Euclid – AI in the Dark Space

M. Poncet (1), M. Huertas-Company (2) & al

BiDS'19

(1) CNES

(2) Observatoire de Paris

Summary

- Euclid project
- Problematics
- Galaxy Zoo Example
- Galaxies Morphology & Structure Use Case
- Next Steps
- Some conclusions

Euclid Project and Consortium

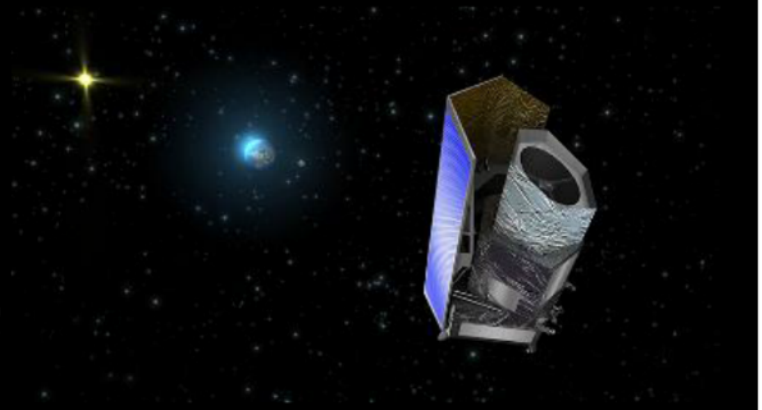
M2 mission in the framework of the **ESA Cosmic Vision Programme**

Euclid mission objective is to map the geometry and understand the nature of the dark Universe **(dark energy and dark matter)**

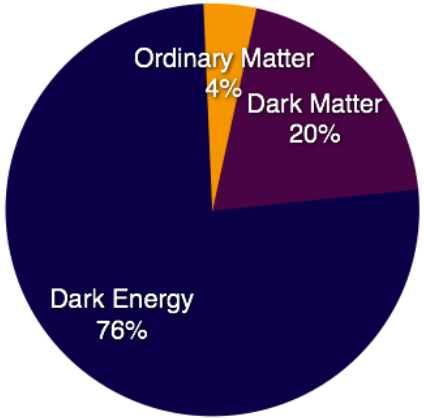
Actors in the mission: **ESA** and the **Euclid Consortium** (institutes from 15 European countries and USA, funded by their own national Space Agencies)

Euclid Consortium:
16 countries
220 labs
~1500 members

One of the biggest on going collaboration !



http://sci.esa.int/science-e/www/area/index.cfm?fareaid=102
http://www.euclid-ec.org

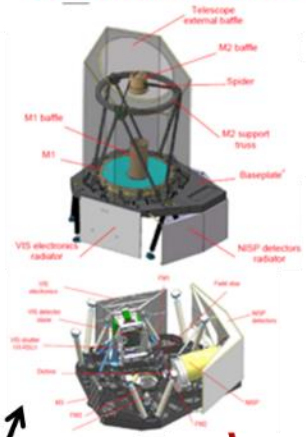


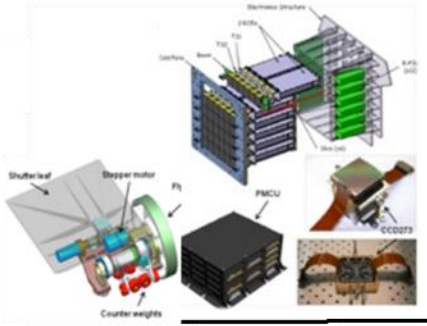
Euclid in a nutshell

Soyuz@Kourou
2022

ears

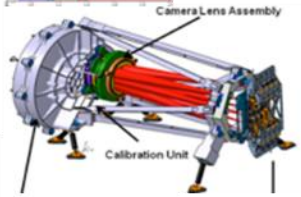
PLM+SVM: 2010-2028+





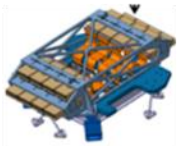
VIS imaging:
2010-2022

(VIS team)



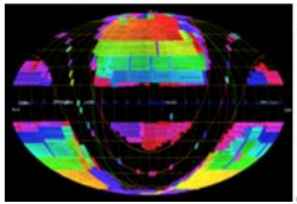
NIR spectro-imaging

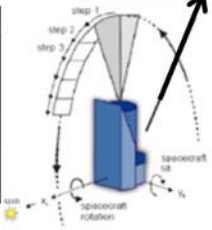
2010-2022 (NISP team)



Avoid Galaxy+Ecliptic

Surveys: 2010-2028+ (Survey WG)





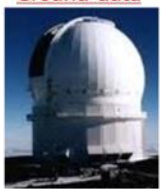
6.25 yrs - 15,000 deg²

Commisionning – SV

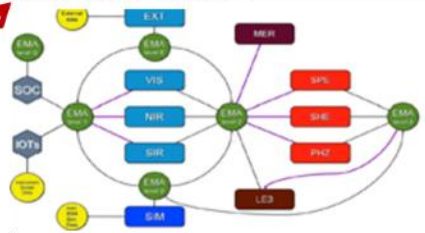
Euclid opération:

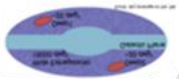
>5.5 yrs:Euclid Wide+Deep

Ground data



SGS: 2010-2029+





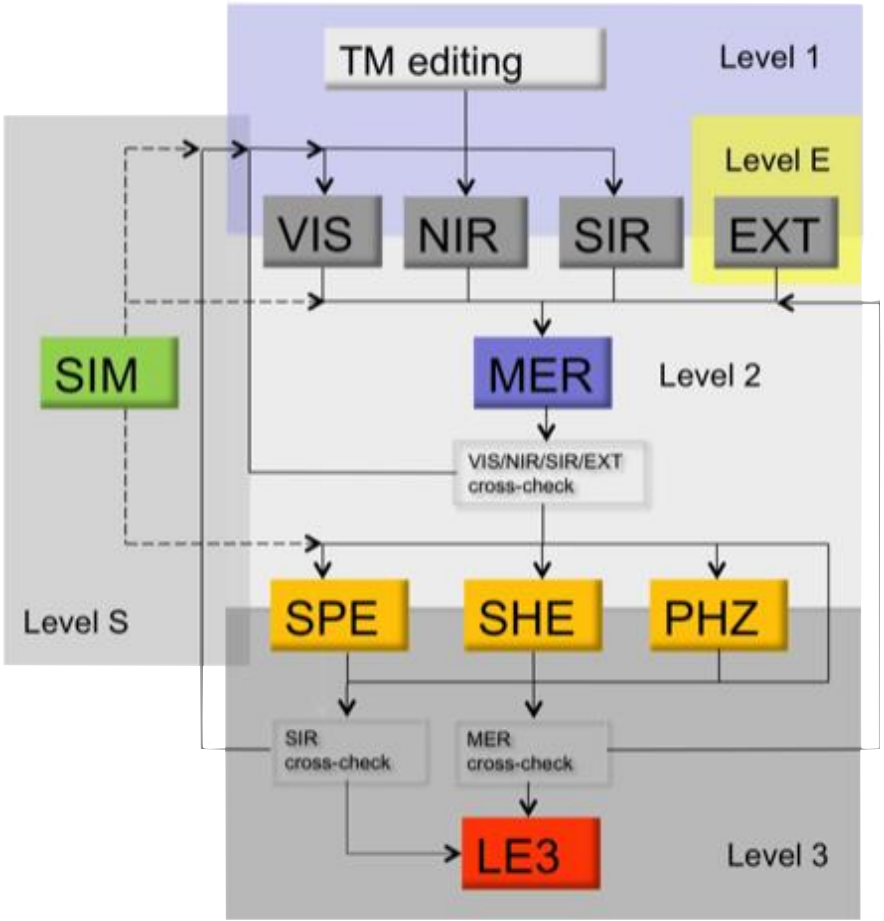
SWG:

2019-2029

+

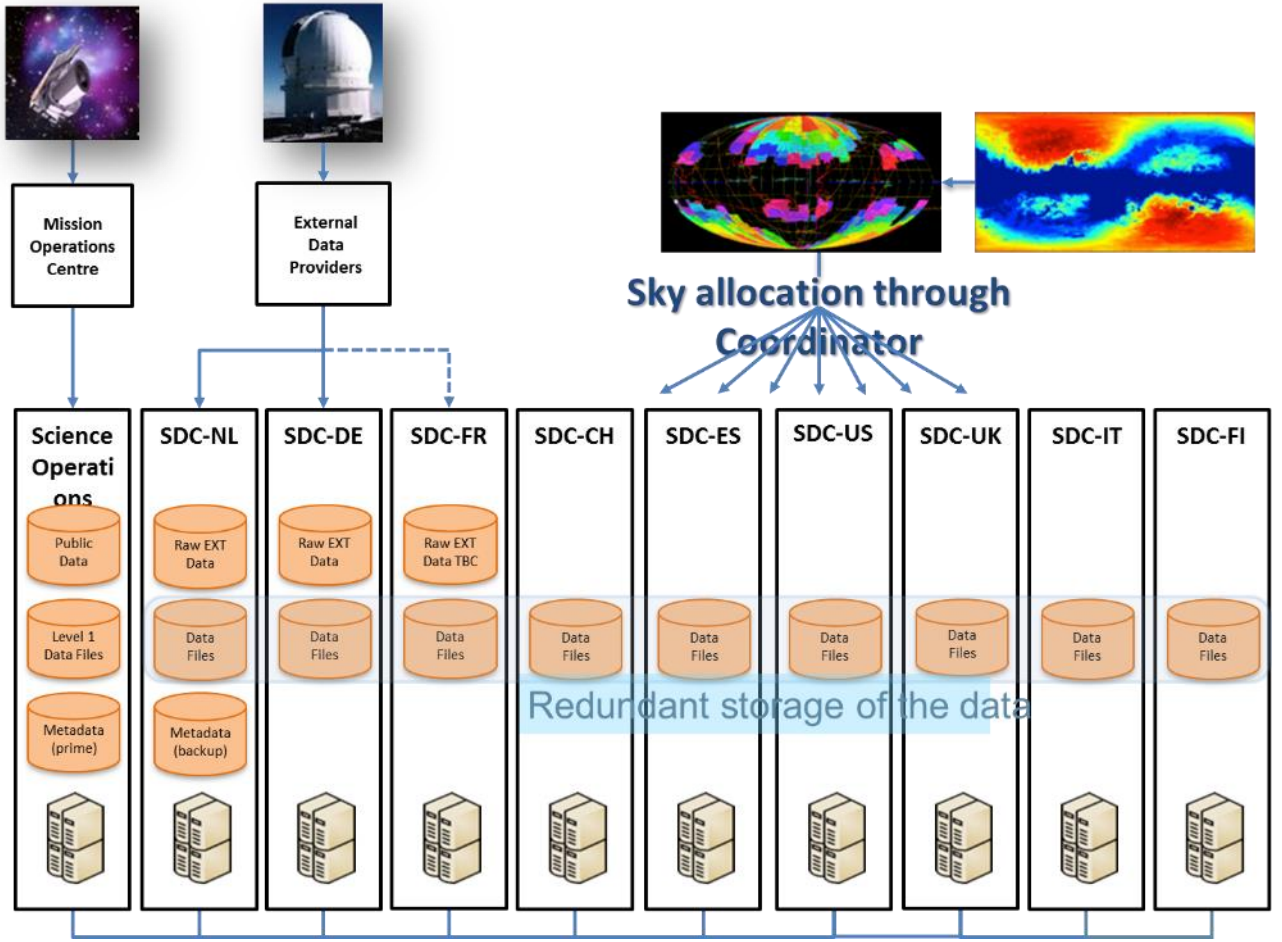
20-30 PB data processing (EC-SGS team) — Science analyses

Euclid Pipeline



- VIS, NIR, EXT: production of fully calibrated photometric exposures from Euclid and ground-based surveys
- SIR: production of fully calibrated 1D spectra extracted from the NISP spectroscopic exposures.
- MER: production of a source catalog containing consistent photometric and spectroscopic measurements.
- PHZ: production of the photometric redshift for all catalogued sources.
- SPE: production of spectroscopic redshifts for all sources with spectra.
- SHE: measurements of galaxy shapes.
- LE3: production of all high-level science products.
- SIM: production of all the simulated data necessary to validate the data processing stages, and to calibrate observational or method biases.

Euclid Ground Segment



Problematics

Some Euclid process cannot be efficiently automated by conventional algorithms and still requires human expertise and manual validation/rejection, e.g.:

- Deblending of galaxy sources (see paper),
- Galaxy structures and classification (see paper),
- Galaxy/Star distinction (see paper),
- Anomalies detections (outliers),
- Redshift assessment (see paper),
- Point Spread Function (PSF) modeling,
- Image deconvolution,
- Modified Gravity Cosmological Model Discrimination.

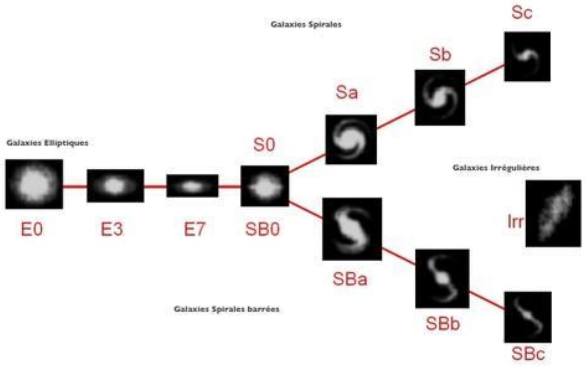
This is not compatible with Euclid constraints, since it would take many years to achieve the whole process for billion of objects.

Legacy solutions – Galaxy Zoo example

The Galaxy Zoo project (https://www.galaxyzoo.org), that aims at galaxy classification, involves ~20 000 volunteers who achieved ~2 million of classifications in months.

Classifications are achieved by human visual inspection after some training. Each galaxy is classified by several people and the majority "wins".

According to the last stats, 21 000 volonteers, 4000 classifications per day in average means ~680 years for 1 billions galaxies !

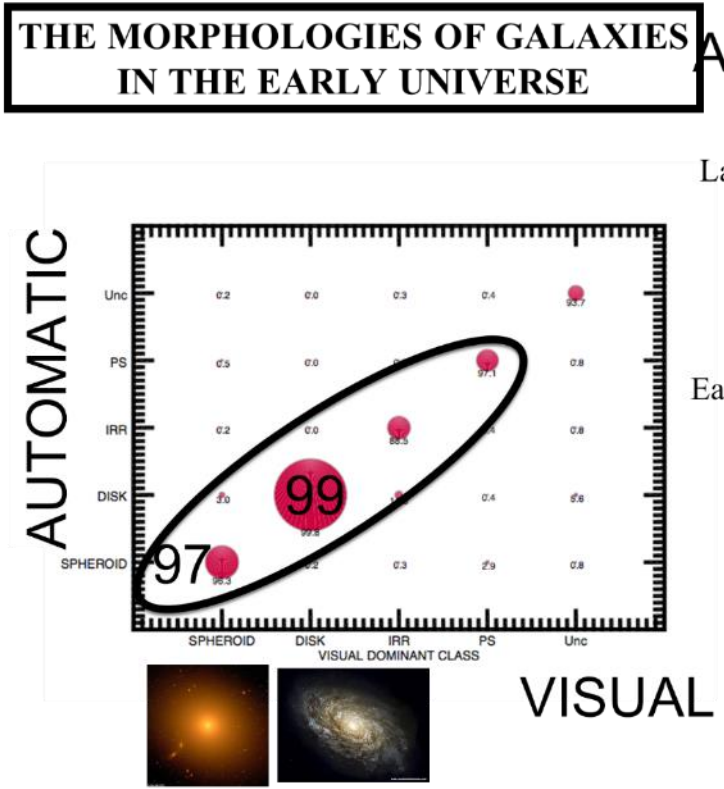


Thus, we have to find innovative ways in order to solve this issue. A segment of the Artificial Intelligence (AI) domain –Machine Learning (ML) and Deep Learning (DL) – is a very promising approach in this case.

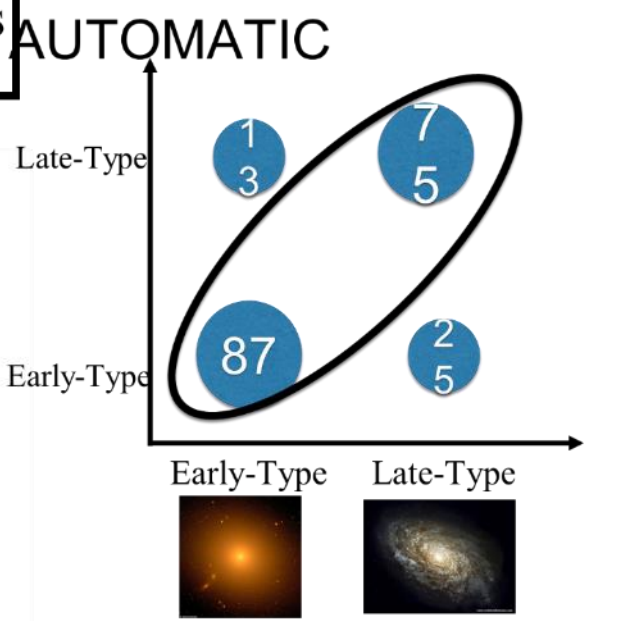
Coming back to the previous example, some tests show that the same amount of galaxy classification could be achieved efficiently with the ML/DL approach after the appropriate training phase.

Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES

THE MORPHOLOGIES OF GALAXIES IN THE EARLY UNIVERSE



CNNs



SVMs *MHC+14*

[BEFORE DEEP LEARNING]

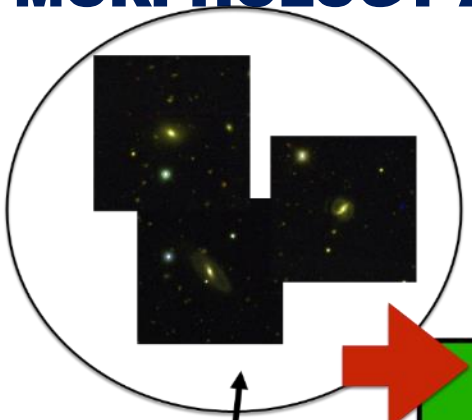
MHC+15b

Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES

DOES DEEP LEARNING SOLVE THE PROBLEM OF GALAXY MORPHOLOGICAL CLASSIFICATION?

…WE STILL NEED TRAINING SETS!

Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES



DATA FROM NEW DATASET

DEEP LEARNING MODEL

Classifications for for the entire sample with reduced human labelling

Transfer knowledge?

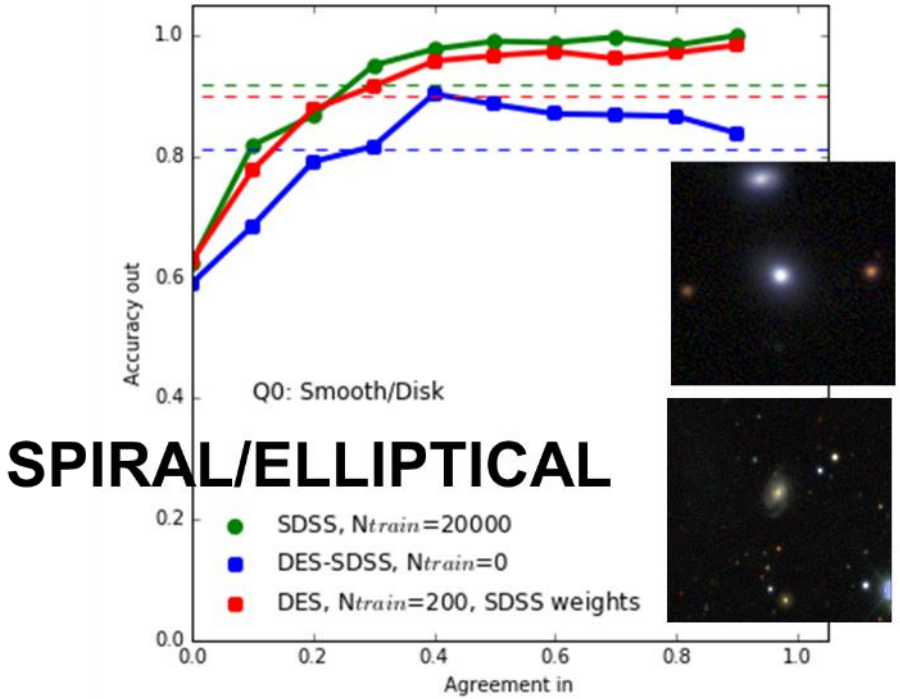
Human classifications from existing dataset

DOMINGUEZ-SANCHEZ, MHC et al. 18b

Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES

Knowledge transfer for galaxy morphology

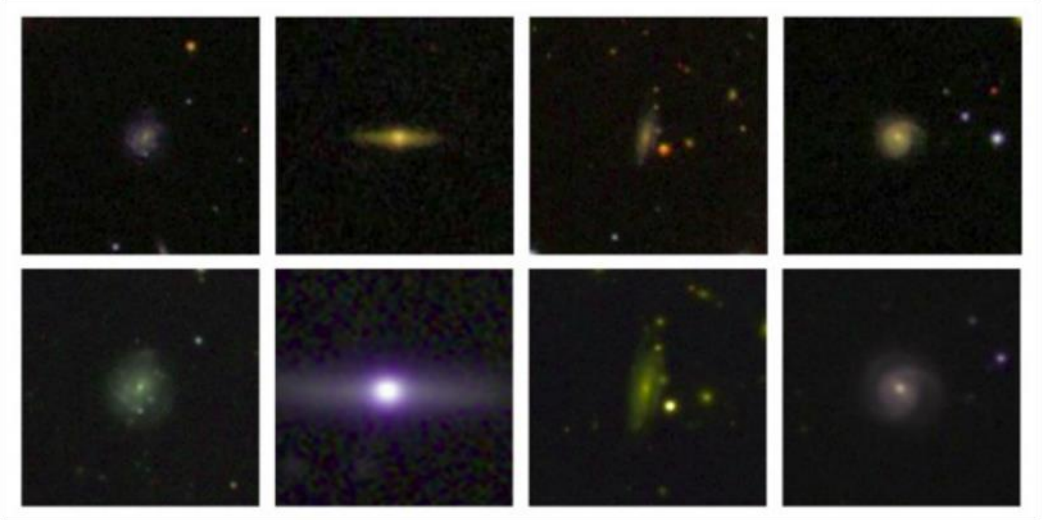
DOMINGUEZ-SANCHEZ, HUERTAS-COMPANY, BERNARDI et al. 18b



Only 200 (**1%!**) of new classifications are needed to reach an accuracy >90% if a machine trained previously is used

Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES

OPEN QUESTION: HOW TO EFFICIENTLY TRANSFER KNOWLEDGE BETWEEN DIFFERENT DATASETS ?

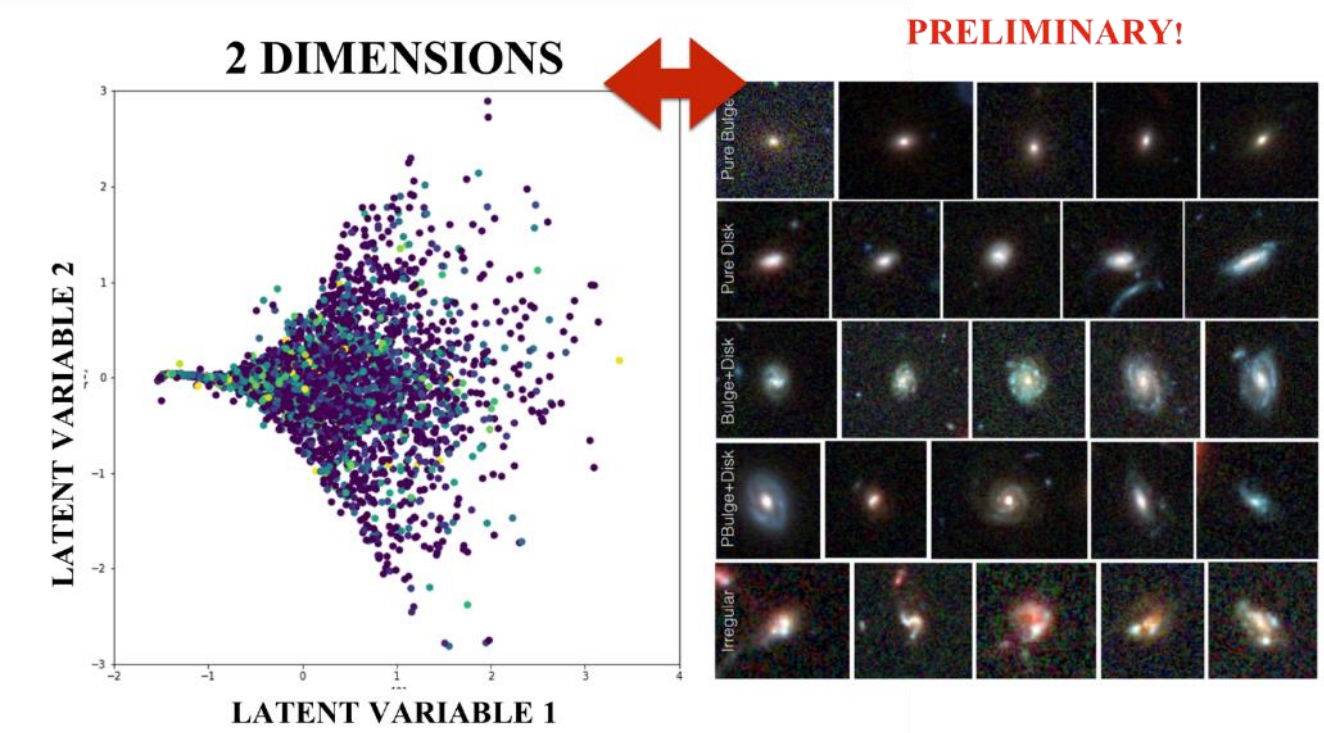


ANY WAY TO SELECT THE OPTIMAL TRAINING SET? ANY WAY TO OPTIMALLY SELECT THE LAYERS TO TRAIN / FREEZE?

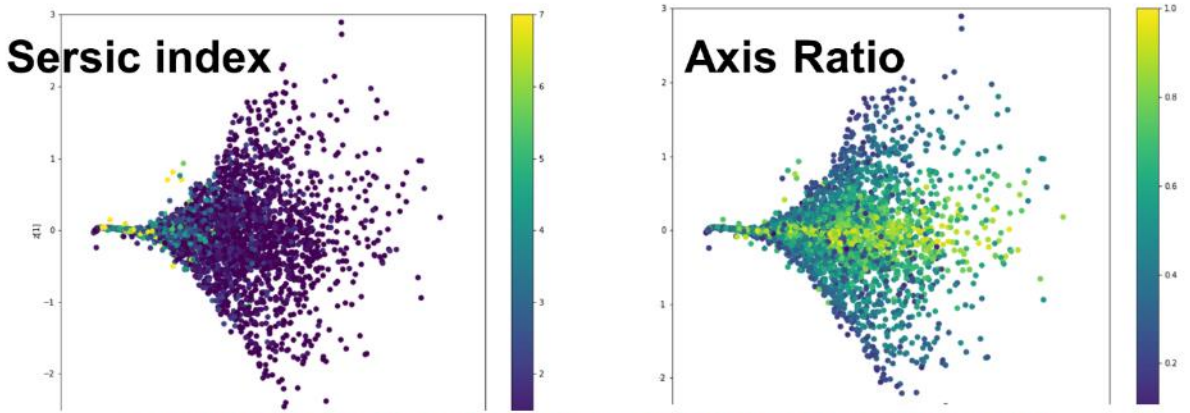
Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES

ALTERNATIVES USING UNSUPVERVISED APPROACHES

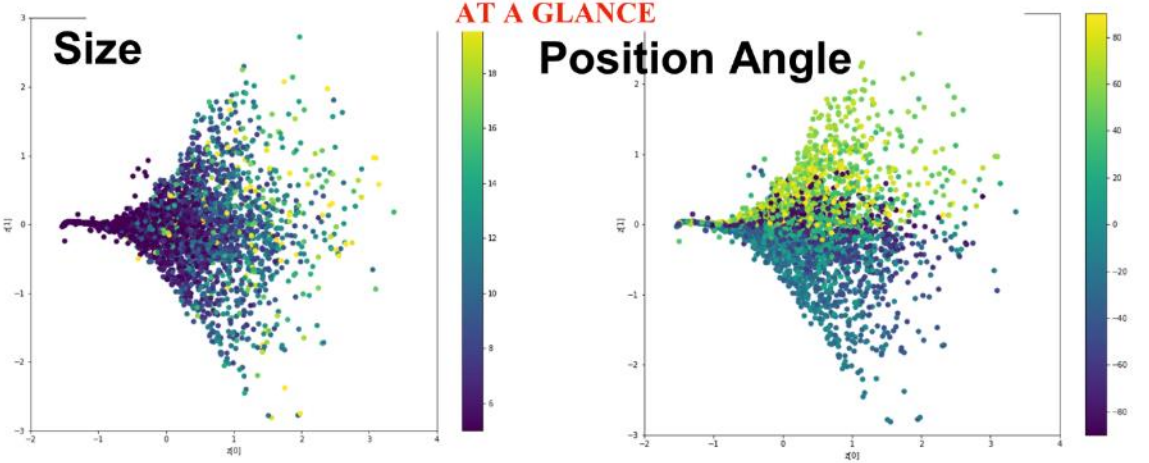
VAE DERIVED LATENT SPACE FOR CANDELS GALAXIES [H BAND]



Use Case: MORPHOLOGY AND STRUCTURE OF GALAXIES



EXPLORE THE MORPHOLOGICAL DISTRIBUTION OF GALAXIES AT A GLANCE



Next steps

Much remains to be done…, e.g.:

- **Setting up of a working group dedicated to IA approach for Euclid pipeline,**
- **Gathering of existing initiatives and needs,**
- **Bringing them together to constitute corresponding teams,**
- **Selection of the Euclid ML/DL framework,**
- **Setting up GPGPU cluster(s) for development training phase,**
- **Using SDCs' HTC clusters for production phase.**

Some conclusions

- **Many prototypes already exist inside Euclid Consortium that are kind of Proof of Concept.**
- **Need to promote them to production code and to setup the corresponding developement tools, as well as training and production infrastructures.**
- **ML/DL is key factor of feasibility and success where conventional algorithms and/or manual processing/validation are no more relevant.**

thank you!