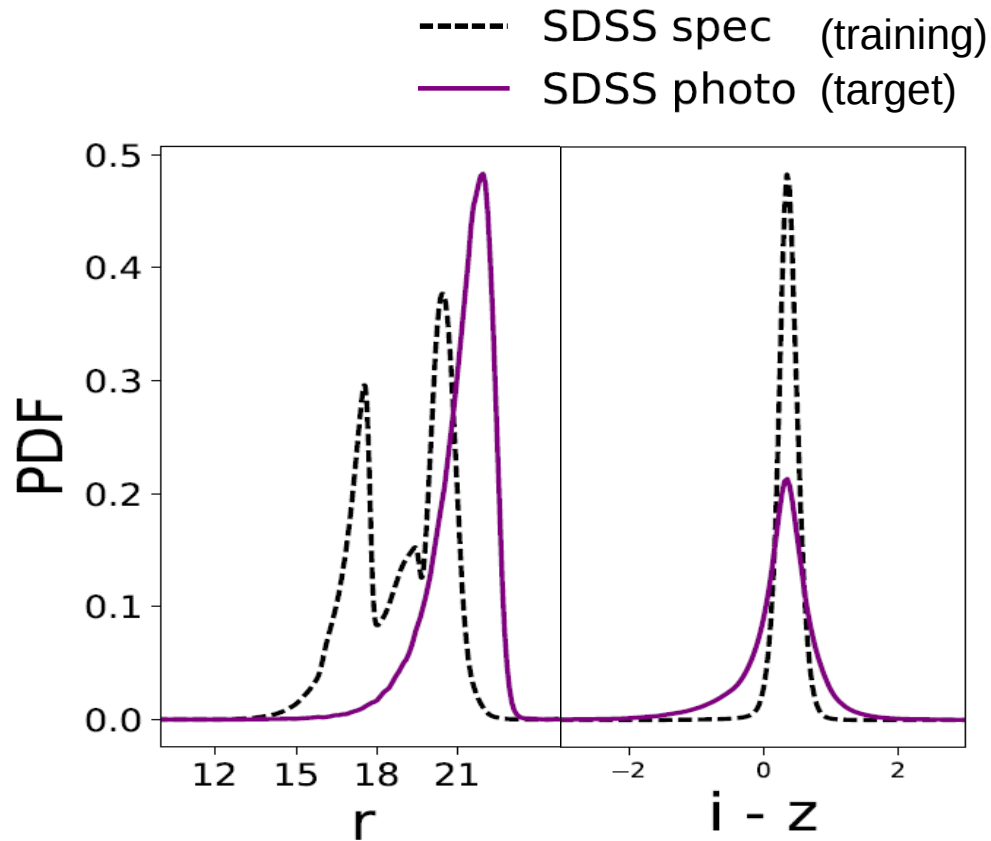


Classification along the network



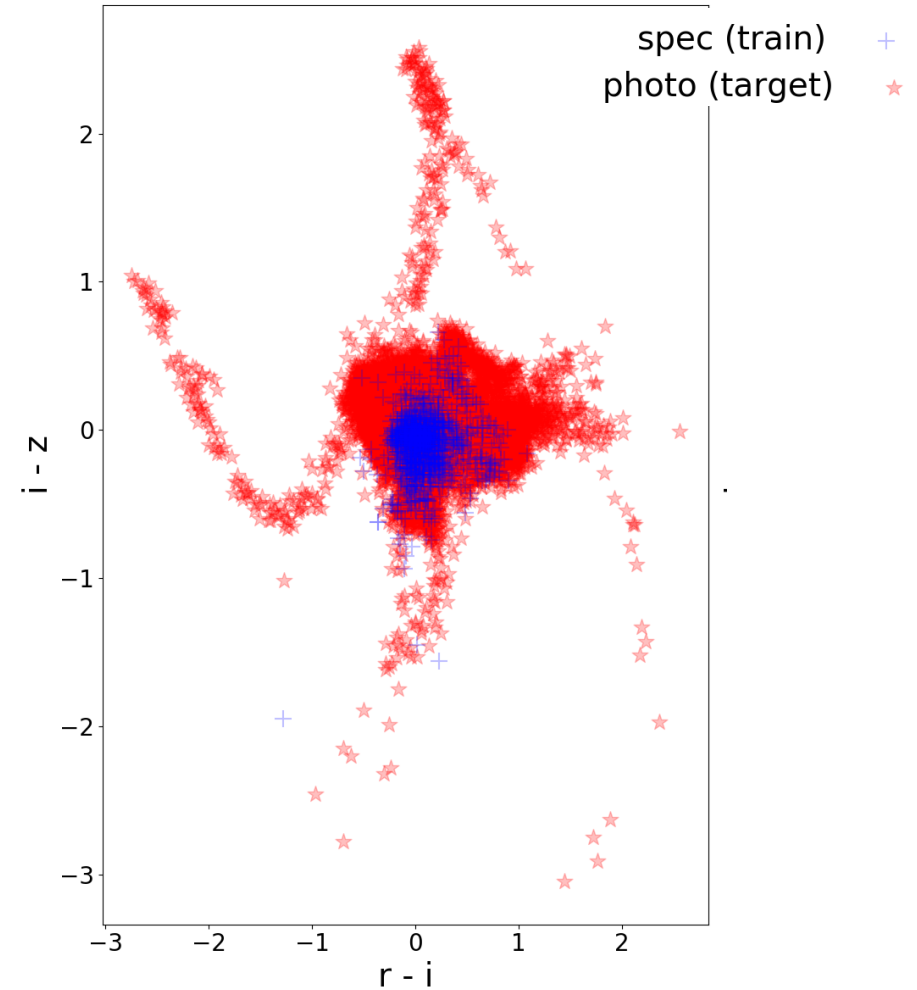
problem: Representativeness

Determination of distances



Beck, Lin, Ishida et al., (2017) MNRAS

Supernova Classification



SNPCC data, Kessler et al., 2010 – plot by Ishida

Representativeness
is a bottleneck that
must be addressed
in order to
optimize scientific
output from LSST
data!

Naturally accommodates

Novelty detection!

Unsupervised Learning



Realistic expectations require realistic simulations



**PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES
CLASSIFICATION CHALLENGE (PLASTICC)**

Realistic expectations require realistic simulations

Goals:

- 1) Increase the participation of non-astronomers
- 2) Facilitate posterior usage of results
- 3) Answer multiple questions



**PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES
CLASSIFICATION CHALLENGE (PLASTICC)**

Realistic expectations require realistic simulations

Goals:

- 1) Increase the participation of non-astronomers
- 2) Facilitate posterior usage of results
- 3) Answer multiple questions

*A public data challenge
built from state of the art
Transient simulations as observed by LSST*



**PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES
CLASSIFICATION CHALLENGE (PLASTICC)**

Realistic expectations require realistic simulations

Goals:

- 1) Increase the participation of non-astronomers
- 2) Facilitate posterior usage of results
- 3) Answer multiple questions

To be released in early 2018

*A public data challenge
built from state of the art
Transient simulations as observed by LSST*



**PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES
CLASSIFICATION CHALLENGE (PLASTICC)**

Machine Learning Challenges and Opportunities

in LSST

- Abolish (or lower the importance of) visual screening in the pipeline!
- Get ready in time!
Algorithm, analysis
Spectroscopic follow-up planning

To be released in early 2018



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

- Diversify your ideas ...
... get more people involved!

Machine Learning Challenges and Opportunities

in LSST

- Abolish (or lower the importance of) visual screening in the pipeline!
- Get ready in time!
Algorithm, analysis
Spectroscopic follow-up planning

To be released in early 2018



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

- Diversify your ideas ...
... get more people involved!

- Large, complex data set available at the catalog level!
- Fertile ground for development of new ML algorithms
- Real, productive interdisciplinarity is not optional!

Knowledge Discovery in Databases

Science + Methods

Extra slides

Alternative approach: **Active Learning**

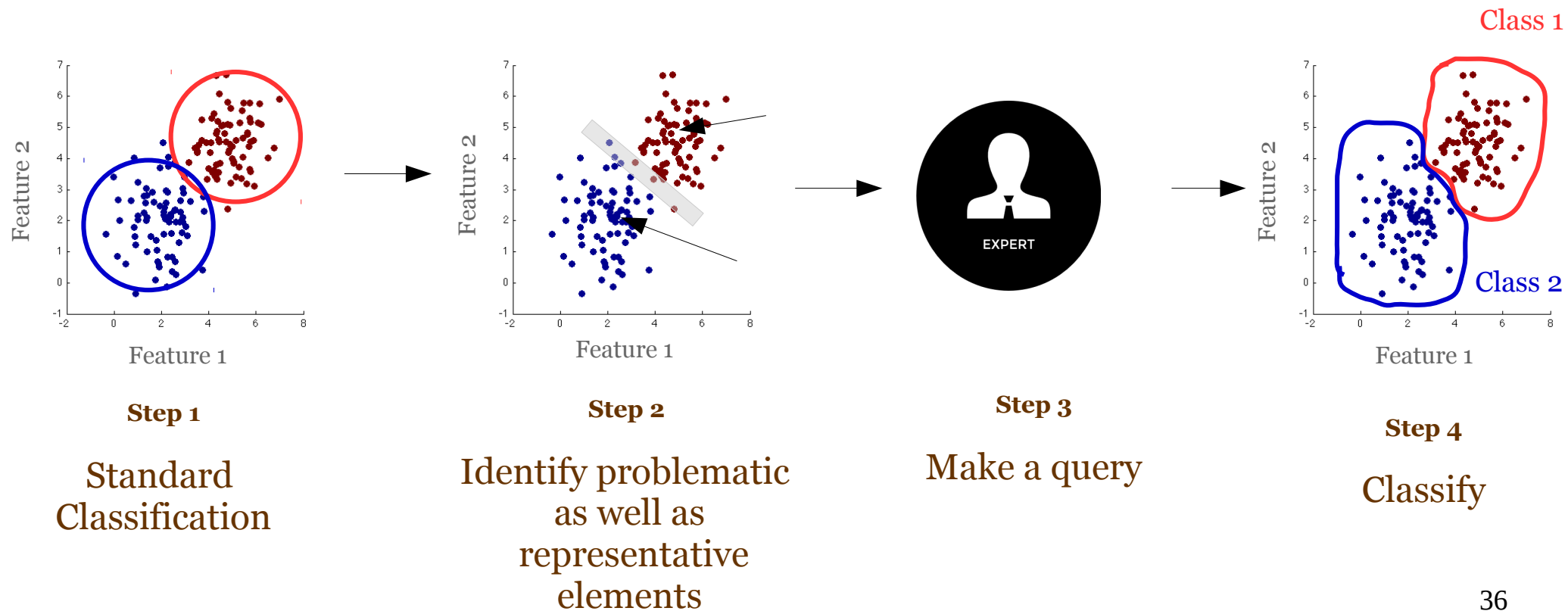
or Optimal Experimental Design

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

Alternative approach: **Active Learning**

or Optimal Experimental Design

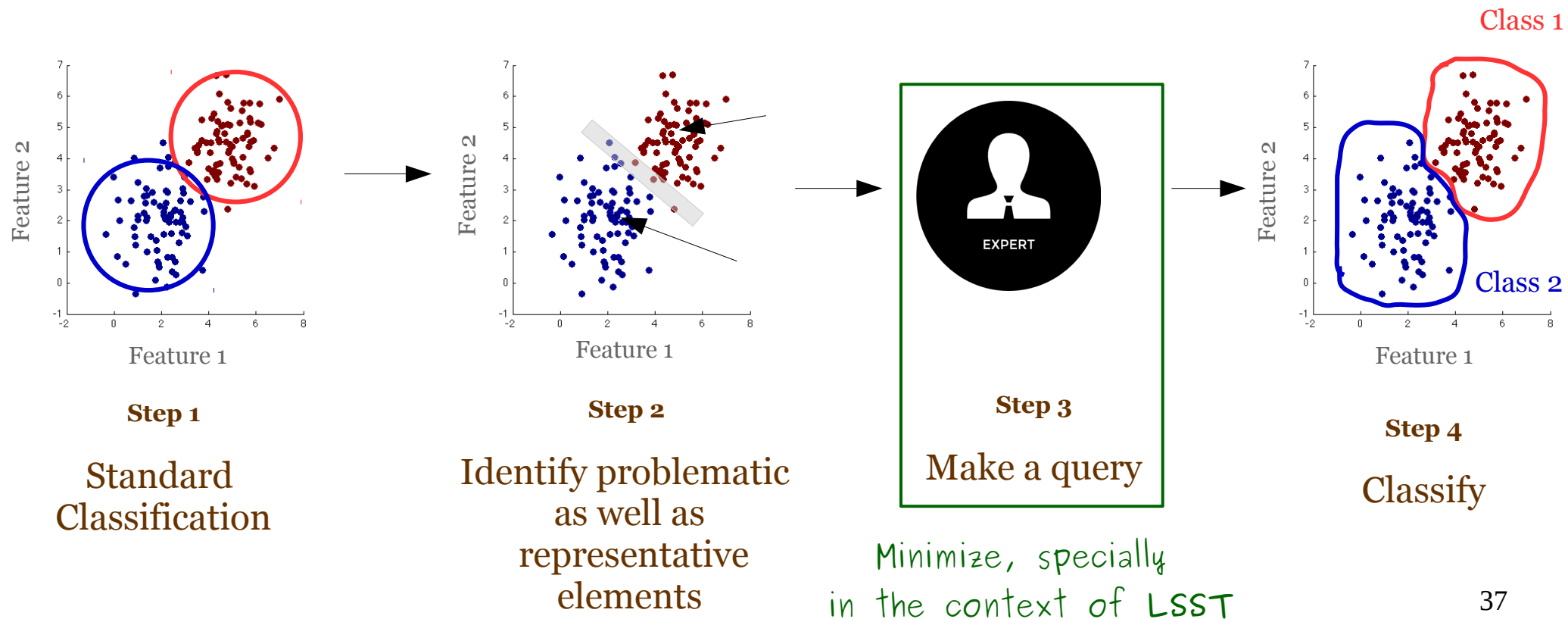
“Can machines learn with **fewer labeled** training instances if they are allowed to ask questions?”



Alternative approach: **Active Learning**

or Optimal Experimental Design

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



Alternative approach: Active Learning

or Optimal Experimental Design

...for Supernova Classification!



Complete data set



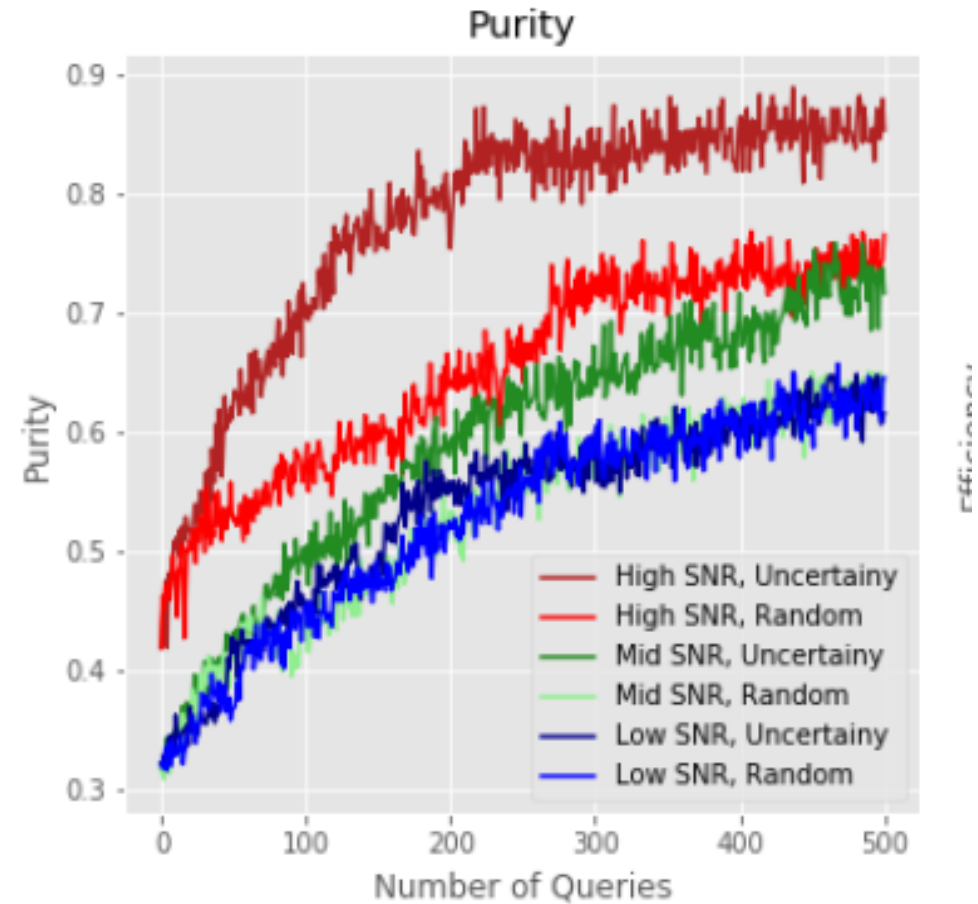
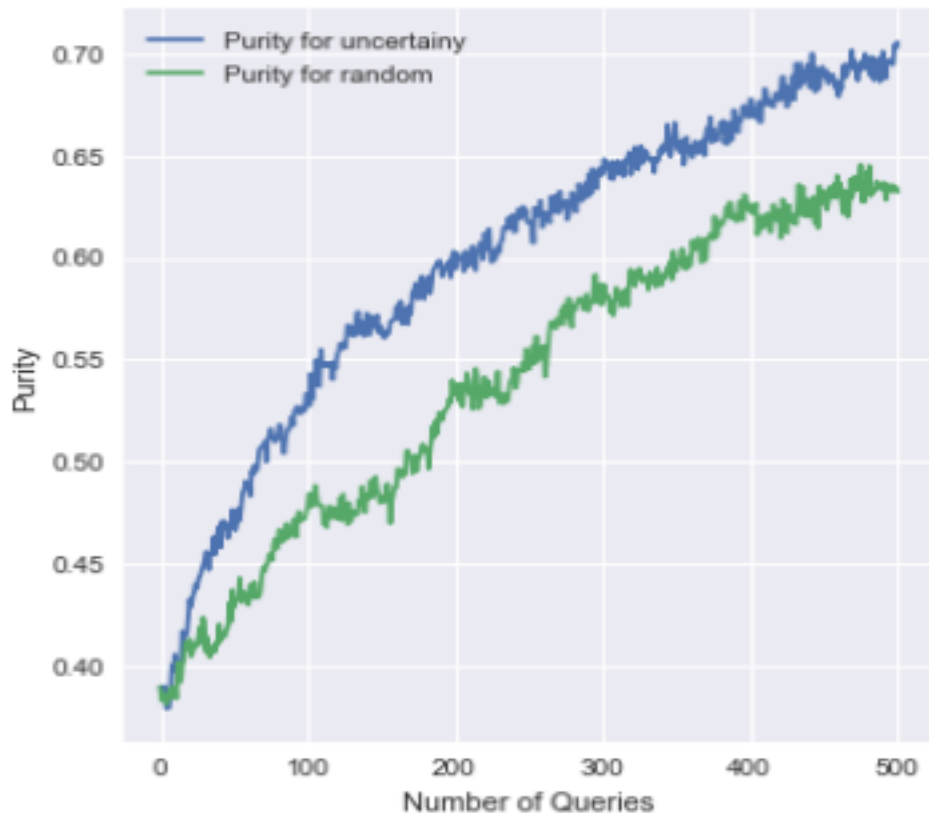
Alternative approach: Active Learning

or Optimal Experimental Design

...for Supernova Classification!



Complete data set



There is no
miracle: lower
quality data
require more
effort in analysis/
design!

Image data: **Galaxy morphology classification**






<https://www.galaxyzoo.org/>

Image data: **Galaxy morphology classification**



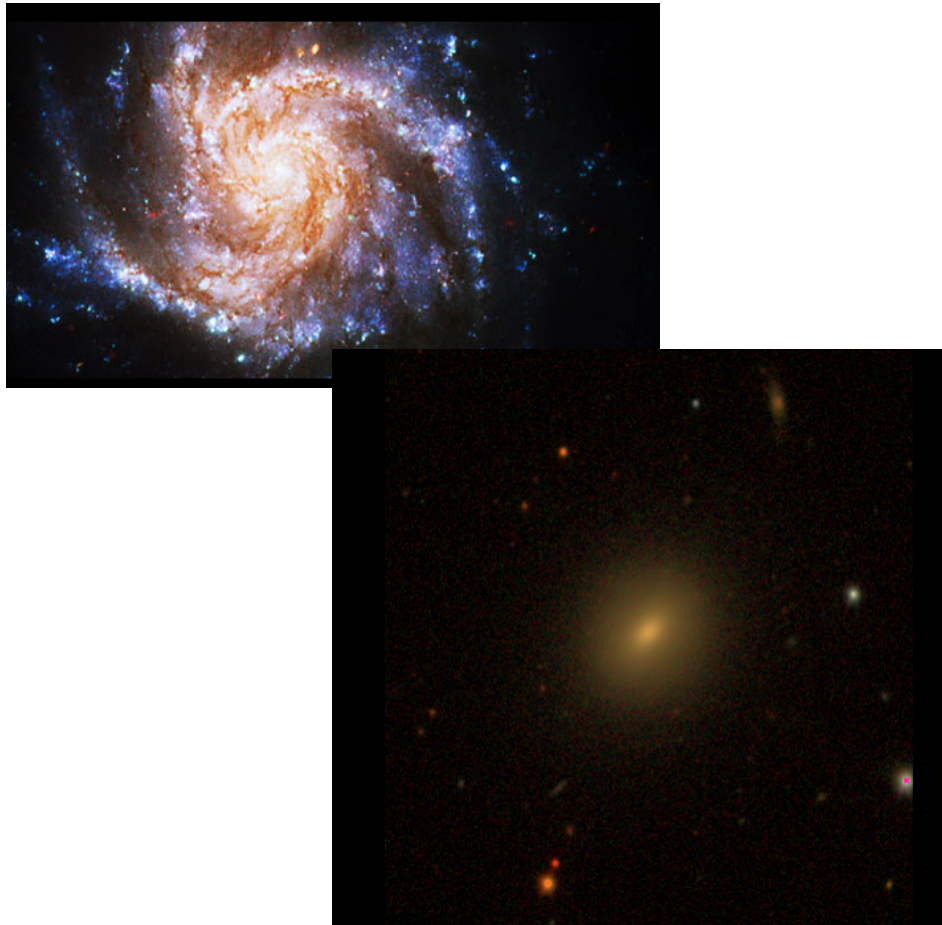
GALAXY ZOO

SHAPE
Is the galaxy simply smooth and rounded, with no sign of a disk?

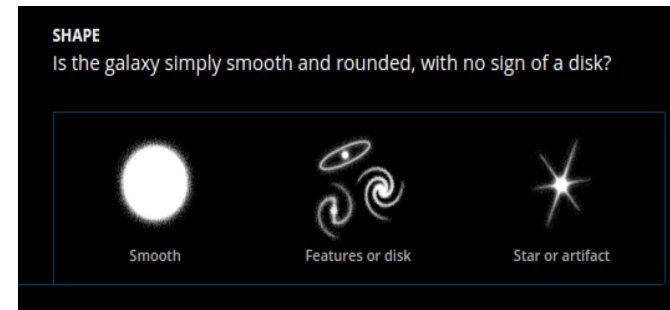
 Smooth	 Features or disk	 Star or artifact
---	---	---

<https://www.galaxyzoo.org/>

Image data: **Galaxy morphology classification**



GALAXY ZOO



Citizen science is merely a way of constructing training sets

<https://www.galaxyzoo.org/>

Image data: Galaxy morphology classification

Example:

Machine Learning for Galaxy morphological classification

Galaxy Classification without Feature Extraction

Kai Lars Polsterer¹

polsterer@astro.ruhr-uni-bochum.de

Fabian Gieseke²

f.gieseke@uni-oldenburg.de

Oliver Kramer²

oliver.kramer@uni-oldenburg.de

