A data-driven approach in Astrophysics

some examples, between "hype" and scepticism, to trigger some discussion

Iary Davidzon (Cosmic Dawn Center / NBI)

Workshop on Perspectives and Applications of Deep Learning for Accelerated Scientific Discoveries in Physics

Copenhagen, May 14th 2020

Introduction

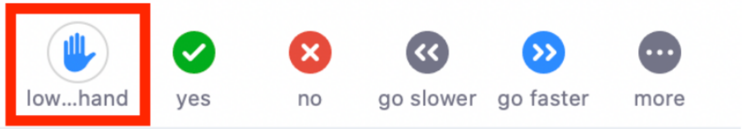
- **I am an observational astronomer** involved in large survey projects (notably, *Euclid* Space Telescope).

- Goal of this talk is to **foster ideas/doubts** for the panel sessions by providing several short examples.

- **“Data driven” is a broad definition**: deep learning, dimensionality reduction, etc… from now on **I will use the acronym ML** (machine learning).

- “Hype vs scepticism” to **set a common ground**

- Interruptions are welcome!



- Let’s clean the air from some misconception…

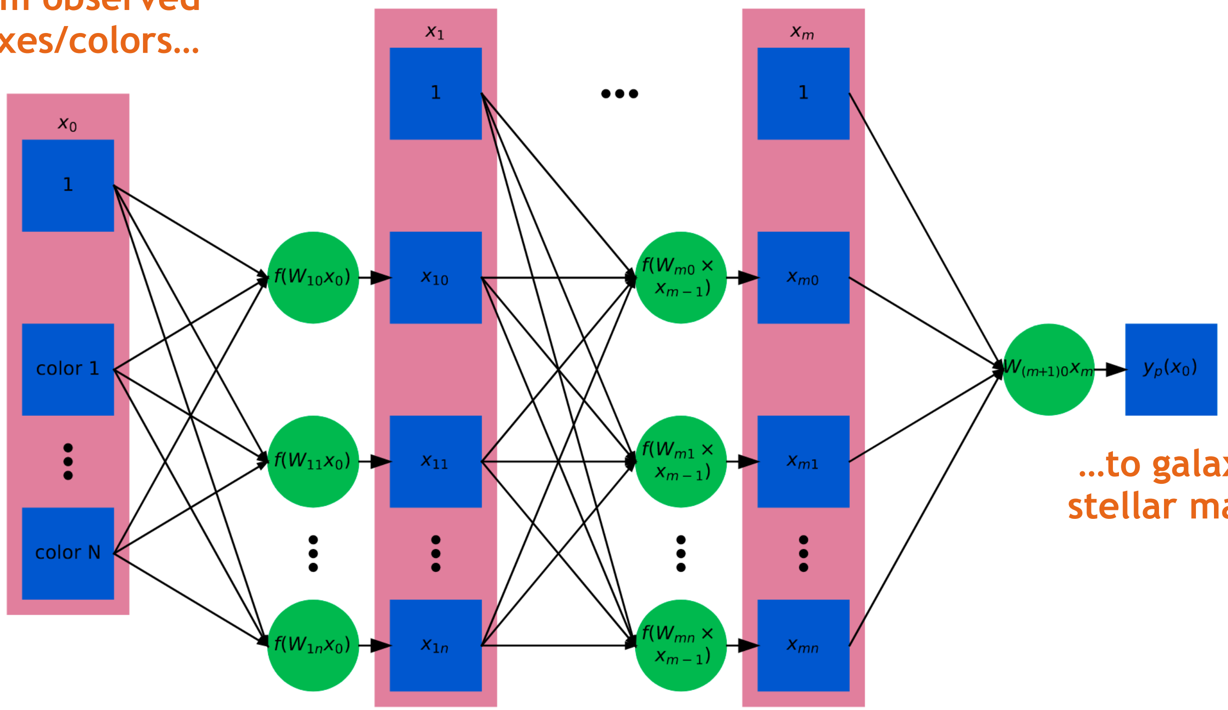
(in my personal view)

For the skeptics…

- ***Black box* doesn't mean *Black magic:*** we can look inside a ML-based tool, although sometimes it is a difficult task.

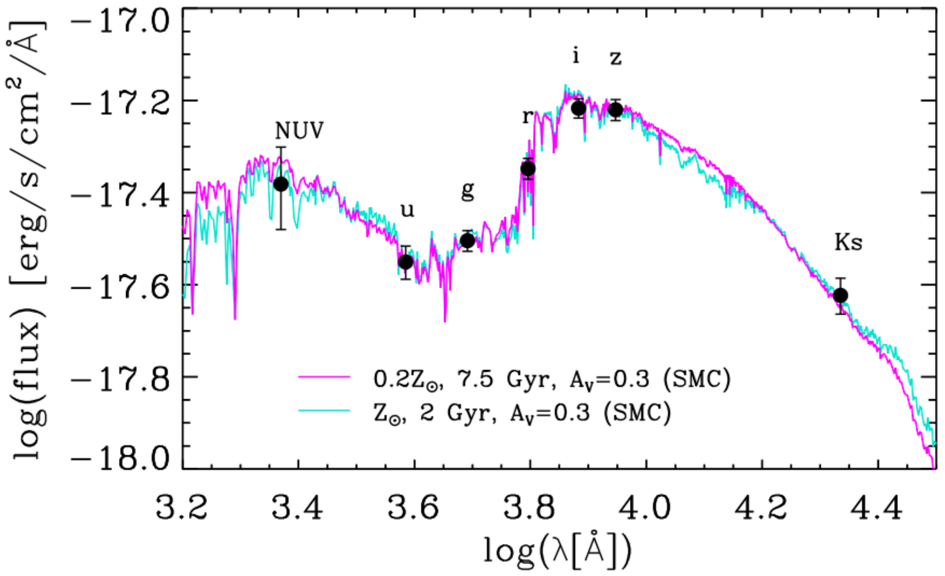
e.g. in *deep learning*, the data structure leading to the prediction is more complex than a set of equations… but we can "empirically" understand it.

from observed fluxes/colors…



…to galaxy stellar mass

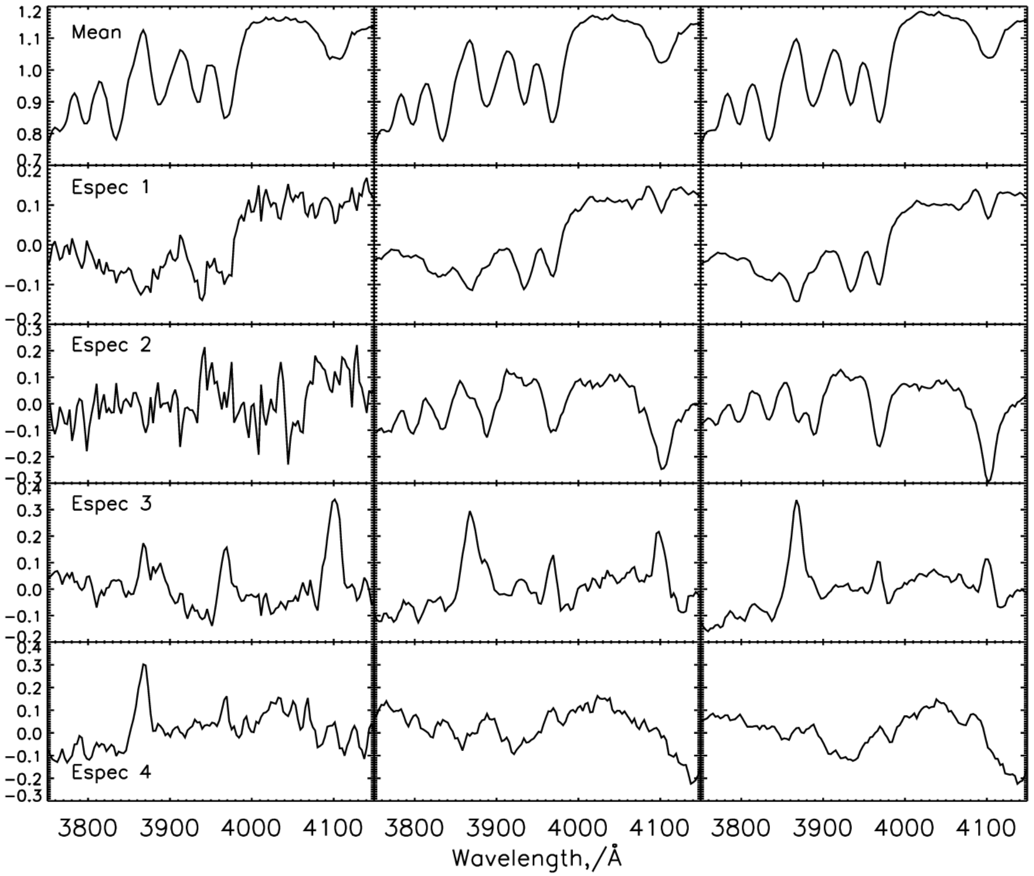
Simet et al. (2019)



classical method: fitting galaxy models to photometric data points

For the skeptics…

- **Training is always a healty exercise:** even a biased training sample can be helpful (if well understood); advantages of unsupervised learning.



Budavári et al. (2009)

"streaming" Principal Component Analysis is less sensitive to outliers and reduces noise in the eigenspectra

For the skeptics…

- **We could really use some help** to digest extremely big data from future astronomical surveys

V. Rubin Observatory
20 TB of data per night

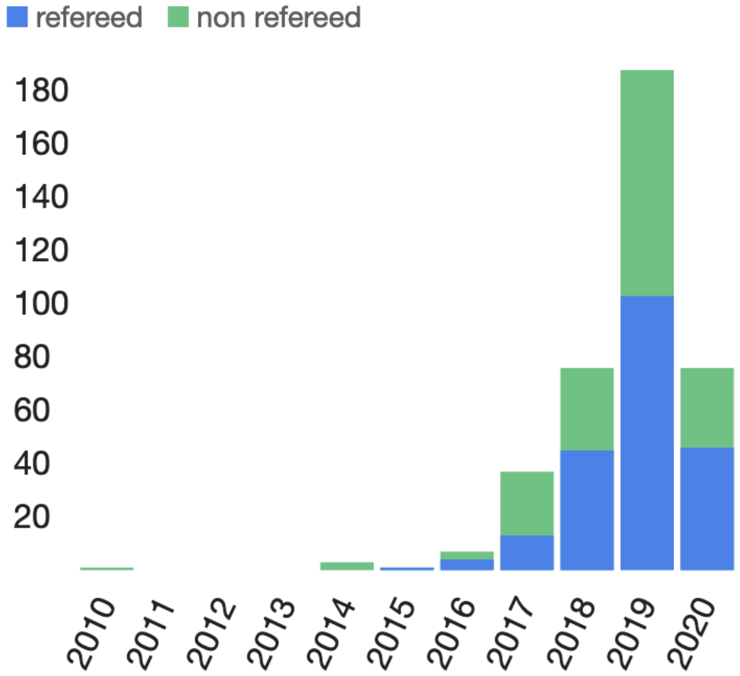
Wide-Field IR Survey Telescope
will have 100x field of view of Hubble

Aside: is ML helping job-wise?

- encouraging transition outside academia?
- is the opposite true?
- new job profiles?
- does it mess up the literature?

The LSST Corporation "is committed to foster and build new modes of **interdisciplinary collaboration** at the interface of astronomy, physics, computer science, mathematics, and information science" [and outreach].

lsstcorporation.org



astrophysical articles including "machine learning" in the abstract

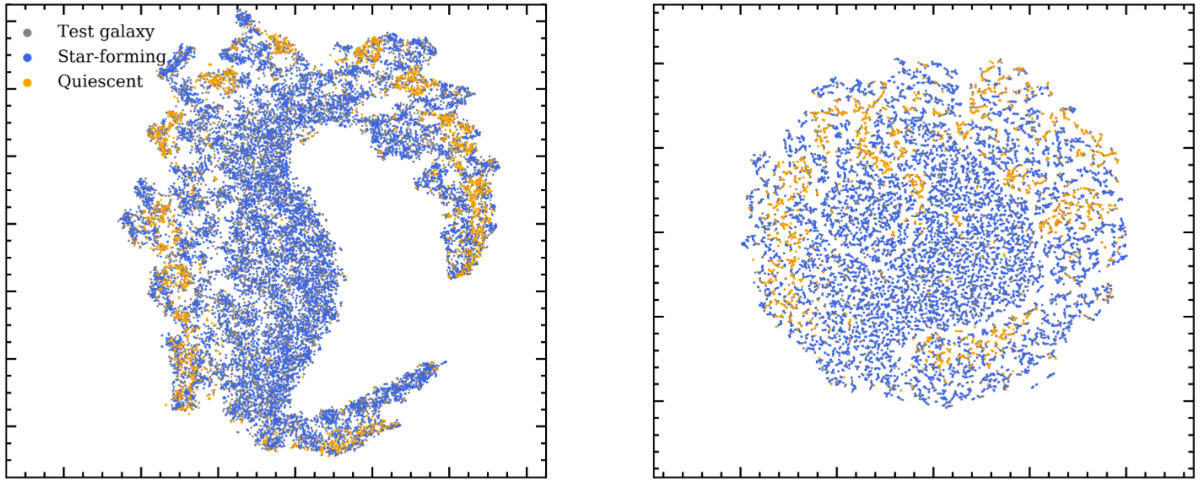
from ADS/NASA

*during this covid isolation
I have watched a lot of TV series

For the others: curb your enthusiasm… *

- **Data doesn't speak for itslef:** *data driven* doesn't mean *assumption free*. Even for unsupervised ML.

Steinhardt et al. (2020)



Dimensionality reduction of a given (galaxy) manifold, e.g. the panchromatic space

Fig.1) two different representations using **t-distributed stochastic neighbor embedding (t-SNE)**

Carrasco-Kind & Brunner (2014)

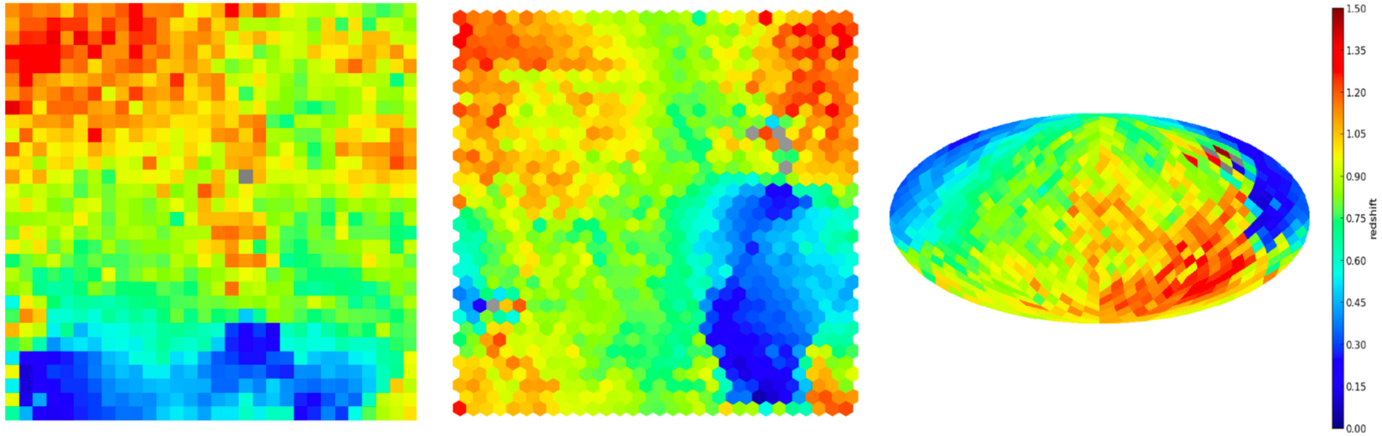


Fig. 2) three different projections using **Self-Organizing Maps (SOM)**

see also Davidzon et al. (2019)

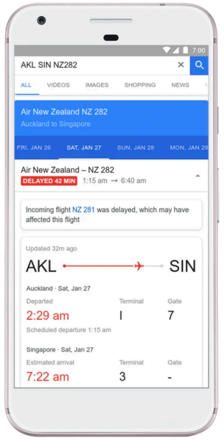
Curb your enthusiasm…

- **Orange vs Apple© comparison:** methods that are well tested in industry may not be suitable for astrophysics. Heads up for students**

- feature and label errors, bias, incompleteness
- deeper/stratified knowledge

Astronomy has a lot to say on that!

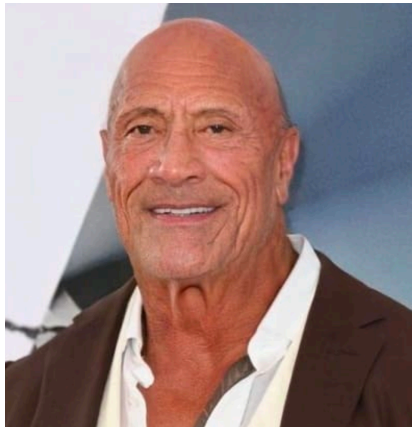
www.google.com/flights



from predicting flight delay…

…to planning your night at the telescope?

Instagram / FaceApp



from aging people…

…to "observing" a galaxy in the future?

** part of this is inspired by working with G. Girelli and T. Vène

ML applied to DIKW levels: different goals and implications

1. Data collection, reduction, management
2. Extracting information from signal (measurements with error bars)
3. Interpretation (e.g. model fitting) to produce codified knowledge
4. From knowledge to "wisdom": causality, big picture, future perspective

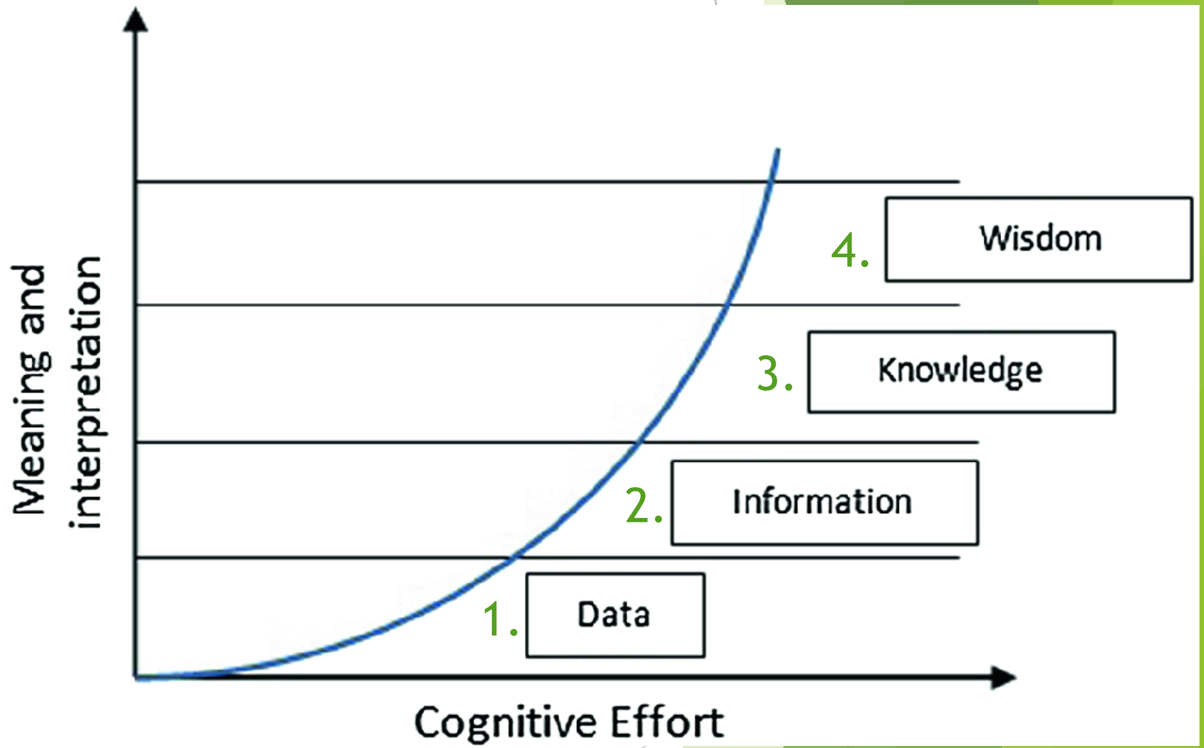


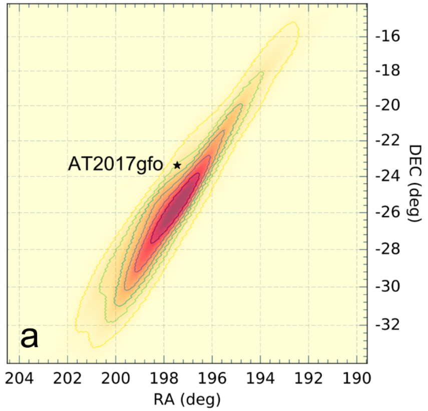
Image found by google.com, no credit was available

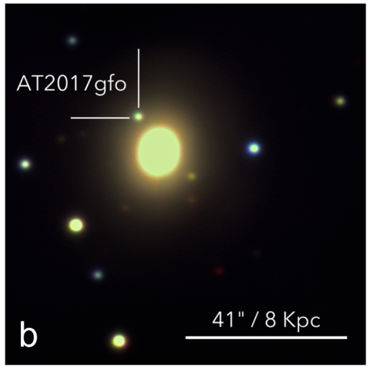
How to combine highly heterogeneous data?

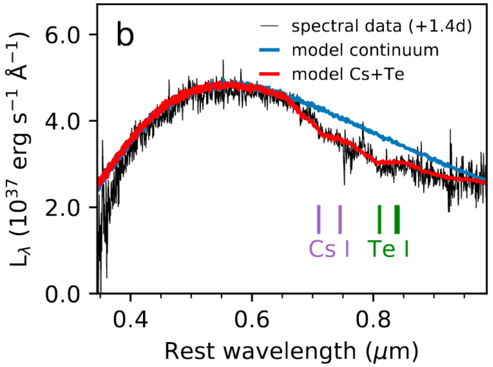
How to label extremely large samples?

“multi-messenger” information for a binary neutron-star merger:

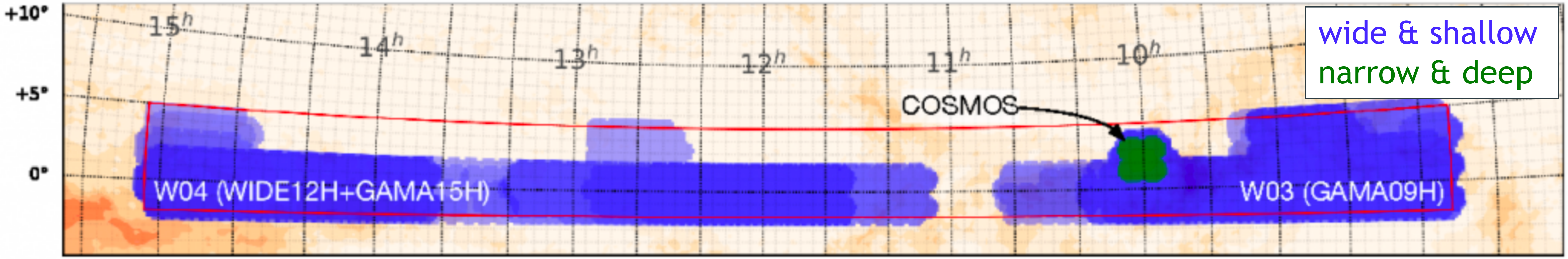
from Smartt+17, Levan+17







“wedding cake” observing strategy:

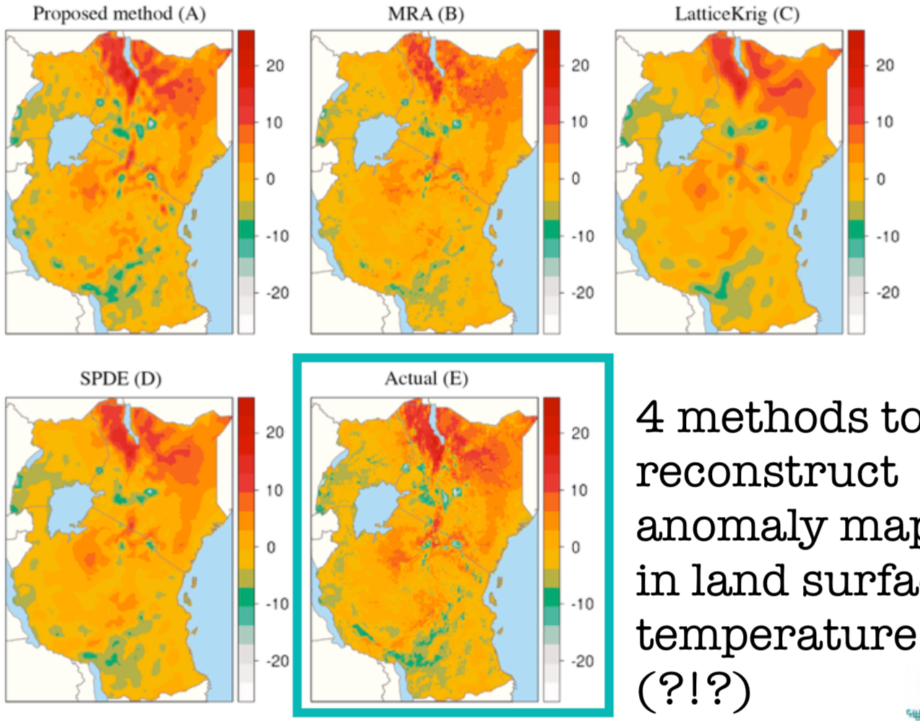


Hyper Suprime-Cam Subaru Strategic Program

Answers: data homogenization, domain adaptation, transfer learning…

No astro-example, sorry! But we can borrow ideas from other fields (Public Health Applications, Remote sensing, etc.)

- conversion to common ref. system and data format
- correct for distortions and other calibration effects
- fill the gap (missing data)
- most software still requires human supervision ☹



4 methods to reconstruct anomaly map in land surface temperature (?!?)

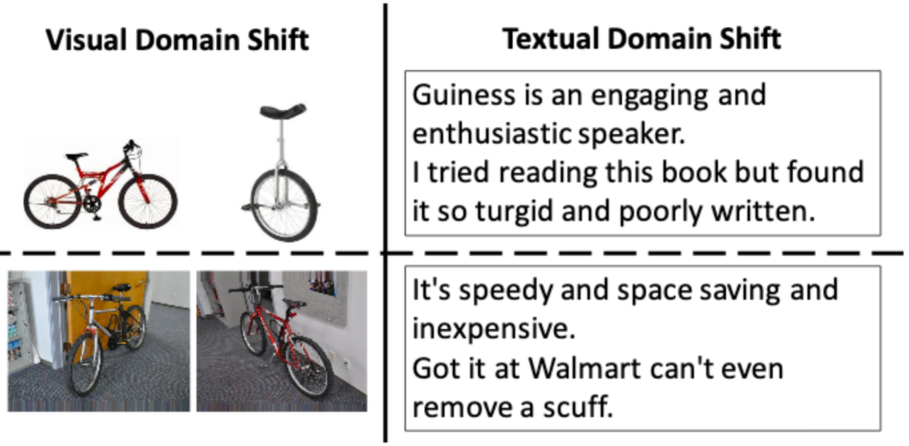
Ton+18 (https://doi.org/10.1016/j.spasta.2018.02.002)

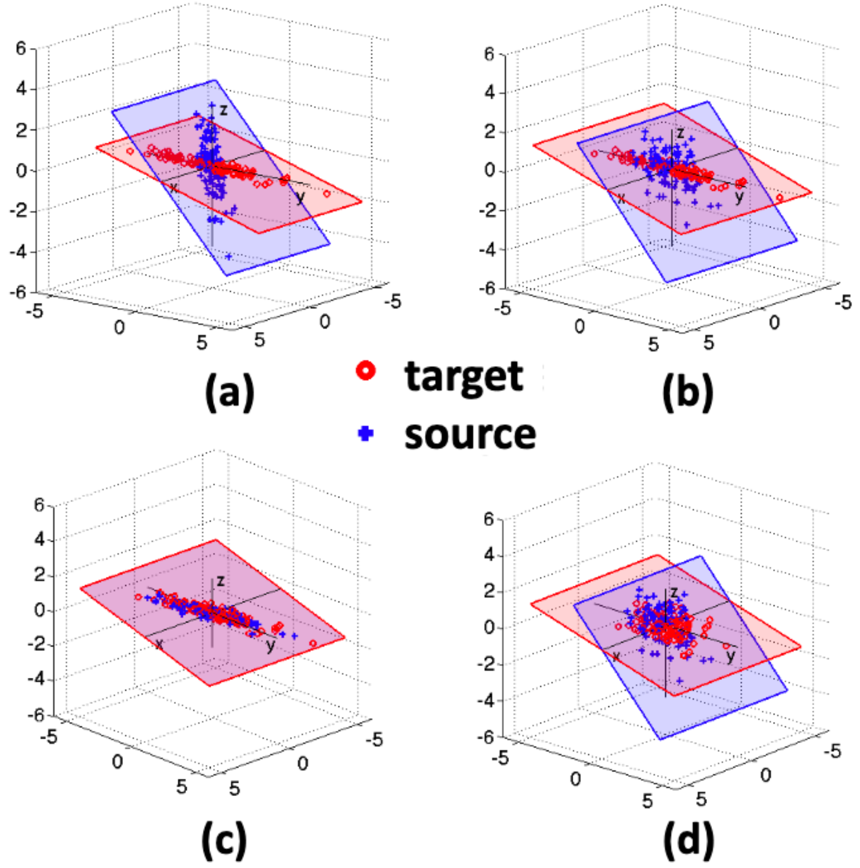
Answers: data homogenization, domain adaptation, transfer learning…

- Un/supervised domain adaptation, as in Daumé III (2007) and Sun et al. (2015)

source for training

target domain





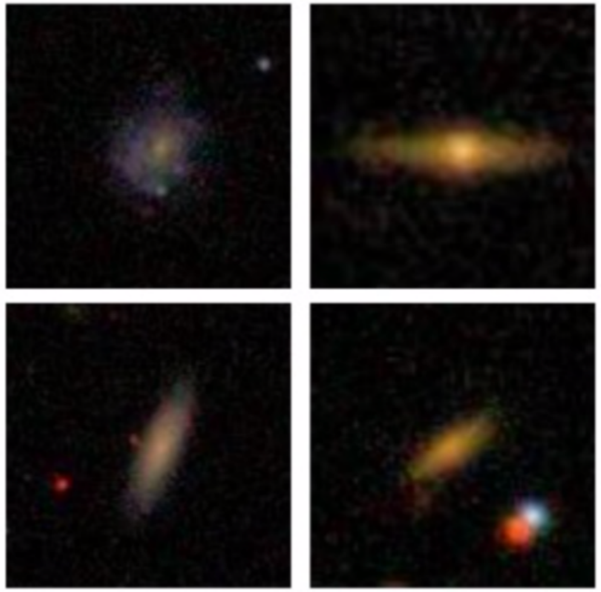
both domains have same features, but different distributions/correlations

many thanks to Viviana Acquaviva for introducing me to this topic

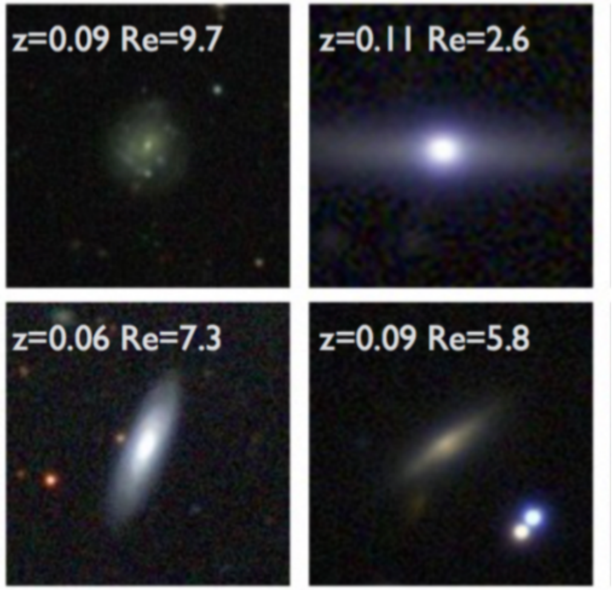
Answers: data homogenization, domain adaptation, transfer learning…

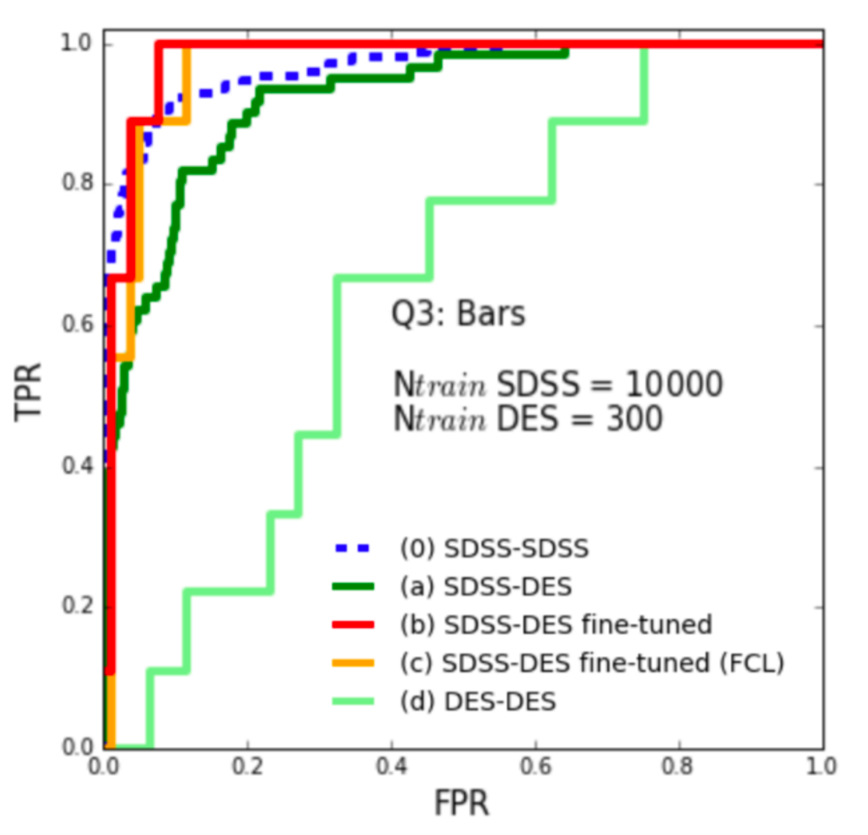
- A trained deep learning model applied to a new (unlabeled) data set. Dominguez-Sanchez et al. (2018):

Sloan Digital Sky Survey (SDSS):
galaxy morphology classified by eye
citizens science!



Dark Energy Survey:
better resolution but <10%
of data has a morphology label



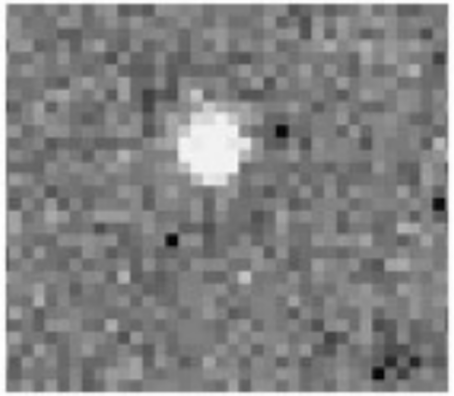


GWs application in George et al. (2018)

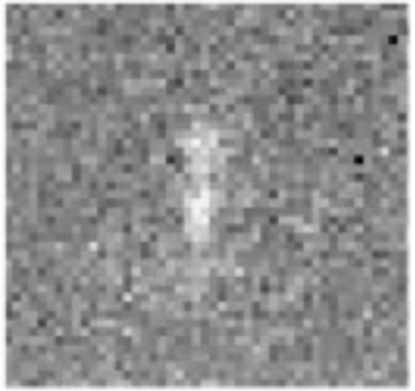
- How to deal with uncertainties?

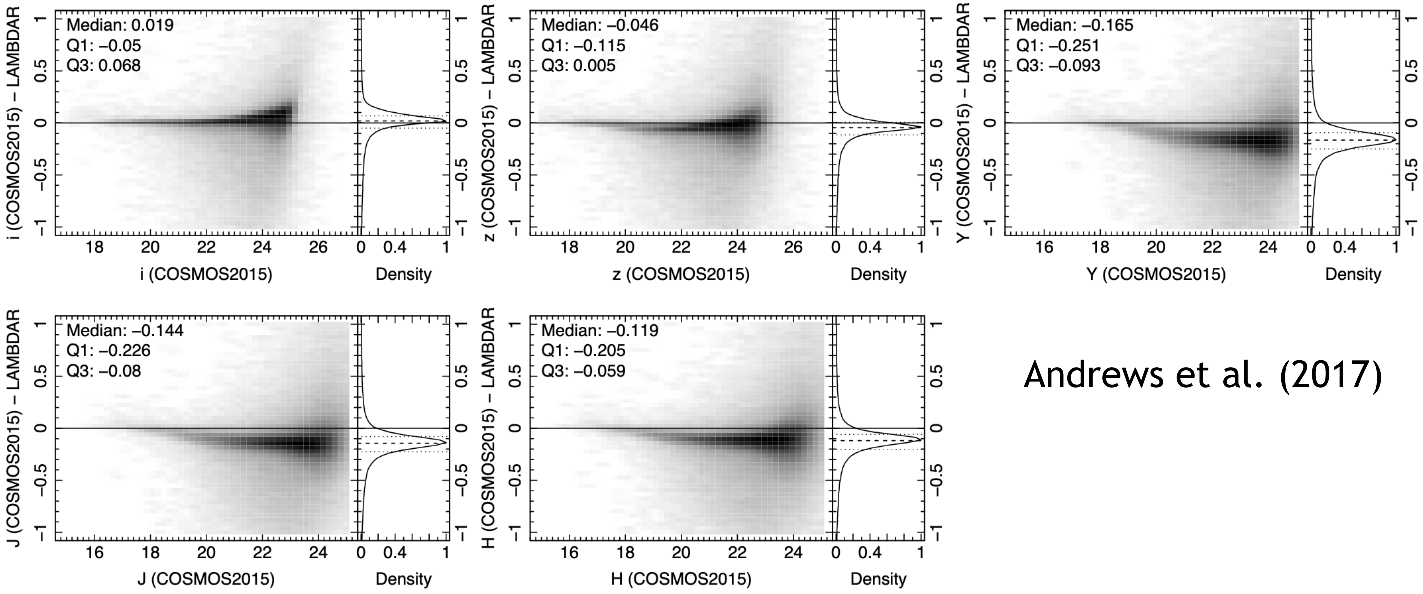
we have problems already at step zero: persistence, reflectance, cross-talk…

Viana & Bagget (2010)



these galaxies don’t exist





Andrews et al. (2017)

two teams reducing the same data
two different answers…

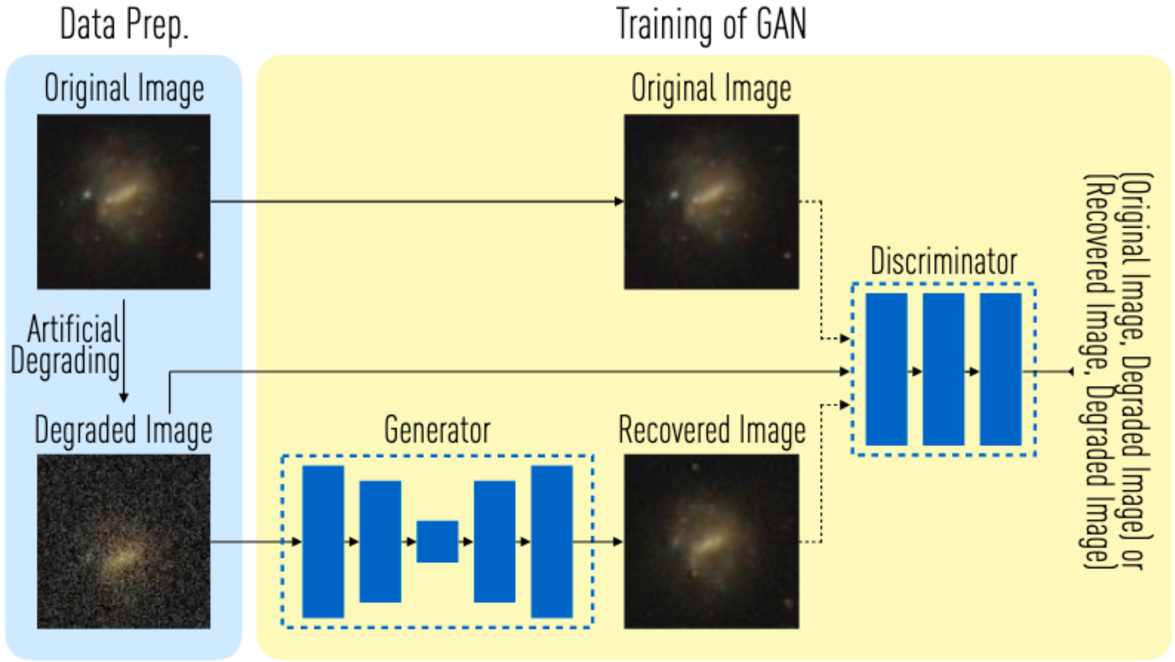
Answers: denoising, GAN, probabilistic ML…

- Extreme Error Deconvolution (see Bovy, Hogg, Roweis 2009)

going to skip this, unless anybody has something to say!

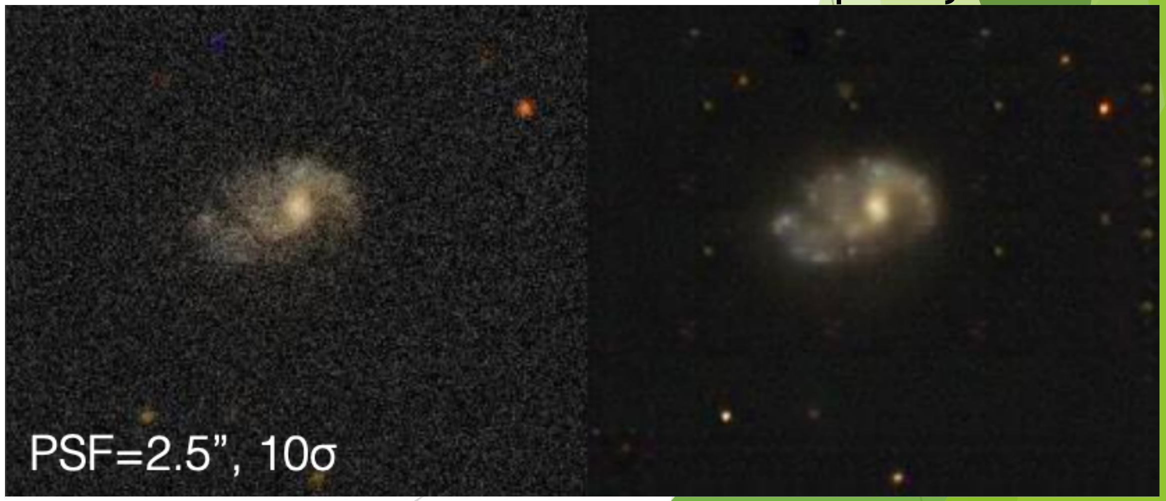
Answers: denoising, GAN, probabilistic ML…

- Generative Adversarial Network to "increase" image resolution, e.g., Schawinski et al. (2017):



Low resolution

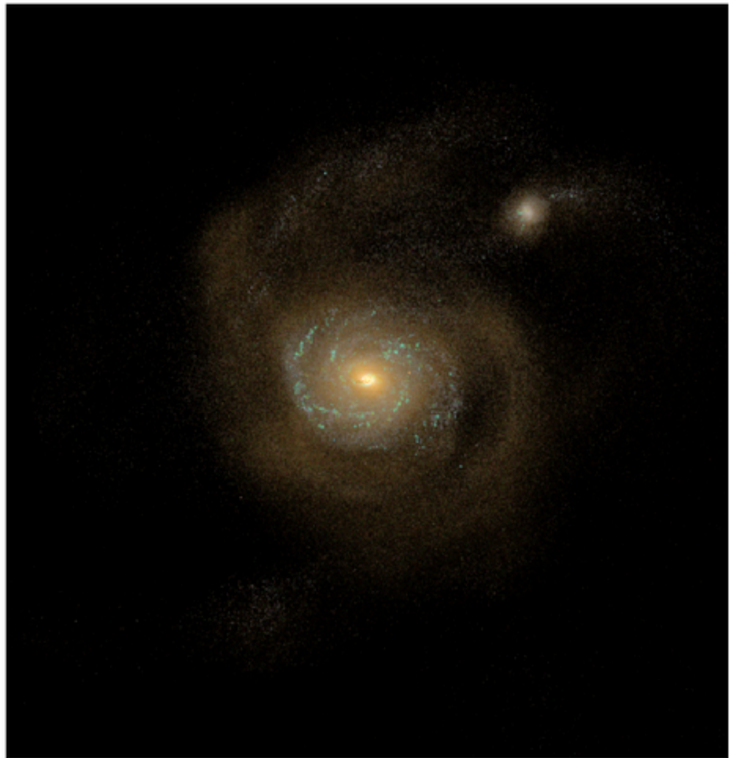
GAN reconstruction to Hubble quality



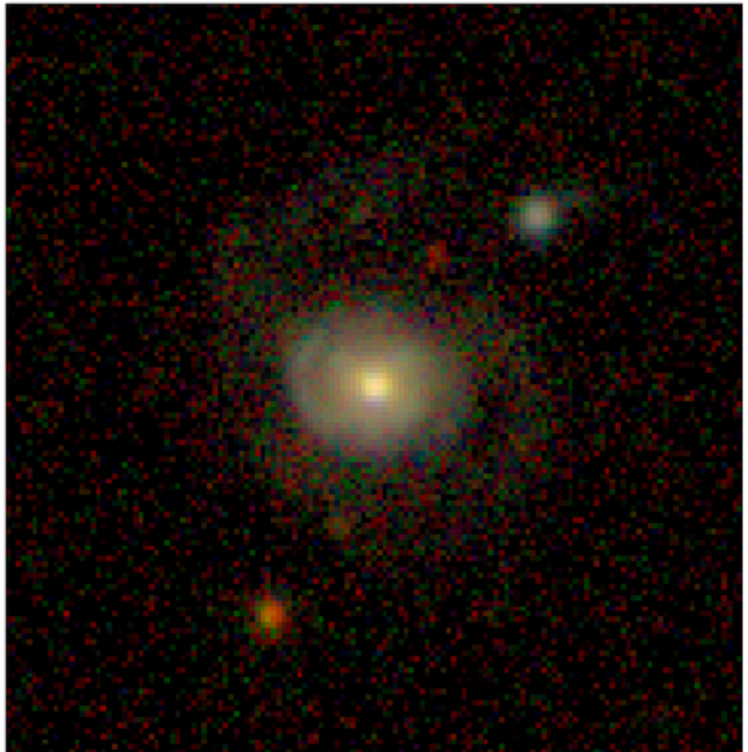
Answers: denoising, GAN, probabilistic ML…

- Another GAN, this time to add noise to a numerical simulation and make the synthetic images more realistic. Bottrell et al. et al. (2017a, 2019b)

idealized simulation



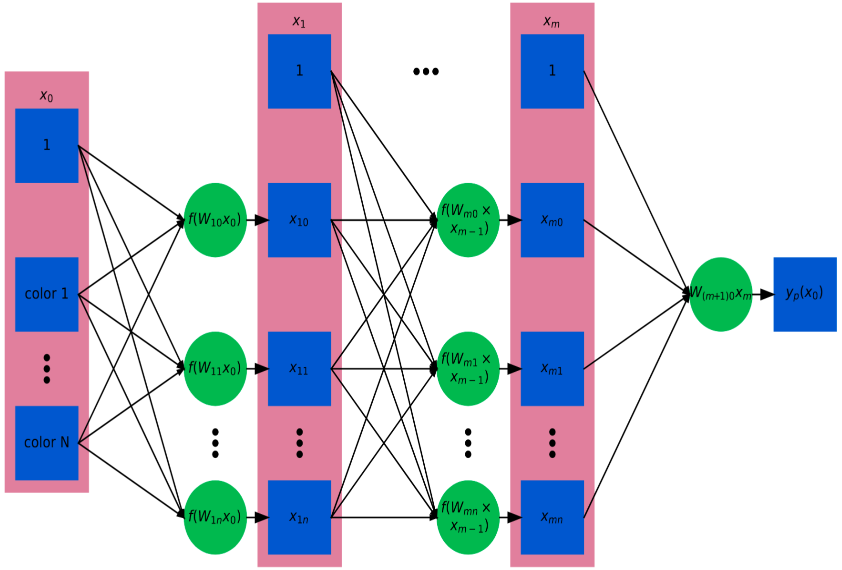
mimicking SDSS



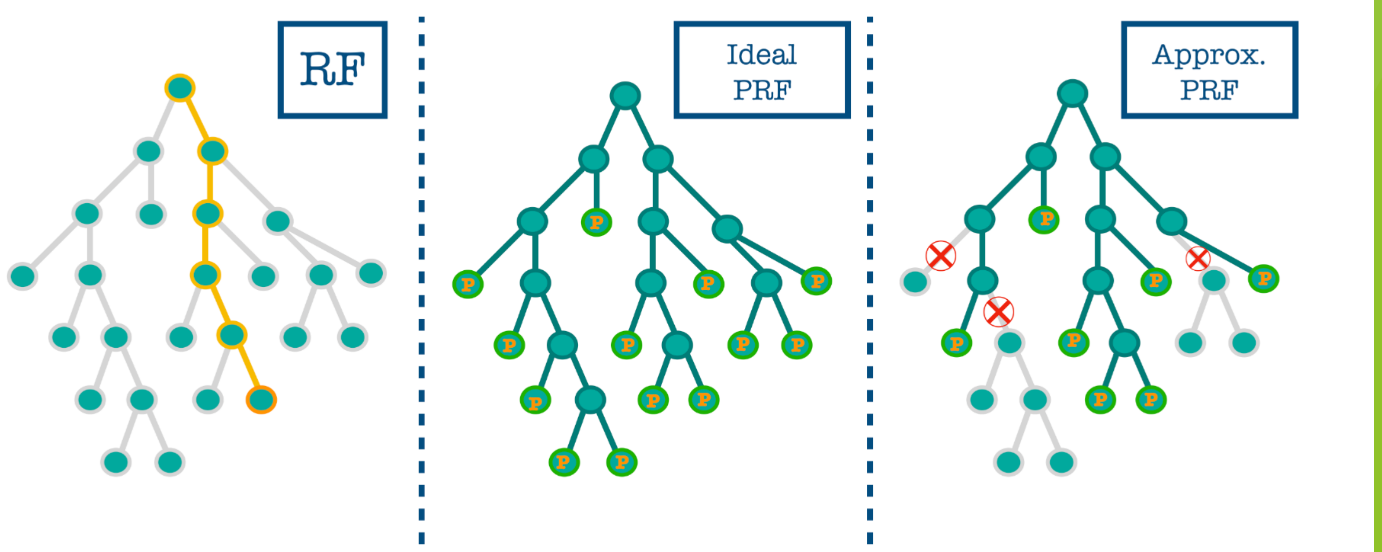
Answers: denoising, GAN, probabilistic ML…

- Treatment of uncertainties in the ML algorithm itself.

Bayesian Neural Networks: instead of fixed values the weights are described by a posterior predictive density



Probabilistic Random Forest
(Reis et al. 2018)



Twitter-size summary

Astrophysics offers exciting challenges at any level, from data collection to physical interpretation. ML can takle them.

As a distinctive feature of Astronomy is the complexity & uncertainty of data, I focused on ML applications dealing with that.

Thanks for watching!