

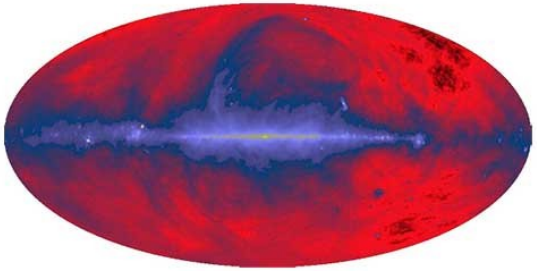
Machine learning in astrophysics: old and new challenges

Agnieszka Pollo

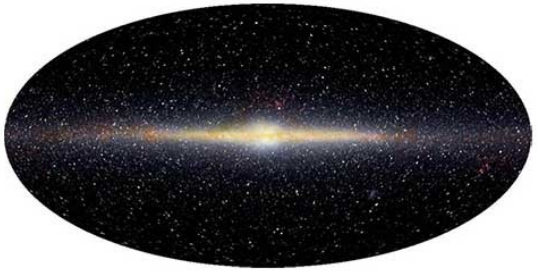
National Centre for Nuclear Research &
Jagiellonian University, Poland

with the team
Kasia Malek, Hareesh Thuruthipilly, Margherita Grespan, Junais, Ola Solarz, Luis Suelves, Will Pearson, Francesco Pistis, Gosia Siudek, Artem Poliszczuk, Szymon Nakoneczny, Aditya Narendra and others

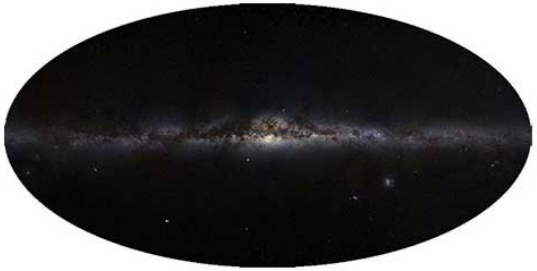
Warsaw, 07.09.2026



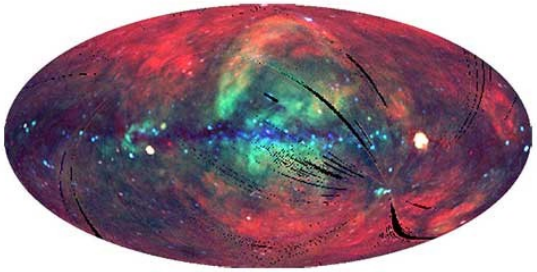
radio



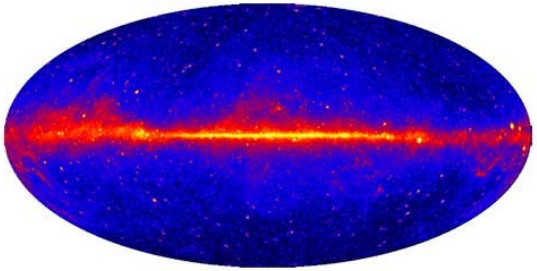
infrared



optical



X-ray

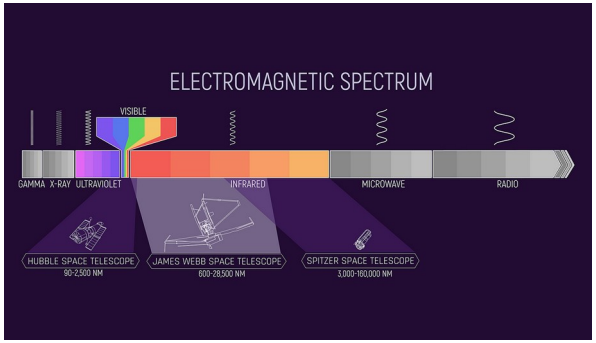


gamma-ray

Astrophysical data: the same sky, but...

- Different kinds of light
  - Radio
  - mm
  - IR
  - optical
  - UV
  - X
  - Gamma rays

multi-wavelength astronomy



NASA, ESA, CSA, Joseph Olmsted (STScI)

Copyright: NASA. Radio: Haslam et al. 1982; infrared: NASA; optical: ESO/S. Brunier; X-ray: Max Planck Institute for Extraterrestrial Physics and S. L. Snowden; gamma-ray: NASA/DOE/Fermi LAT Collaboration)

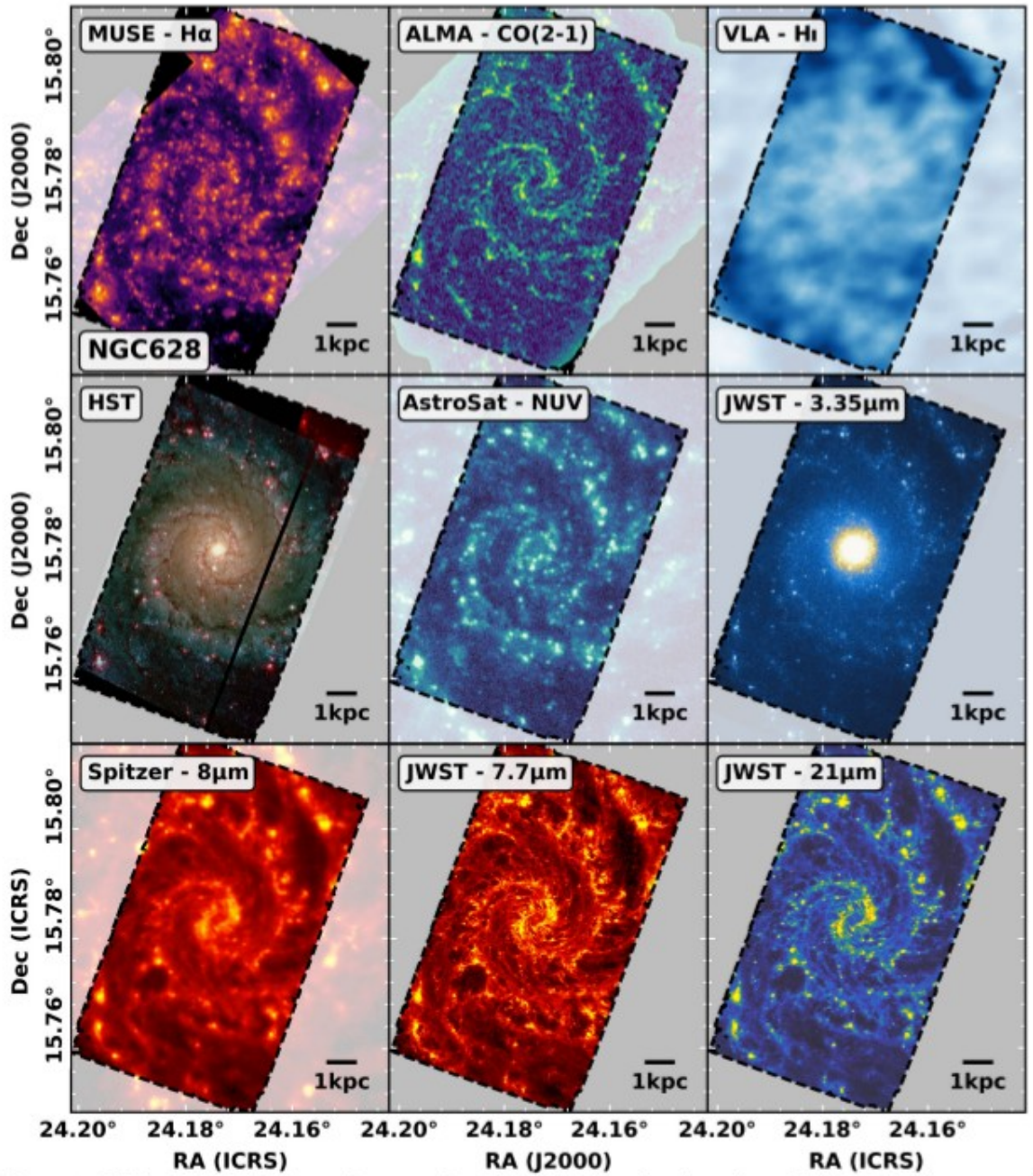


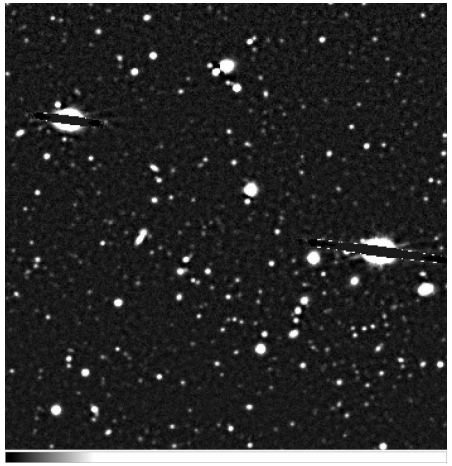
Figure 2. Multi-wavelength observations covering the nearby star-forming galaxy NGC 628. In the top row of panels from left to right, we show the Hα map from PHANGS–MUSE (Emsellem et al. 2022), CO (2-1) peak intensity from PHANGS–ALMA, (Leroy et al. 2021), and the VLA HI moment 0 map from the THINGS survey (Walter et al. 2008). In the middle row, left panel, we show an image produced from the *B* (blue), *V* (green) and *I* (red) band filters from *HST*, and the continuum subtracted *HST*–Hα (658 nm; red, see Lee et al. 2022). In the middle center and right panels, we show the UV emission map from AstroSat (Hassani et al. in prep) and NIRCam F335M emission from *JWST* (Lee et al. subm.). On the bottom row, from left to right, we show the 8 μm map from *Spitzer* (Dale et al. 2009), the F770W map and the F2100W map from MIRI *JWST* (Lee et al. subm.). All panels have been overlaid with a dashed contour showing the coverage of the *JWST* observations.

M74 (Phantom Galaxy) - different appearance at different wavelengths

Watkins et al. 2022

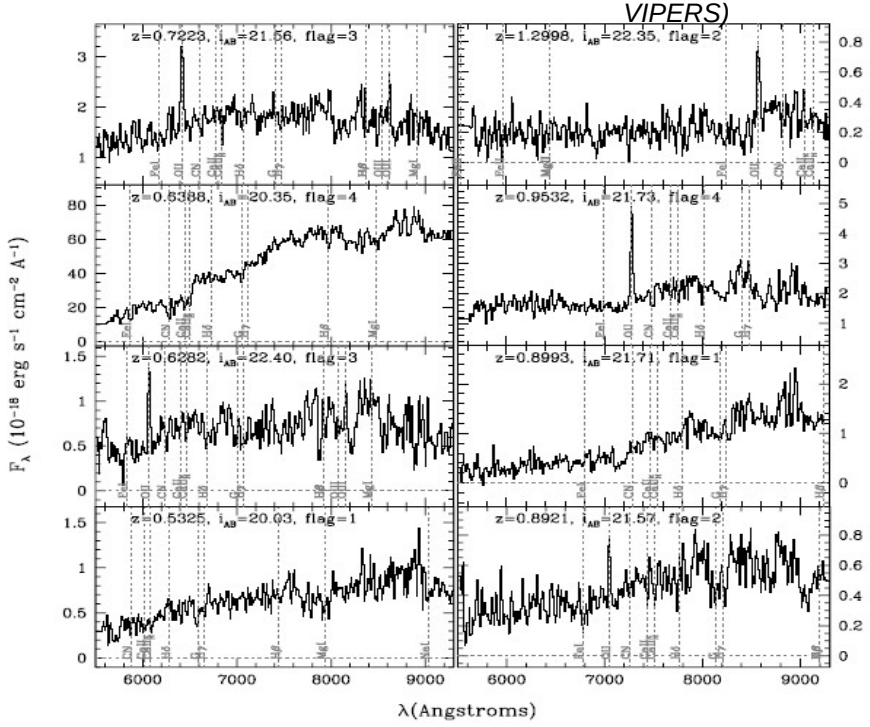
Different light properties

(Copyright: Steven Mihos



Position/brightness

(Copyright: VIPERS)

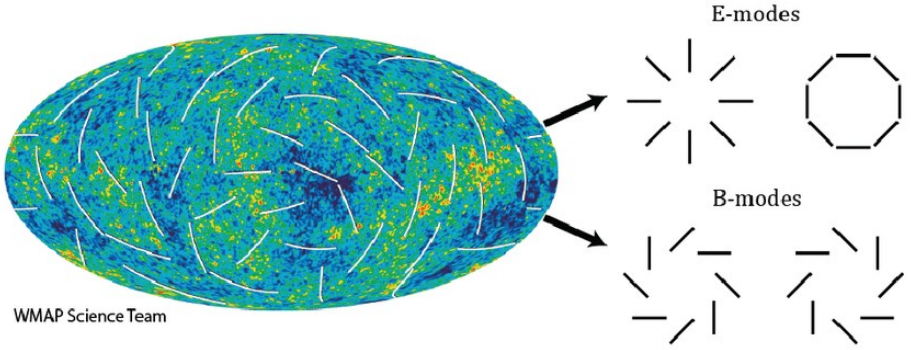


spectroscopy



colour

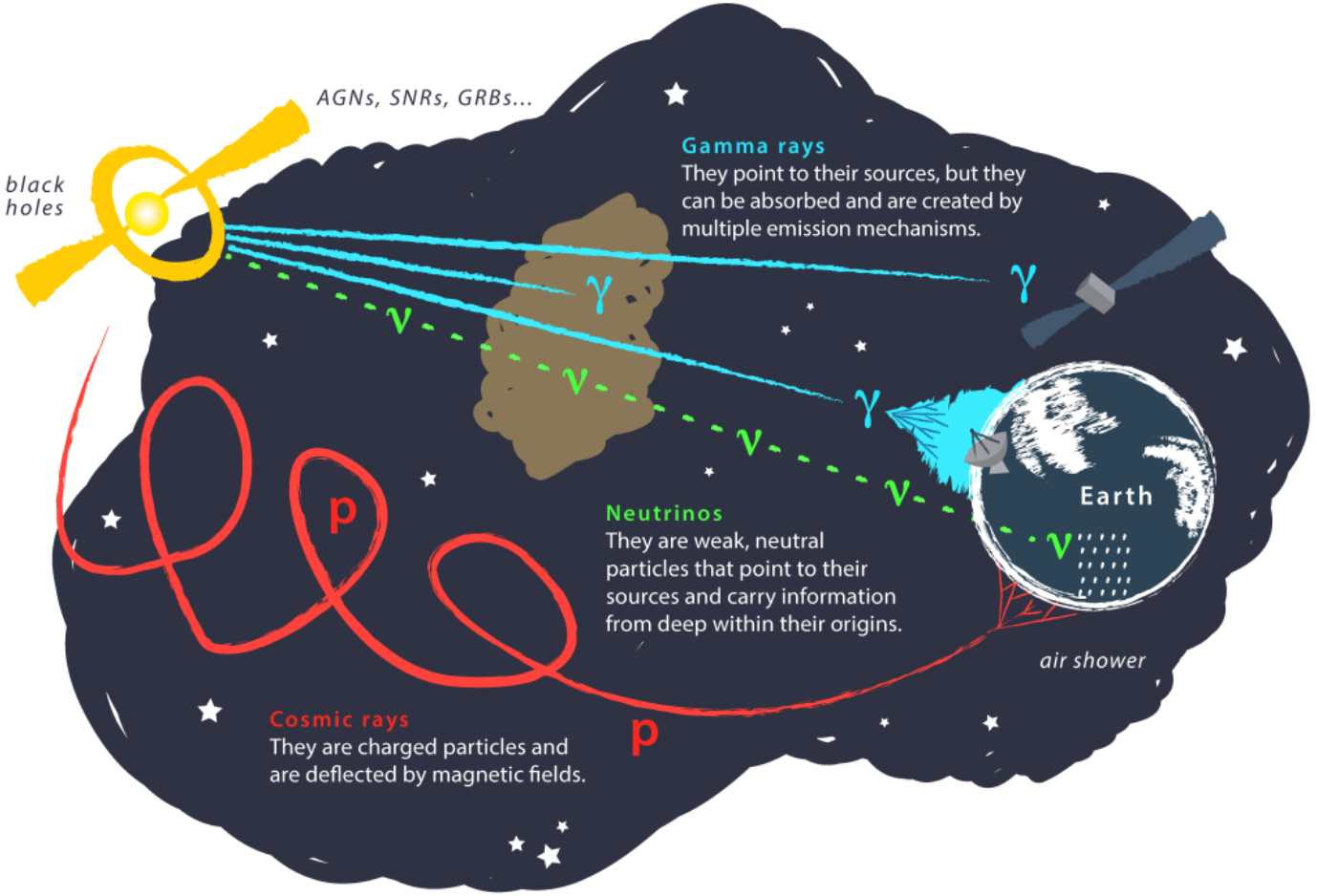
polarization



Credit: NASA, ESA, CSA, and STScI.

Astronomy: other observables

- Particles
  - Neutrinos
  - Cosmic rays

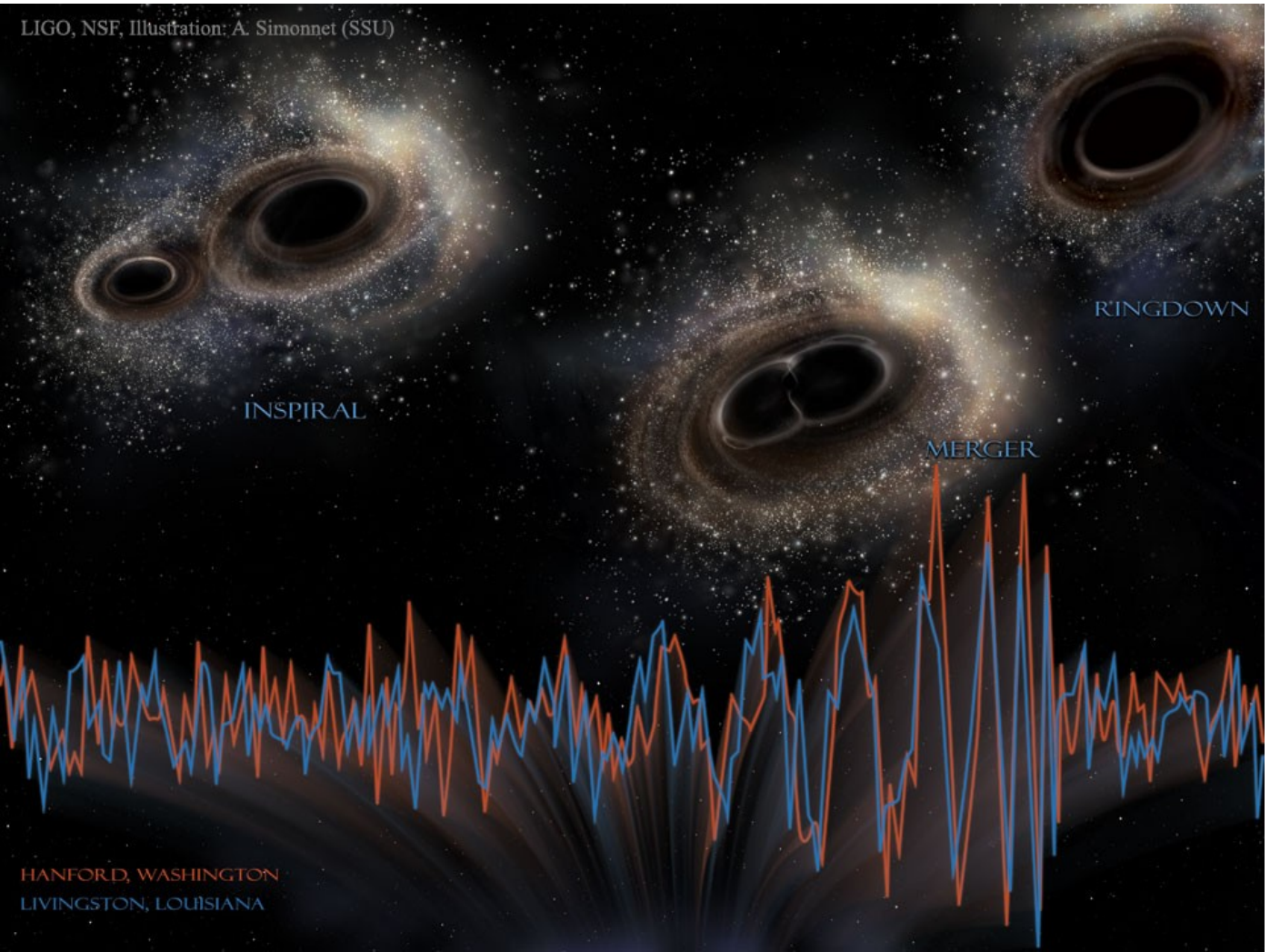


Astroparticle physics

Image: Juan Antonio Aguilar and Jamie Yang.
IceCube/WIPAC

Astronomy: other observables

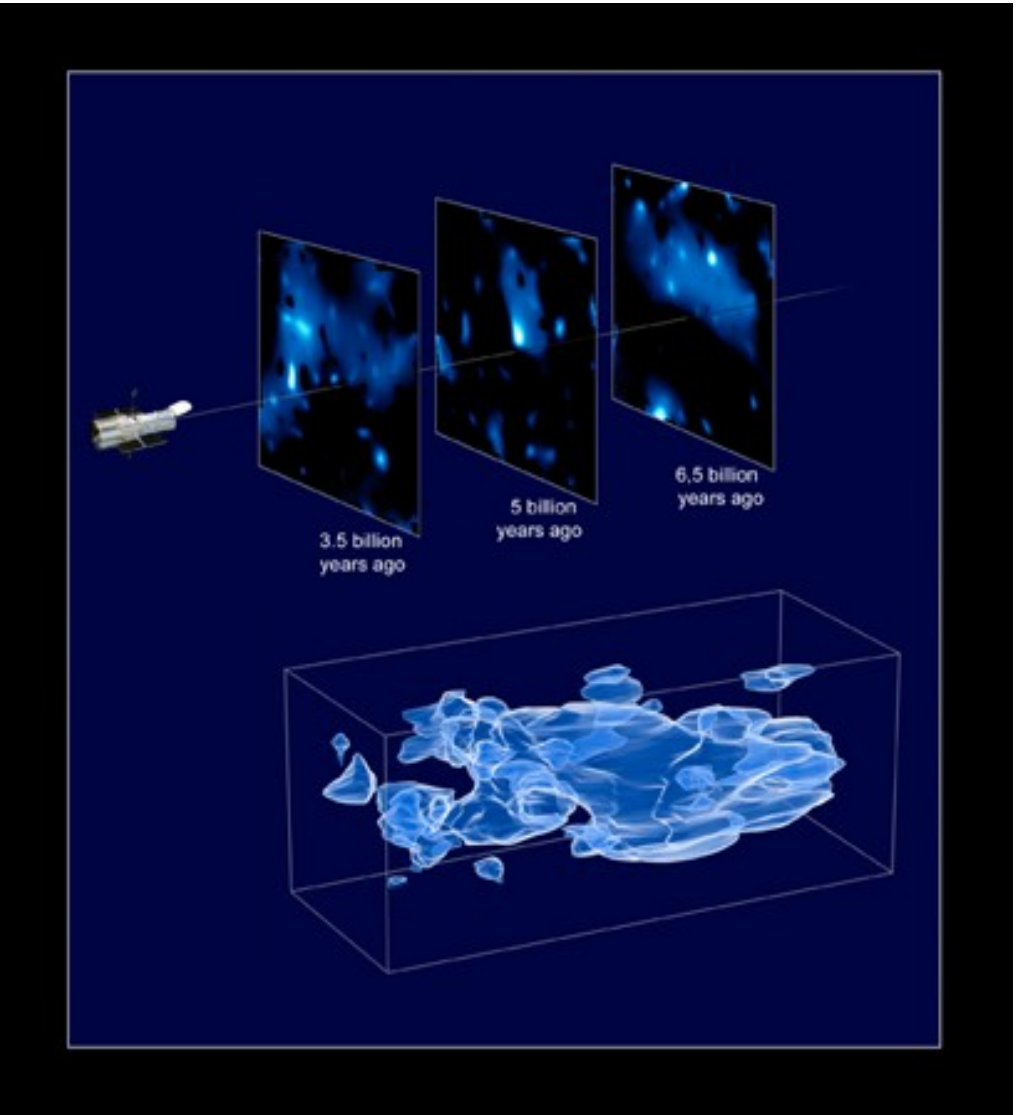
- Gravitational waves



gravitational wave astronomy

Credit: LIGO, NSF, Aurore Simonnet (Sonoma State U.)

Astronomy: other observables



- Indirect measurements
  - e.g. cosmological tomography through weak gravitational lensing

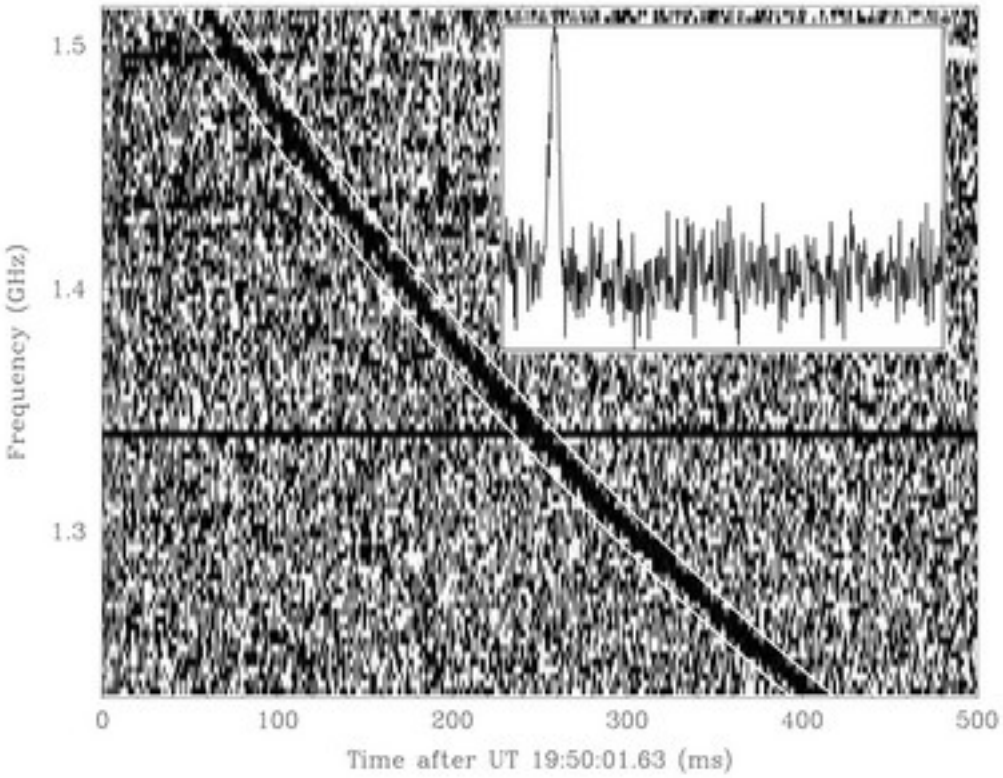
precision cosmology

Credit: ESA

Credit: Swinburn. Pierwszy szybki rozbłysk radiowy.

Astronomy: other observables

- Variability
  - Variable objects: stars, quasars...
  - Moving objects (comets, asteroids, stars)
  - Transient events (supernovae, kilonovae, GRBs, FRBs...)



time domain astronomy

Modern (near future) astronomy

- Radio
- mm
- IR
- optyka
- UV
- X
- gamma

Multiwavelength astronomy

Multimessenger astronomy

Particle astrophysics

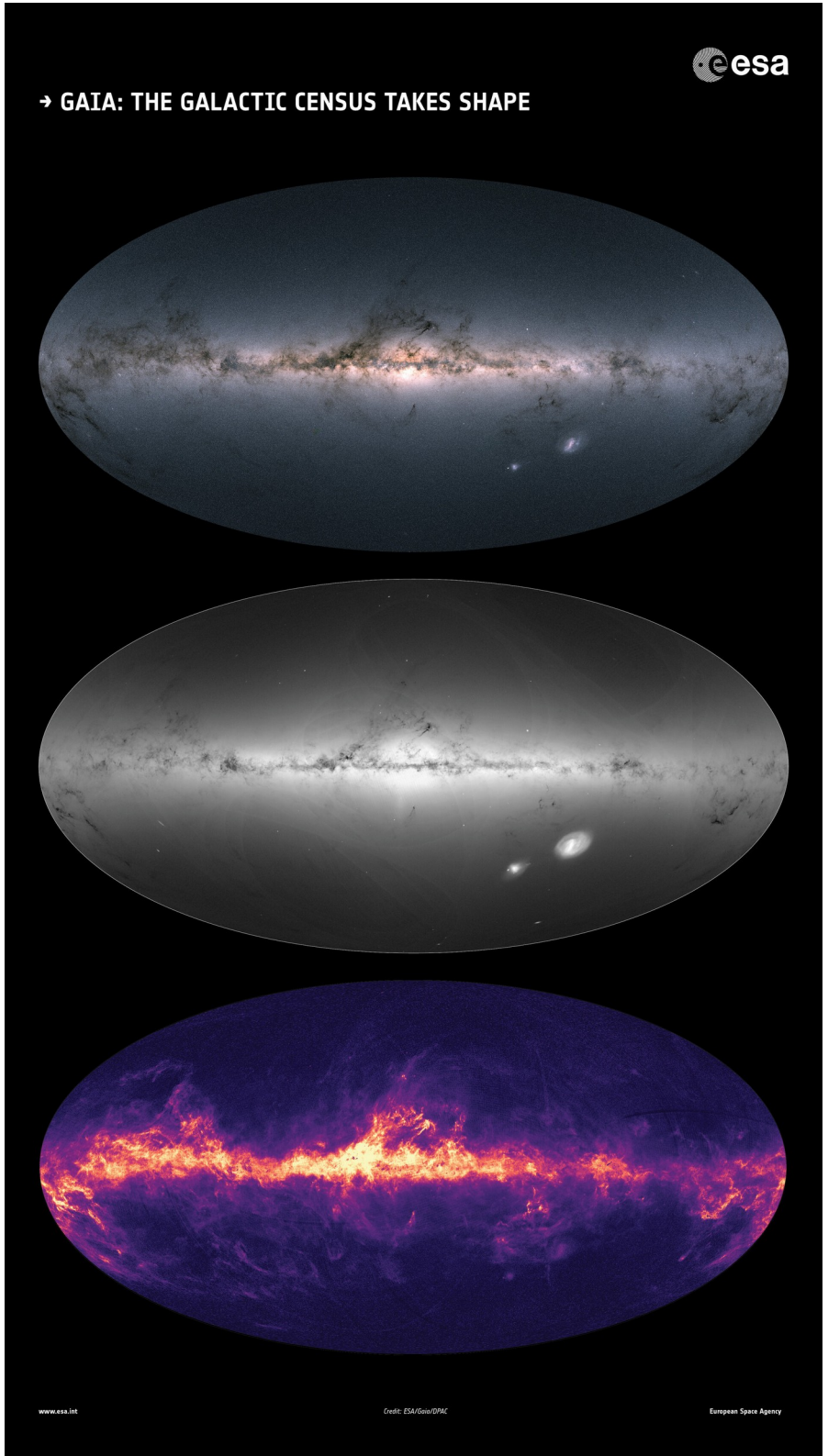
GW astrophysics

Other data

Time domain

Changes in the data domain: Milky Way

- ~5,000 stars visible to the naked eye
- ~1004 stars in the Tycho Brahe/Johannes Kepler catalog (1627)
- 1993: Hipparcos Catalogue: 118 218 stars
- 2022: Gaia EDR4: 1,811,709,771 = 10^9 sources - (mostly) stars
- (total in Milky Way: 100 thousand million = 10^11 stars)



Presently we know

~2 bln (2*10^9) Milky Way stars

estimated 100-400 bln = 1-4*10^11, i.e. we know ~1% of all

near future: Very Rubin expected to observe ~17 bln of stars gwiazd

Acknowledgement: Gaia Data Processing and Analysis Consortium (DPAC); Top and middle: A. Moitinho / A. F. Silva / M. Barros / C. Barata, University of Lisbon, Portugal; H. Savietto, Fork Research, Portugal; Bottom: Gaia Coordination Unit 8; M. Fouesneau / C. Bailer-Jones, Max Planck Institute for Astronomy, Heidelberg, Germany.

Changes in the data domain: extragalactic world

~3 galaxies seen by naked eye

~110 „nebulae” (among them 40 galaxies) in the Messier catalogue (1774)

1888-1908: New General Catalogue of Nebulae and Clusters of Stars (NGC): 7,840 (+5,386)

~1990: APM (Automated Plate Measurement): 14,681 nearby galaxies

~1990: CfA2 Redshift Survey: 18,000 nearby galaxies

1995: CFRS – Canada-France Redshift Survey – of 700 galaxies

~2000: SDSS - ~150,000 of nearby galaxies and quasars

2000-2010: deep surveys -> ~ ~1000 to ~30,000 galaxies

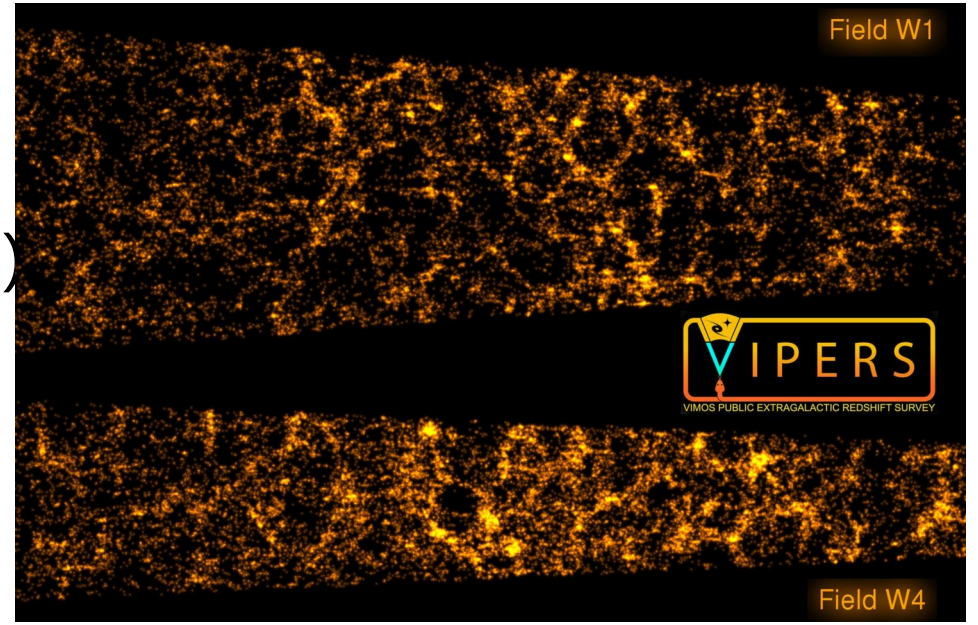
2010-2020: deep surveys -> ~100,000 galaxies; local surveys: SDSS i and its next editions: several mln of galaxies

2026: DESI – 100,000 galaxies in one night; 47 mln in the last data release (two years ago: it was 1 mln galaxies)

Presently we know

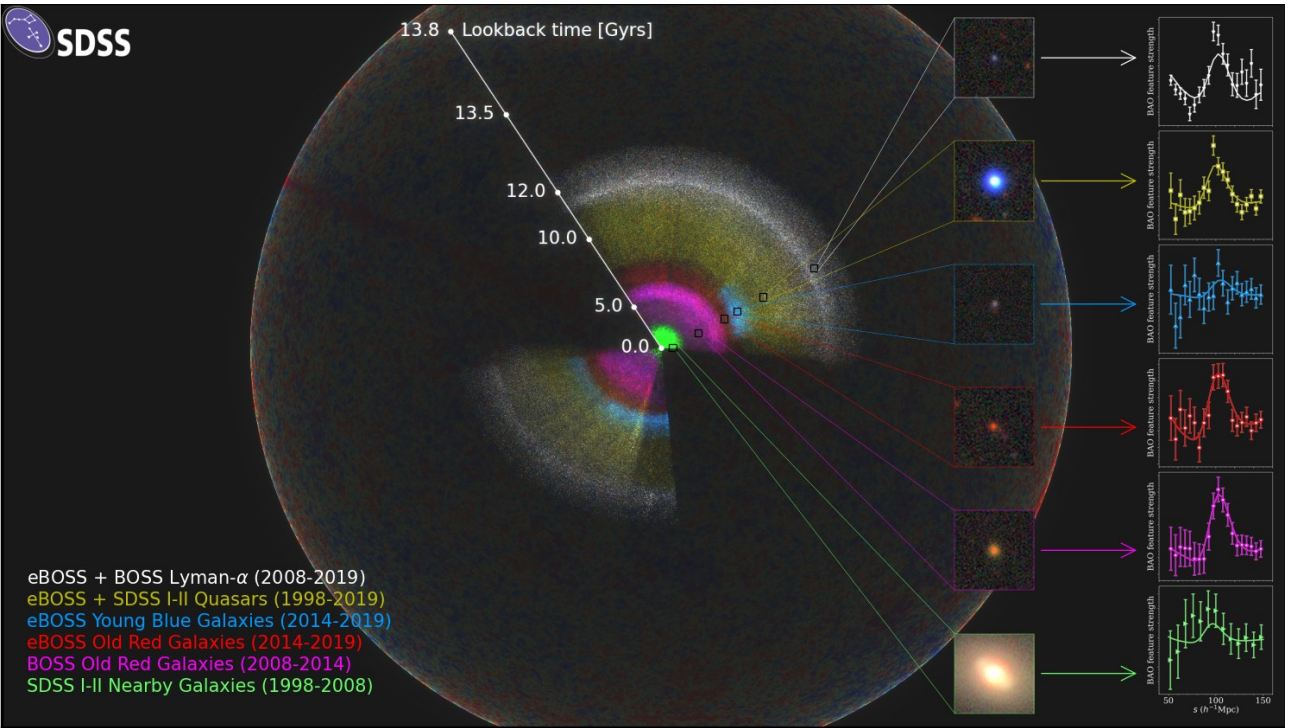
~ 47 mln of (confirmed and catalogued) local and distant galaxies

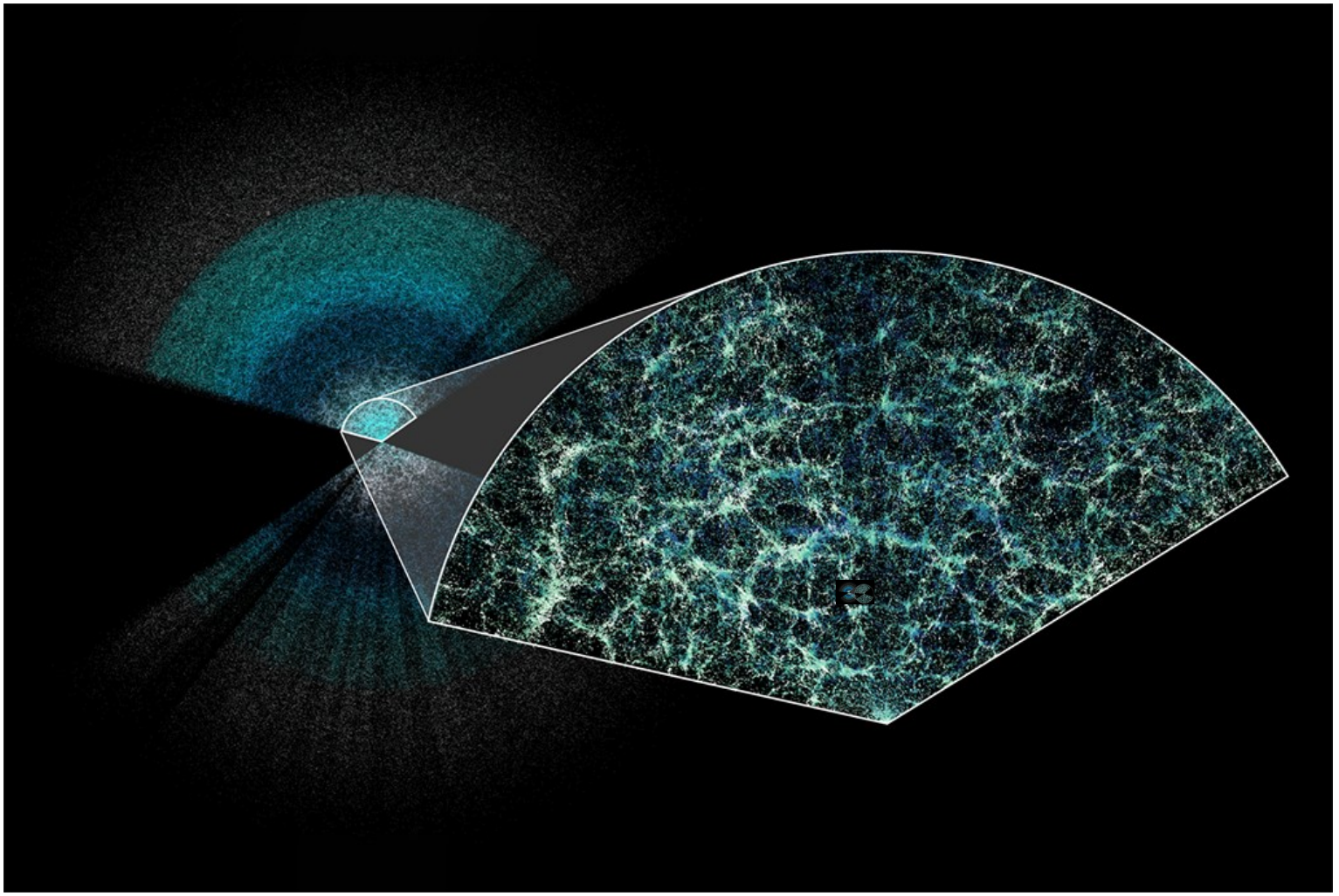
and this number is growing fast



(Copyright: VIPERS)

(Copyright: SDSS)



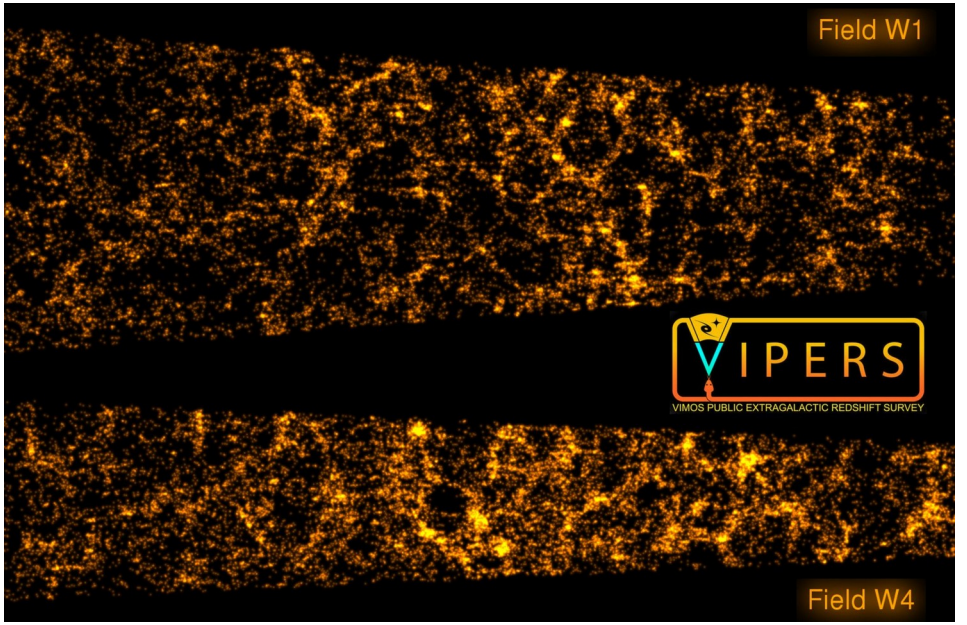


First **DESI (Dark Energy Spectroscopic Instrument)**: release (2026): ~ 47 mln of distant galaxies

Presently we know

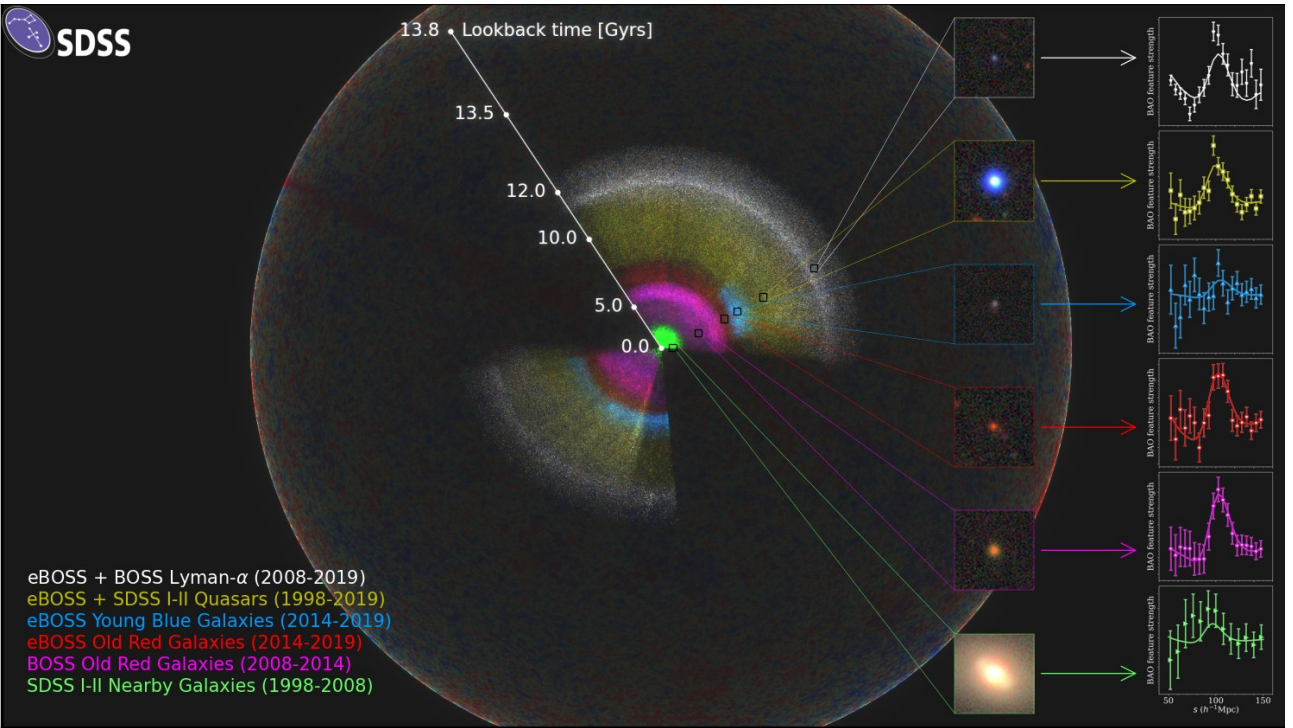
~47 mln of (confirmed and catalogued) local and distant galaxies

and this number is growing fast



(Copyright: VIPERS)

(Copyright: SDSS)



Near future: Vera Rubin Observatory will catalog (photometrically) 20 bln galaxies during the next ~10 years

Presently we know

47 mln of (catalogued and confirmed) galaxies in local and distant Universe but

Soon ~10^9

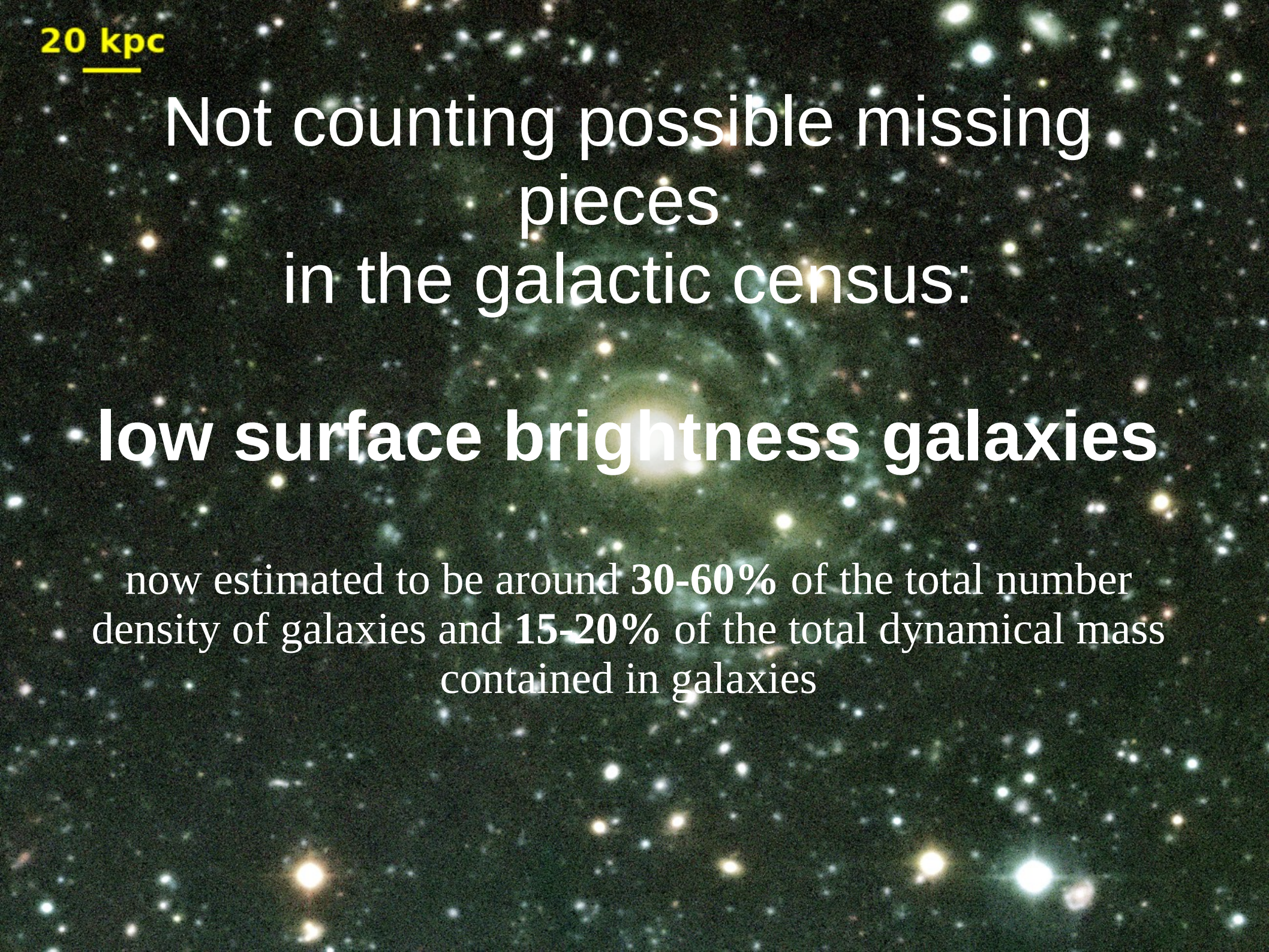


Credit: NASA, ESA, R. Ellis (Caltech), and the HUDF 2012 Team



Credit: NASA, ESA, CSA, and STScI. JWST

The estimated number of galaxies in observable Universe is several hundreds of bln (several×10^11) galaxies (with huge error bar)



Not counting possible missing pieces in the galactic census:

low surface brightness galaxies

now estimated to be around **30-60%** of the total number density of galaxies and **15-20%** of the total dynamical mass contained in galaxies

Data of near future

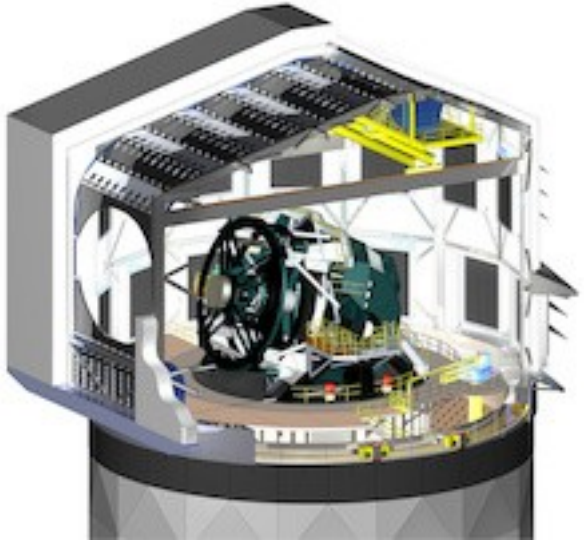
Large astronomical data of near future: Vera Rubin Observatory

First light in June 2025



- Large Survey of Space and Time (LSST)
- The deepest and widest sky survey ever + time domain
- mirror 8.4-m; 3200 megapixel camera
- **37 mld** stars and galaxies
- 10-yers survey
- **15-30 Terabytes (~all SDSS) per night**
- After 10 years:

~200 petabytes of data





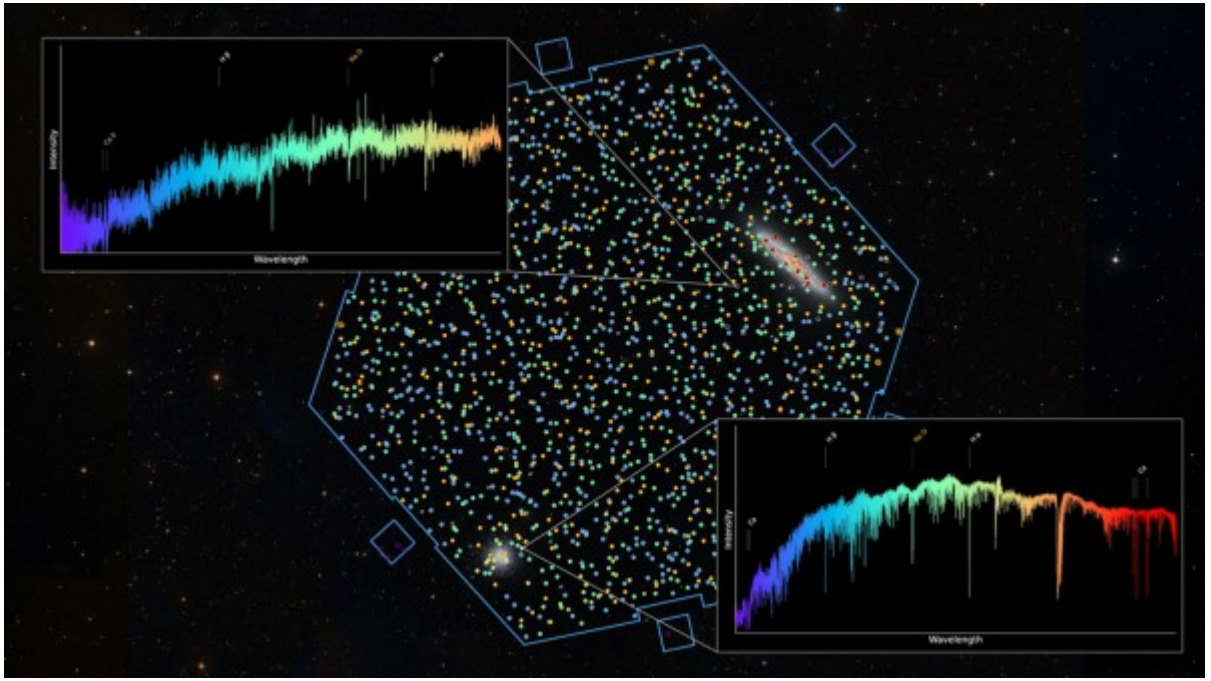


Large astronomical data of near future: 4MOST

First light in October 2025

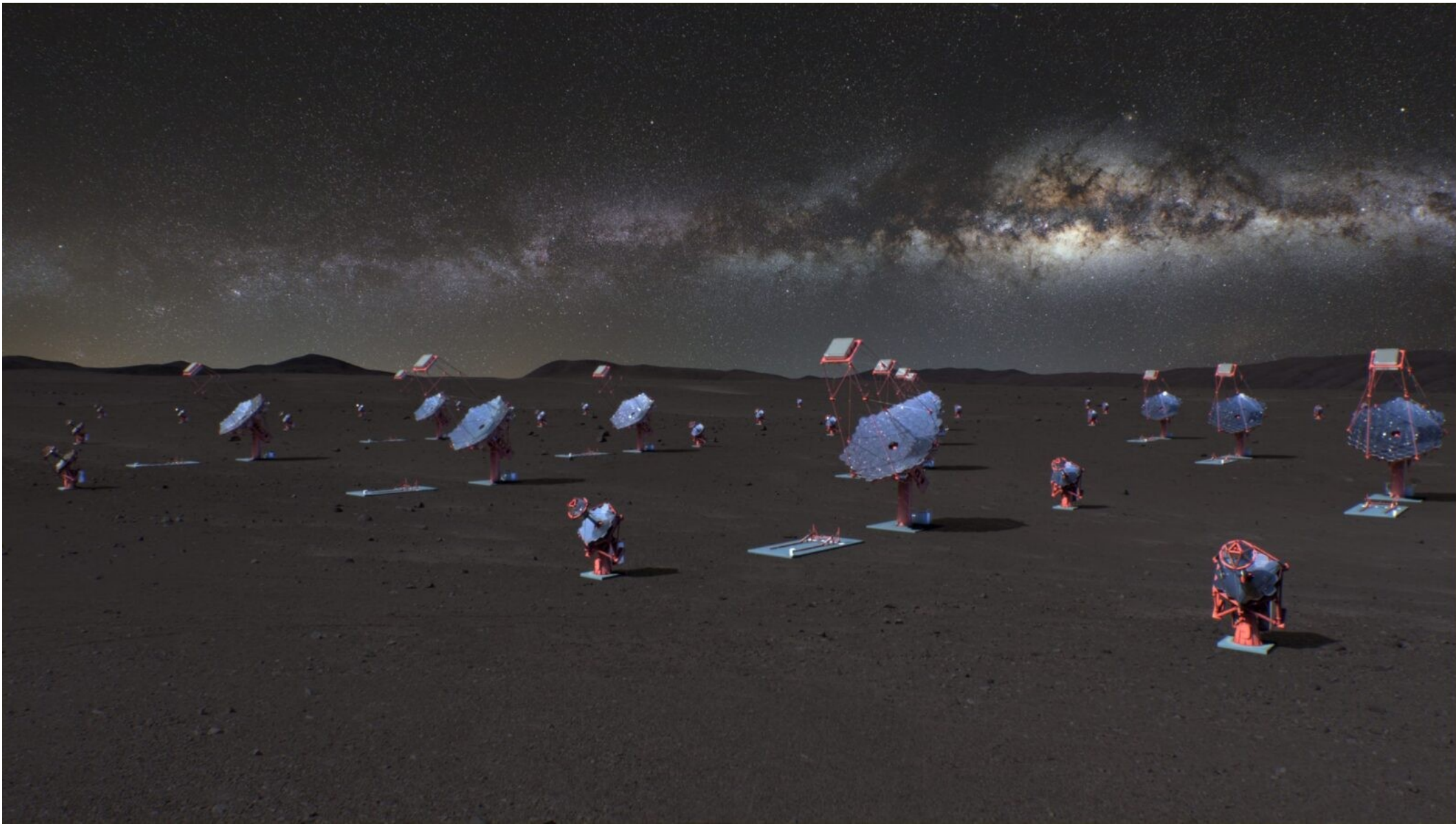


Credit: Y. Beletsky (LCO)/ESO



Credit: /ESO

High energy astrophysics - Cherenkov Telescope Array Observatory (CTAO)



Several petabytes yearly

Radio - Square Kilometre Array Observatory (SKAO)



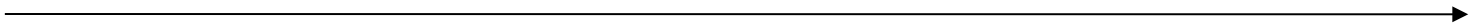
157 terabytes per second, 710 petabytes yearly

(Astronopmicaly) big data

To compare:
human genom - 3.1 Gb

1980 1990 2000 2010 2020 2030

MB GB TB PB EB



CCD → larger and deeper sky surveys → -> LSST, SKA...
tools for join analysis of different types of data
(Virtual Observatory)

Catalogs (and images) → Images (and catalogs)

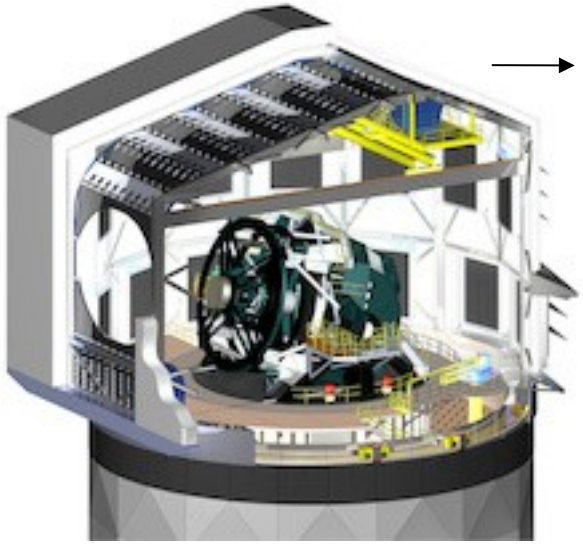


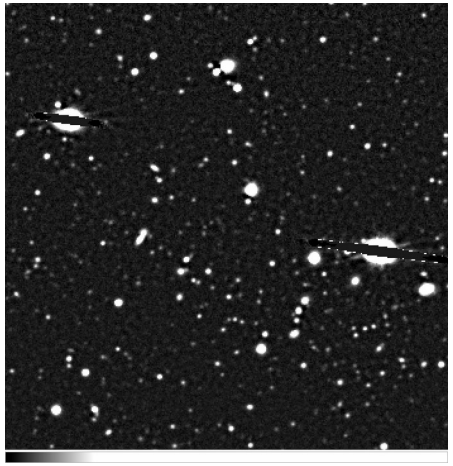
Servers of theŚwierk Compuiting Center (fot. NCBJ)

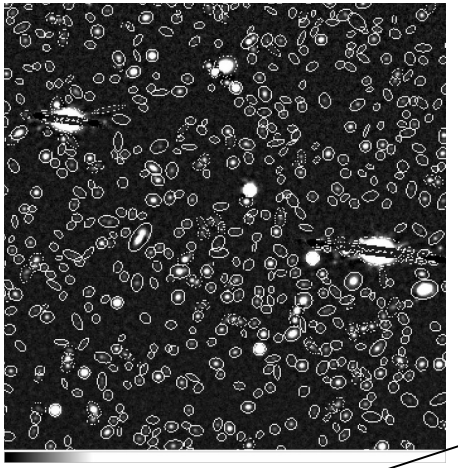
Astronomical data lifecycle

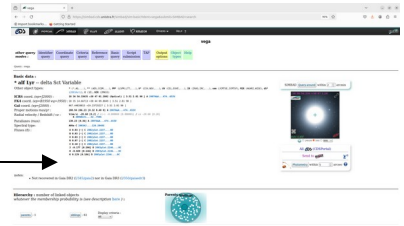
Big sky surveys

Databases → crosscorrelation

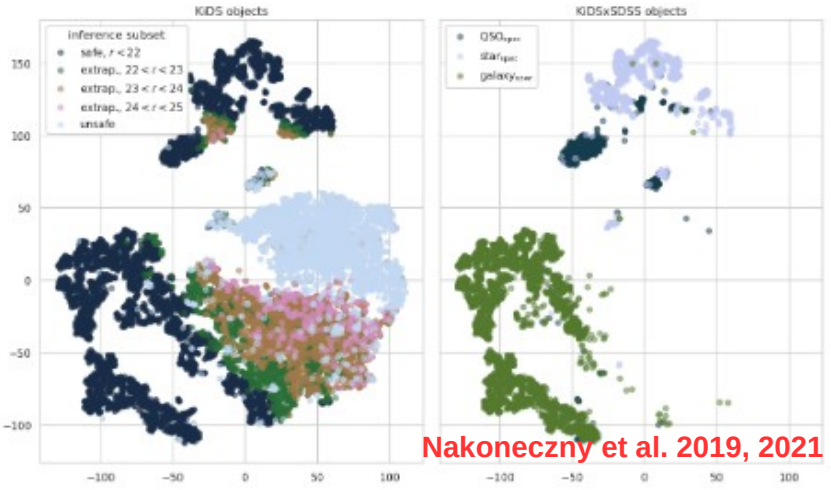








Data reduction (callibration, source extraction, background subtraction etc.)



Nakoneczny et al. 2019, 2021

Data selection

Follow up observations



Analysis (ML's key role)

Machine learning for (mostly) extragalactic science

- Huge and soon much larger „big data” in the era of „precision cosmology”
- Goal(s):
  - source classification
  - source identification
  - reconstruction of properties
  - Novelty/anomalies search
  - - .... (speeding up codes; generation of mock samples etc.)
- Supervised → when we know a priori what sources we expect to find and we can use some datasets for training
    → classification (for separate groups) or
    → regression (for smooth transition/source properties)
- Unsupervised (+semi-supervised) → clustering of sources into previously unknown and unexpected classes

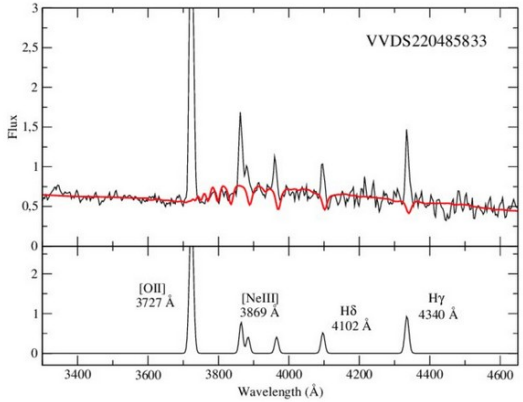
Machine learning for astronomy – challenges

- Problems and challenges
  - Extrapolation (small and biased training samples)
  - Physical interpretability (do trends we see really mean something? No to black box approach – we would like to learn new physics)
  - Reproducibility
  - Resources

Machine learning for astronomy – challenges

- Problems and challenges
  - (relatively) small parameter space





- photometry + imaging (in different spectral ranges)
- spectra
- time variability
- polarization

„Multimessenger time domain astronomy”

Machine learning for astronomy – challenges

- Problems and challenges
  - transfer learning

JWST (space)

CFHT (ground-based)







Illustris (simulation)

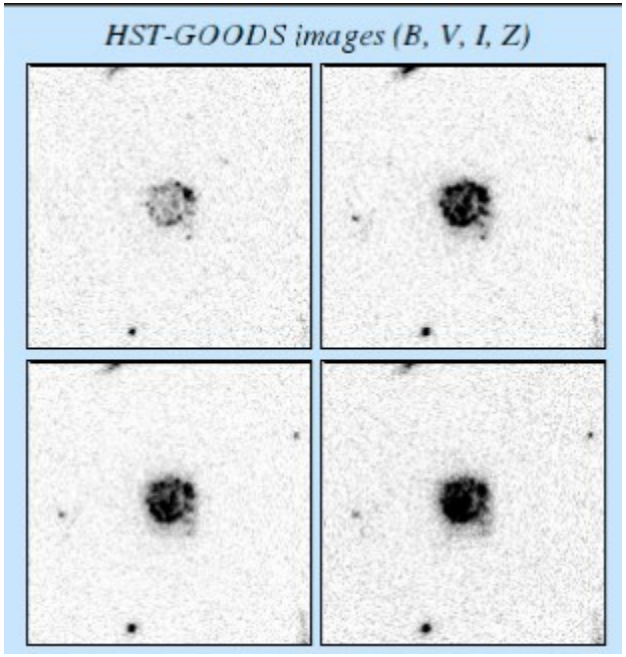
Machine learning for astronomy – challenges

- Problems and challenges: data representability

HST (M31)



what we would like to see



what we usually do see

Machine learning for astronomy – challenges

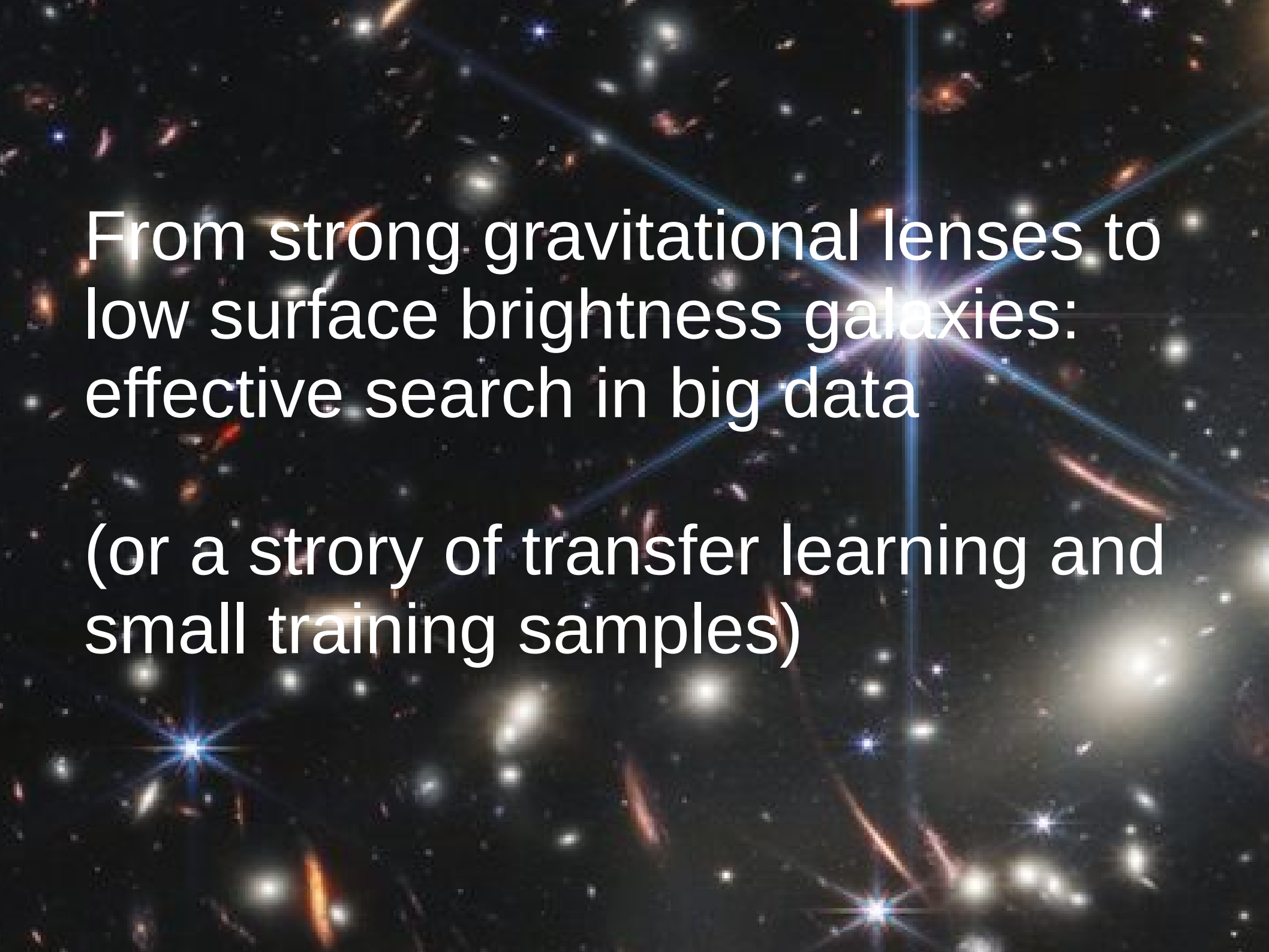
- Problems and challenges: data representability

  - training based on brighter objects to generalize over faint ones
  - different distributions of properties of training and generalisation samples
  - fainter objects are
    - intrinsically fainter – having different physical properties
    - more distant –> if in space, also in time – representing different evolutionary stages
    - more distant -> different rest frame

Machine learning for astronomy – challenges

- Problems and challenges: model interpretability
  - I get a model but does it have any physical meaning?
  - But also: maybe I can find new physical information in the ML-based model
- Physically informed models: if I inform the model, do I skew the interpretation?
- Foundation models for astronomy: how much control do I have over the process (how much of the black box approach do I accept?)

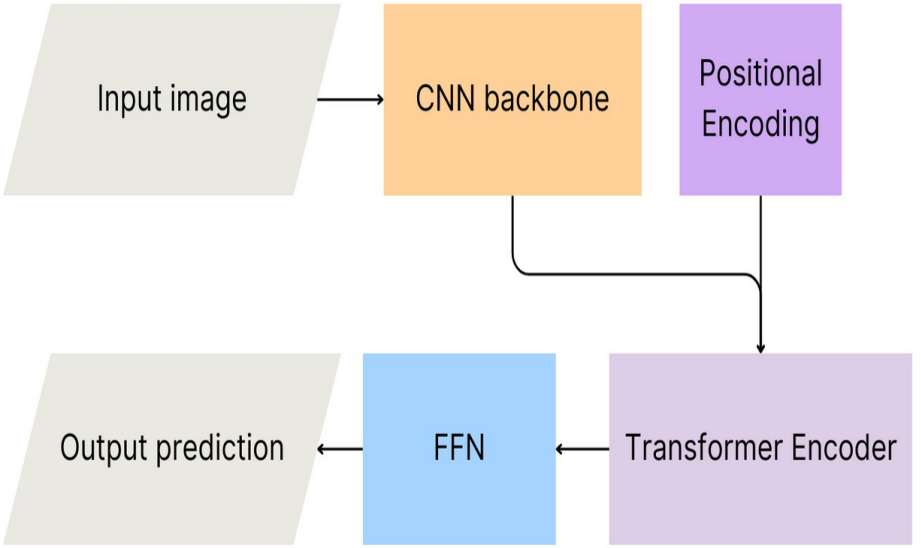
Some examples of what ML can actually be used for (and how challenges can be met)


From strong gravitational lenses to low surface brightness galaxies: effective search in big data
(or a strory of transfer learning and small training samples)

Thuruthipilly et al. 2022: trained a transformer ecoder that achieved 93% accuracy in finding strong lenses in the simulated Strong Lens Finding Challenge dataset (made to simulate KiDS data).

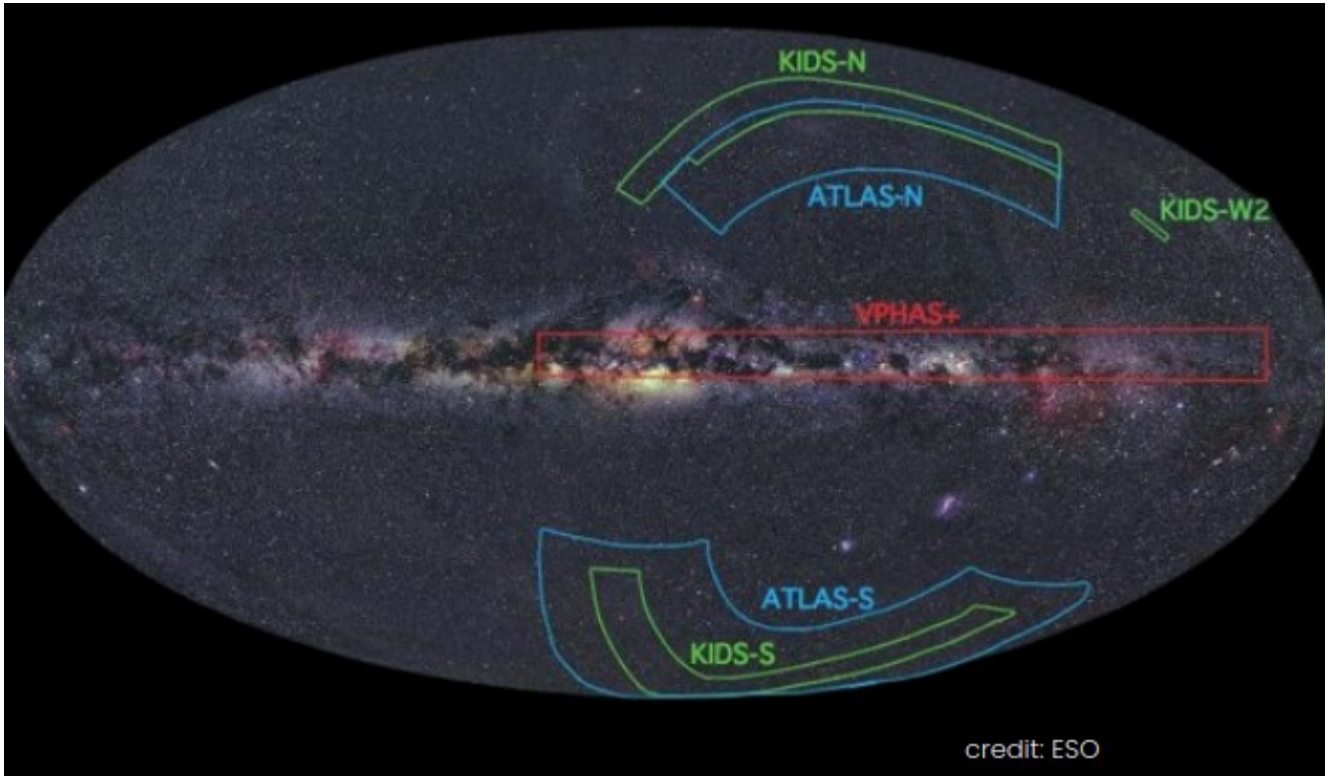
Out of 21 different models tested, we selected the three with the best performance on four-bands images and one model trained only on r, i, and g bands.





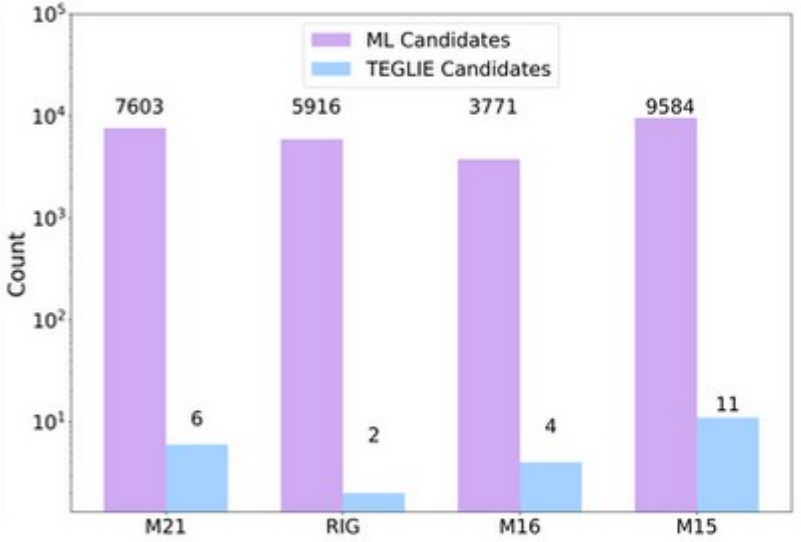
Name	Model structure	Accuracy	AUROC	TPR_0	TPR_{10}
15	8 CNN Layers + 8 H_128 + 4 (E)	92.99	0.978	0.140	0.48
16	16 CNN Layers + 8 H_128 + 8 (E)	90.97	0.962	0.225	0.24
21	8 CNN Layers + 8 H_128 + 4 (E)	92.79	0.98	0.00	0.64
RIG	8 CNN Layers + 8 H_128 + 4 (E)	91.15	0.974	0.11	0.28

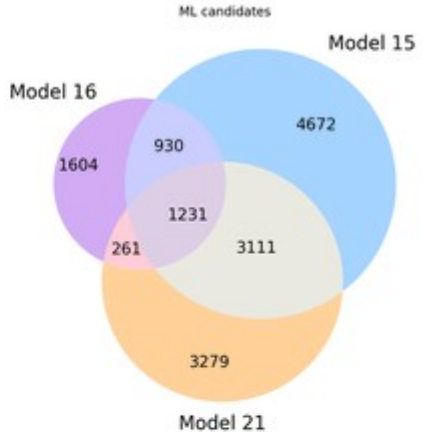
Can we transfer this model easily to real data?

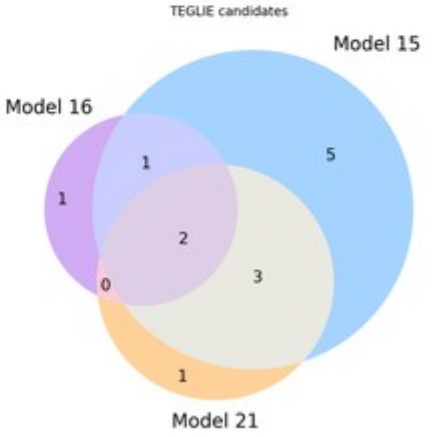


- KiDS is a large optical imaging survey in the Southern sky
- Using the VLT Survey Telescope (VST), located at the ESO Paranal Observatory, KiDS has mapped 1350 deg^2 of the night sky in four broad optical bands filters (u, g, r, i).
- It is a good testing ground for next-generation wide area surveys.

T+22's model performance on real KiDS data







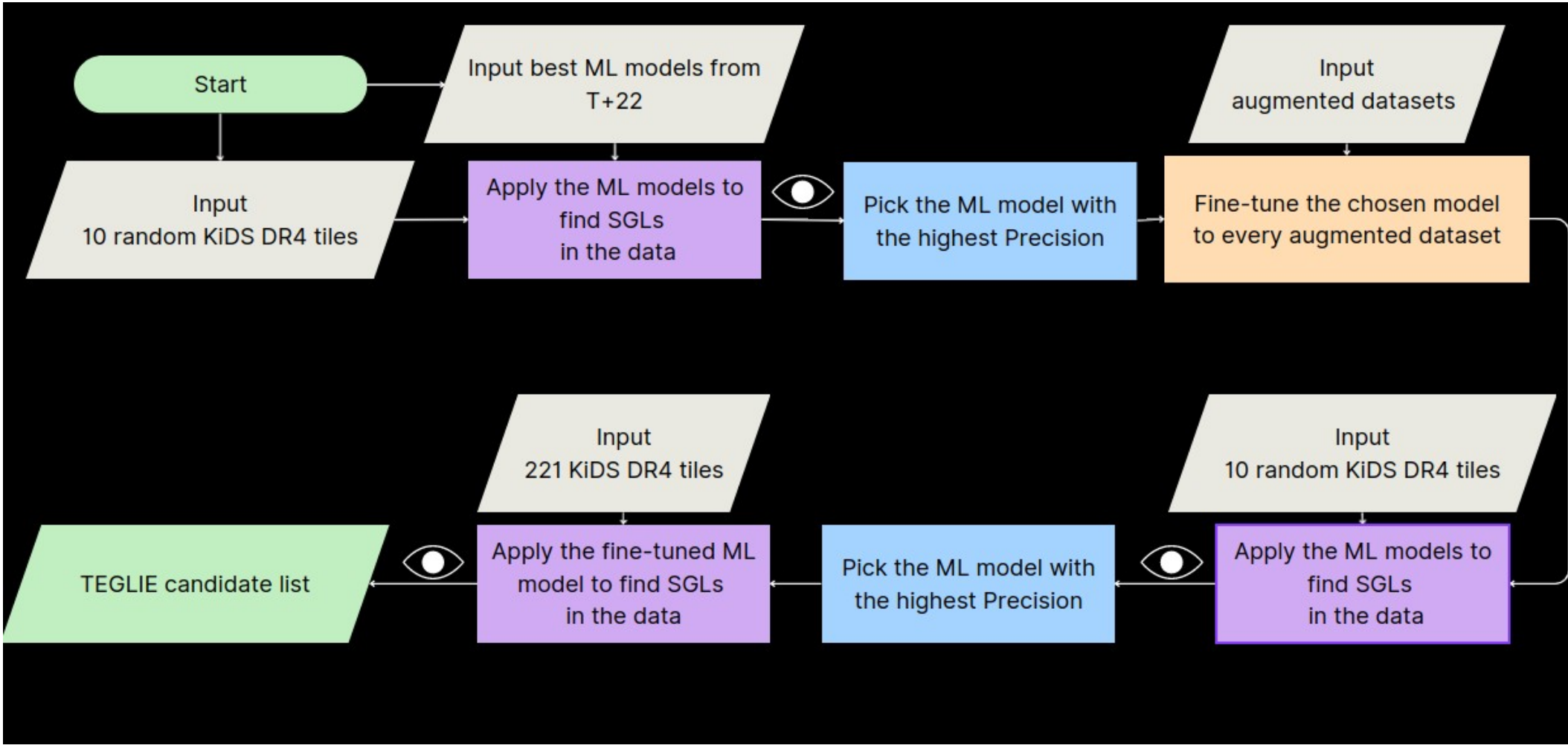




- Why so many false positives?
  - Simulations are as realistic as possible but not sufficiently
  - Essential: we need a good training sample not only of lenses, but also of non-lenses

Grespan, Thuruthipilly et al. 2024

Grespan, Thuruthipilly et al. 2024





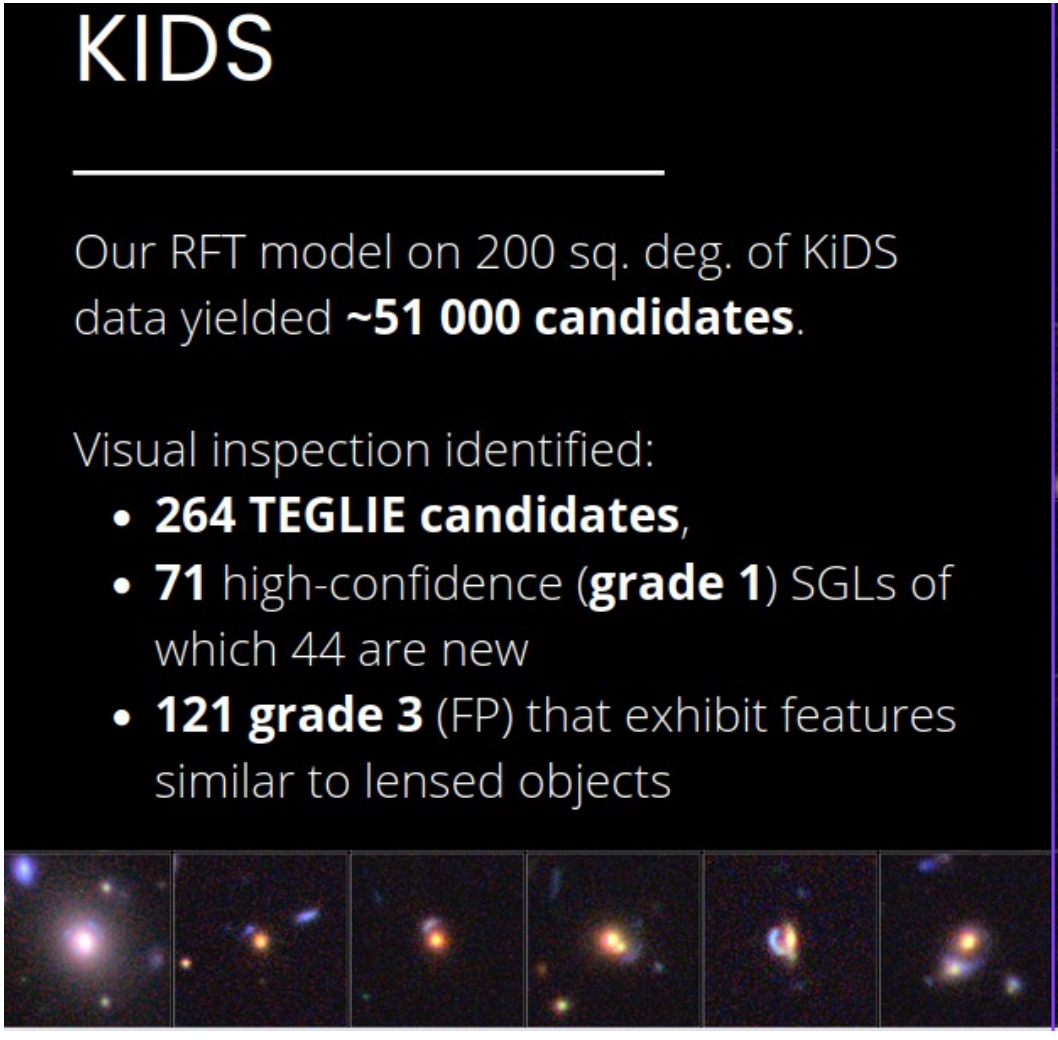


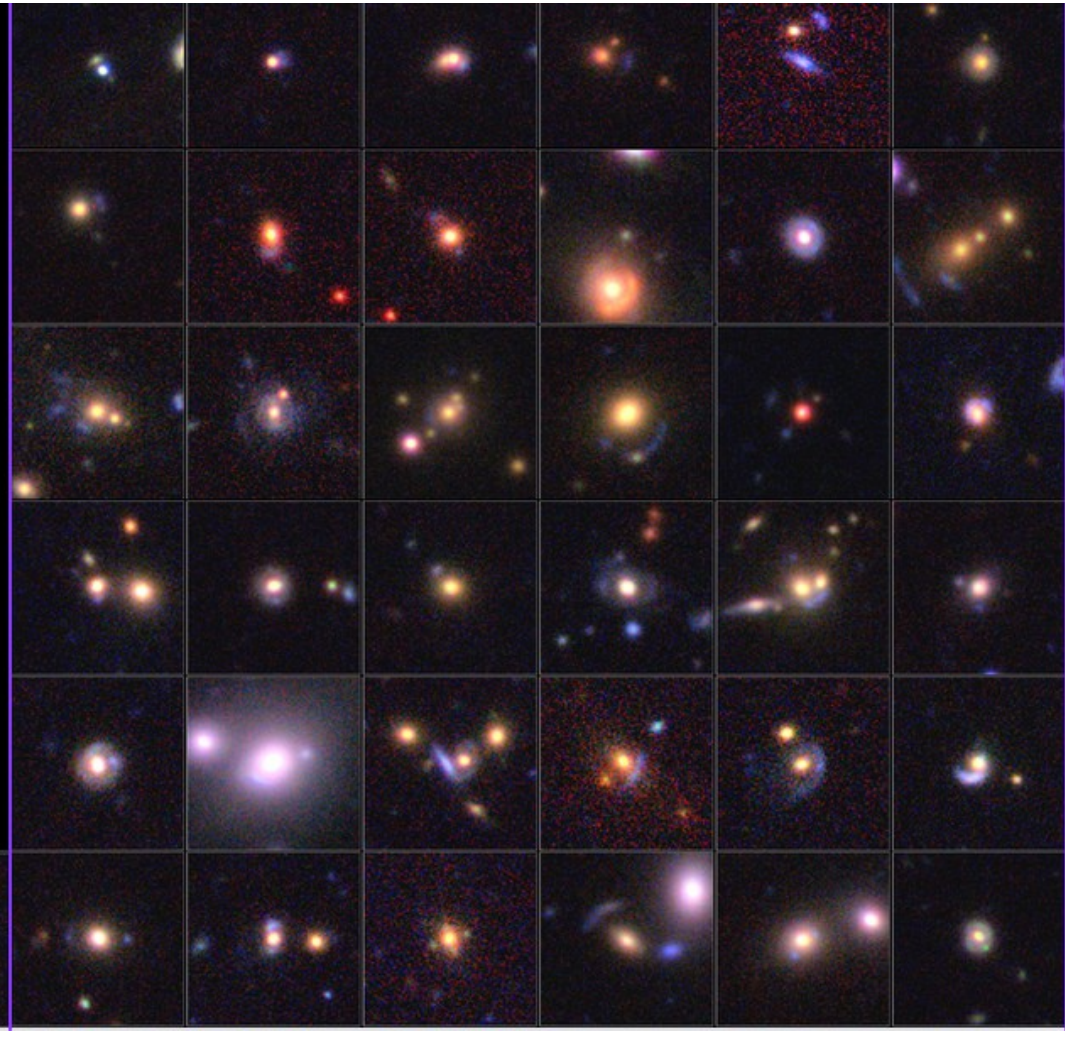
KIDS

Our RFT model on 200 sq. deg. of KiDS data yielded **~51 000 candidates**.

Visual inspection identified:

- **264 TEGLIE candidates**,
- **71** high-confidence (**grade 1**) SGLs of which 44 are new
- **121 grade 3** (FP) that exhibit features similar to lensed objects





Model now being used for the LSST data

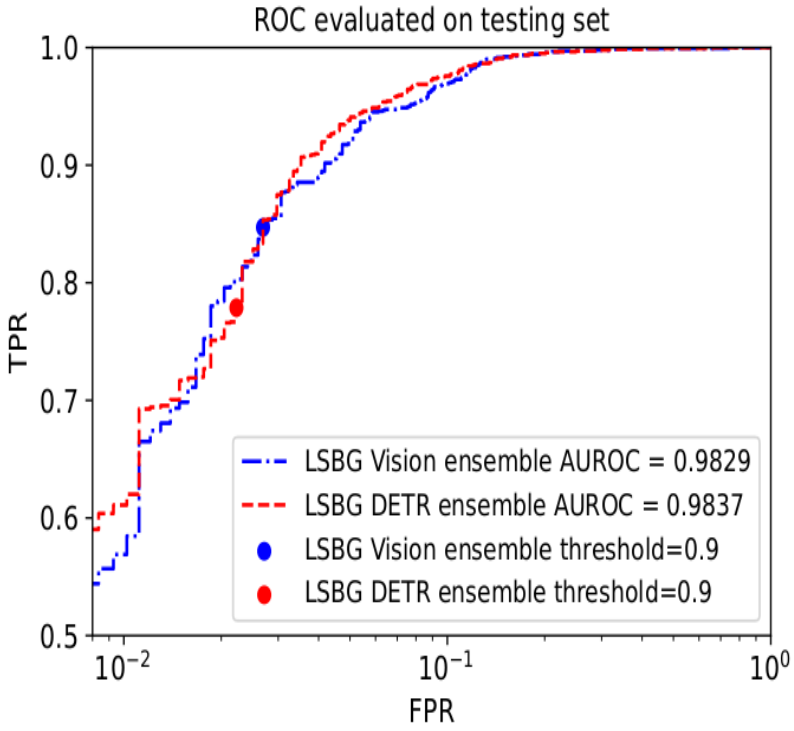




Grespan, Thuruthipilly et al. 2024

Can we transfer this to other **types** of data?

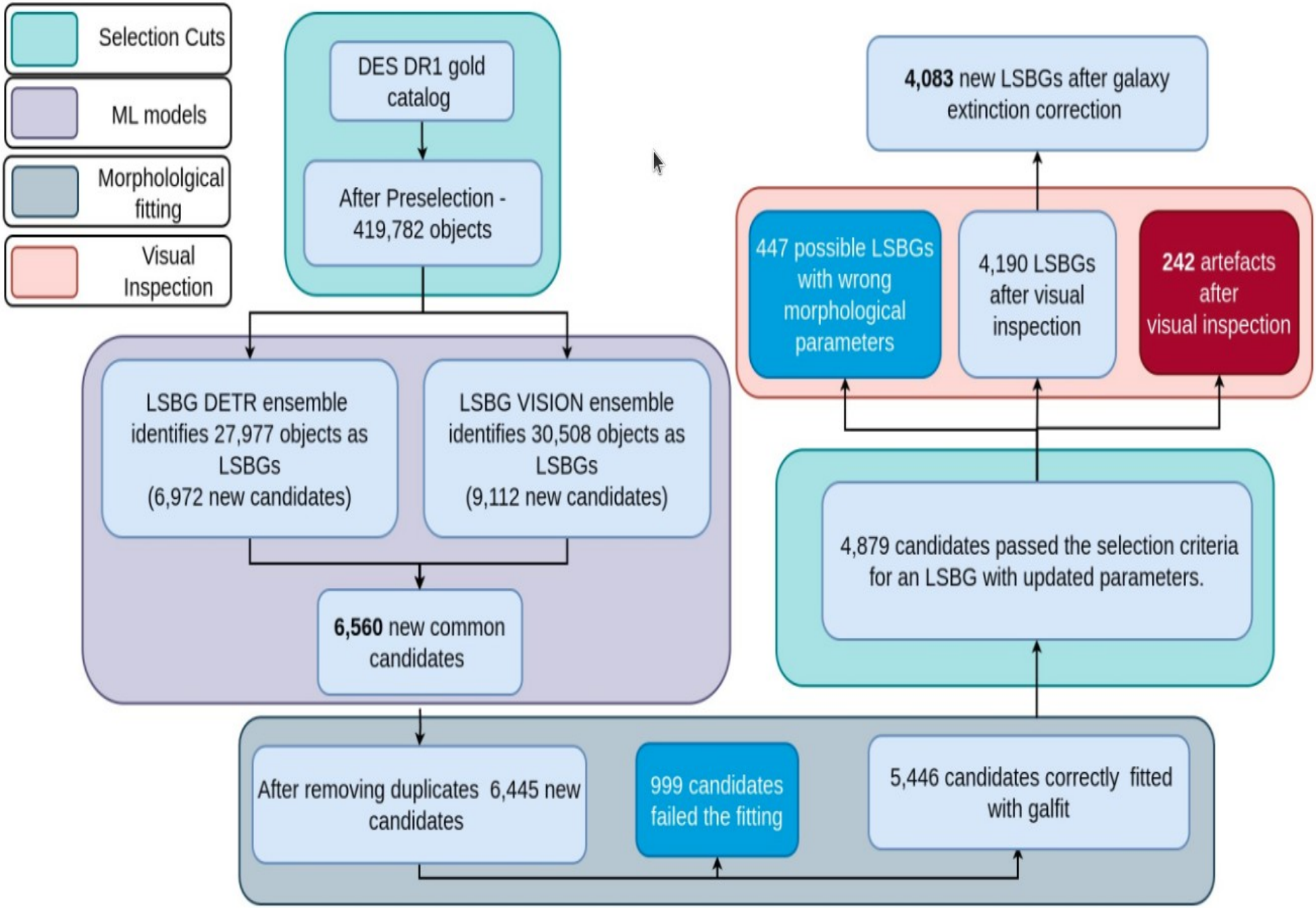
- Low surface brightness galaxies
- Data: Dark Energy Survey Y3 (Sevilla-Noarbe 2021).
- Goal: better classification of LSBGs and artifacts.
- Starting point (and training data): a catalog of > 20 000 LSBGs. by Tanoglidis et al. (2021; based on a support vector machine (SVM) + visual inspection) → a catalog of > 20 000 LSBGs.
- Final point → a new improved catalog.





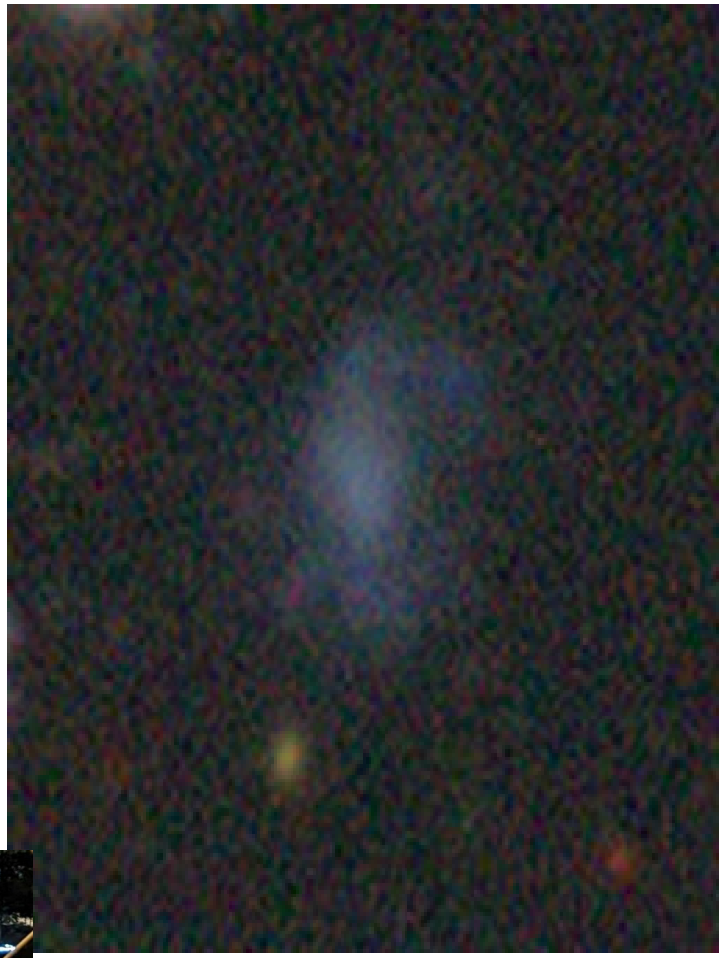


Thurutupilly, Junais et al. 2024


Selection Cuts
ML models
Morphololgical fitting
Visual Inspection
DES DR1 gold catalog
After Preselection - 419,782 objects
LSBG DETR ensemble identifies 27,977 objects as LSBGs (6,972 new candidates)
LSBG VISION ensemble identifies 30,508 objects as LSBGs (9,112 new candidates)
6,560 new common candidates
After removing duplicates 6,445 new candidates
999 candidates failed the fitting
5,446 candidates correctly fitted with galfit
4,879 candidates passed the selection criteria for an LSBG with updated parameters.
447 possible LSBGs with wrong morphological parameters
4,190 LSBGs after visual inspection
242 artefacts after visual inspection
4,083 new LSBGs after galaxy extinction correction

DES Y3 Gold: new catalog of LSBGs

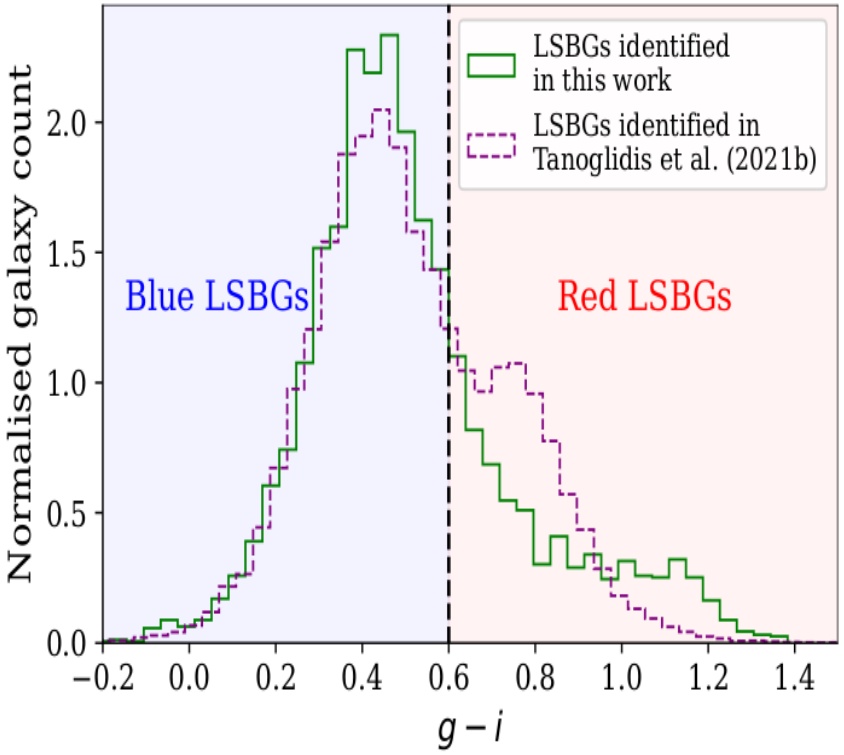
- method: self-attention-based encoder models coupled with CNN (note: with big data 1% accuracy improvement translates thousands of new detections)
- 27,000 LSBGs, among them 4083 new (mostly blue + extreme red, as compared to previous works)
- among them, 317 UDG candidates, including 276 new ones

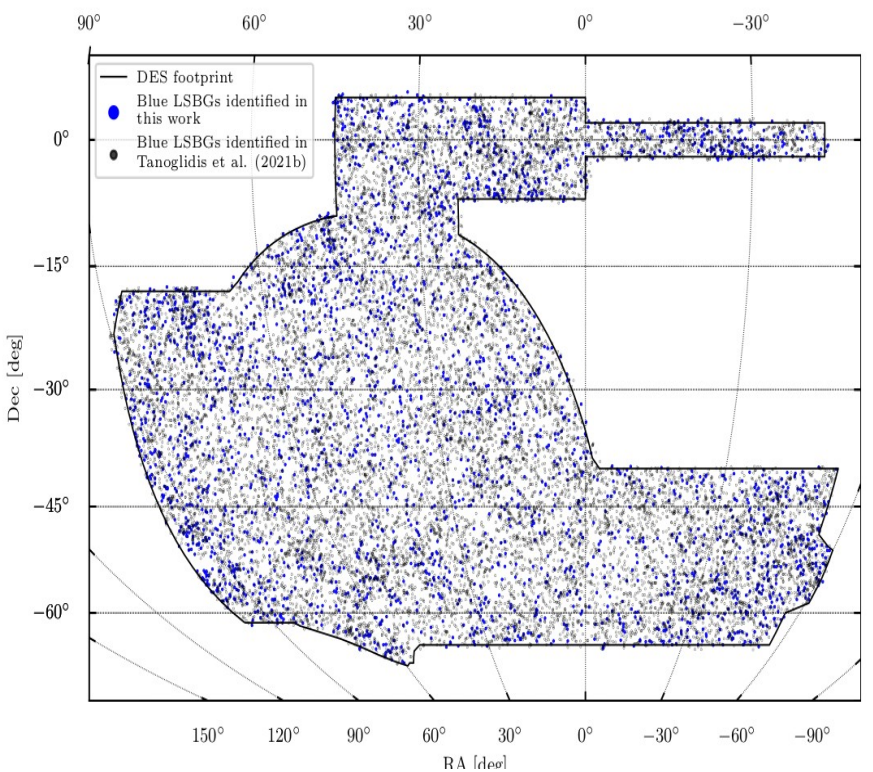




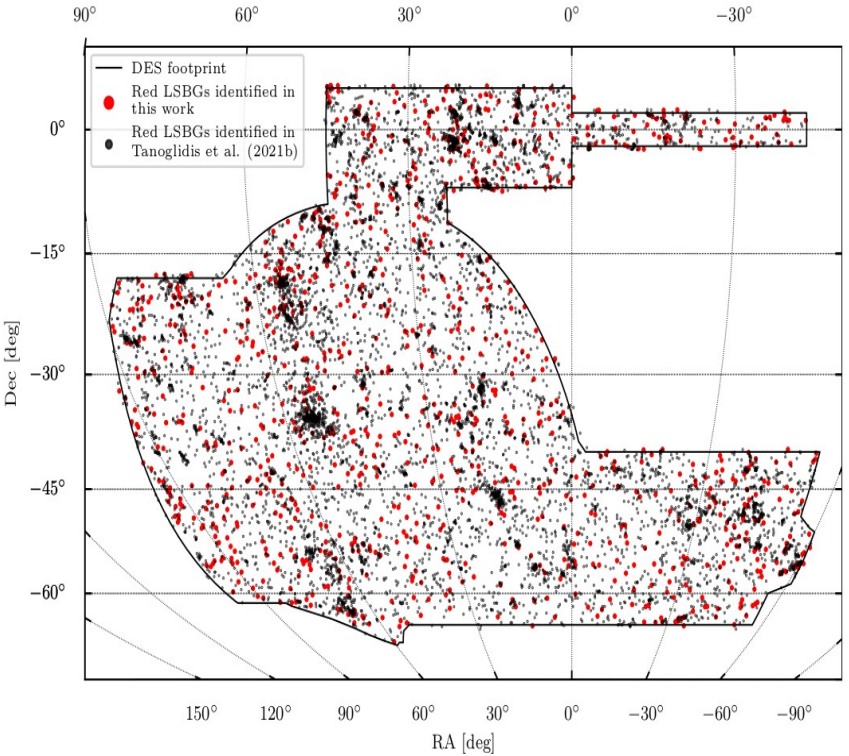
Thurutupilly et al. 2024

DARK ENERGY SURVEY





Maps of blue and red LSBGs (old and new) in the DES field





Thurutupilly et al. 2024

Transfer Learning to find LSB galaxies



Learns the features of one dataset to identify them in a different one

DES to HSC: Detecting low-surface-brightness galaxies in the Abell 194 cluster using transfer learning

H. Thuruthipilly[1,*], Junais[2,3,*], J. Koda[4,*], A. Pollo[1,5], M. Yagi[6], H. Yamanoi[7], Y. Komiyama[6,8,9], M. Romano[10,11,1], K. Małek[1,12], and D. Donevski[1]

From DES to KiDS: Domain adaptation for cross-survey detection of low-surface-brightness galaxies

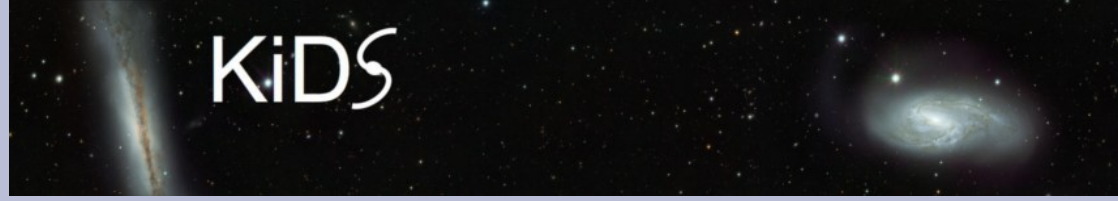
Hareesh Thuruthipilly[1,2,3], Krzysztof Lisiecki[1], Junais[4,5], Katarzyna Małek[1], Agnieszka Pollo[1], William J. Pearson[1], Antonio Vanzanella[1], Saptarshi Pal[1], Miguel Figueira[1], Pratik Dabhade[1], Anna Durkalec[1], Aidan P. Cotter[1], Unnikrishnan Sureshkumar[1,6], Nandini Hazra[1], Patryk Matera[1], Subhrata Dey[1], Michal Vrábel[1], Anirban Dutta[1], Henry Willems[1], Nicola Principi Cavaterra[1], Natalia Dobrowolska[1], and Wojciech Knop[1]

Model name	Accuracy (%)	TPR
LSBG VISION 1	95.58	0.96
LSBG VISION 2	95.48	0.96
LSBG VISION 3	95.58	0.97
LSBG VISION 4	95.14	0.96
LSBG VISION Ensemble	**95.62**	**0.96**
LSBG DETR 1	95.68	0.96
LSBG DETR 2	95.36	0.95
LSBG DETR 3	95.48	0.96
LSBG DETR 4	95.54	0.97
LSBG DETR Ensemble	**95.62**	**0.96**

Combination of models, LSBG Detection Transformer (DETR) and LSBG Vision Transformer (ViT)
Layers: CNN extracts features, Encoder filters them, FFN “learns” the features

PRO: Successfully identifies DES LSBGs in DP1
CON: Sensitive to pre-selection of potential LSBGs which rely on standard pipeline photometry

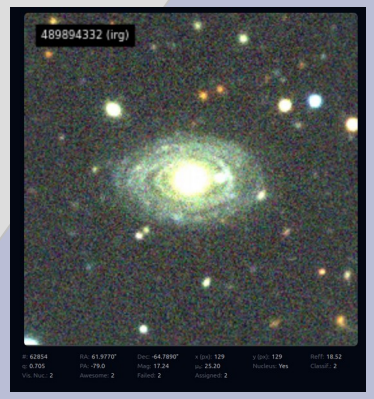
From DES to KiDS DR5: sucessful domain adaptation (and results)

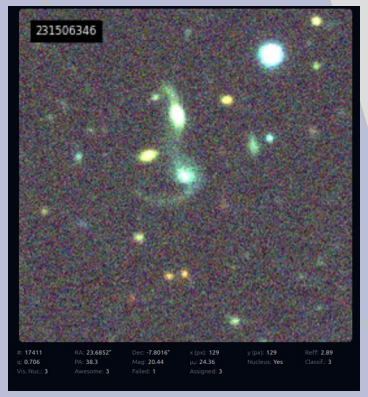


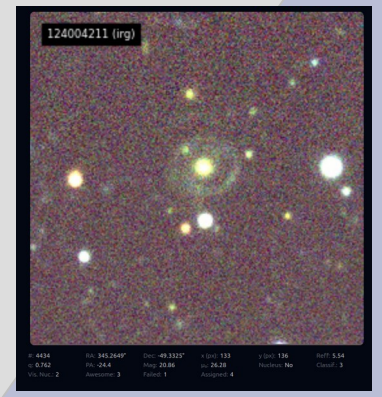


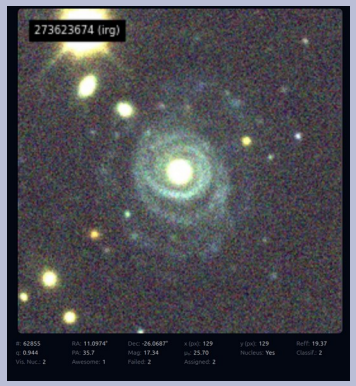


20 180 low surface brightness galaxies, 434 UDGs with morphological classifications







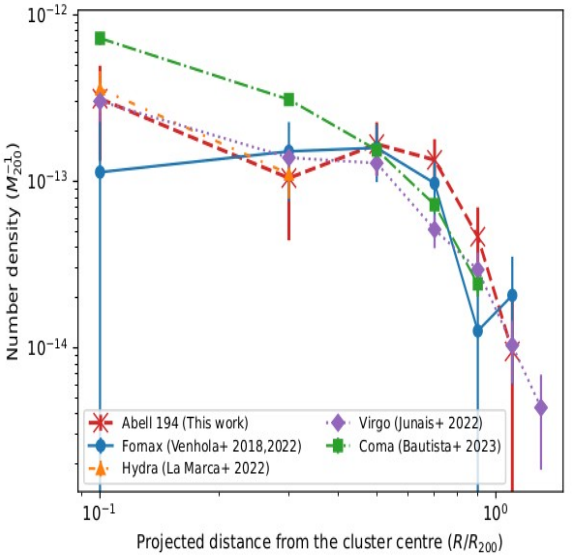


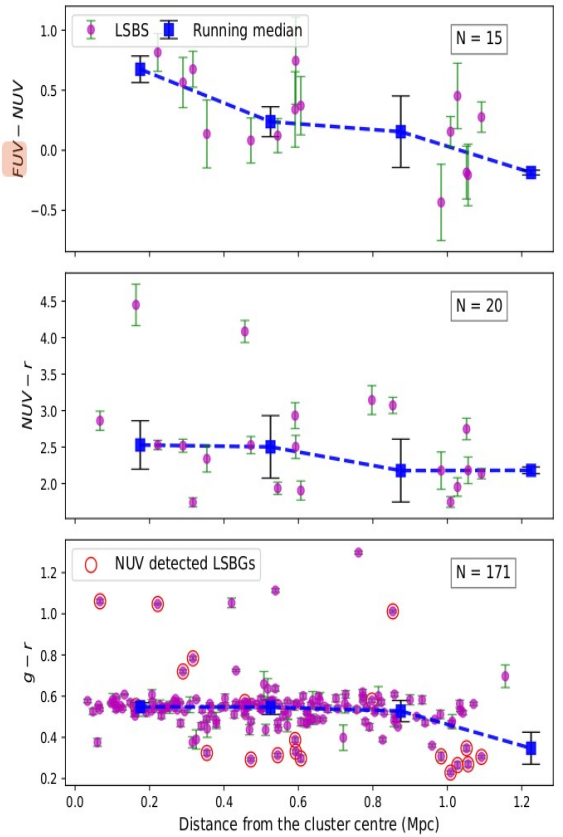


Thurutupilly et al. 2026

DES to HSC (on the way to LSST): the case of Abell 194

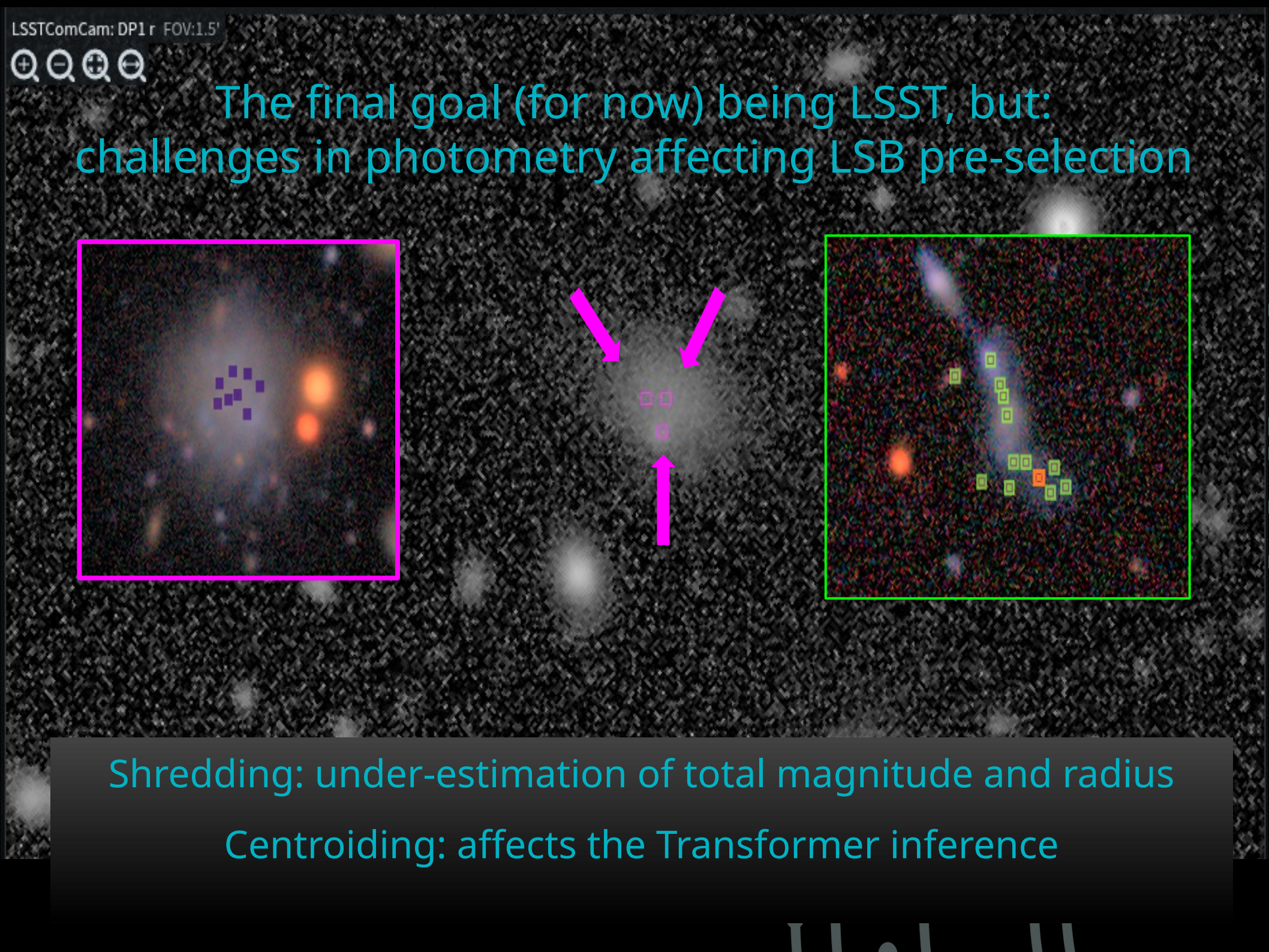
- 171 low surface brightness galaxies (LSBGs) in the Abell 194 cluster using HSC data, including 87 new discoveries.
- The transformer model achieved a true positive rate (TPR) of 93% in HSC data without any fine-tuning.
- Among the LSBGs, 28 were classified as ultra-diffuse galaxies (UDGs).



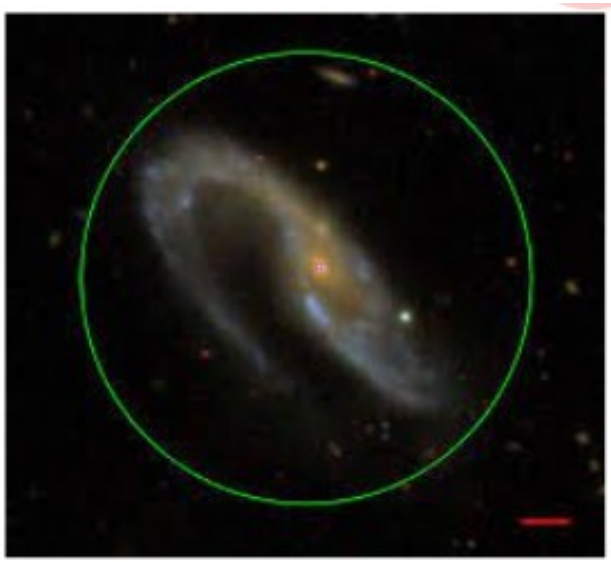




Thurutupilly al. 2025


LSSTComCam: DP1 r FOV:1.5'
The final goal (for now) being LSST, but:
challenges in photometry affecting LSB pre-selection
Shredding: under-estimation of total magnitude and radius
Centroiding: affects the Transformer inference

Low surface brightness features
of normal galaxies
(see Subhrata's talk)
or ML-based parameters
for better inrepretability of statistical
results



Galaxy mergers

→ Very important in shaping galaxy evolution (each galaxy like Milky Was has undergone several major mergers in the past

→ Very common at high redshift

→ 10-15% of merging galaxies today

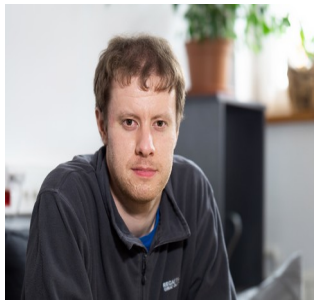
→ At high(er) redshift: activators of starbursts and AGNs

→ Today: weaker/questionable contribution to star formation enhancement but still important in galaxy mass assembly



Credit: HST, NASA/ESA

Galaxy and Mass Assembly Survey: mergers in the large scale structure, or where do mergers happen?



Pearson et al. 2019

→ Galaxy merger catalogs in the GAMA survey (selected → by ML (CNN) and → according to the Gini parameter)

→ Method: correlation function and marked correlation function (again)

→ concept: probability of a galaxy to be a merger (according to CNN) can be regarded as a measure of galaxy “mergeriness” and then used as a weight (“mark”)

→ 0.1<z<0.16, volume limited sample(s)

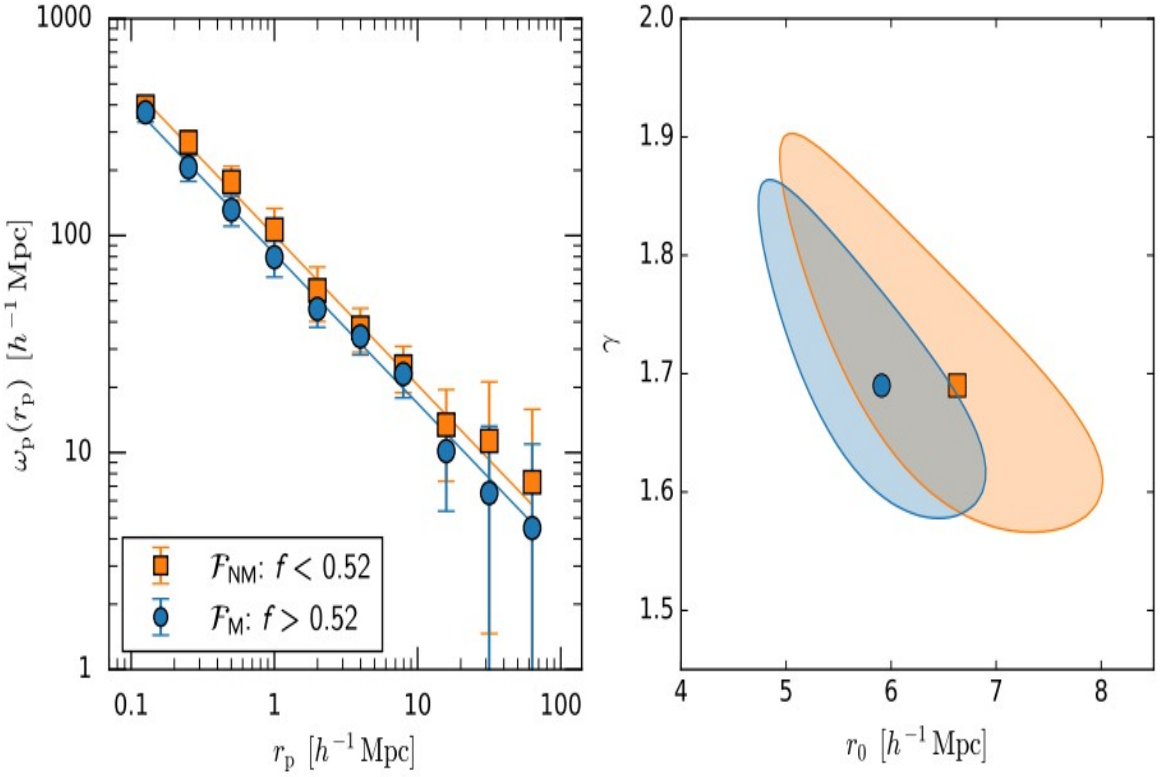


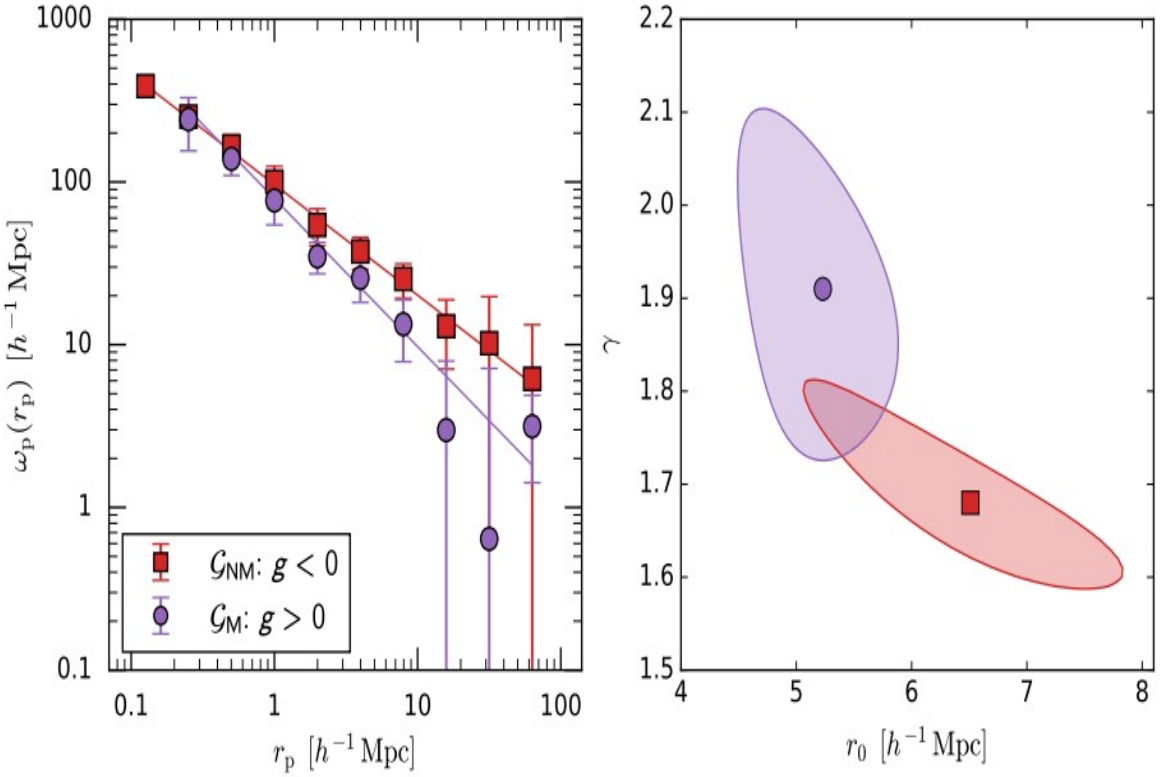
Sureshkumar et al. 2024

Galaxy and Mass Assembly Survey: mergers in the large scale structure, or where do mergers happen?



Sureshkumar et al. 2024





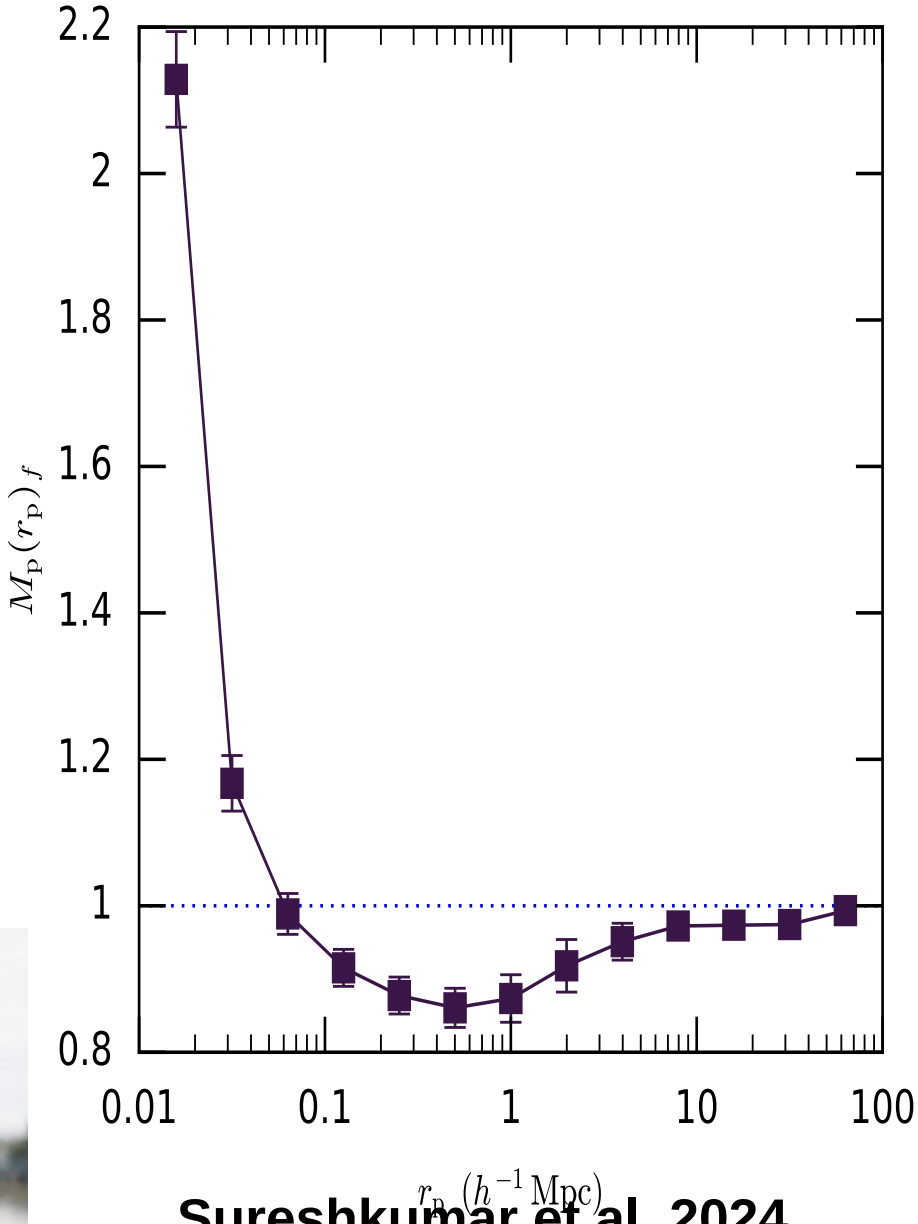
Galaxy and Mass Assembly Survey: where do mergers happen?

→ Merging galaxies in the present day Universe prefer <u>underdense</u> environments – galactic outskirts rather then centers (GAMA: Sureshkumar et al. 2024, NEP: Pearson et al. 2024)

→ No significant rise in SFR w/r to similarly massive galaxies (Pearson et al. 2019, Pearson et al. 2024)

→ Most important is the invisible (i.e. low surface brightness features around).





Sureshkumar et al. 2024

Humans in ML world

Algorithms are at most as smart as people who use (and teach) them

But also: another type of distributed computing network (or: citizen science).

Galaxy Zoo — Zooniverse

https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

Import bookmarks... Getting Started

PROJECTS ABOUT GET INVOLVED TALK BUILD A PROJECT NEWS SIGN IN REGISTER

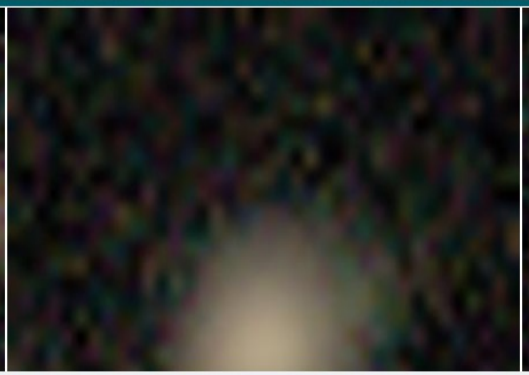
Language English

Galaxy Zoo

ABOUT CLASSIFY TALK COLLECT

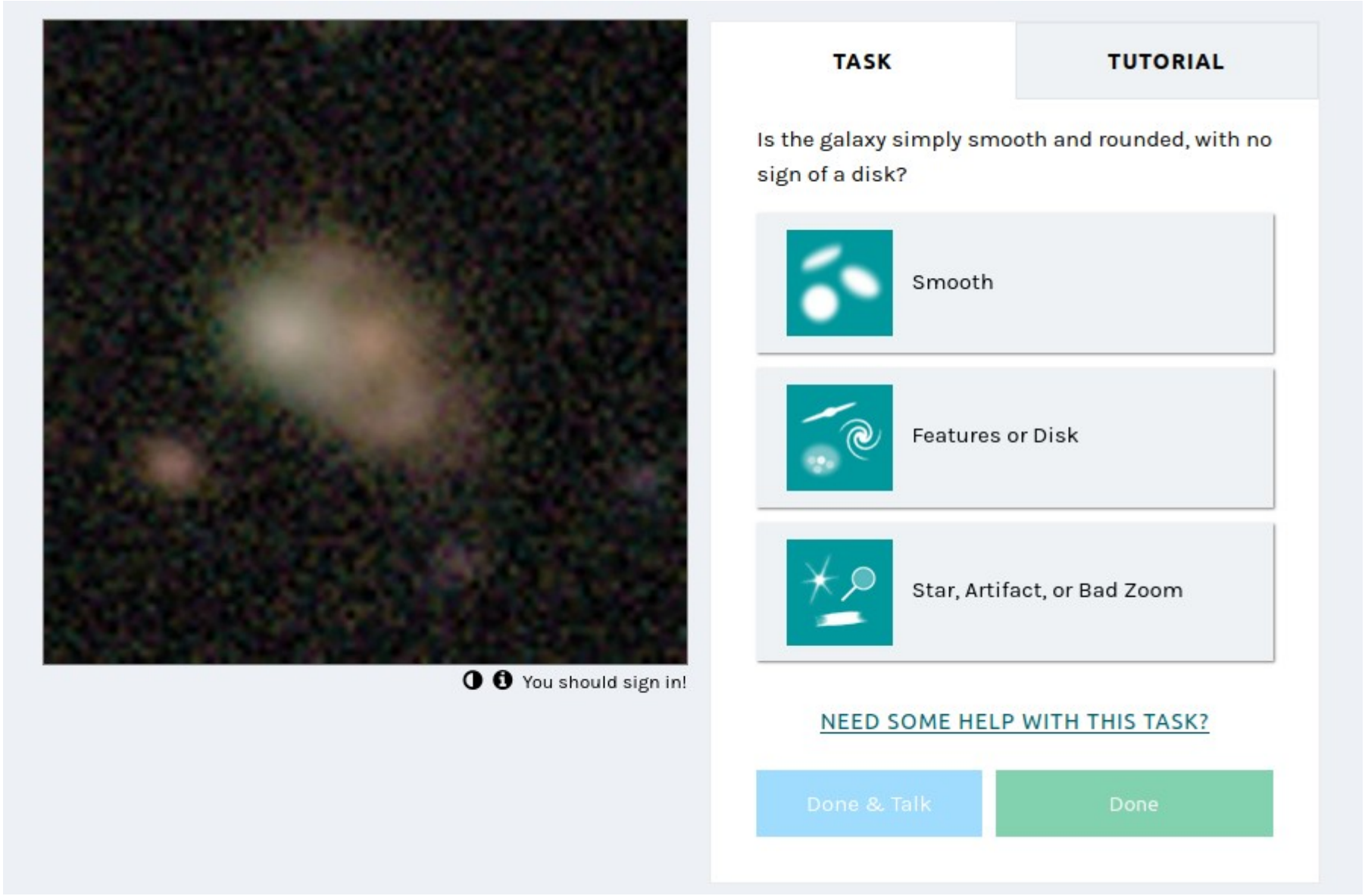
Few have witnessed what you're about to see

Learn more Get started


6 people are talking about **Galaxy Zoo** right now.

Join in



→ may be the best source of training data for future ML in astronomy (but how reliable?)

www.zooniverse.org/projects/zookeeper/galaxy-zoo

Welcome to Galaxy Zoo : Rubin, our first look at early data from the Vera C. Rubin Observatory. Click the Rubin workflow in orange below to dive in.



GALAXY ZOO

Few have witnessed what you're about to see

LEARN MORE >

FROM THE ORGANIZATION:

RUBIN OBSERVATORY

GALAXY ZOO

GET STARTED!

You can do real research by clicking to get started here!

6% COMPLETE
JWST COSMOS >

0% COMPLETE
Rubin DP1 Classic >

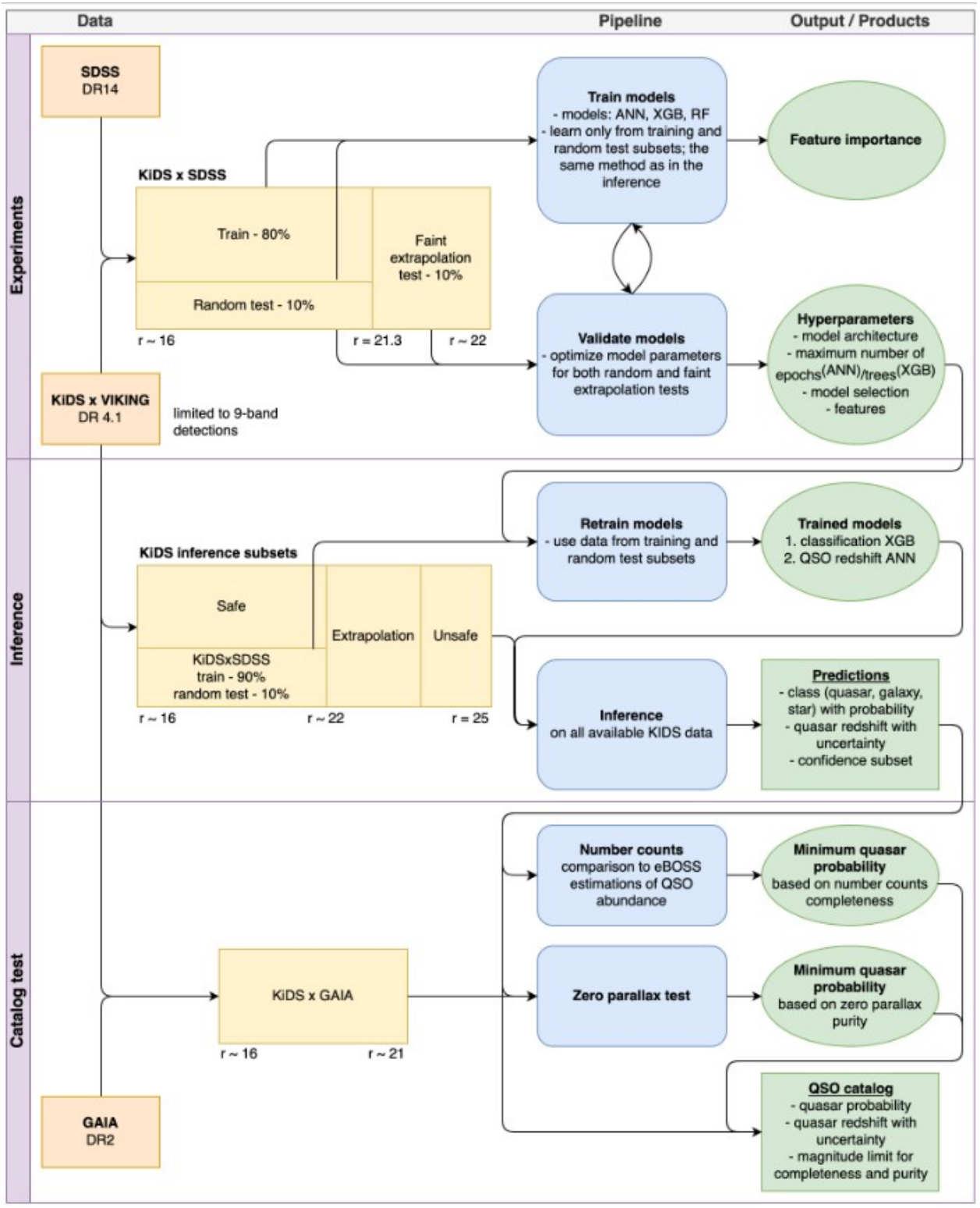
Summary

- Extragalactic Big Data
  - now more and more necessary to introduce new automated methods to study new large data (LSST)
- Problems and challenges
  - Extrapolation (small and biased training samples; limited parameter spaces)
  - **Physical interpretability** (do trends we see really mean something?)
  - Reproducibility
  - Foundation models – how much of the black box approach can we safely add
  - **Resources (costs of GPUs)**

The story of extrapolation

What is an AGN and how to find (and measure) them

- Challenges: different types and different diagnostics
- Big Data, big search: need for large and varied training samples!
- AGN properties, including photo-zs are tricky to recover even using „traditional” techniques
- Bright but training data available for low z/the bright end of LF: extrapolation problem.
- However, if we have a big training sample and are smart to use ML methods, we can get a reliable AGN sample and its properties (KiDS: Nakoneczny et al. 2019, Nakoneczny et al. 2021)

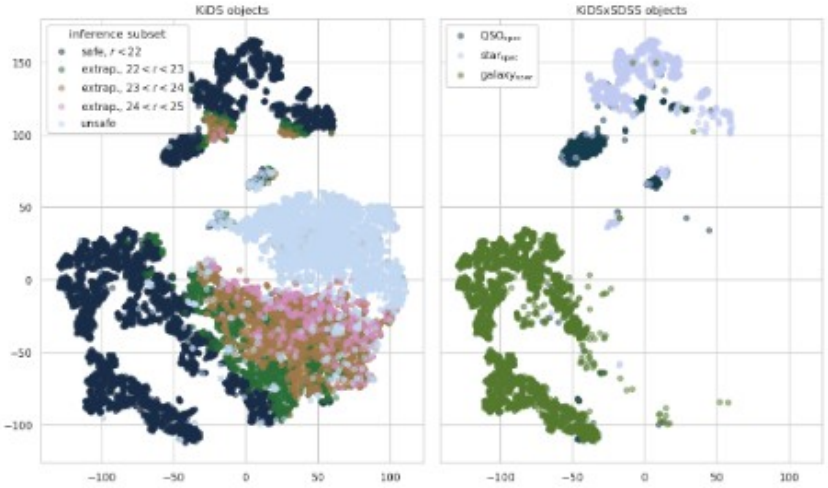


Nakoneczny et al. 2019

Nakoneczny et al. 2021

45 million objects of the KiDS photometric data limited to 9-band detections

-> 158,000 quasar candidates in the safe inference subset (r < 22) and an additional 185,000 candidates in the reliable extrapolation regime (22 < r < 23.5)



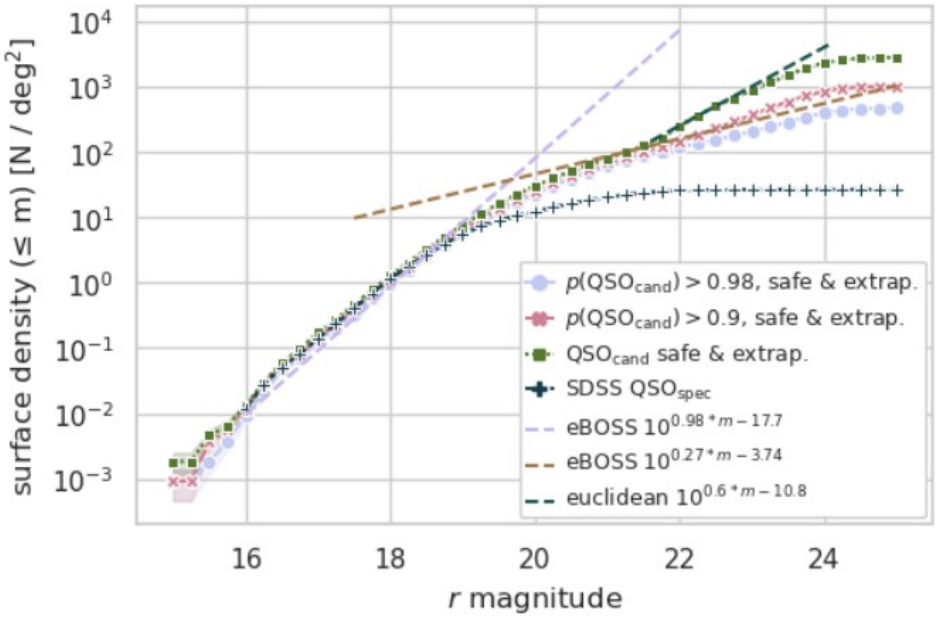
KiDS quasar candidates: how to make sure they are what we think they are

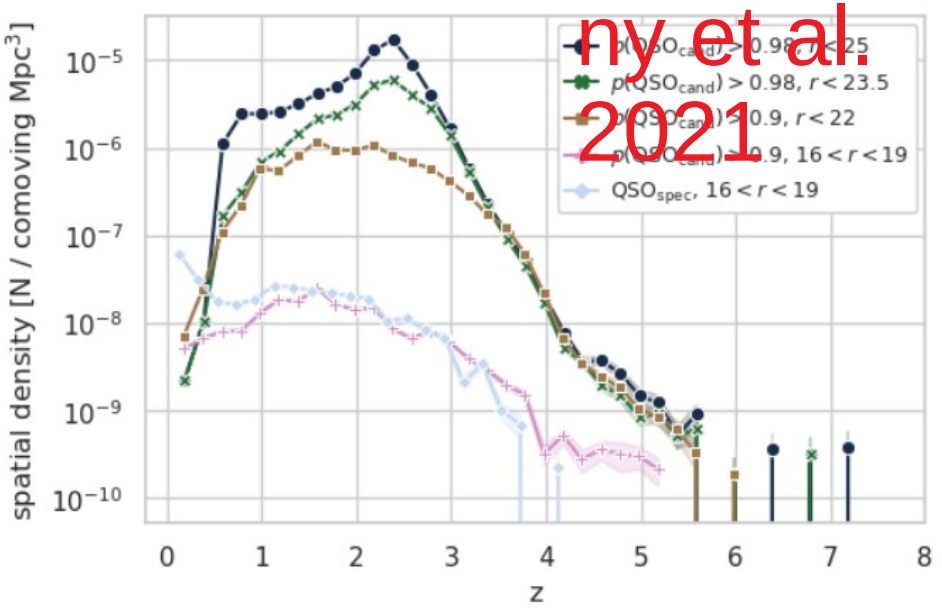
Nakoneczny et al. 2019

ID: Xboost
photo-zs: ANN

	safe $r < 22$	safe & extrap. $r < 23.5$	safe & extrap. $r < 25$
$\mathrm{QSO_{cand}}$	266k (100%)	1.6M (100%)	3M (100%)
$p(\mathrm{QSO_{cand}}) > 0.90$	**158k (59%)**	637k (39%)	1.1M (36%)
$p(\mathrm{QSO_{cand}}) > 0.98$	127k (48%)	**311k (19%)**	507k (17%)

Nakoneczny et al. 2021





(Previously) unknown classes inside known data and long history of interpretability

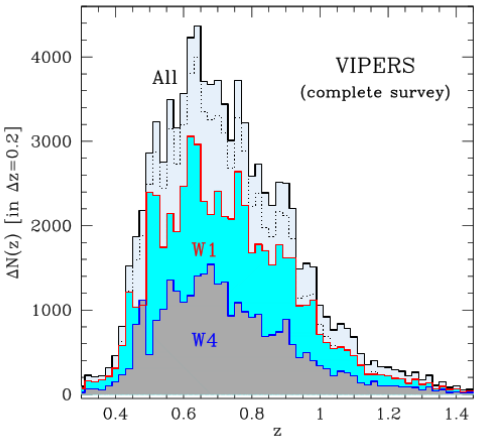
VIPERS

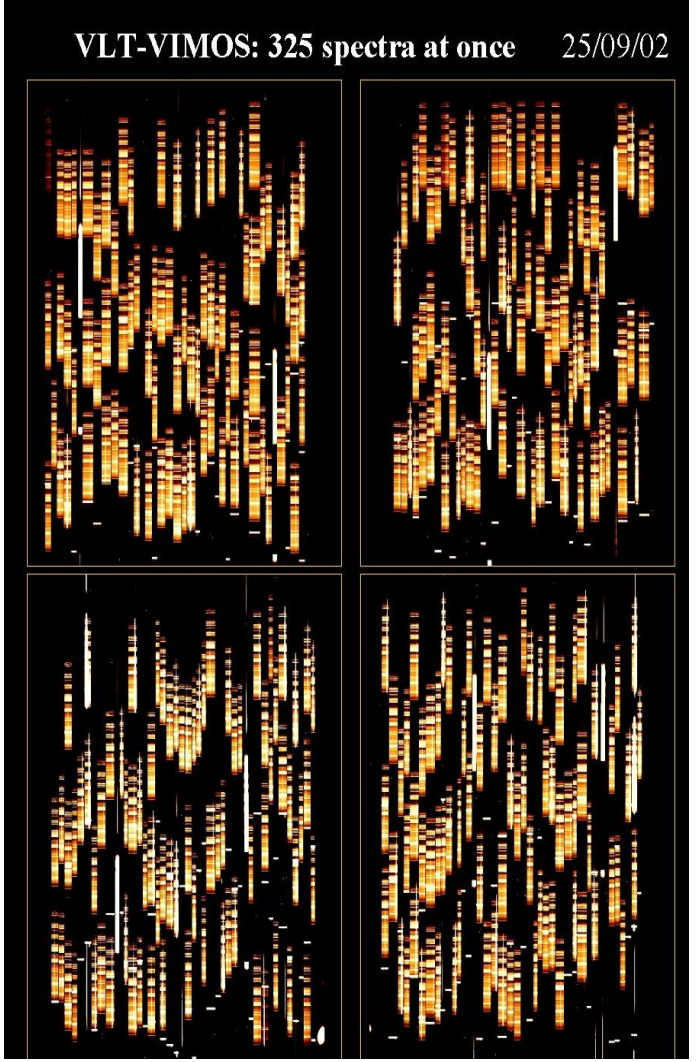
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

SURVEY STATUS AS OF 06/11/2016

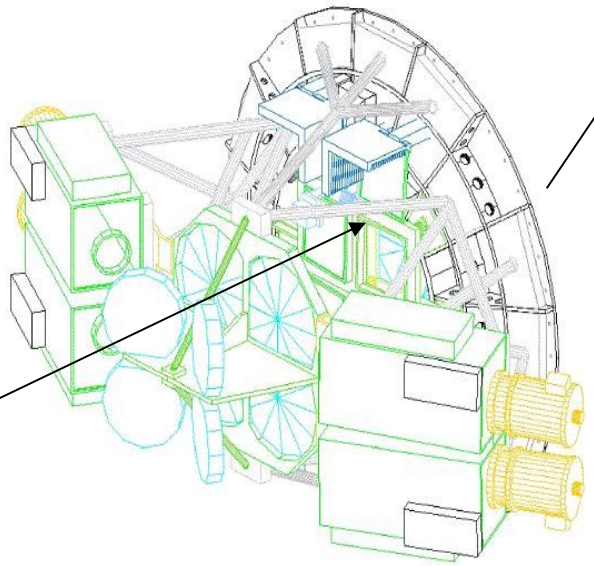
EFFECTIVE TARGETS	MEASURED REDSHIFTS	STELLAR CONTAMINATION	COVERED AREA
93252	88901	2265 (2.5 %)	100.0 %

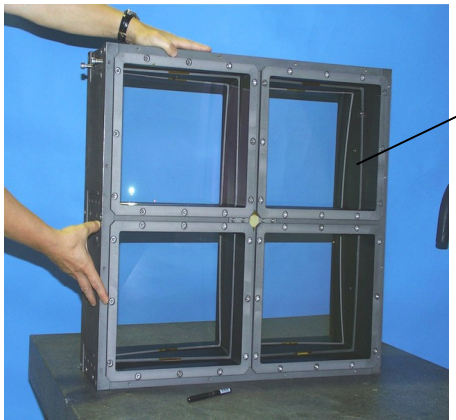
EFFECTIVE TARGETS (ET) are all the primary targeted objects with the exclusion of the ones flagged as -10 (undetected). MEASURED REDSHIFTS (MR) are the fraction of ET for which a redshift has been measured. STELLAR CONTAMINATION are the MR objects which have been identified as stars.





Large ESO Programme, 2008-2016



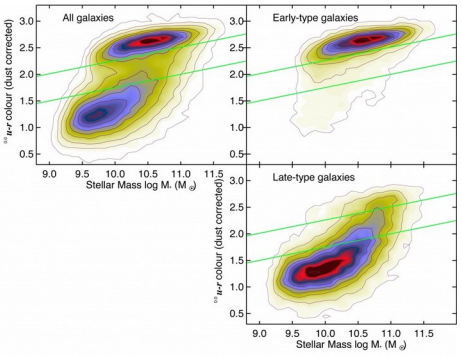


Goal: **100 000** spectra
of galaxies
at 0.5<z<1.2
2 fields on the sky, 24 deg^2

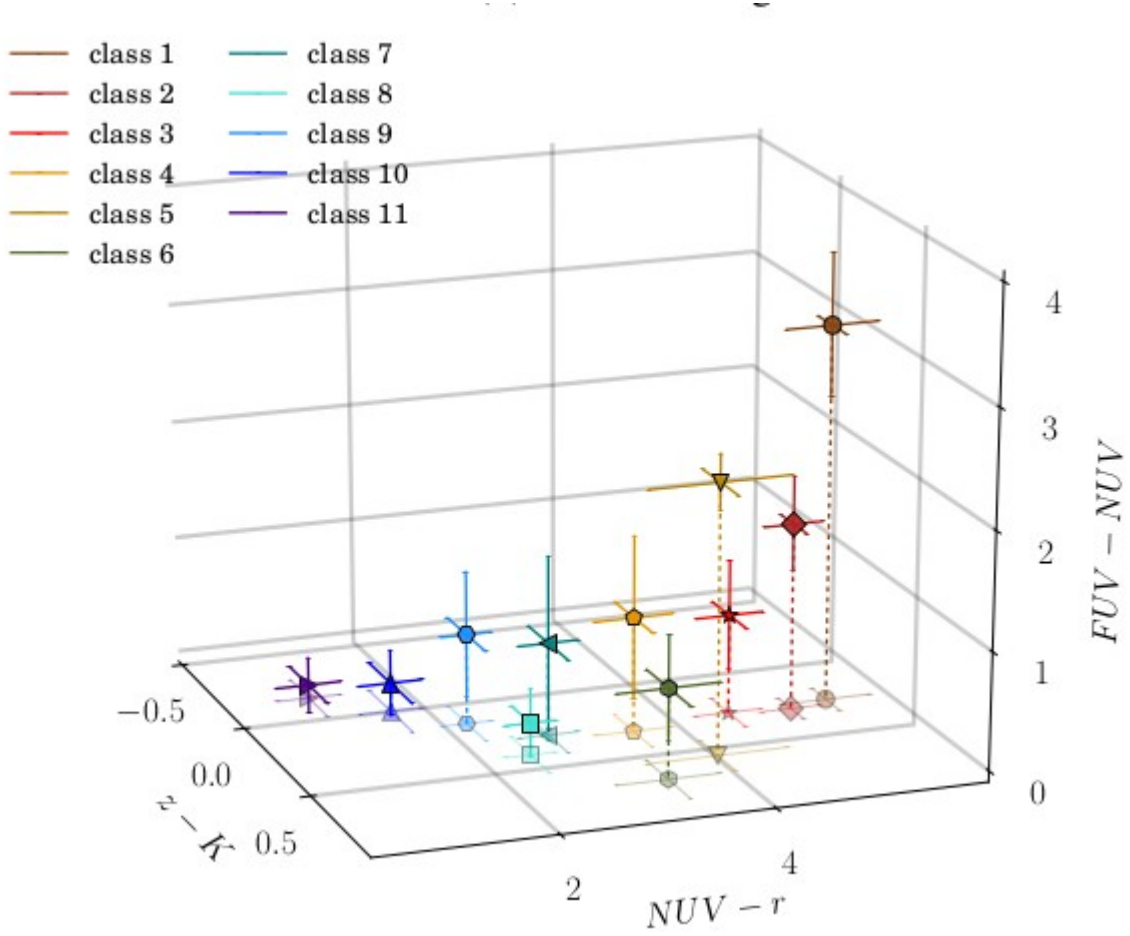
Guzzo et al. 2014, 2017, Scodeggio et al. 2018

VIPERS
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

Schavinski et al. 2016



- Method: unsupervised - FEM - Fisher Expectation-Maximization (Bouveyron & Brunet 2011);
- Parameter space: of 12 rest-frame optical magnitudes and a spectroscopic redshift



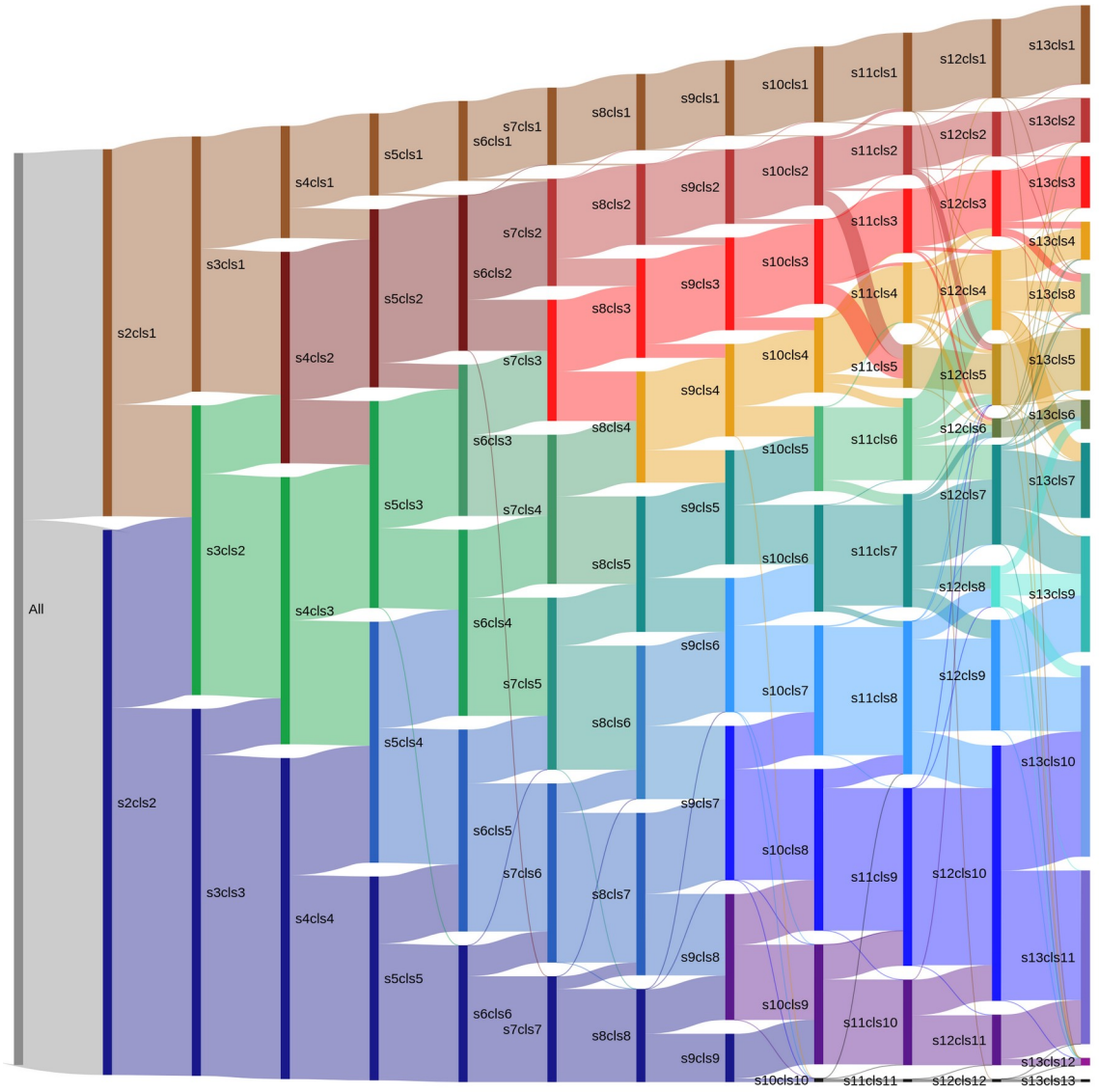
Beyond bimodality:
how many galaxy populations can be blindly selected at z~1?

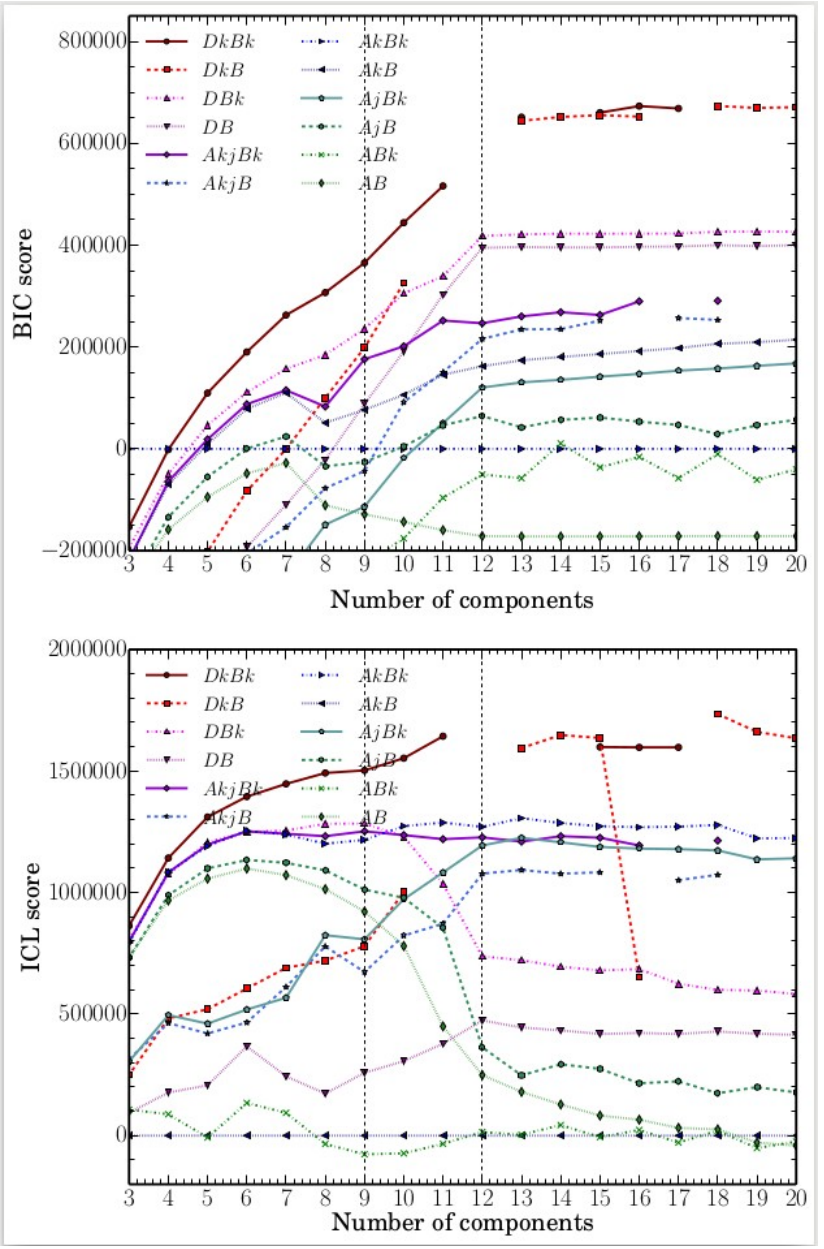
11 *well separated* classes of galaxies at 0.5<z<1
(+ a 12th class of outliers), forming the sequence of: 3, 3, and 5 subclasses of early, intermediate and late types, respectively.
well reproduced in SDSS (local Universe)

Siudek et al. 2018
Turner et al. 2021

VIPERS
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

How many galaxy populations can be blindly selected at z~1?





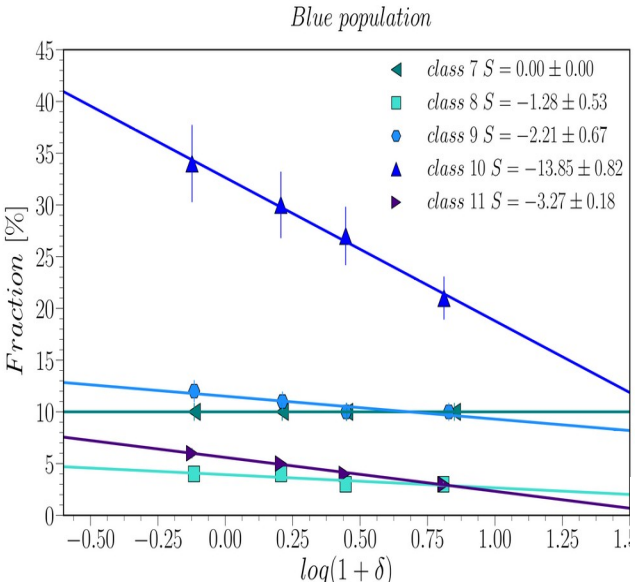
Siudek et al. 2018

VIPERS
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

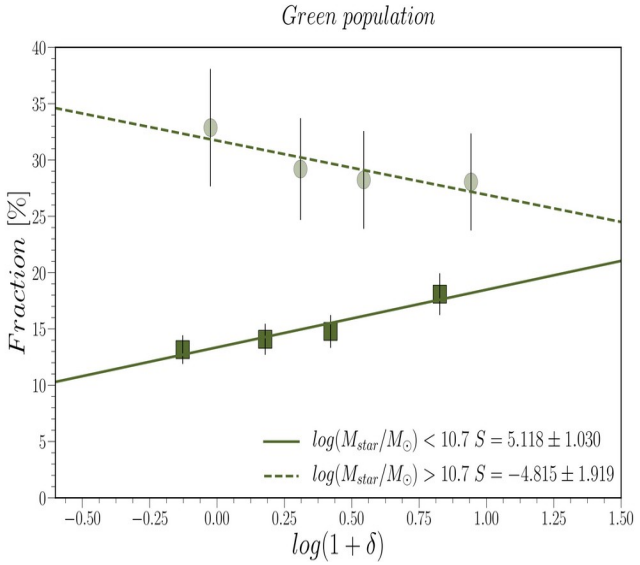
Does this 11 class division reflect actual physical information?

→ Traces of different galaxy evolutionary paths seen in multi-color space?

→ See what happens when quantities not related to classification are introduced (environment!)...



For blue galaxy populations: the downsizing trend is mostly driven by only one (admittedly, the largest) subpopulation (consistent with mass-driven passive evolution)
while the fractions of other blue SF galaxies are much less mass/environment-dependent



For intermediate and dusty populations the environmental trends are reversed depending on stellar mass: low mass ones behave like passive galaxies; high mass ones - like active galaxies

- a variety of galaxy populations is physical, and indicates a variety of their evolutionary paths

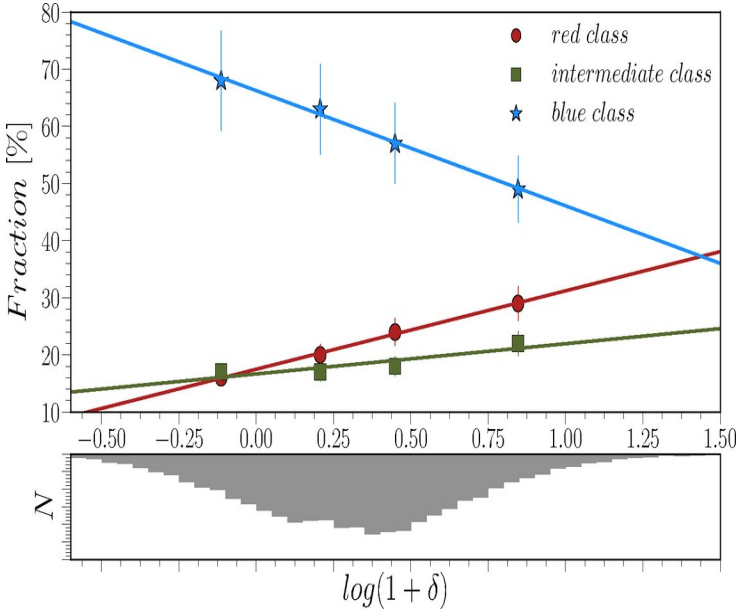
(Siudek et al. 2022)

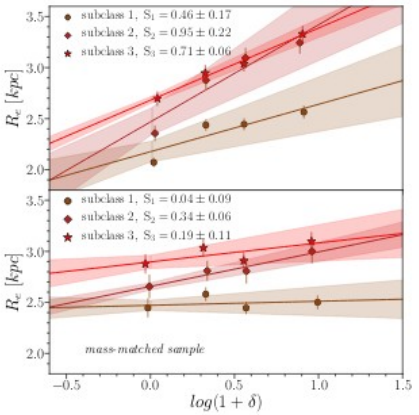
VIPERS
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

Does this 11 class division reflect actual physical information?

→ Traces of different galaxy evolutionary paths seen in multi-color space?

→ See what happens when quantities not related to classification are introduced (environment!)...



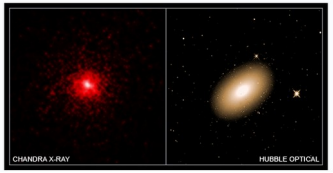


...the reddest red class: small and size does not depend on environment (independently on stellar mass): a product of early fast quenching (while the other two might have grown also through mergers)

a catalog of 77 „red nuggets” (relic galaxies which never merged in their lives) at z~0.7 (**Lisiecki et al. 2022**)

- a variety of galaxy populations is physical, and indicates a variety of their evolutionary paths

(**Siudek et al. 2022**)



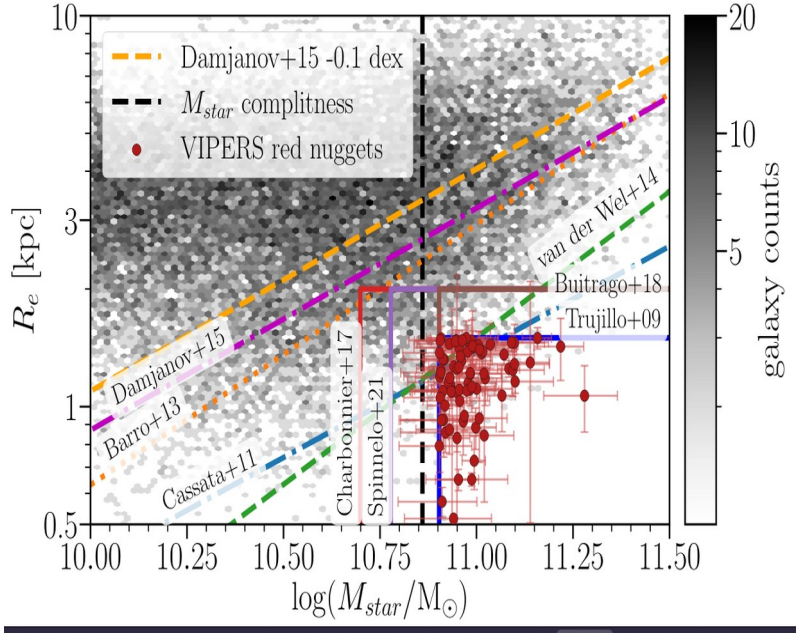
Mrk12 16 – not one of ours but as ours would look „today”

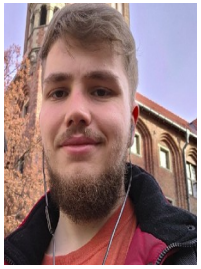
VIPERS
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

→ the first mass complete catalog of 77 “**red nuggets**” at z~0.7

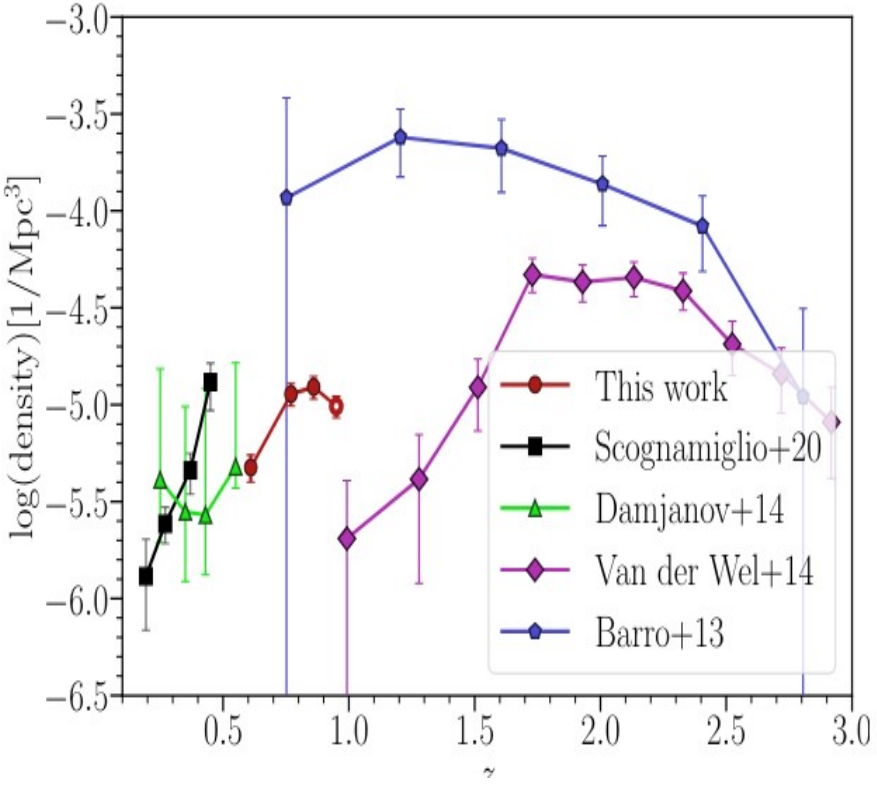
→ filling the gap between high z “red nuggets” and low-z “relics”

→ properties only weakly dependent on environment

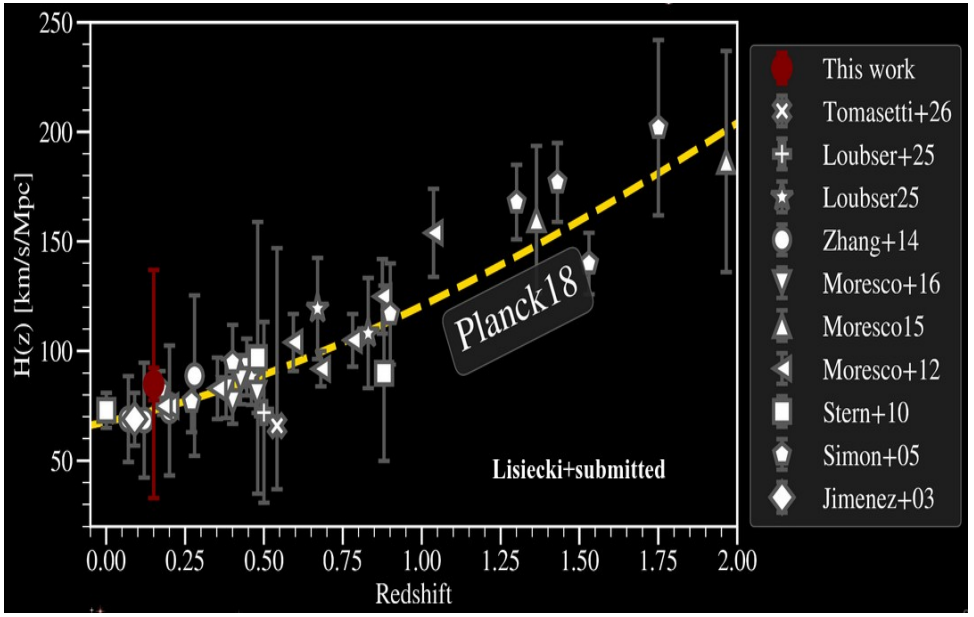




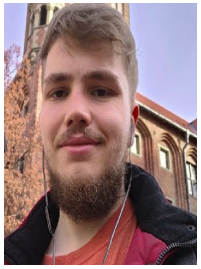
Lisiecki et al. 2023
Siudek, Lisiecki et al. 2023



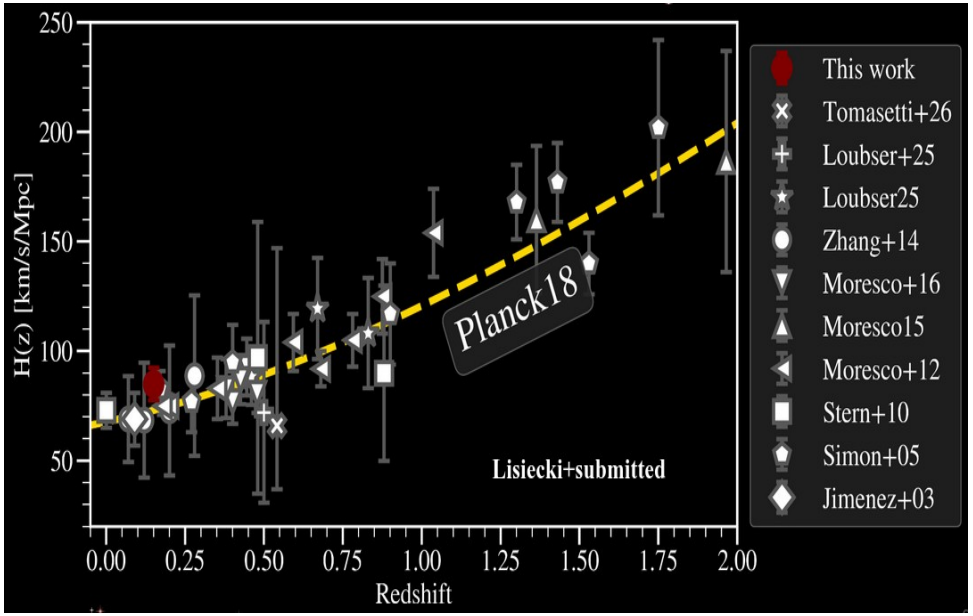
First measurement of H(z) from red nuggets as cosmic chronometers (low z – SDSS based E-INSPIRE - for starters)



With present error



Lisiecki et al. 2026



With error expected from the ~LSST statistics

Merger in the background
or a history
of unexpected interpretability

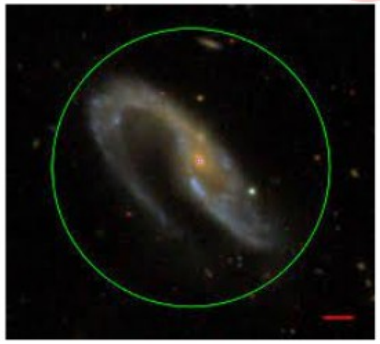


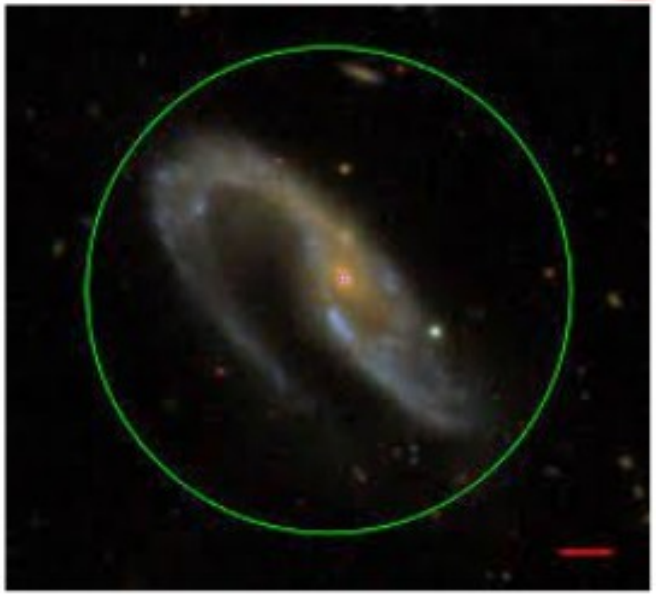
Credit: HST, NASA/ESA

How to automatically find merging galaxies?

→ People very often use Deep Learning (with moderate success)

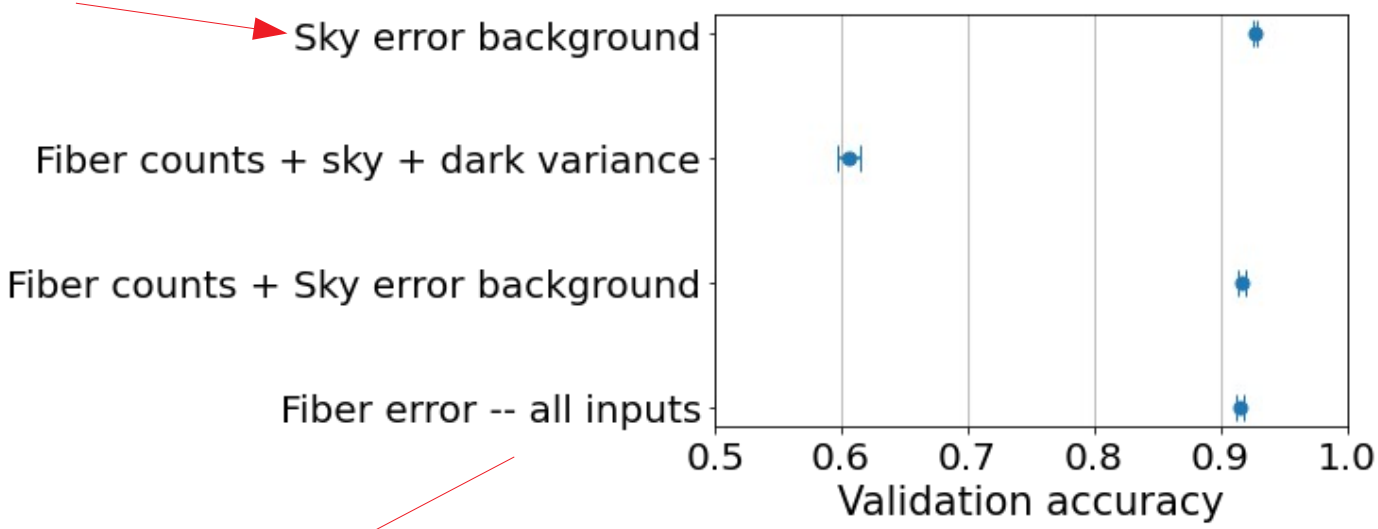
→Concept: see if we can do any good (but faster/easier/more interpretable) with photometry only (fluxes, colours, errors)

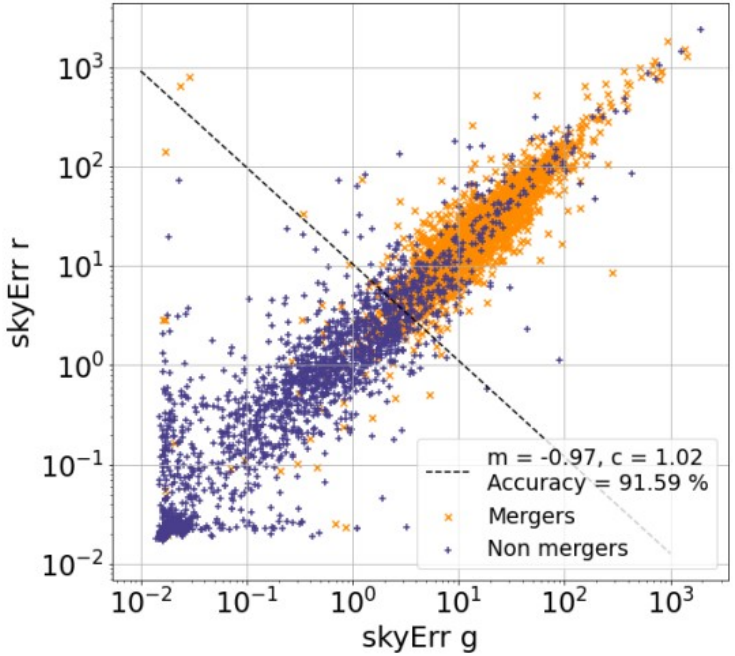




How to automatically find merging galaxies?

→ What is a magical ingredient of fiber errors?





→ We do not need any ML do get ~92% accuracy – it was just about finding the key data

→ Physics: merging galaxies (today) do not differ that much from other galaxies – what makes them different are their surroundings (tidal tails etc.)

→ new generation of DL for background only (never forget the power of differential analysis...)