Machine Learning Tutorial

III - Beyond textbook ML

ADA IX Summer School
20-22 May 2018, Valencia - Spain

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

In collaboration with
Alexandre Boucaud
Paris-Saclay Center for Data Science

Summary

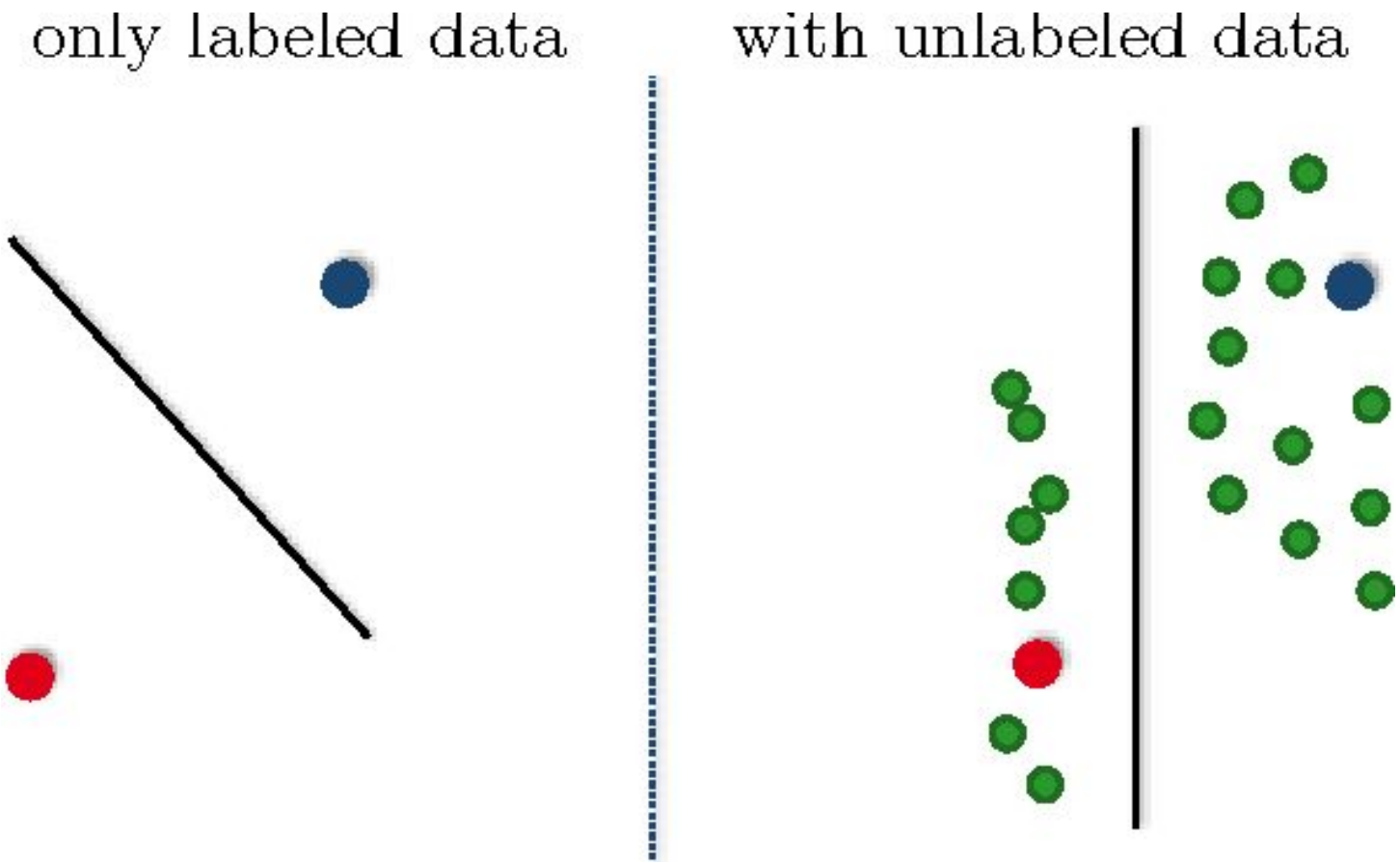
I. More than Supervised x Unsupervised

II. Challenges ahead

III. Human learning

Semi-supervised learning

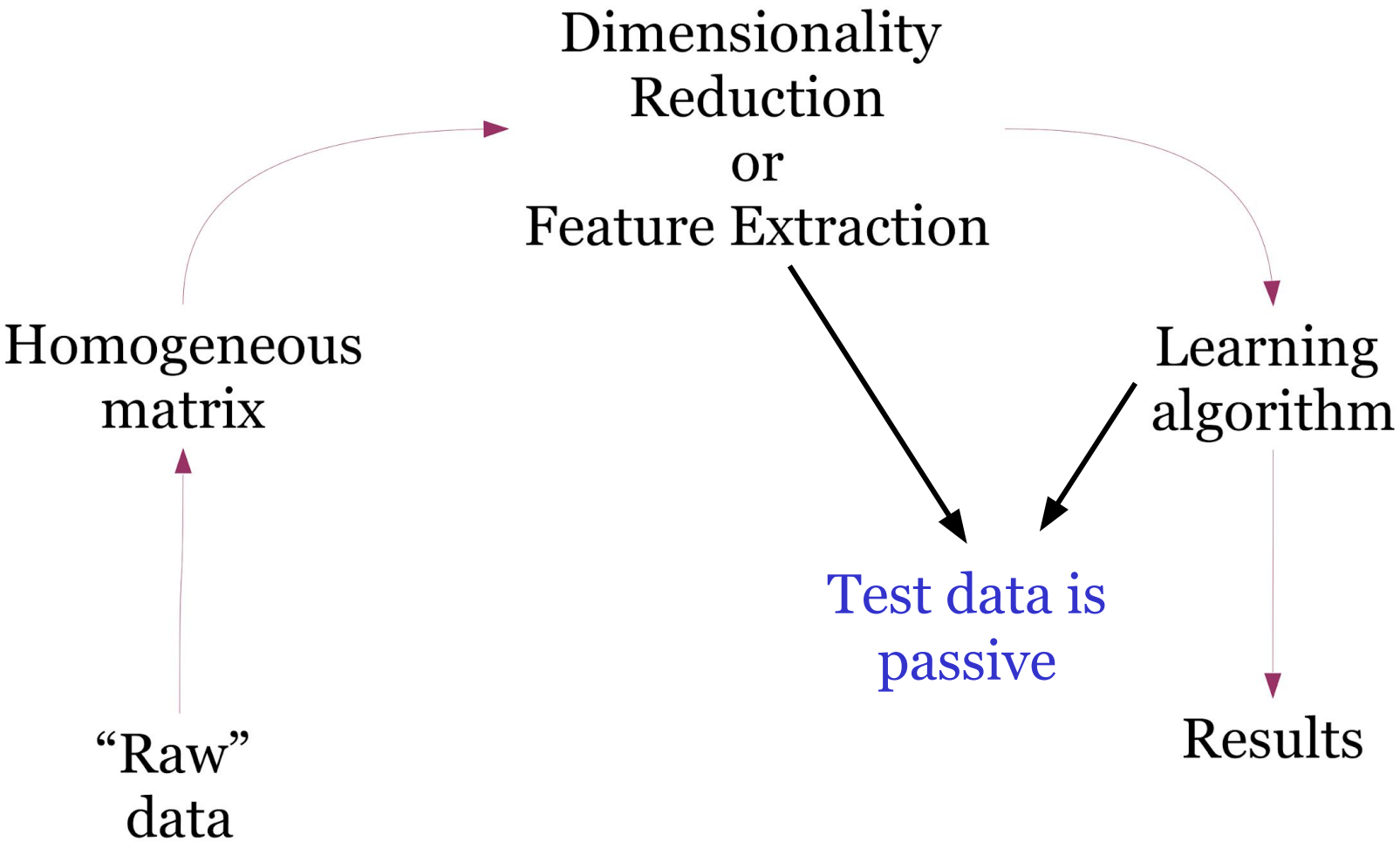
Getting partial information from the unlabelled sample



Valko et al., Semi-Supervised Apprenticeship Learning, 2012, JMLR,

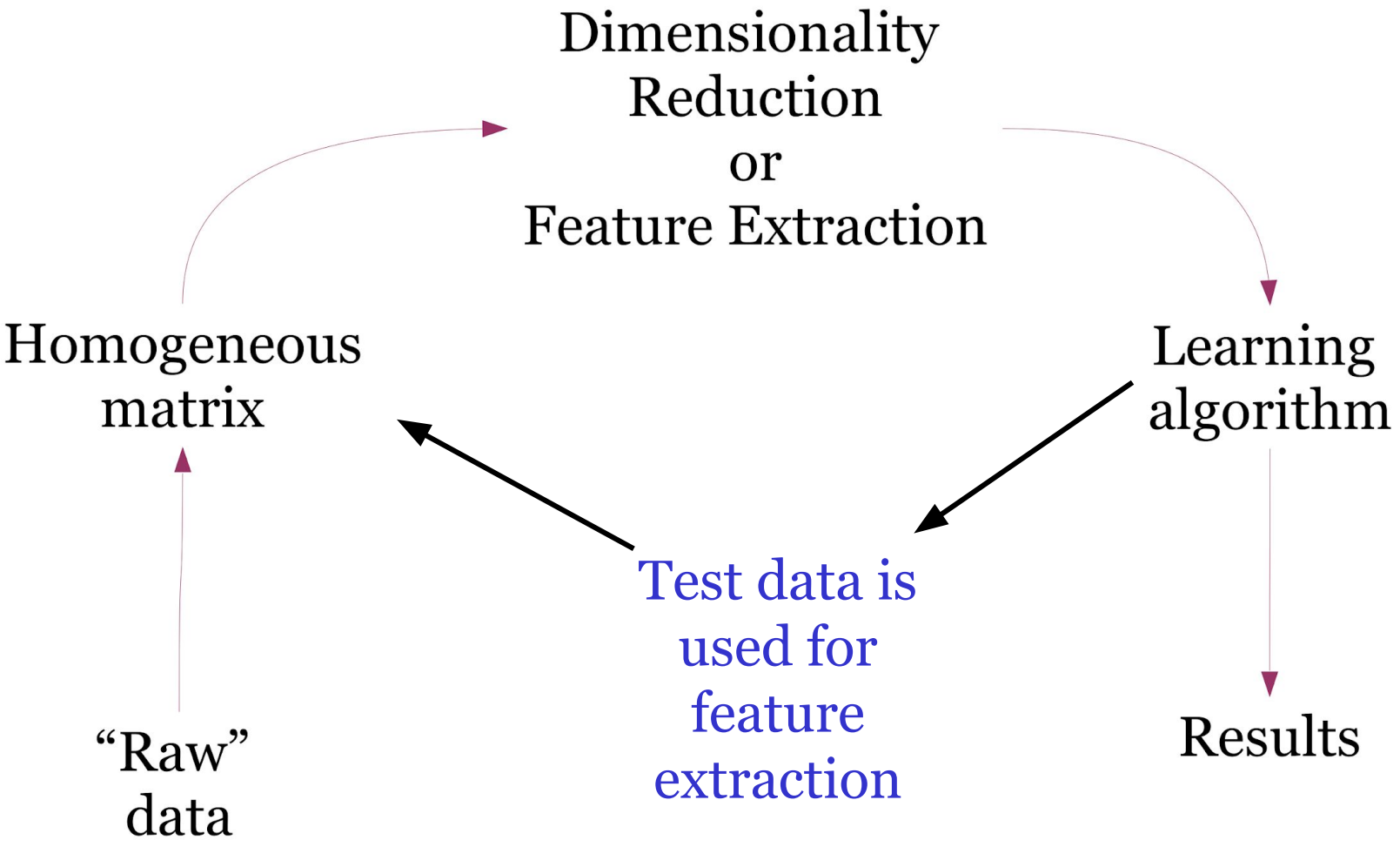
Machine Learning

Work-flow



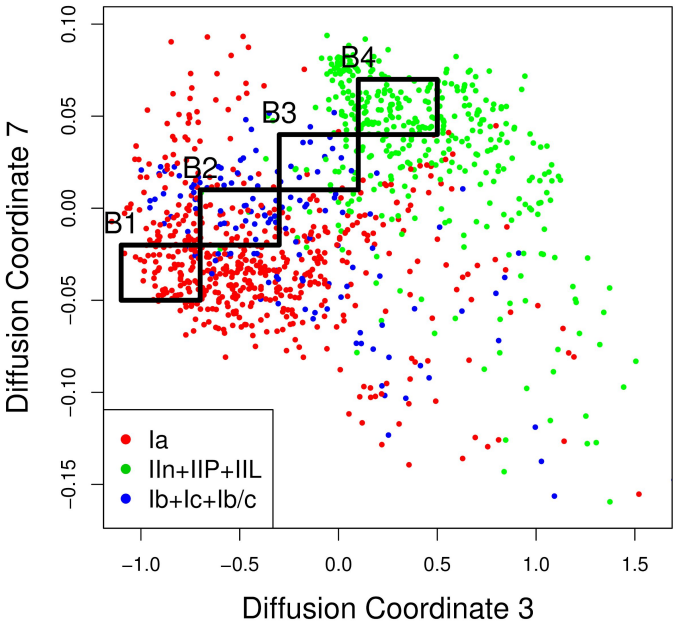
Semi-supervised learning

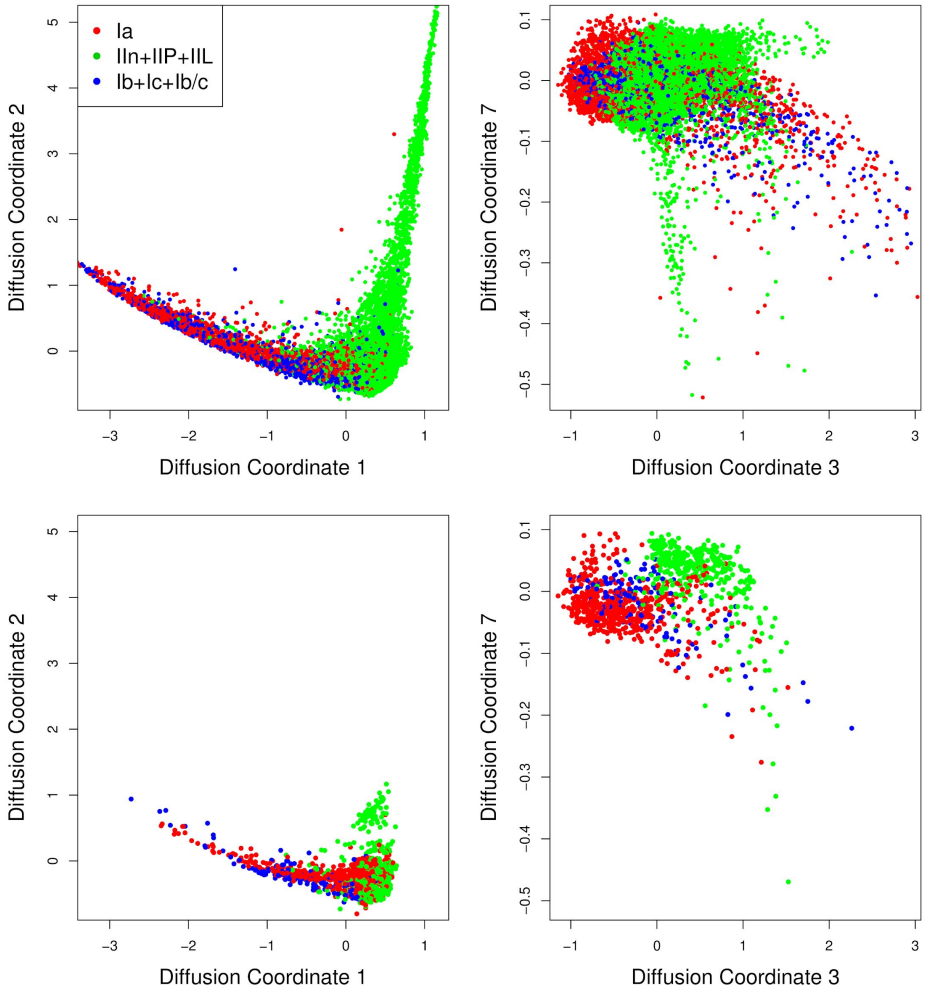
Work-flow



Semi-supervised learning

For Supernova Photometric Classification

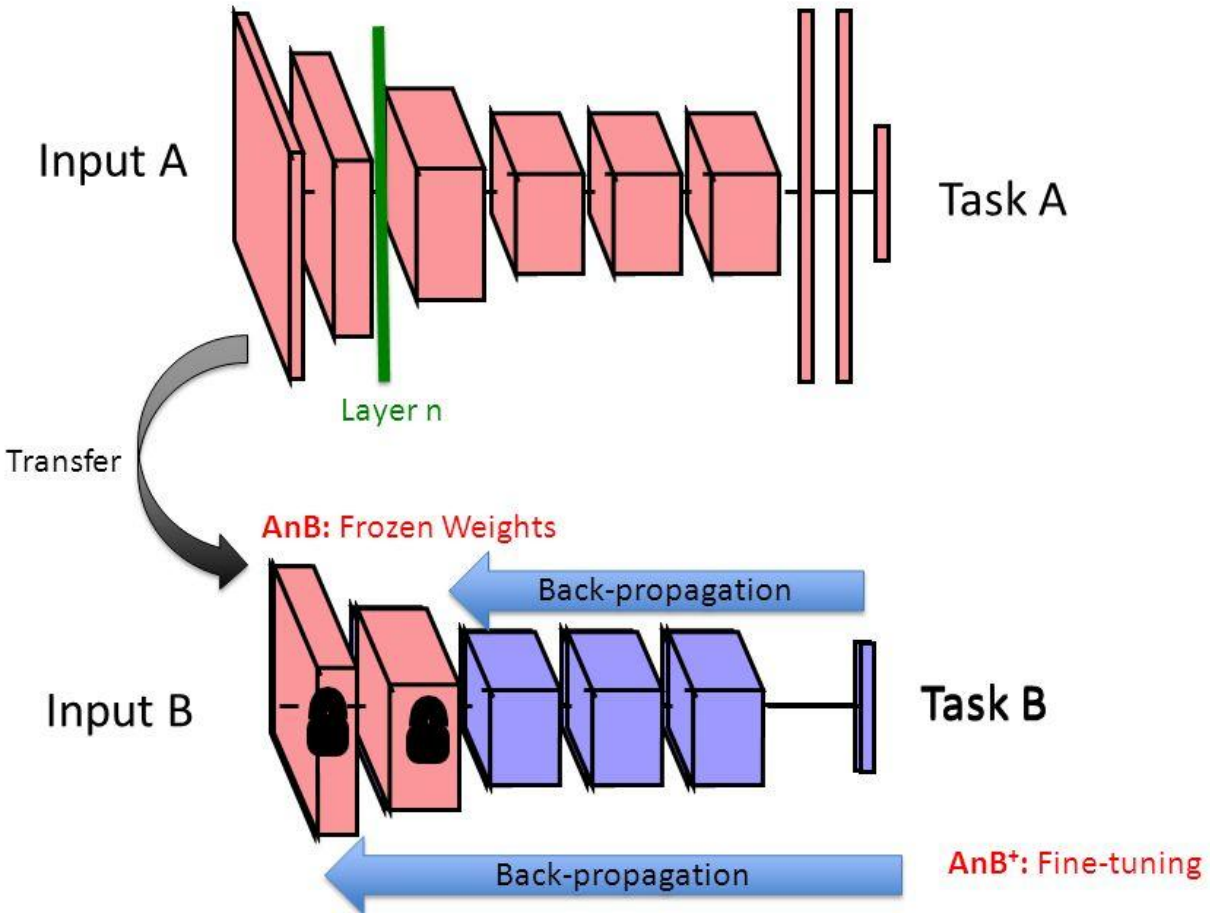




Richards et al., 2011, MNRAS

Transfer Learning

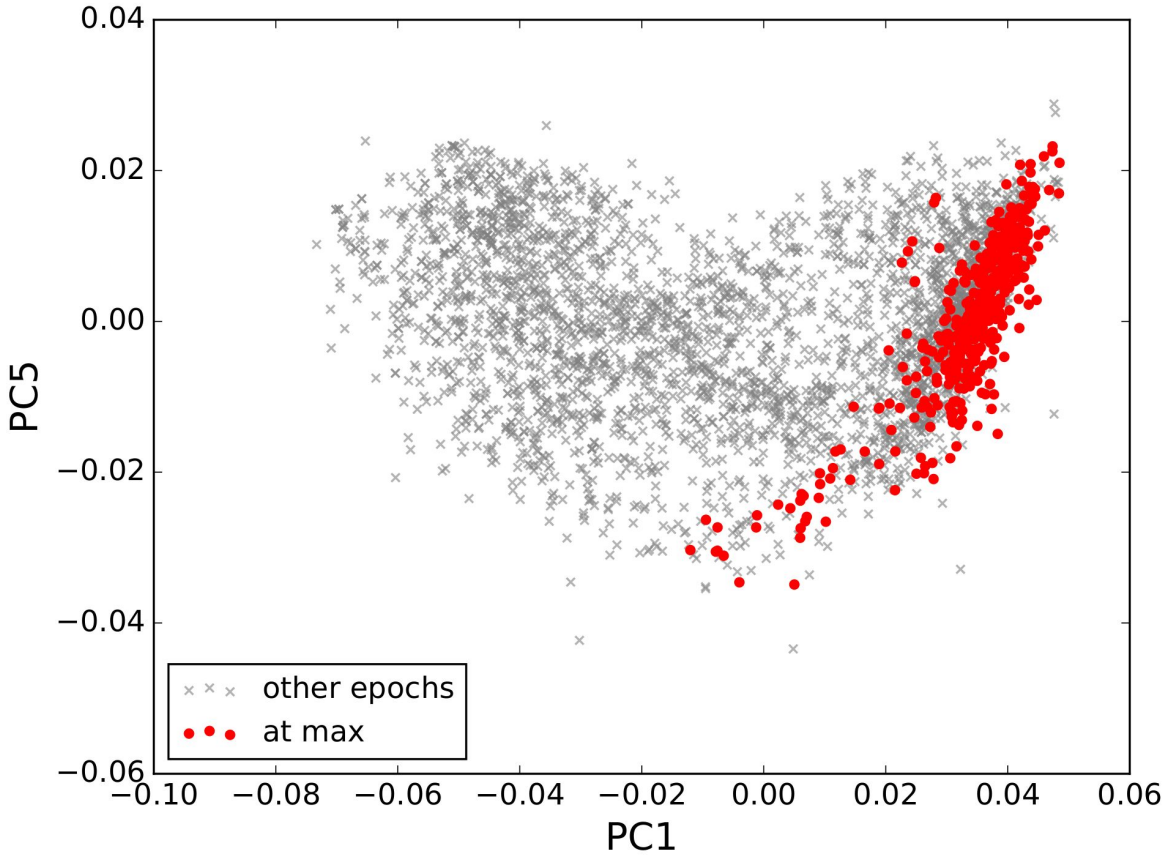
Borrowing information from somewhere else



http://slideplayer.com/slide/8370683/

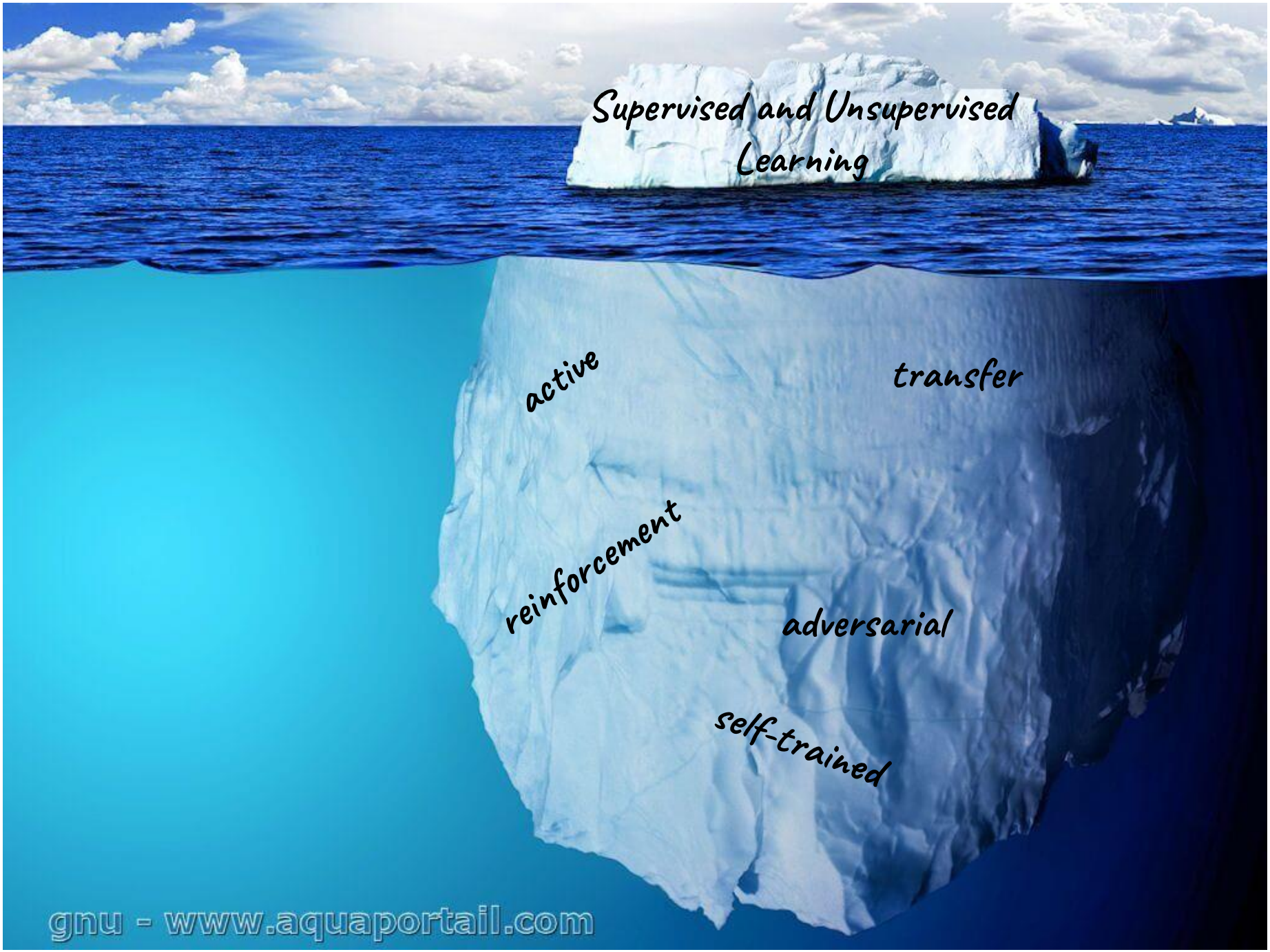
Transfer Learning

In Astronomy



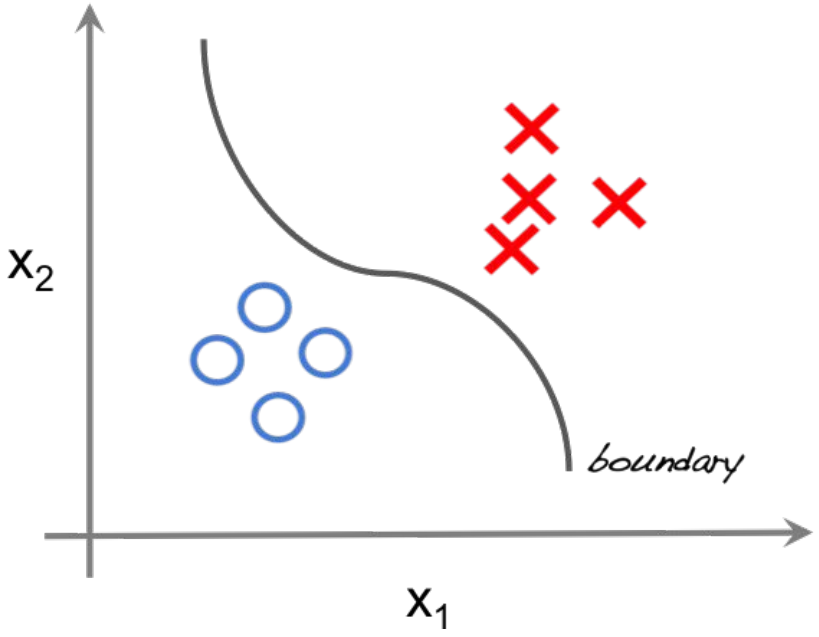
*Sasdelli, **Ishida** et al., 2016, MNRAS, 461, Issue 2, p.2044, from **CRP #2***


Supervised and Unsupervised Learning
gnu - www.aquaportail.com


Supervised and Unsupervised Learning
active
transfer
reinforcement
adversarial
self-trained
gnu - www.aquaportail.com

Categories of Machine Learning:

Supervised x Unsupervised

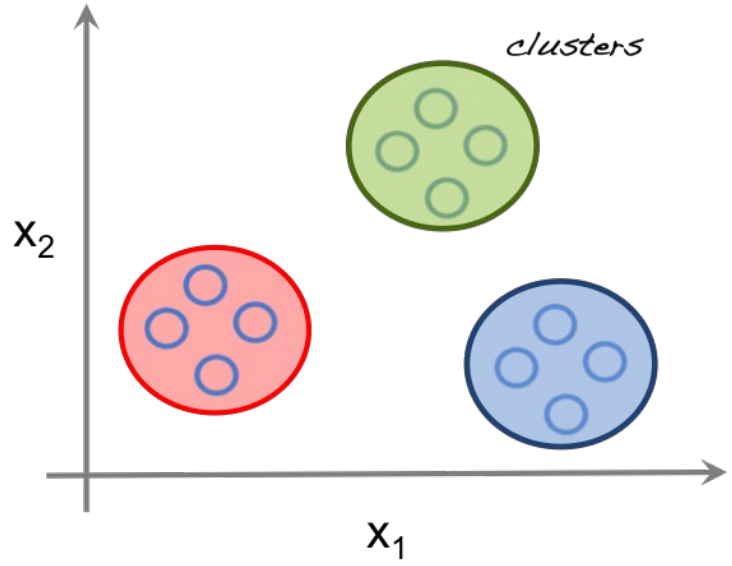


Tranining sample:

features + labels

Target sample:

features



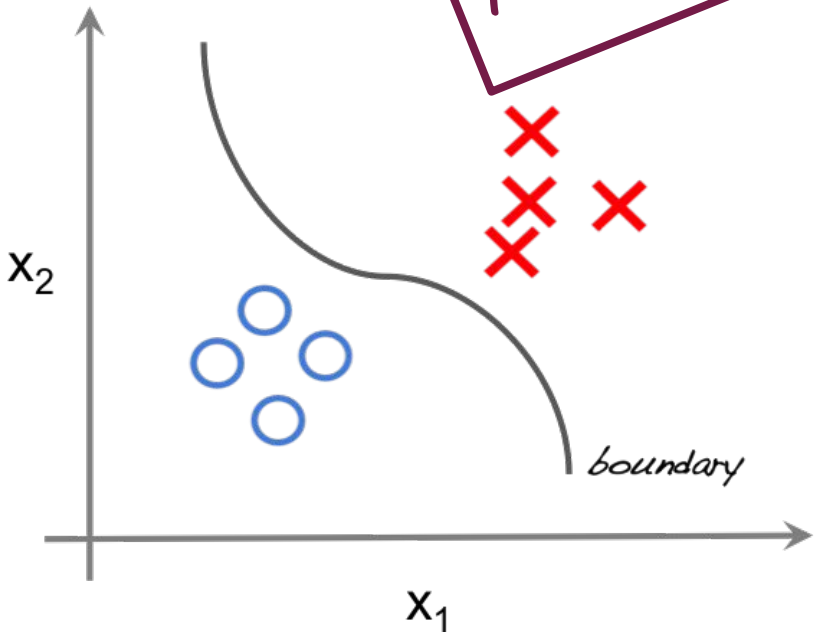
Data sample:

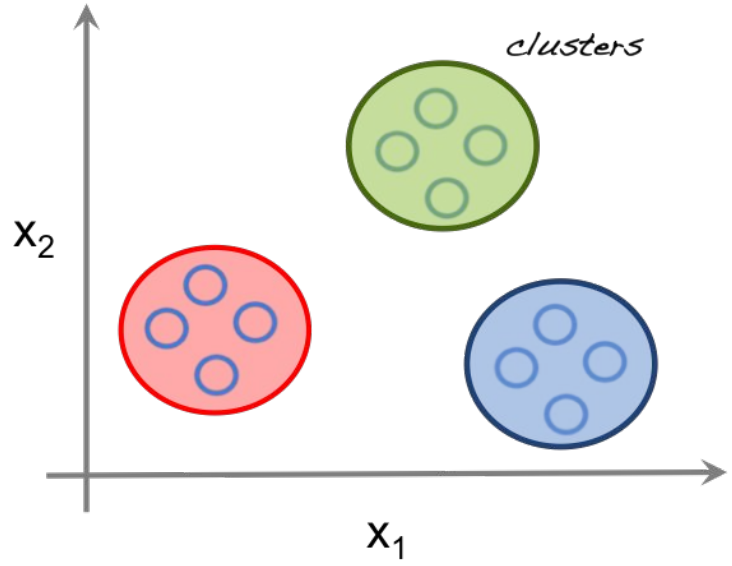
features

Categories of Machine Learning:

Supervised [illegible] pervised

Passive Learning





Tranining sample:

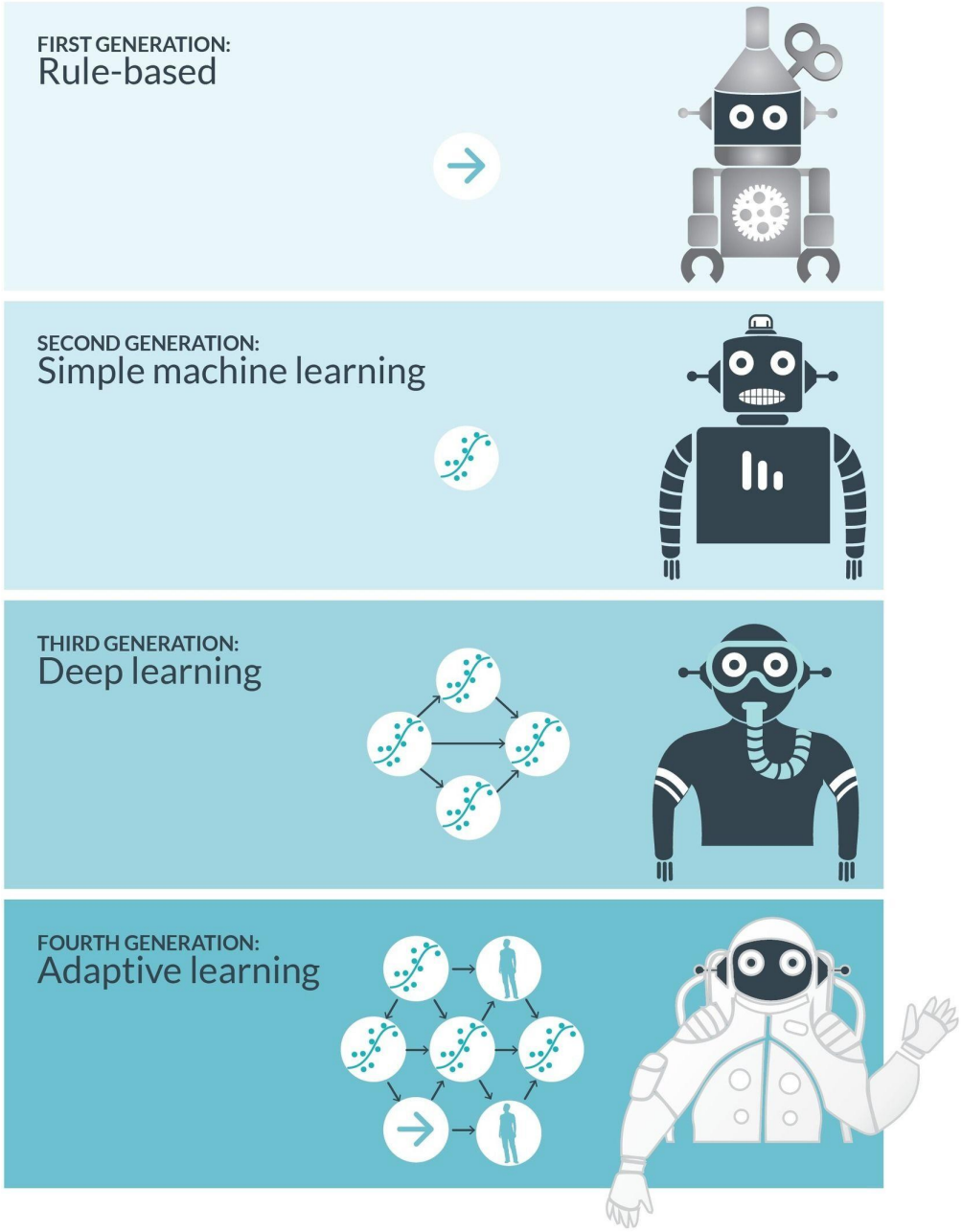
features + labels

Target sample:

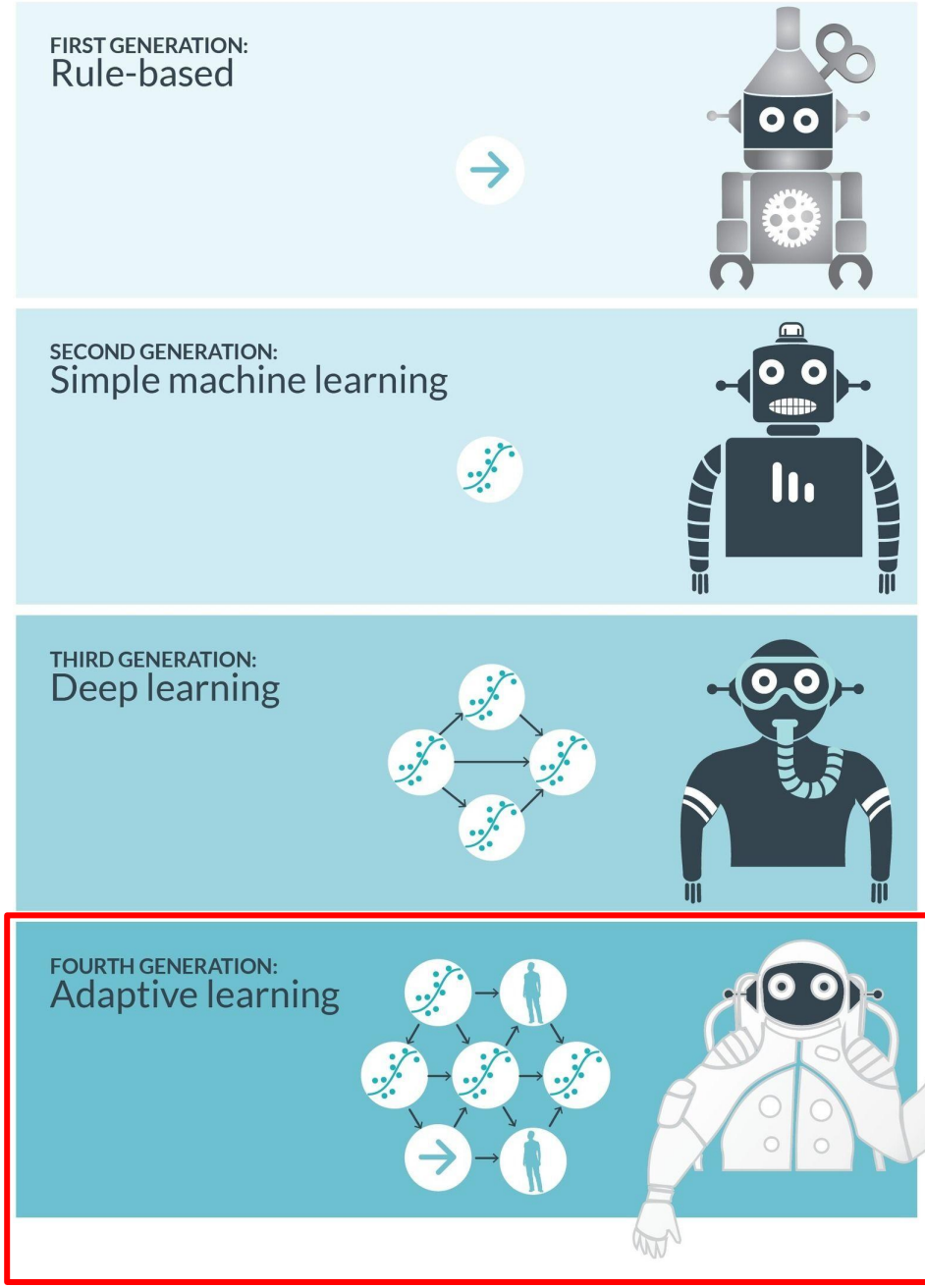
features

Data sample:

features



Machines need to evolve...

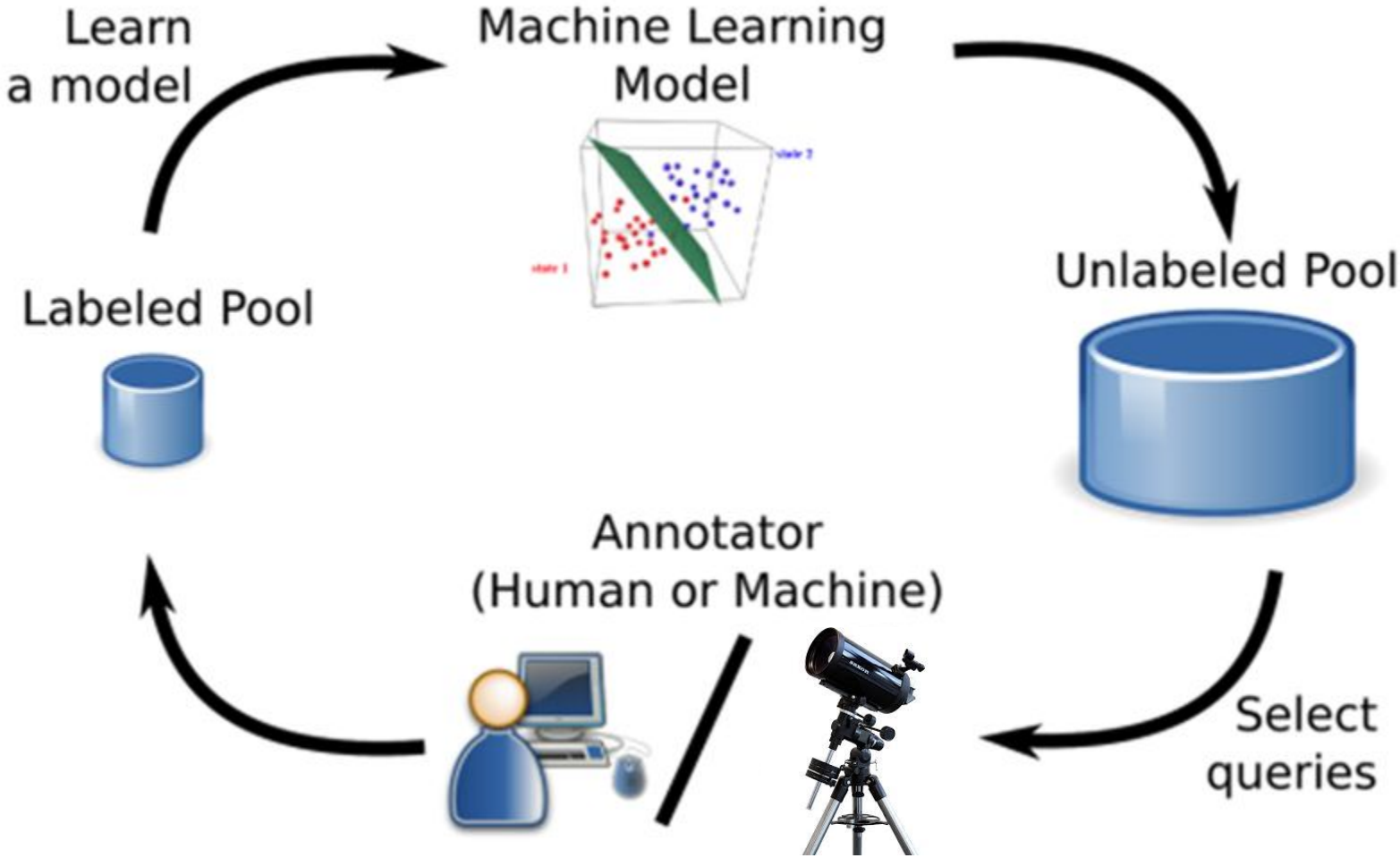


Machines need to evolve...

so they need to adapt!

Active Learning

Automatizing the construction of the training sample



Active Learning

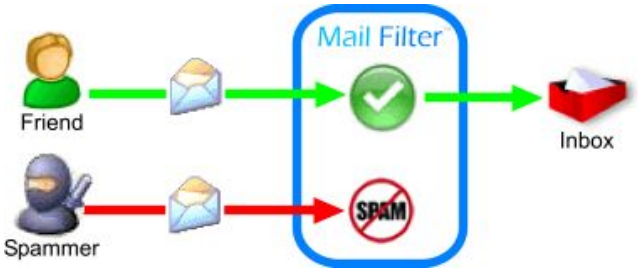
Automatizing the construction of the training sample

Can machines learn **better**, with **fewer** labelled examples, if they are carefully chosen?

Active Learning

Automatizing the construction of the training sample

Can machines learn **better**, with **fewer** labelled examples, if they are carefully chosen?



Active Learning

Automatizing the construction of the training sample

amazon

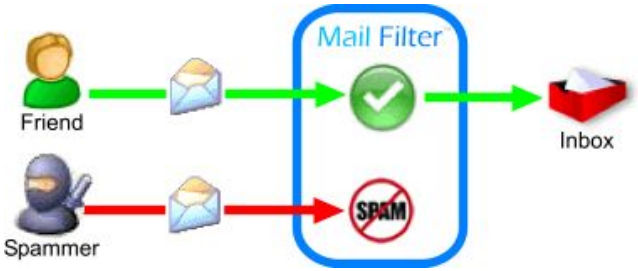
35% of Amazon's revenue are generated by it's recommendation engine.

NETFLIX

75% of users select movies based on Netflix's recommendations.

Can machines learn **better**, with **fewer** labelled examples, if they are carefully chosen?

YES!



Active Learning

Automatizing the construction of the training sample

amazon

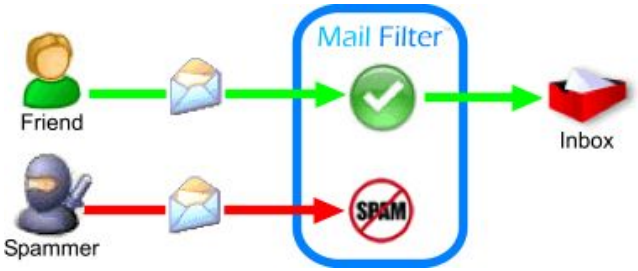
35% of Amazon's revenue are generated by it's recommendation engine.

NETFLIX

75% of users select movies based on Netflix's recommendations.

Can machines learn **better**, with **fewer** labelled examples, if they are carefully chosen?

YES!



Active Learning

Automatizing the construction of the training sample

amazon

35% of Amazon's revenue are generated by it's recommendation engine.

NETFLIX

75% of users select movies based on Netflix's recommendations.

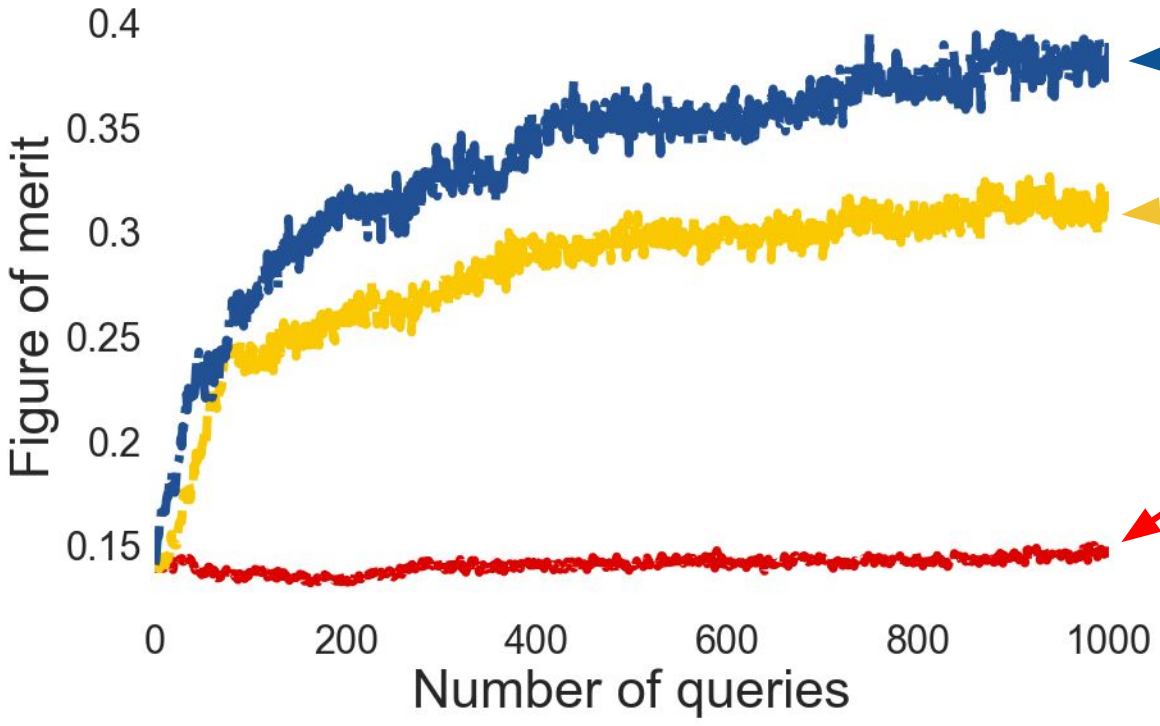
Can machines learn **better**, with **fewer** labelled examples, if they are carefully chosen?

YES!

Active Learning

Better performance, fewer training

In Astronomy



Active Learning

Passive Learning

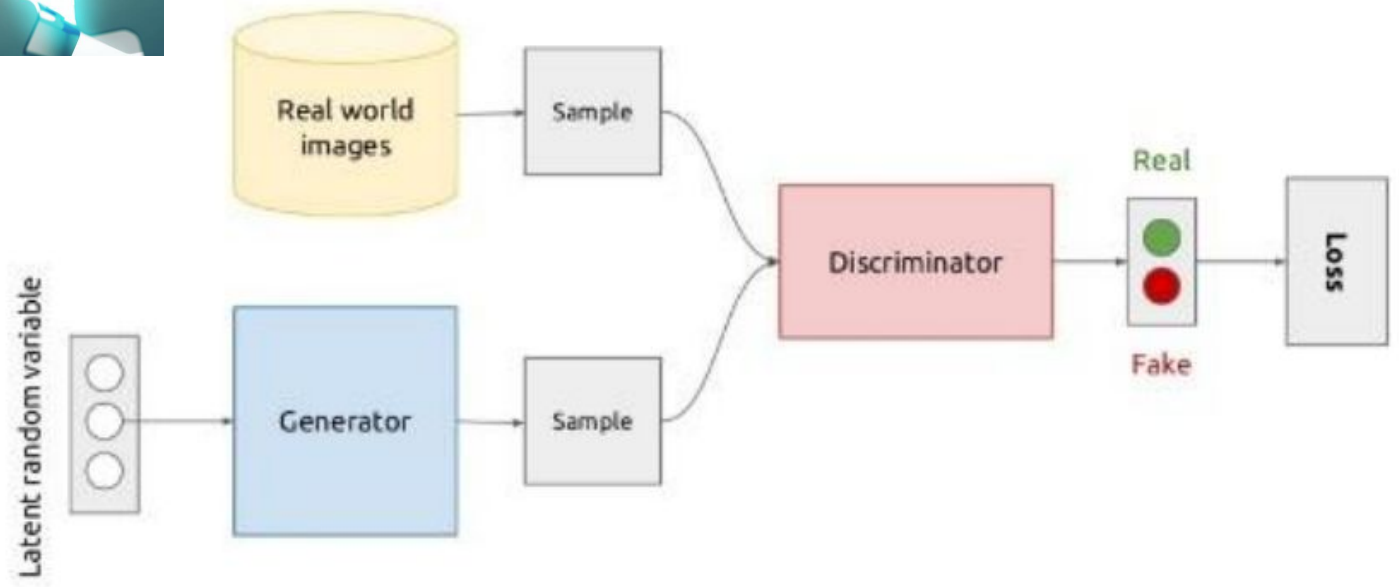
Canonical strategy

Cosmo21 Session 5 Wednesday

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

Adversarial Learning

The benefits of a worthy opponent



http://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

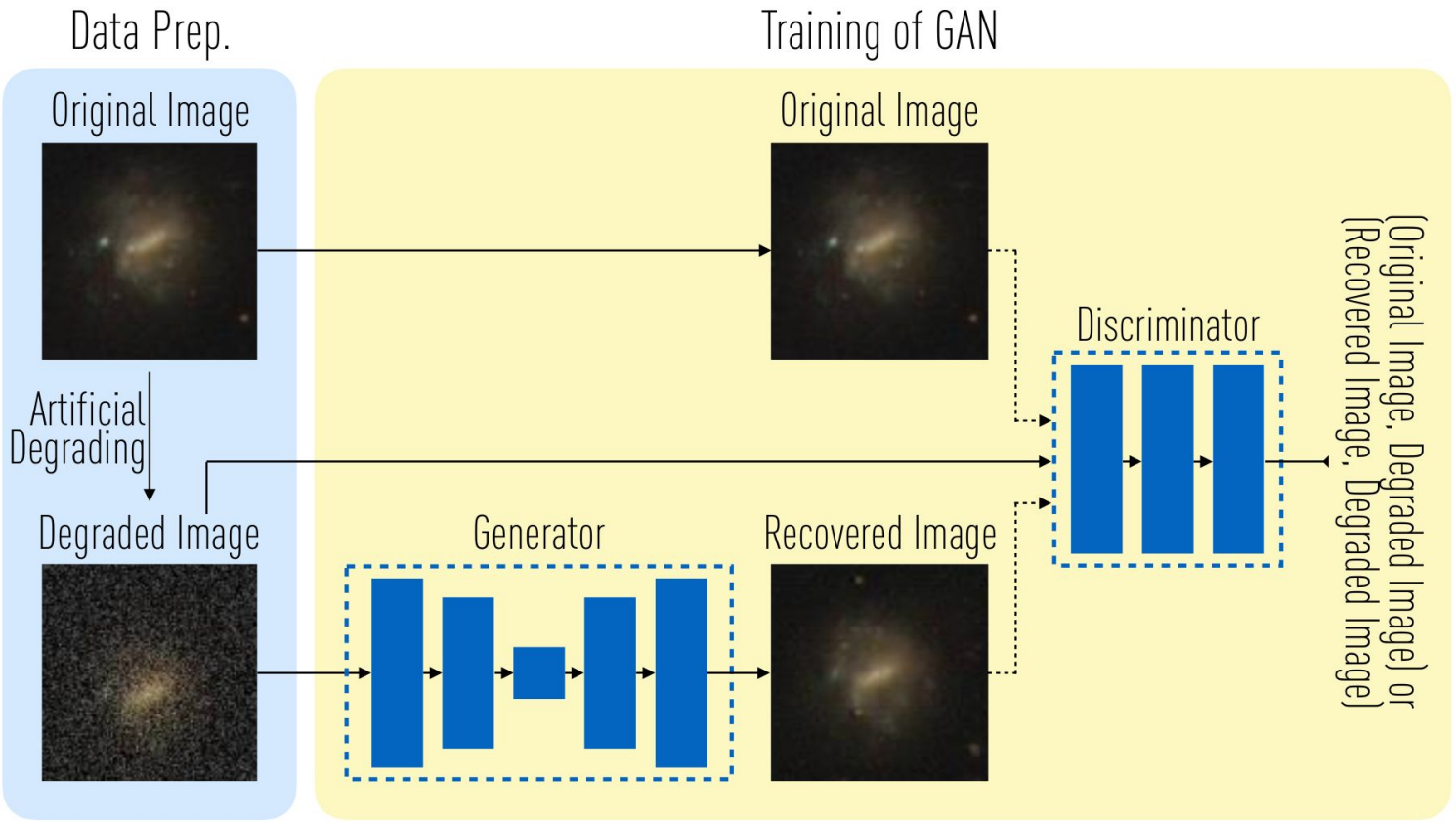
https://mascherari.press/introduction-to-adversarial-machine-learning/

Adversarial Learning

The benefits of a worthy opponent

K. Schawinski et al, 2017

In Astronomy

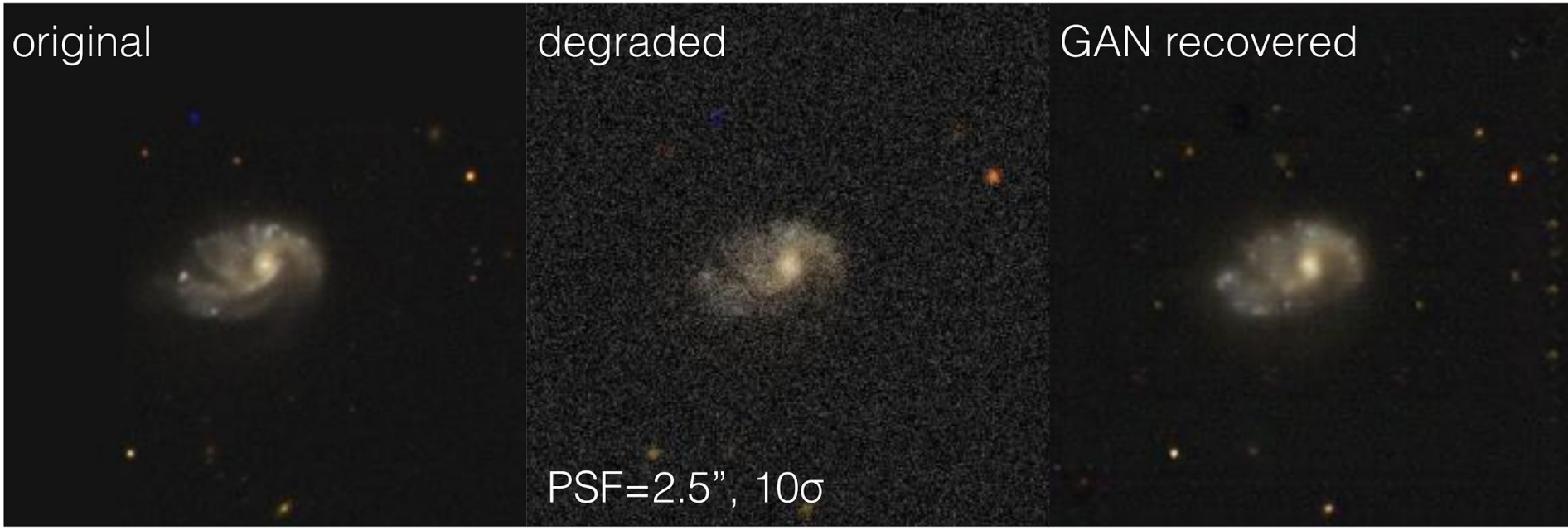


Adversarial Learning

The benefits of a worthy opponent

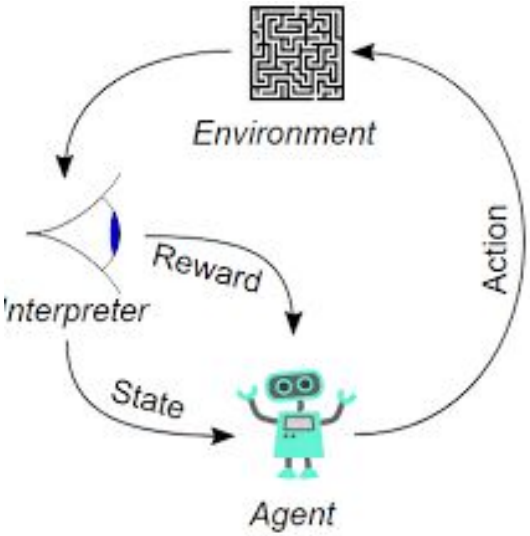
K. Schawinski et al, 2017

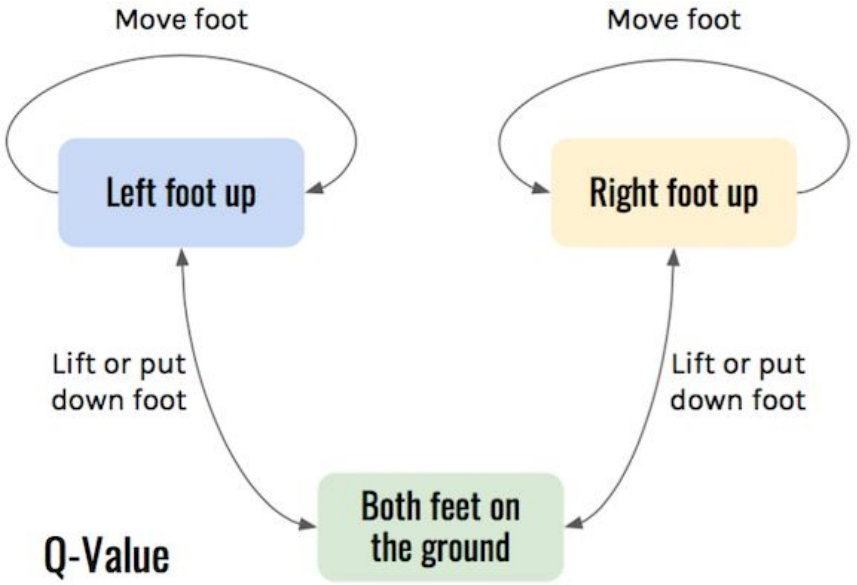
In Astronomy



Reinforcement Learning

The importance of feedback





State	Action	Q-Value
Left foot up	Move foot forward	+ 0.5
Left foot up	Put foot down	+ 0.0
Left foot up	Move foot backward	- 0.5
...	...	...

https://hackernoon.com/using-q-learning-to-teach-a-robot-how-to-walk-a-i-odyssey-part-3-5285237cc3b1

This is only the beginning…

… moreover, Astronomical data is tough!

Representativeness

Supervised ML model

data **training**, target

χ set of all samples, x

Y set of possible labels, y

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Only works if test is representative of training

Data generation model:

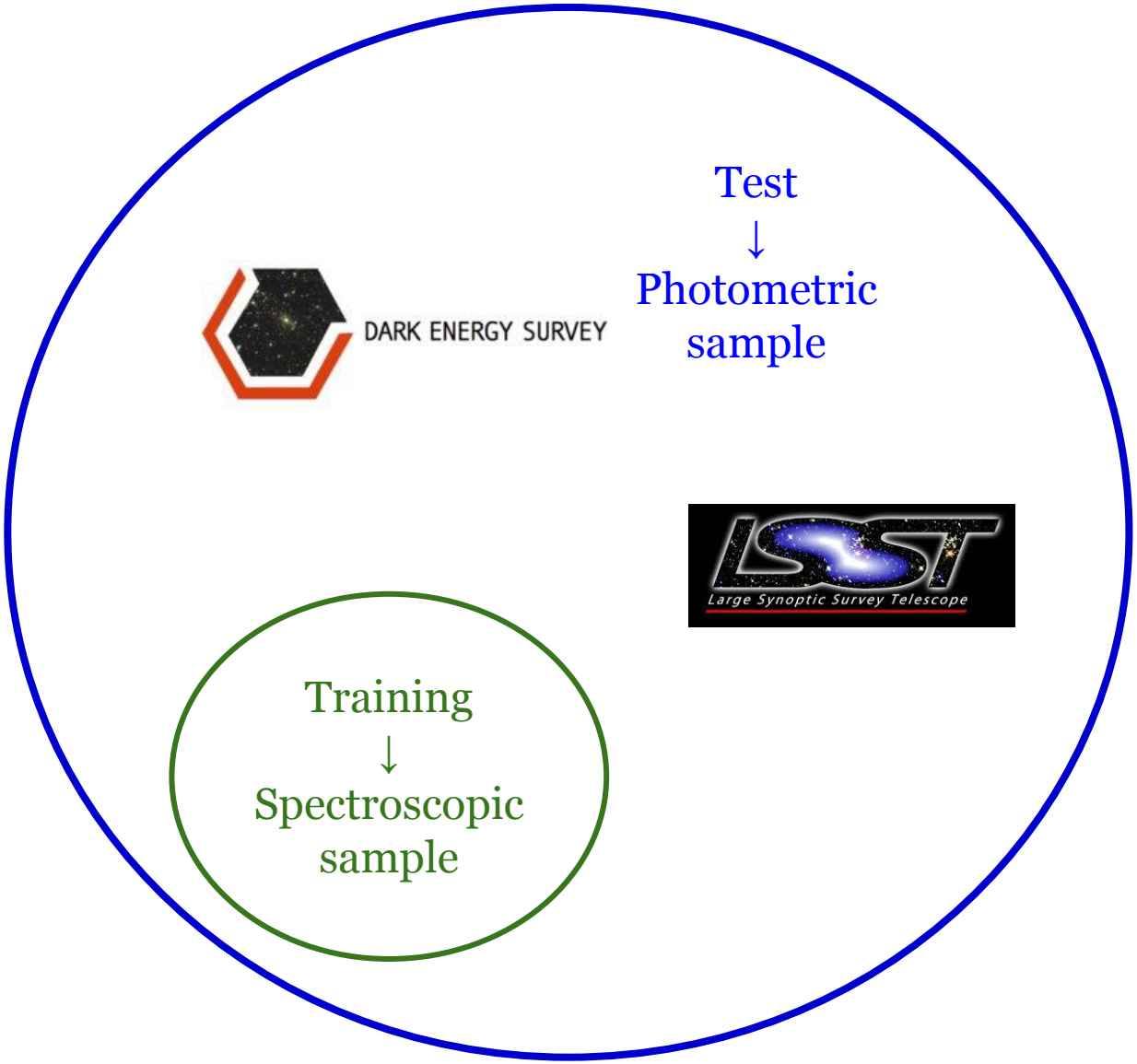
$x_i \sim P_\chi$

f true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x \sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

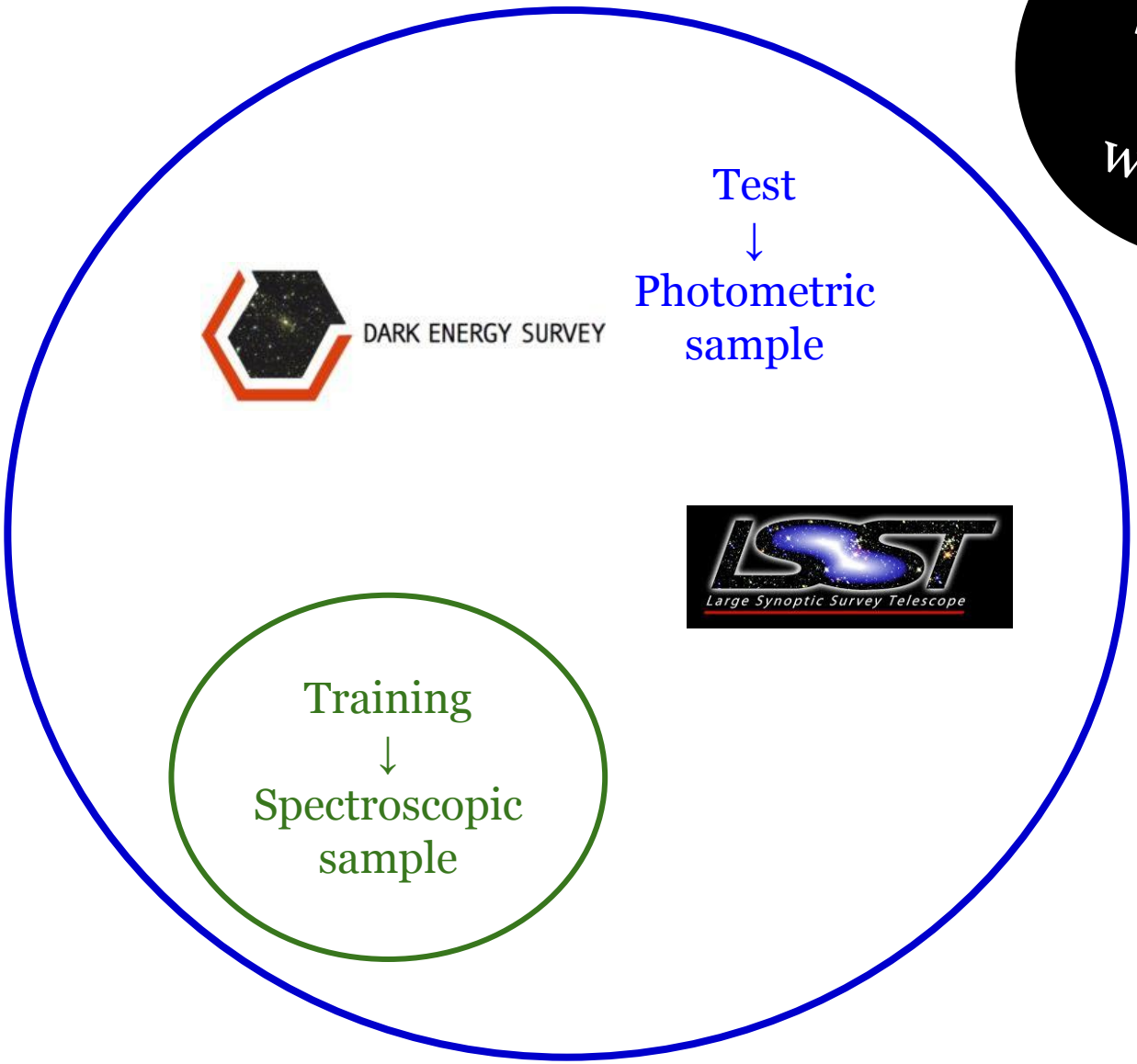
In Astronomy …



… it is NOT the case!

In Astronomy …

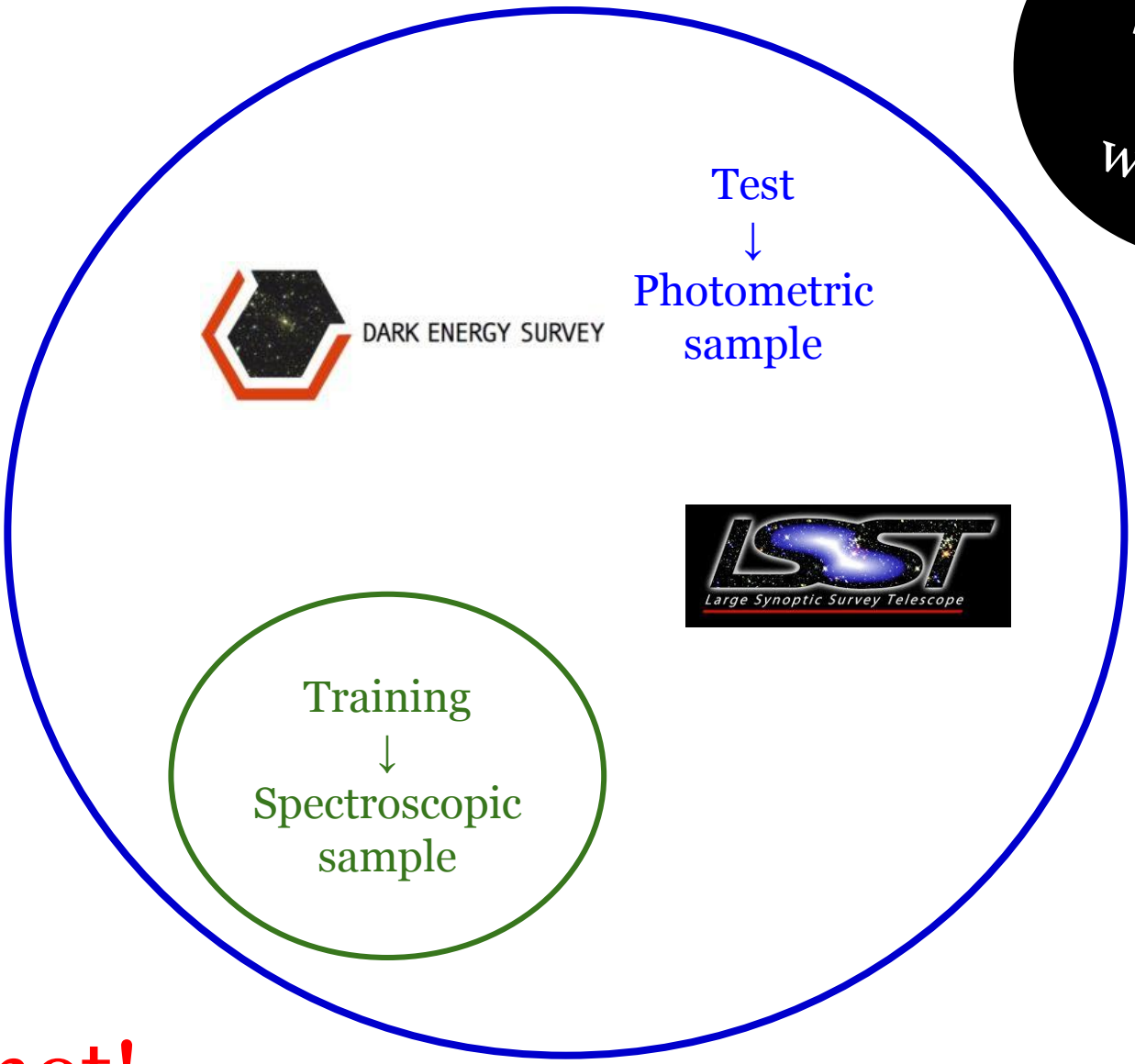
Why should this work?



… it is NOT the case!

In Astronomy …

Why should this work?



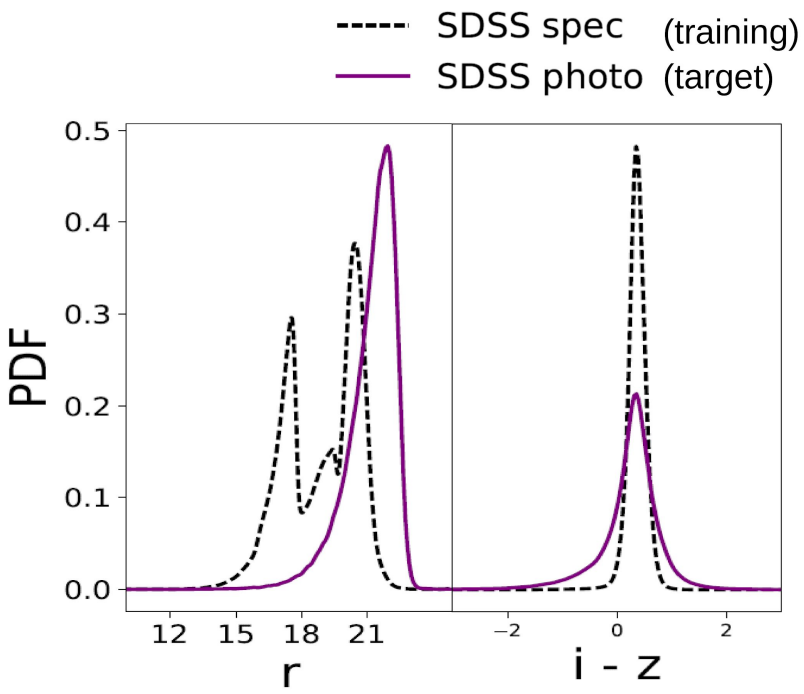
It does not!

… it is NOT the case!

Representativeness

(or the lack of) in Astronomy

Photometric redshift estimation



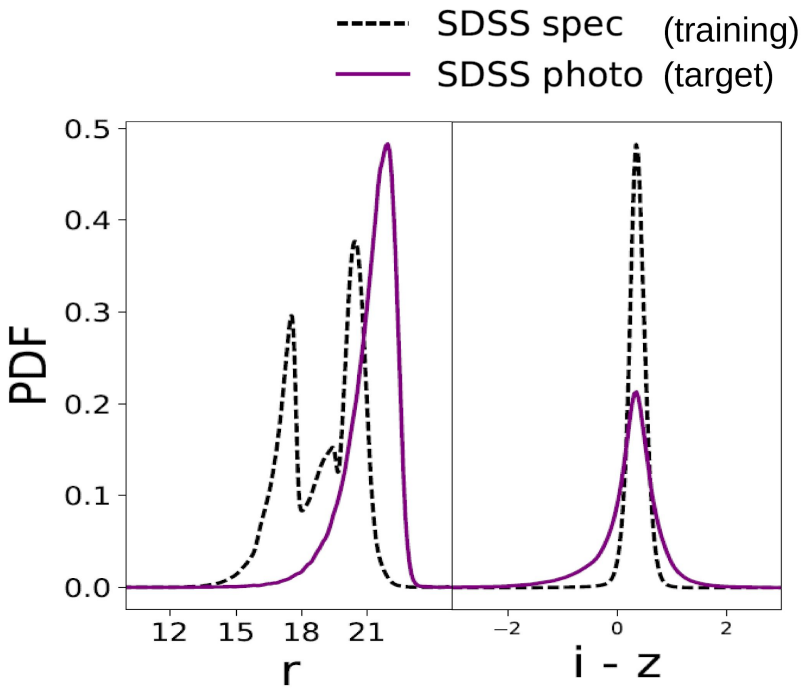
Beck et al., 2017, MNRAS - from **CRP #3**

Check **Linc's talk:** Cosmo21 - session 10
Thursday afternoon

Representativeness

(or the lack of) in Astronomy

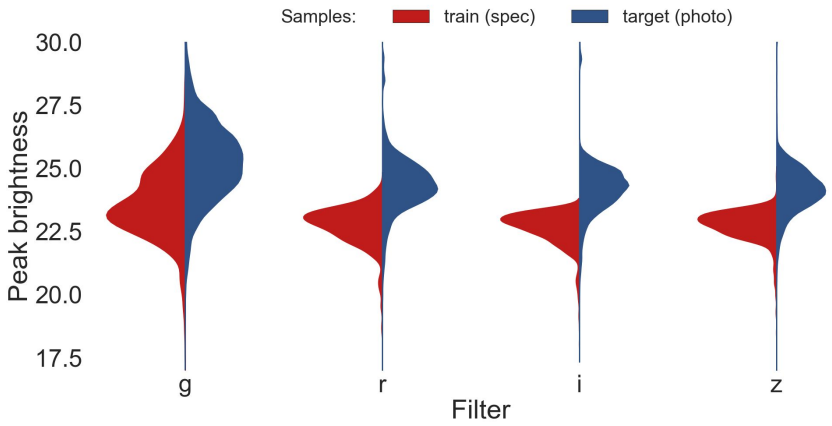
Photometric redshift estimation



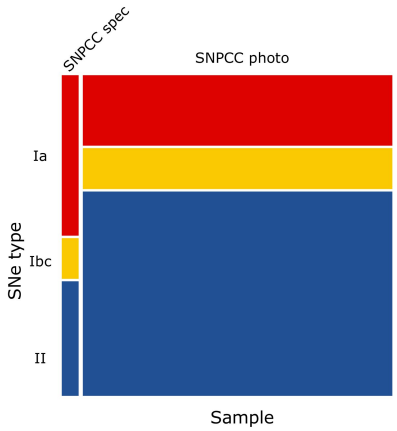
Beck et al., 2017, MNRAS - from **CRP #3**

Check **Linc's talk:** Cosmo21 - session 10 Thursday afternoon

Supernova Classification



Emille's talk: Cosmo21 - session 5 Wednesday



Measurement Errors

Bayesian Machine Learning

Time Domain

SNe over LSST sky

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

We will have to adapt!

We are getting there…

Developments in human learning

Developments in human learning

Community code development

Developments in human learning

kaggle

Community code development

RAMP

Check **Alex's talk:** Cosmo21 - session 10
Thursday afternoon

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
- 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)

Expected release date:

Summer/Fall 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

Get ready!

Play with the first SNPCC data!

Here is a peak at the data.

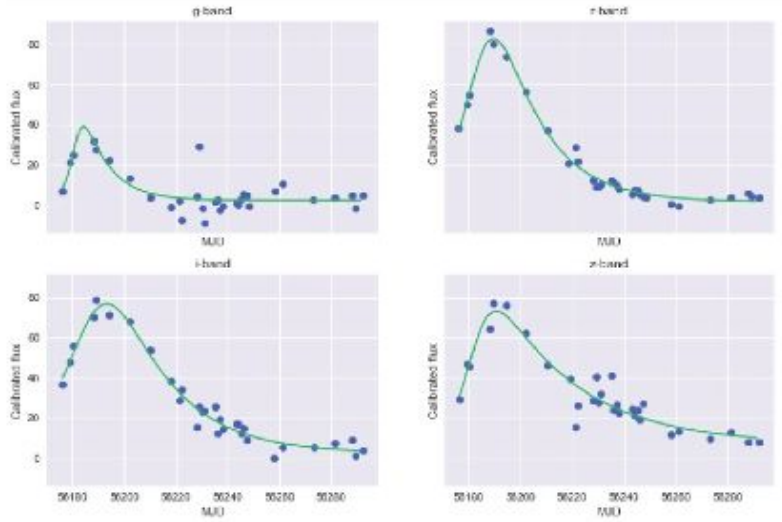
```
In [2]: X, y = read_data('data/des_train.pkl')
```

```
In [3]: # 5 first rows of the dataframe
        X.head()
```

Out[3]:

	mjd_g	fluxcal_g	fluxcalerr_g	mjd_r	fluxcal_r	fluxcalerr_r	mjd_i	fluxcal_i	f
1186	[56283.203, 56289.199, 56292.098, 56304.051, 5...	[3.182, -13.07, -7.177, -0.4645, 0.2908, 16.63...	[6.285, 11.79, 11.65, 2.479, 7.569, 6.195, 3.6...	[56283.211, 56295.215, 56292.102, 56304.055, 5...	[-4.973, -6.427, 9.881, -1.683, 6.724, 23.82, ...	[3.328, 6.181, 4.83, 1.857, 2.966, 2.553, 3.07...	[56283.215, 56289.109, 56292.191, 56304.07, 56...	[-0.815, -9.294, 0.6371, 0.6895, -1.583, 24.51...	
...	[56177.172, 56179.172, ...	[27.19, 34.93, 24.15, ...	[4.837, 2.372, 1.805, ...	[56177.188, 56179.312, ...	[33.02, 34.65, 26.92, ...	[5.375, 3.109, 1.365, ...	[56177.203, 56179.328, ...	[49.59, 42.79, 30.72, ...	

```
In [14]: plot_lightcurves_with_fit(2)
```



Constructing the data matrix for classification

Now that you can fit each individual light curve, you are ready to build your low dimension representation of the data.

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

PLAsTiCC - RAMP: classification of astronomical transients

Alexandre Boucaud (Paris Saclay Center for Data Science)

Emille E. O. Ishida (Universite Clermont Auvergne)

https://github.com/ramp-kits/supernovae

Build an interdisciplinary, people-centric community!

https://www.cosmostatistics-initiative.org

Summary

- There is much more than supervised/unsupervised learning

- Not only machines, but humans will be required to adapt

- Astronomical data can lead to important development in computer science

- We should also think about updating the academic model

Thank you!

Cosmostatistics Initiative

http://cointoolbox.github.io/