BENCHMARKING FOUNDATION MODELS FOR UNSUPERVISED DISCOVERY IN LARGE MULTIMODAL ASTROPHYSICAL DATASETS

Maxime RONCERAY

Supervisors : Marc Huertas-Company, Alexandre Chanson, Malgorzata Siudek

<u>VERMEER Johannes</u> - 1668

The Astronomer

Vermeer's *Astronomer* depicts the study of the celestial sphere, reflecting humanity's long-standing fascination with the cosmos

A curiosity that continues to drive modern space missions such as Euclid.

Table of contents

Context and Motivation

- Euclid mission overview
- Why anomalies matter in astronomy
- Scientific motivations

Foundation Models

- What are foundation models?
- Why use embeddings for discovery?

Anomalies and Detection Methods

- Classical methods
- Deep generative models
- Bayesian and probabilistic approaches

Application to Euclid Q1

- AstroPT, AstroCLIP, AION
- Multimodality Dataset Overview

Challenges and Perspectives

- Domain Shift and interpretability
- Human validation of anomalies
- Next steps

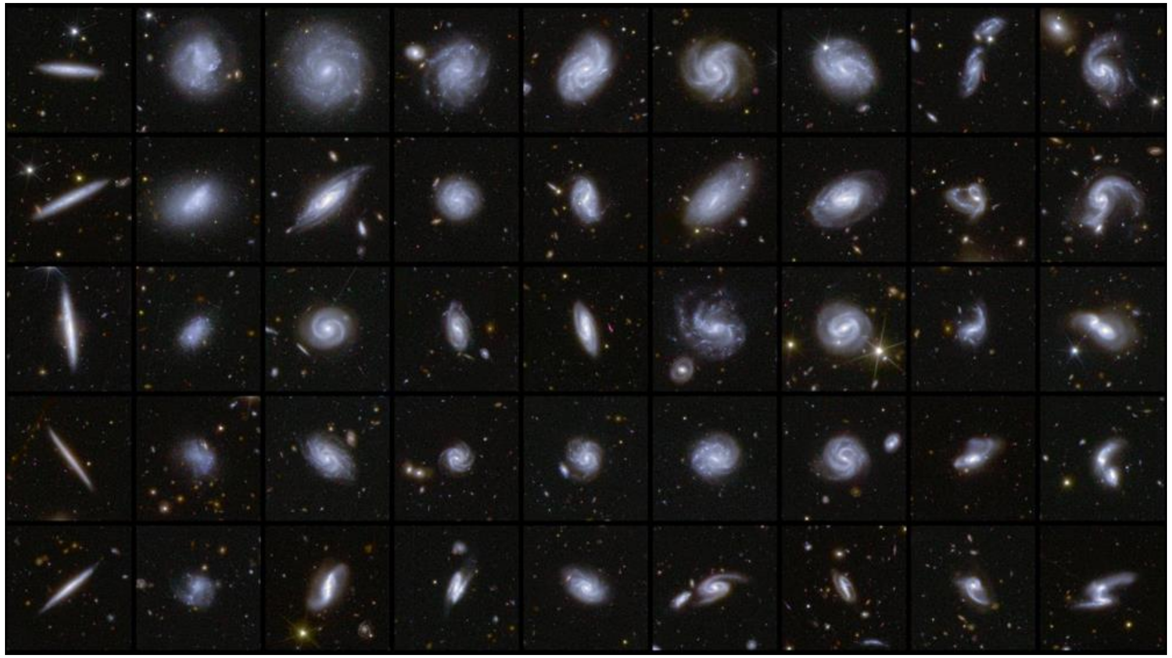
Motivations

- Many discoveries (quasars, galaxies, exoplanets) came from anomalies

- Euclid = billions of galaxies and stellar objects→ rare events lost in “normal” data
- Large variety of rare transiant phenomenon

- Challenge: how to detect the unexpected in massive datasets?

Automated methods and machine learning

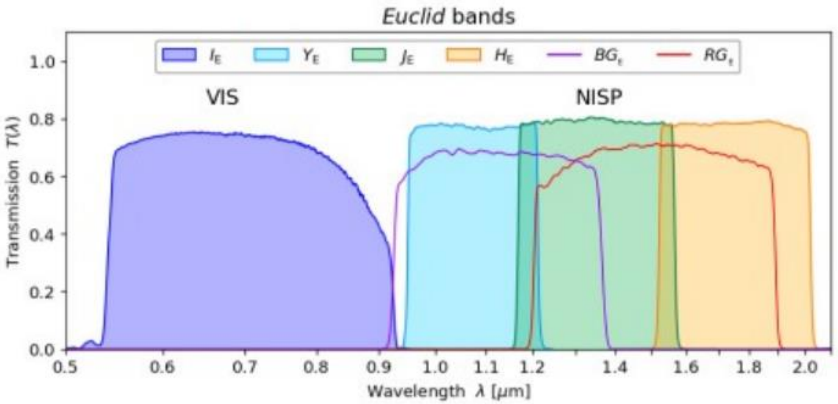


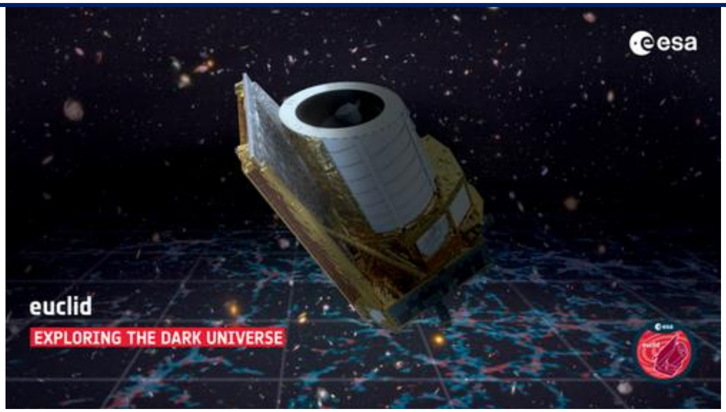
Sample of galaxies observed by Euclid in the Q1 data release.

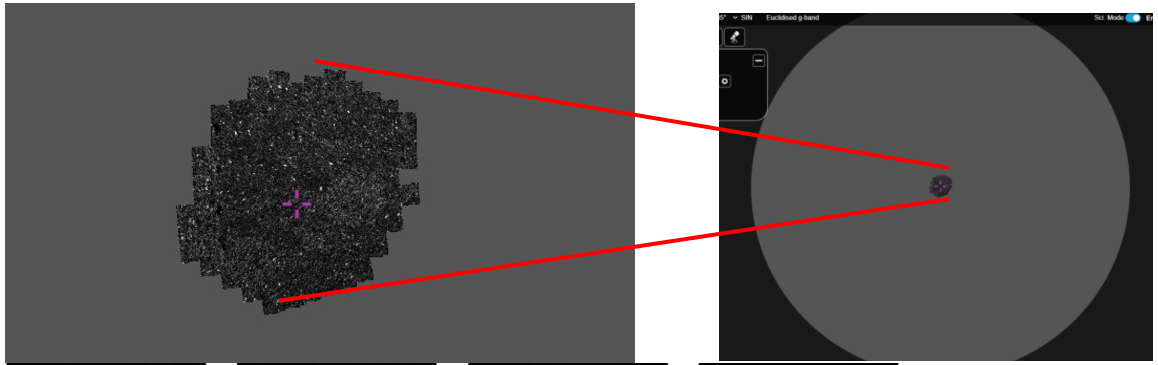
Source: ESA / Euclid / Euclid Consortium / NASA, “Euclid Galaxy Zoo galaxies to classify” (ESA, 2024), under CC BY-SA 3.0 IGO / ESA Standard Licence.

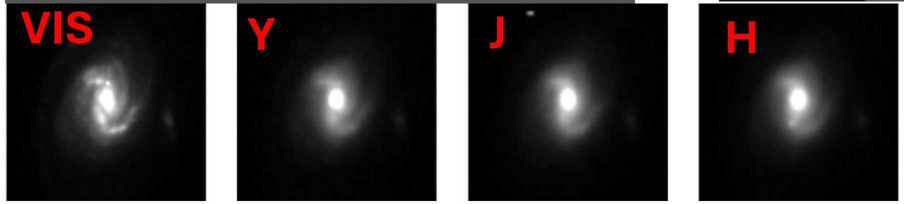
Euclid Mission Overview

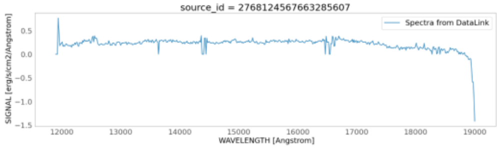
- ESA mission (2025) to map the dark Universe (Find Dark Mater Origin)
- Instruments: VIS (optical, 550–900nm) & NISP (NIR, Y-J-H) + Spectrescopy
- Q1 = first public data (63 deg², 30M sources)
- Redshift range* : **z ≃ 0.7 to z ≃ 1.8, up to** z ≲ 4 for some specialized sudies
- Exposure time: ◆ Photo: 87 s ◆ Spectro: 550 s









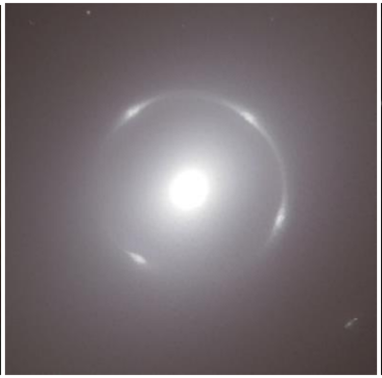


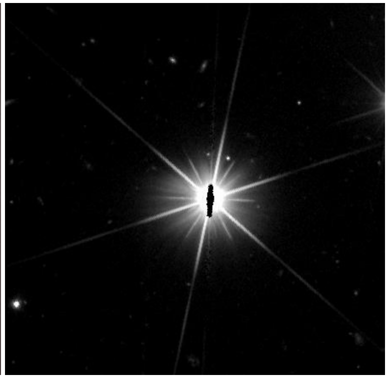
* redshift measures *how much the Universe has expanded* since the light was emitted

Typology of Anomalies

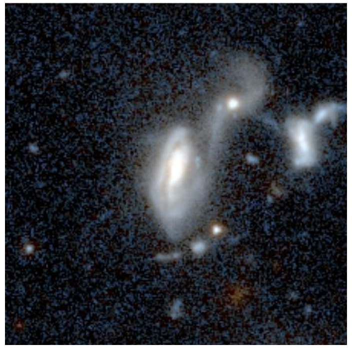
Strong gravitational lenses

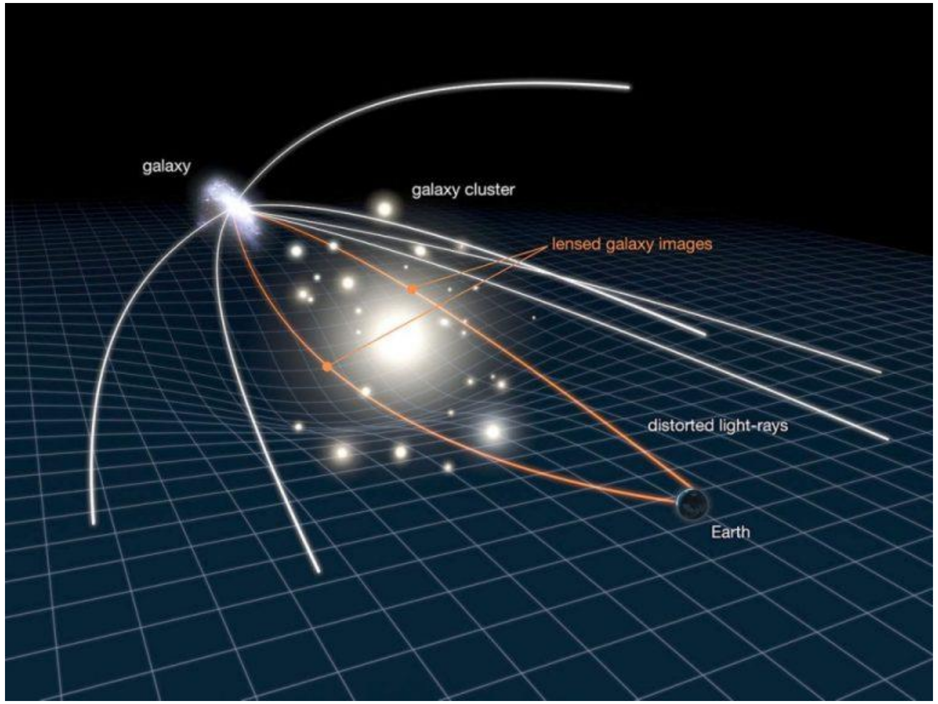
- Multiple images or arcs of a background source caused by massive objects bending light





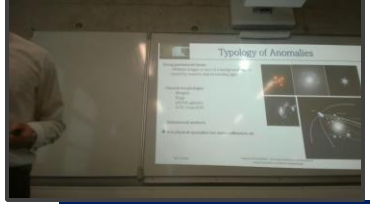
- Unusual morphologies
  - Mergers
  - Rings
  - jellyfish galaxies
  - AGN, Dual-AGN





- Instrumental artefacts

➔ non-physical anomalies but useful calibration set



Why Are Anomalies Important?

Strong gravitational lenses :

- Test general relativity
- Lensing measures *total* mass (including dark matter), independent of light.
- Time delays between multiple images give the *time-delay distance*:
- The *abundance* of strong lenses depends on the halo mass function

$$n(M, z) \propto f(\sigma_8, \Omega_m)$$

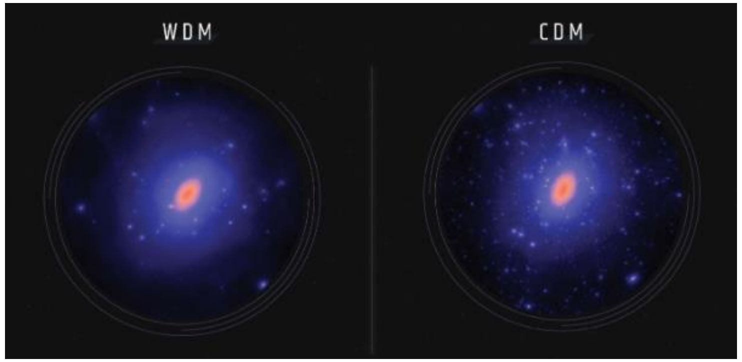


Fig. Simulations of a warm dark matter halo with embedded galaxy (left) vs. a cold dark matter halo version of the same amount of mass in dark and normal matter. *(Credit: R. Wechsler adapted from Bullock and Boylan-Kolchin, 2017, based on simulations by V. Robles, T. Kelley, and B. Bozek, et al.)*

where

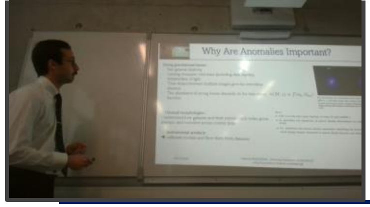
- $n(M, z)$ is the halo mass function at mass M and redshift z
- σ_8 quantifies the amplitude of matter density fluctuations on scales on 8 h^{-1} Mpc scales
- Ω_m represents the matter density parameter, describing the fraction of the Universe's total energy density contained in matter (both baryonic and dark)

- Unusual morphologies :

- understand how galaxies and their central black holes grow, interact, and coevolve across cosmic time.

- **Instrumental artefacts**

➔ calibrate models and filter them from datasets



Anomaly Detection Methods

Classical Anomaly Detection

- Early approaches: Clustering, Isolation Forest
  - Rely on tabular, hand-crafted features
  - *e.g.: SDSS/WISE outlier detection via SVM & Random Forests*

Deep Generative Models

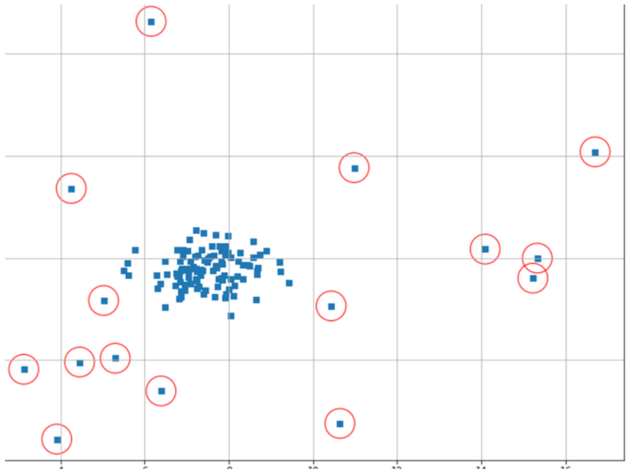
- Learn to reconstruct input → high loss = anomaly
- *e.g. : AE for galaxy morphologies, lens recovery (Cheng+ 2020)*
- Limits: anomalies can be "learned" as inliers

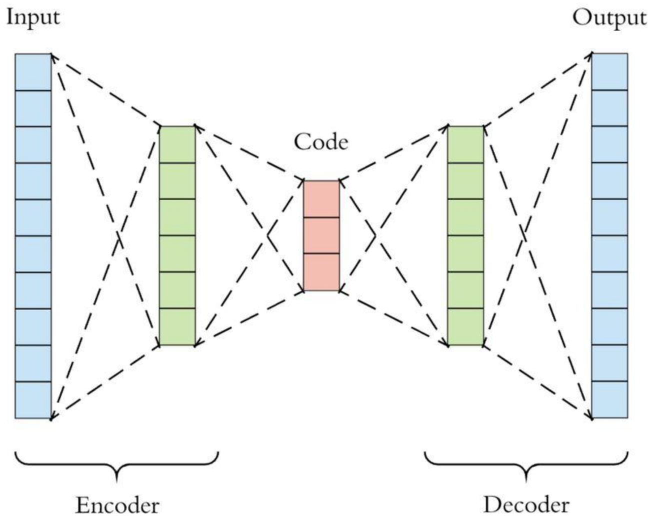
Advanced methods

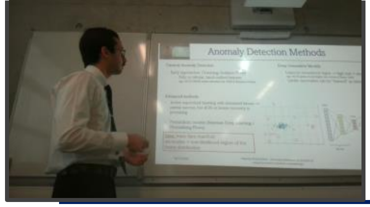
- Active supervised learning with simulated lenses → partial success, but AGN or lenses recovery is promising

- Probabilistic models (Bayesian Deep Learning / Normalizing Flows)

<u>**Idea:**</u> learn data manifold
anomalies = low-likelihood region of the learnt distribution

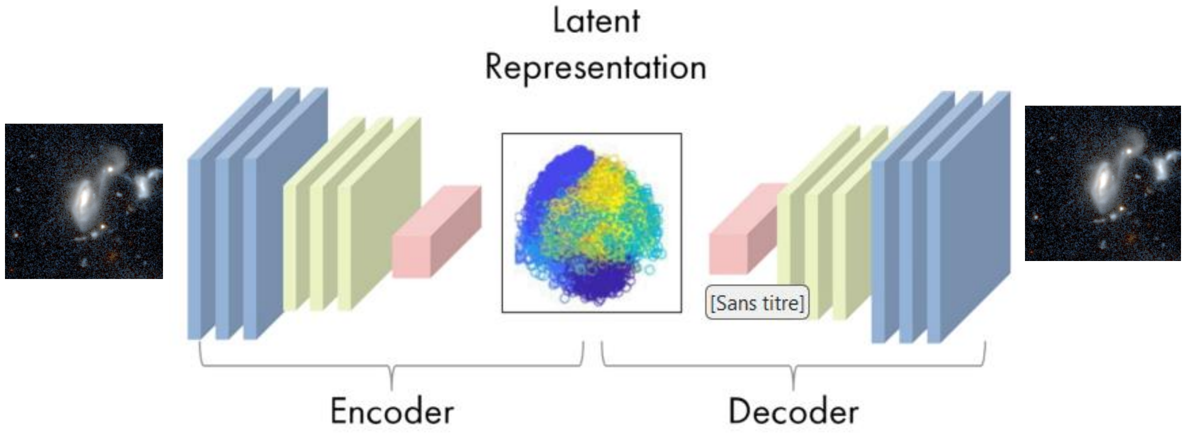






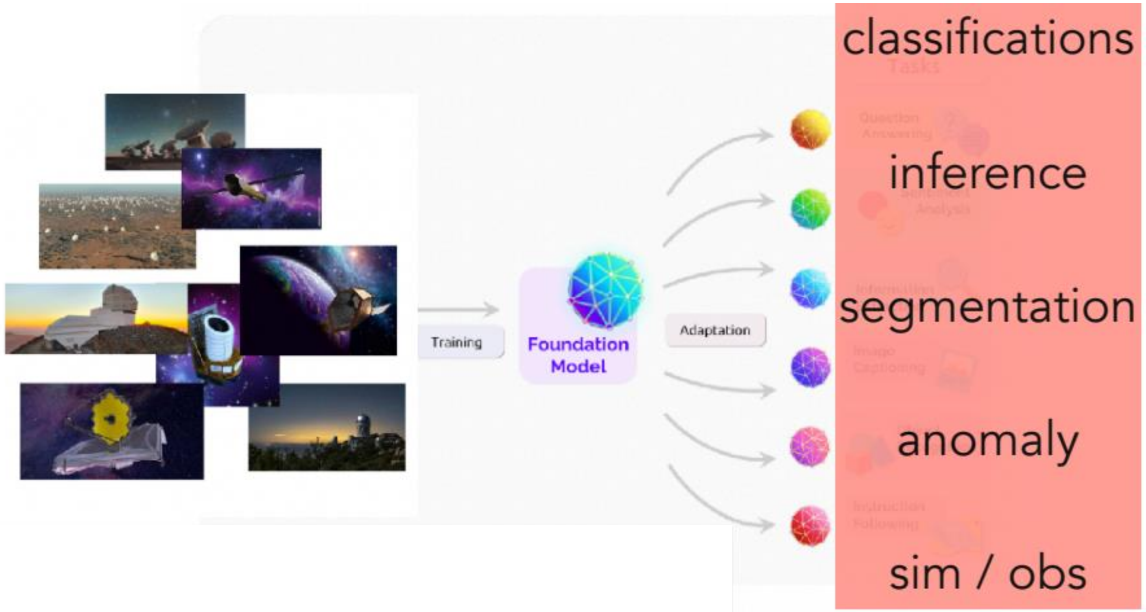
What is a Foundation Model ?

- Large, self-supervised (encoders – decoders) trained on massive datasets
- Learn transferable embeddings (universal representations)
- Allow reuse for many downstream tasks: classification, regression, outlier detection



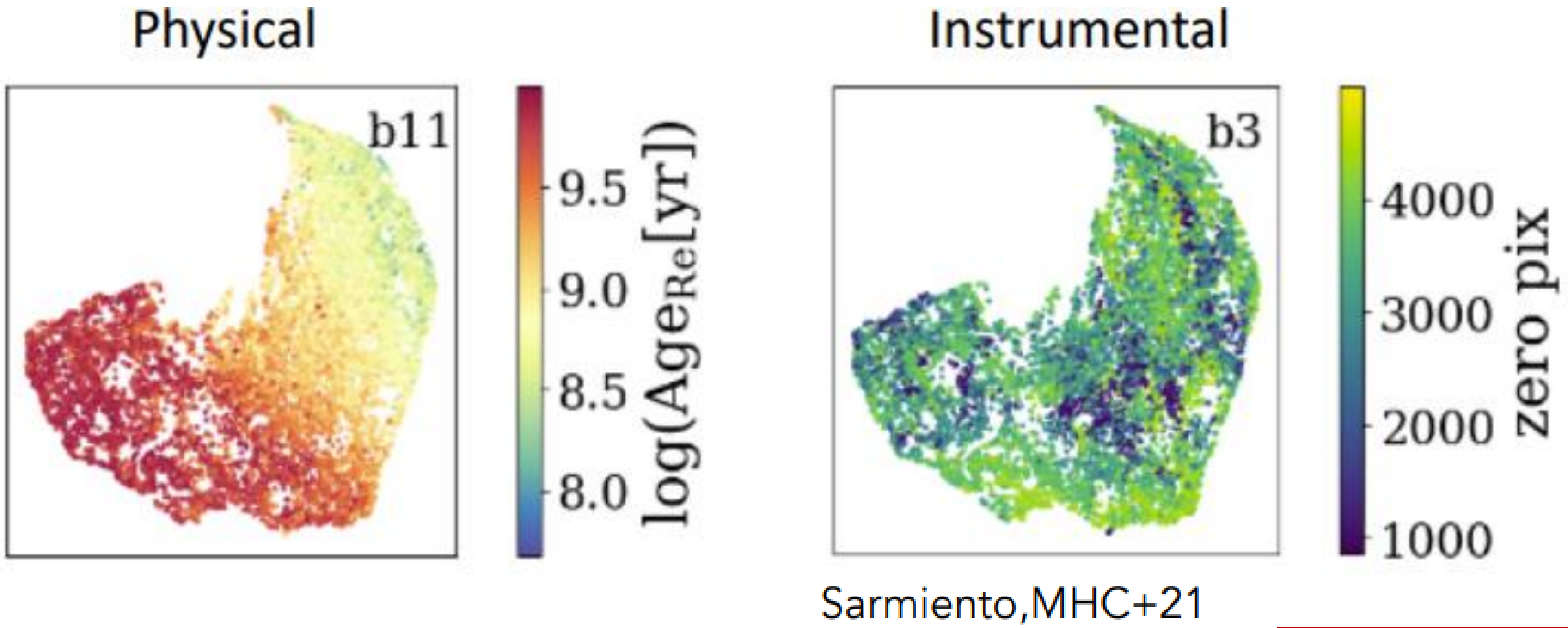
Most of the time : transformers

1. never again train a neural network from scratch
2. no labels required (or very few)
3. integrates all data beyond images
4. very simple analysis tools for downstream tasks



Why Use Embeddings for Discovery ?

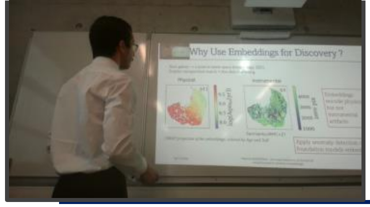
- Each galaxy → a point in latent space (morphology, SED)
- Enables unsupervised search + few-shot fine-tuning



UMAP projection of the embeddings, colored by Age and SnR

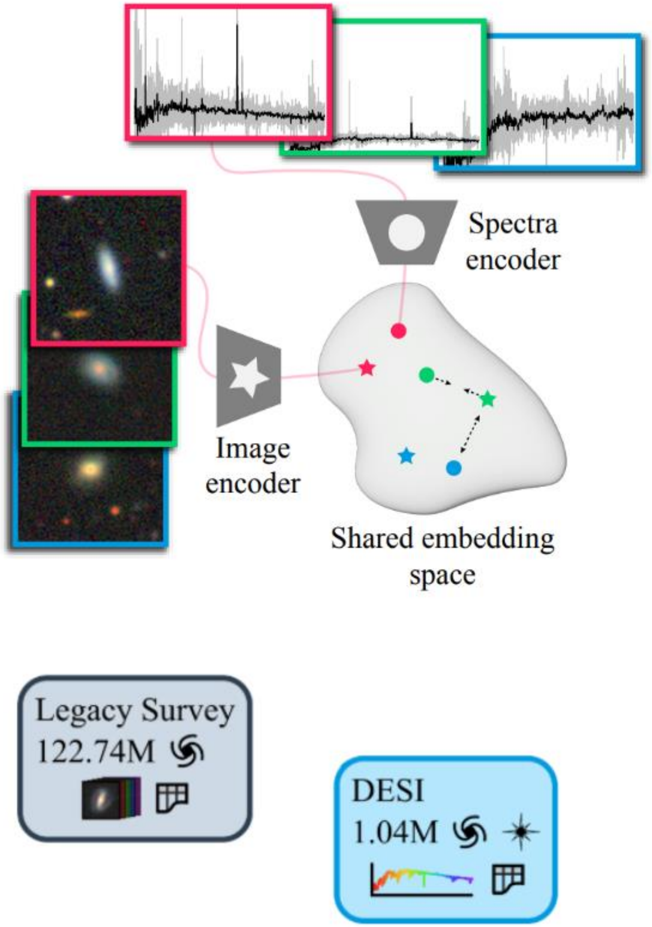
Embeddings encode physics but not instrumental artifacts

Apply anomaly detection on foundation models embeddings

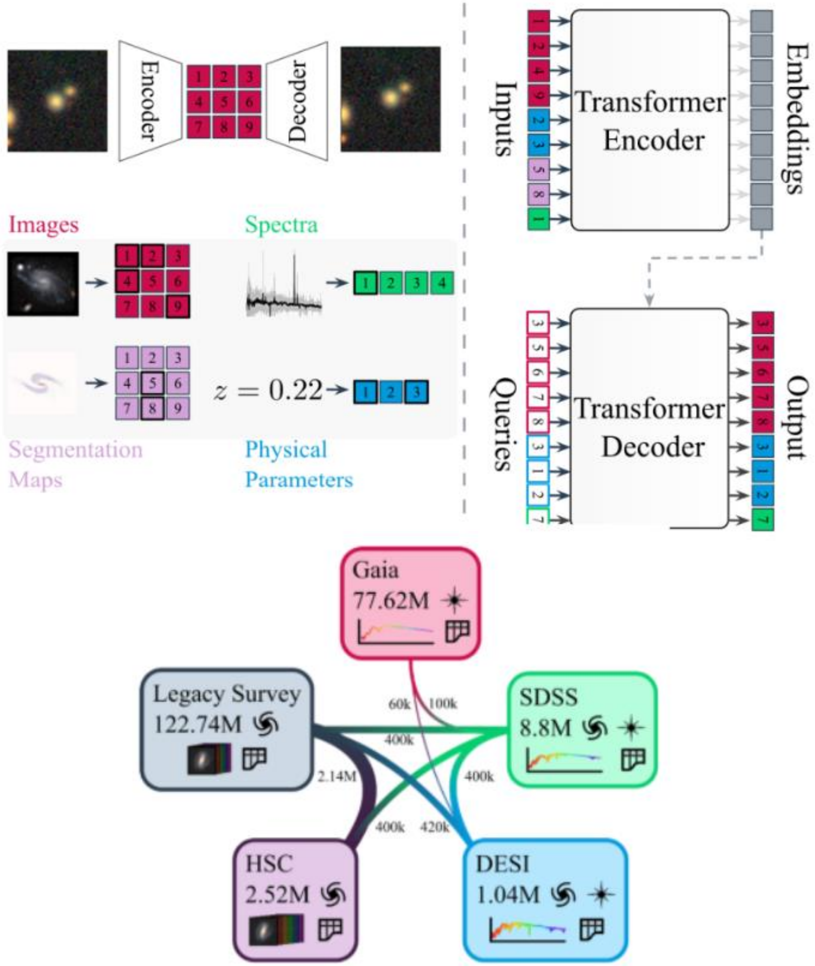


Astronomical Foundation Models

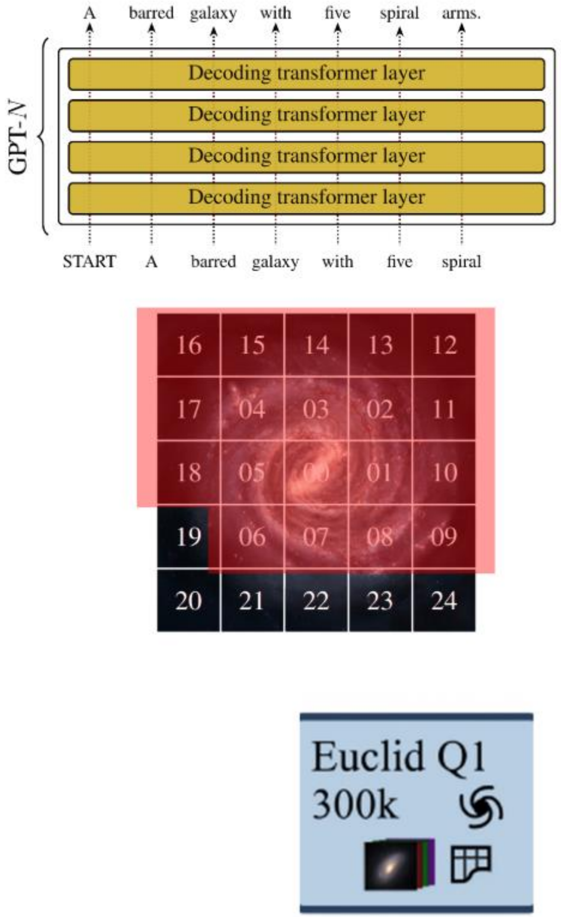
AstroClip

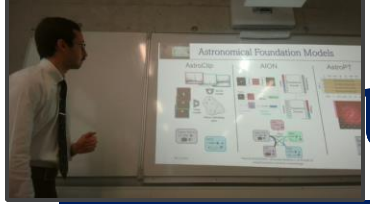


AION



AstroPT

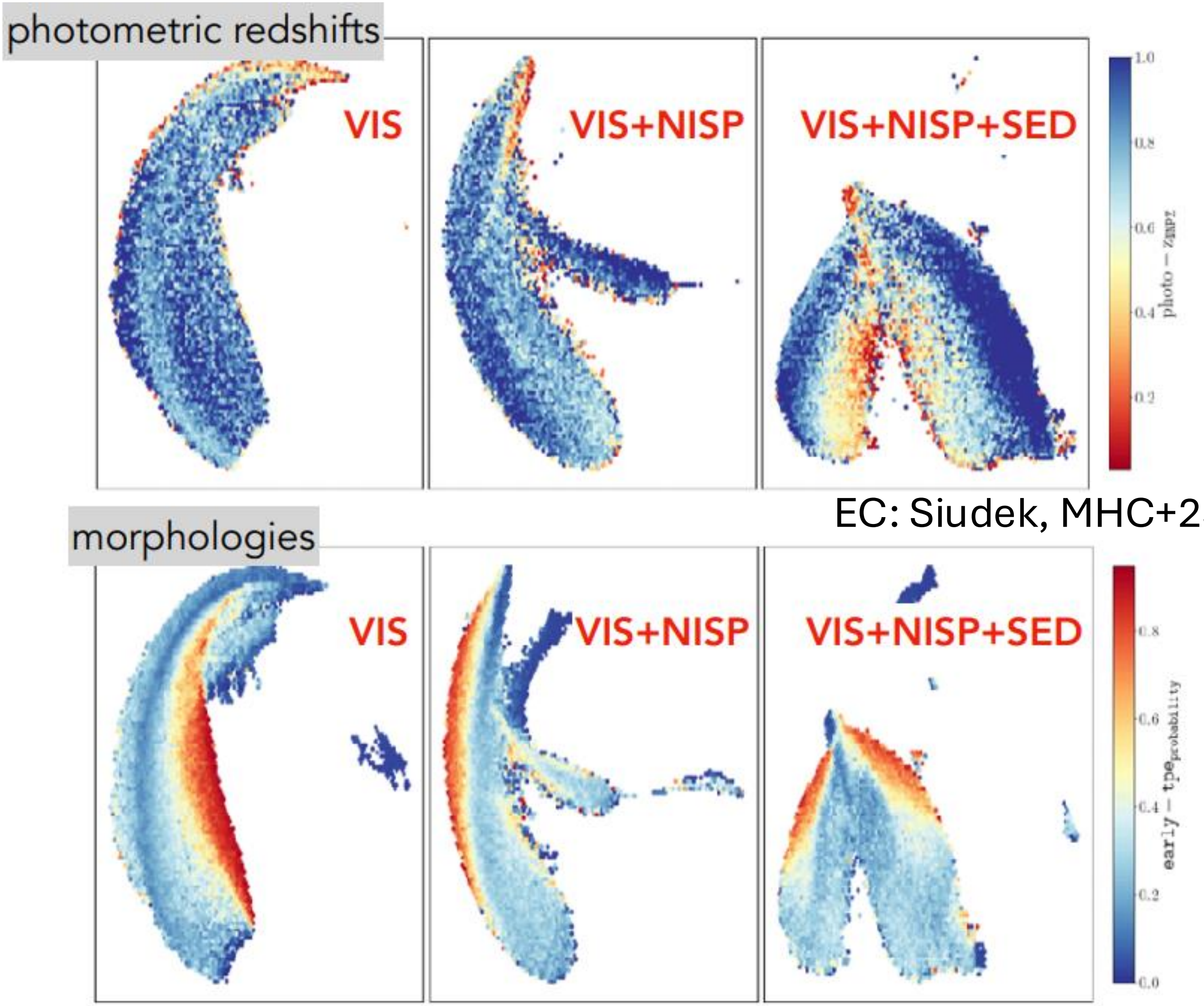




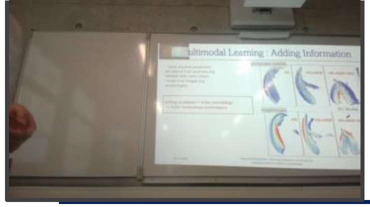
ultimodal Learning : Adding Information

- Some physical parameters are infered from spetrums *(e.g. redshift)* while some others comes from images *(e.g morphologies)*

Adding modalities = richer embeddings
→ better downstream performance

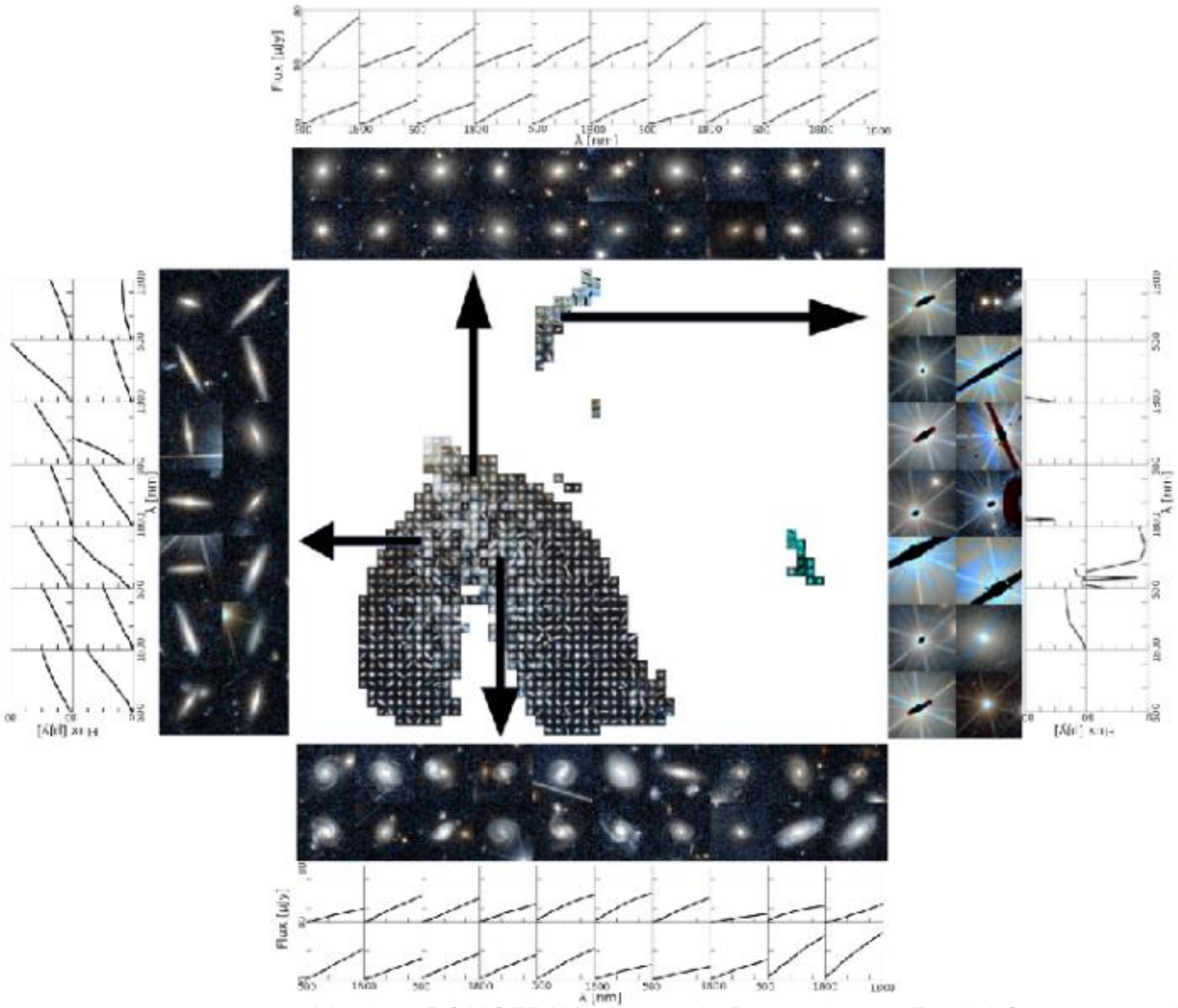


EC: Siudek, MHC+25

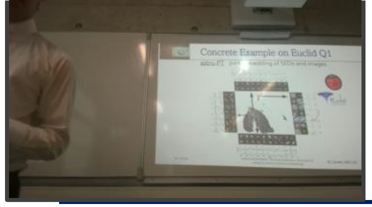


Concrete Example on Euclid Q1

astro-PT joint embedding of SEDs and images



EC: Siudek, MHC+25



Current Challenges

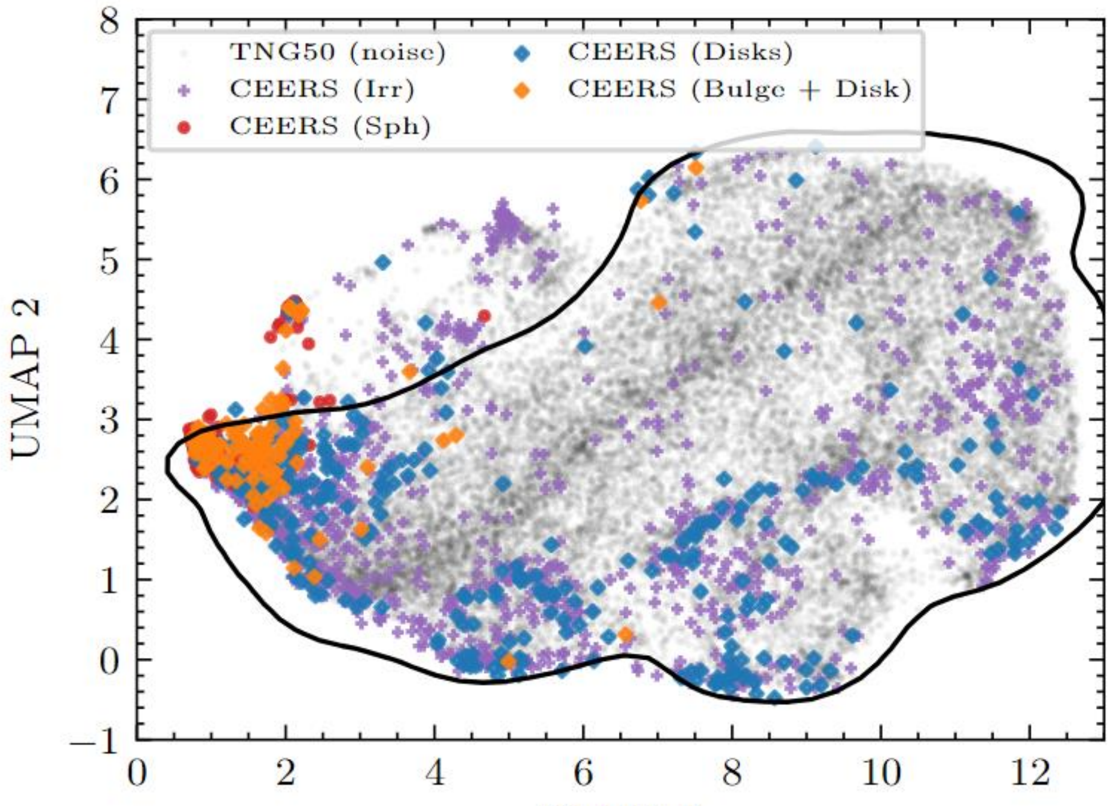
- **Domain shifts:**
  - Simulated vs. Real,
  - Euclid vs. other surveys

➔ *Can a foundation model not trained on Euclid still generalize and produce sufficiently good embeddings*

- **Interpretability:** why is an object flagged as outlier?

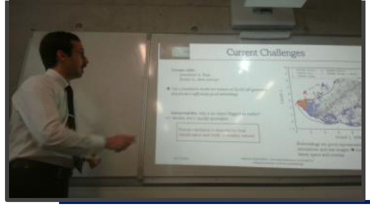
=> Models don't classify anomalies

Human validation is essential for final classification and verify anomalies natures



V.Ferrero

Embeddings are good representation of both simulations and real images ➔ both share same latent space and overlap



Realisation Plan

Use images from Euclid cutouts and crossmatch with DESI

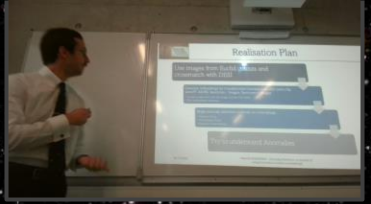
Generate embeddings for 3 multimodals foundation models (astroclip astroPT AION). Spectrum – Images, Spectrum x Images

- For aion maybe fine tune the image encoder (VQ-VAE)
- Test embeddings robustness

Apply anomaly detection methods on embeddings

- Isolation forest
- Normalising Flows
- Bayesian Deep learning

Try to understand Anomalies



Thank you for your attention.

Any questions?

Papers

- Dalya Baron. Machine Learning in Astronomy: a practical overview. 2019. arXiv: 1904. 07248 [astro-ph.IM]. url: https://arxiv.org/abs/1904.07248.

- Euclid Collaboration et al. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model. 2025. arXiv: 2503 . 15312 [astro-ph.GA]. url: https://arxiv.org/abs/2503.15312.

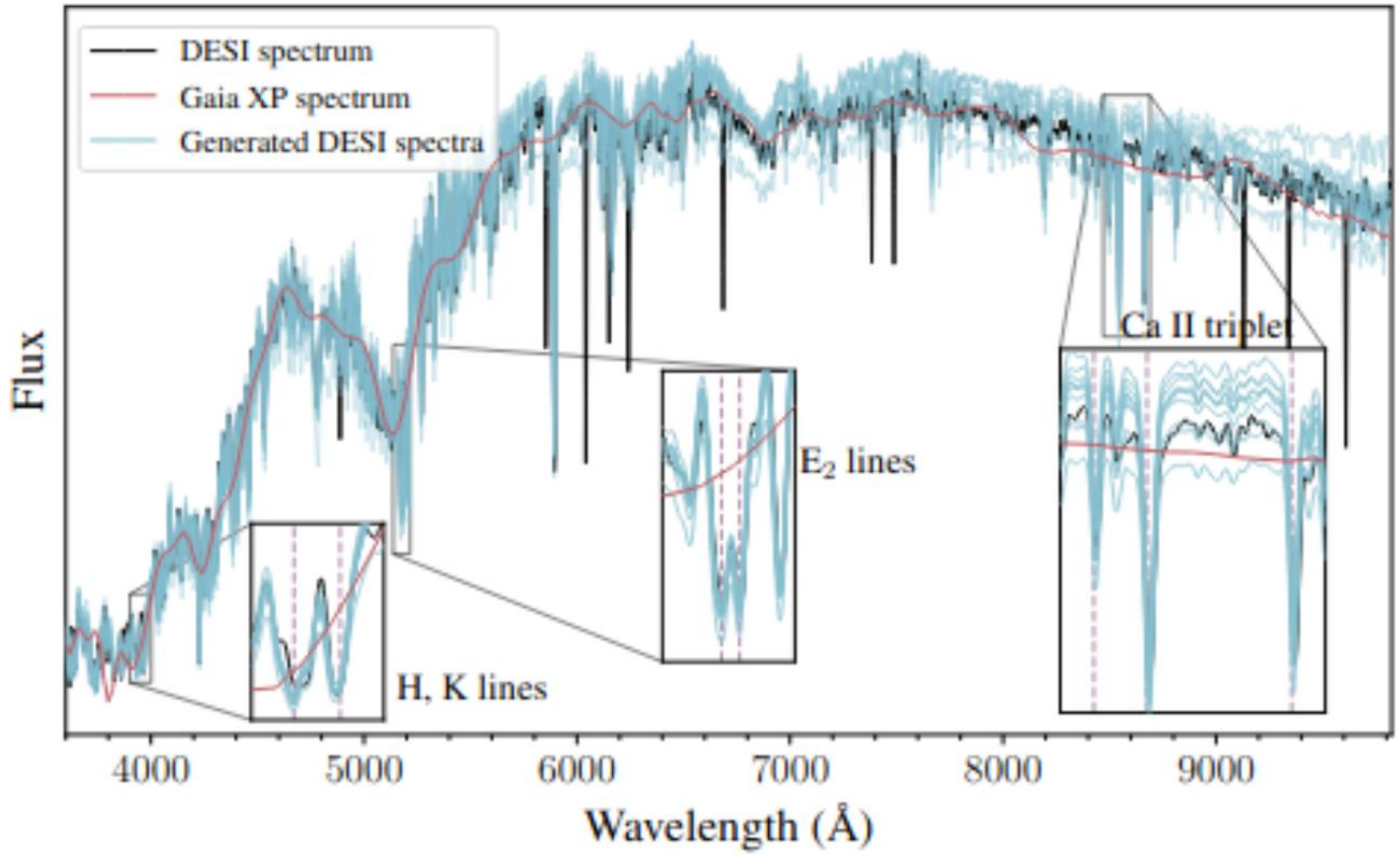
- Jing Ren et al. Foundation Models for Anomaly Detection: Vision and Challenges. 2025. arXiv: 2502.06911 [cs.LG]. url: https://arxiv.org/abs/2502.06911

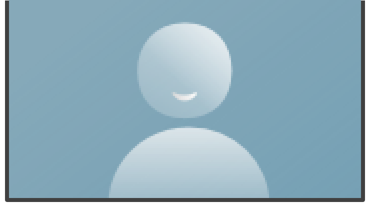
- Liam Parker et al. AION-1: Omnimodal Foundation Model for Astronomical Sciences. 2025. arXiv: 2510.17960 [astro-ph.IM]. url: https://arxiv.org/abs/2510.17960.

- Berta Margalef-Bentabol et al. “Detecting outliers in astronomical images with deep generative networks”. In: Monthly Notices of the Royal Astronomical Society 496.2 (June 2020), pp. 2346–2361. issn: 1365-2966. doi: 10 . 1093 / mnras / staa1647. url: http : //dx.doi.org/10.1093/mnras/staa1647.

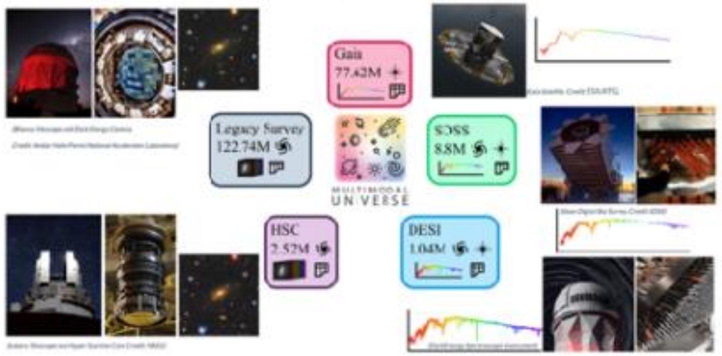
Multimodal Dream

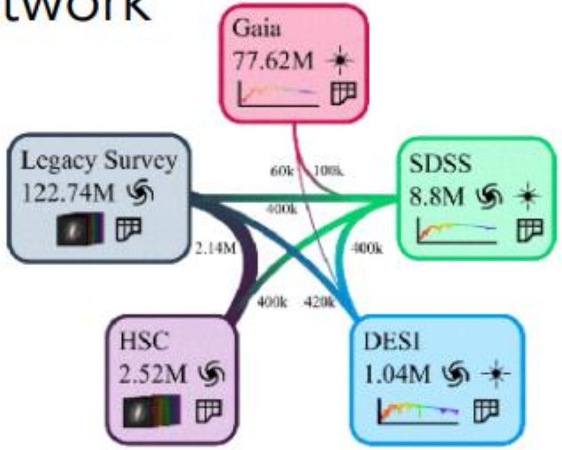
Cross modality : the dream

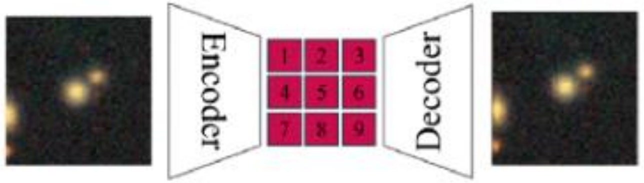


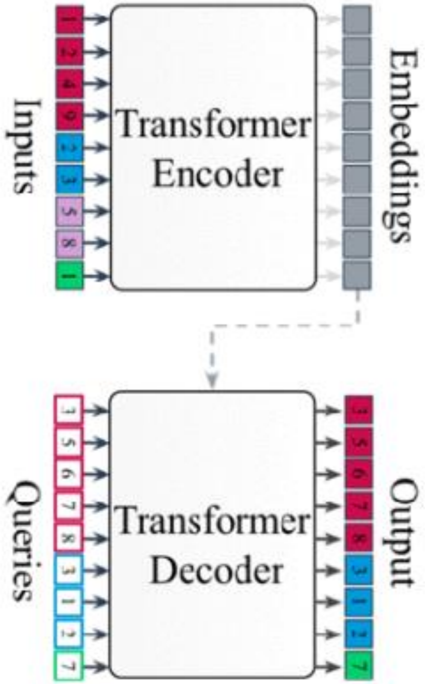


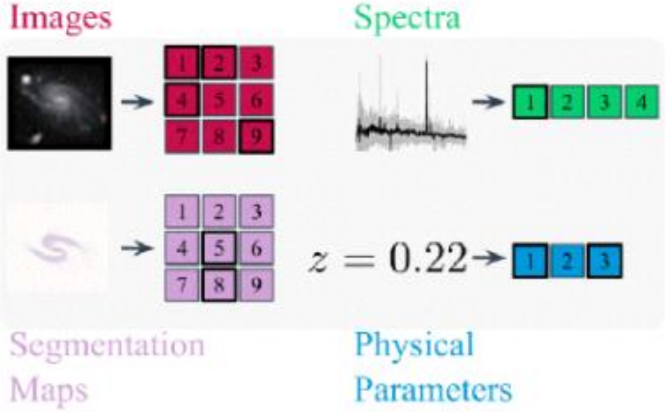
AION-1: AstronomIcal **Omnimodal** Network
(*polymathic team)





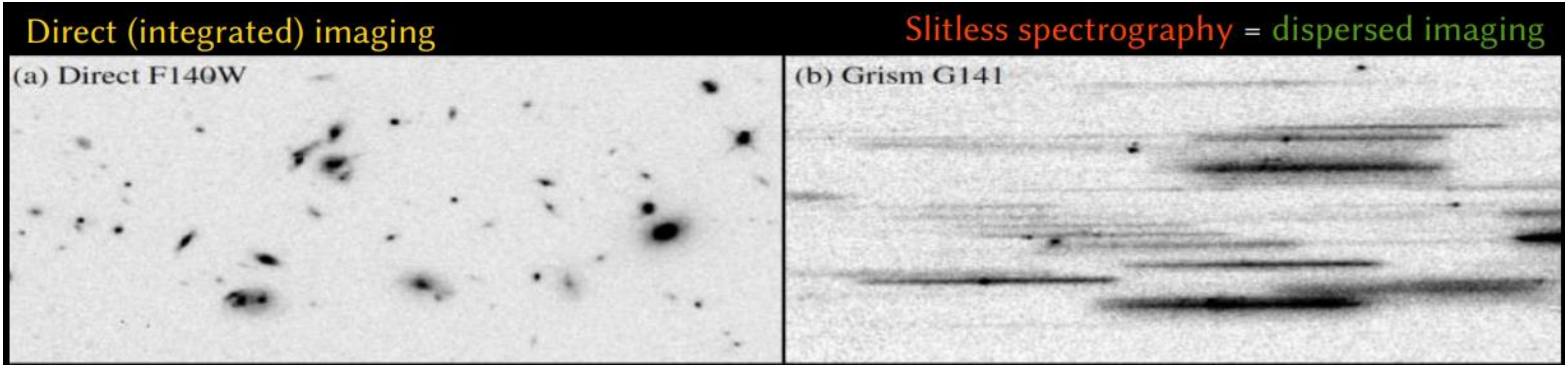






Spectrums example

SLITSLESS SIR euclid ➔ Low quality spectrums



Spectrum from DESI and crossmatch

Redshift scale

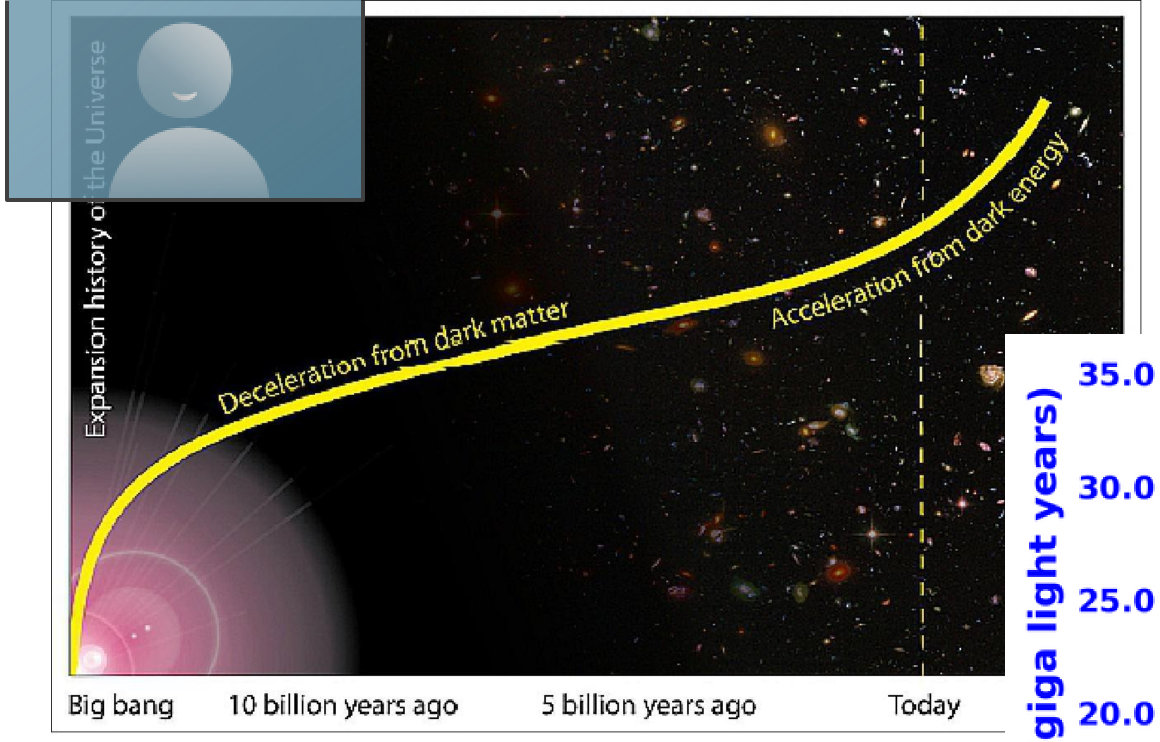
$$1 + z = \frac{a(t_0)}{a(t_{\text{émission}})}$$

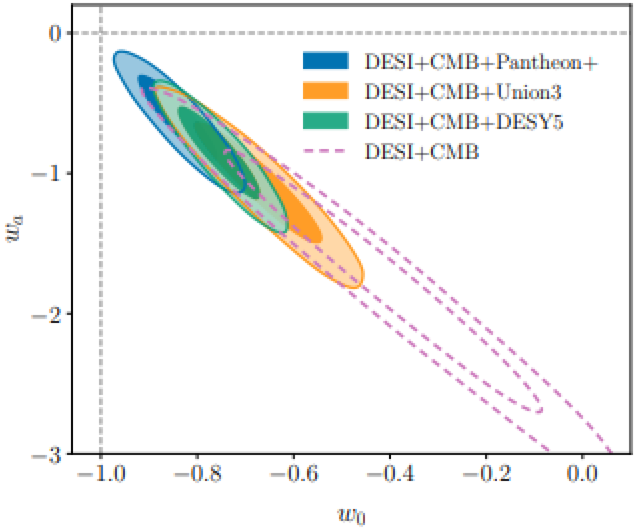
Z = 1000 = CMB = 380 000 yo

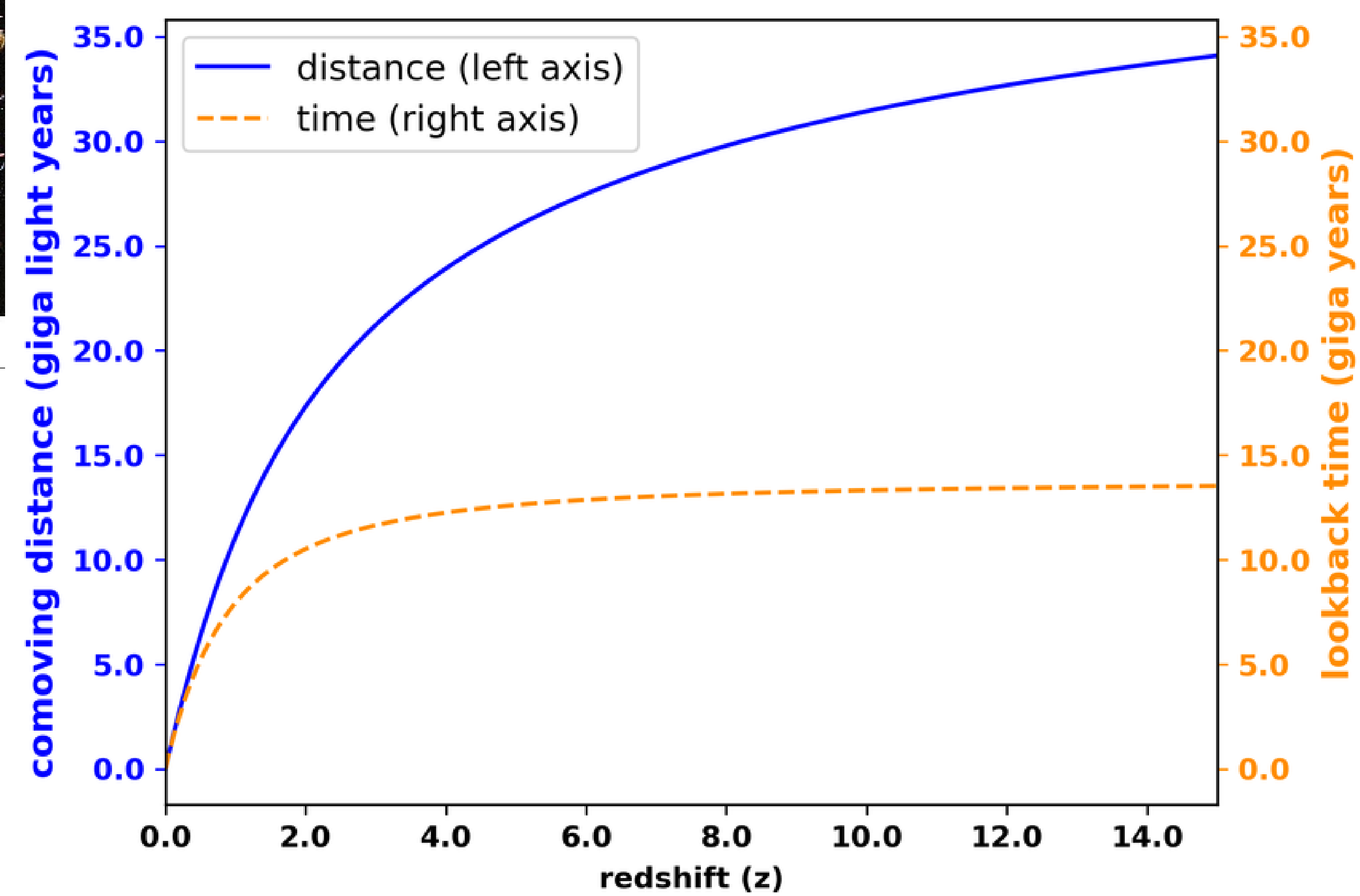
Z = 13 = 320 Million year after Big Bang

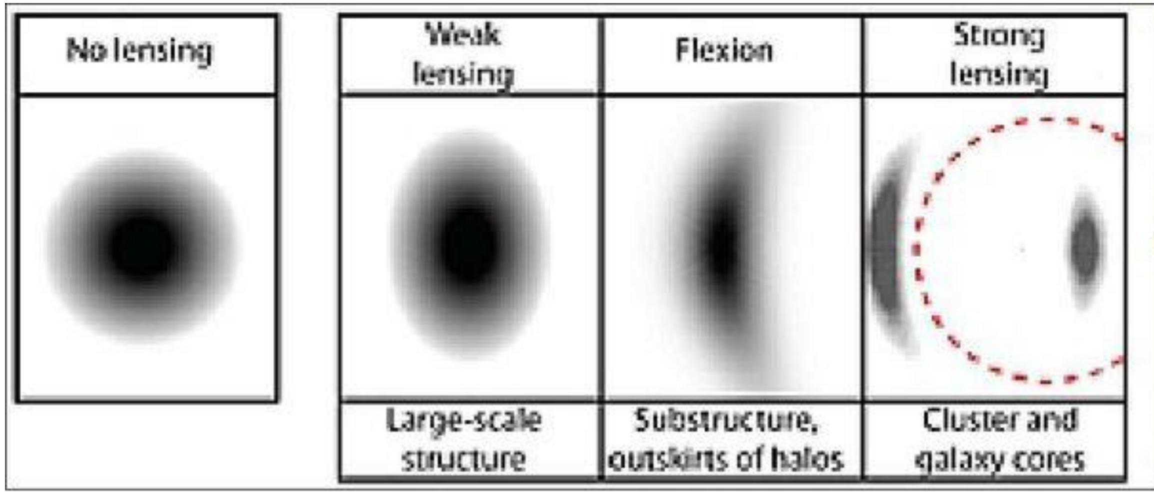
z ≈ **5 – 7,** universe was only 5% of its actual size

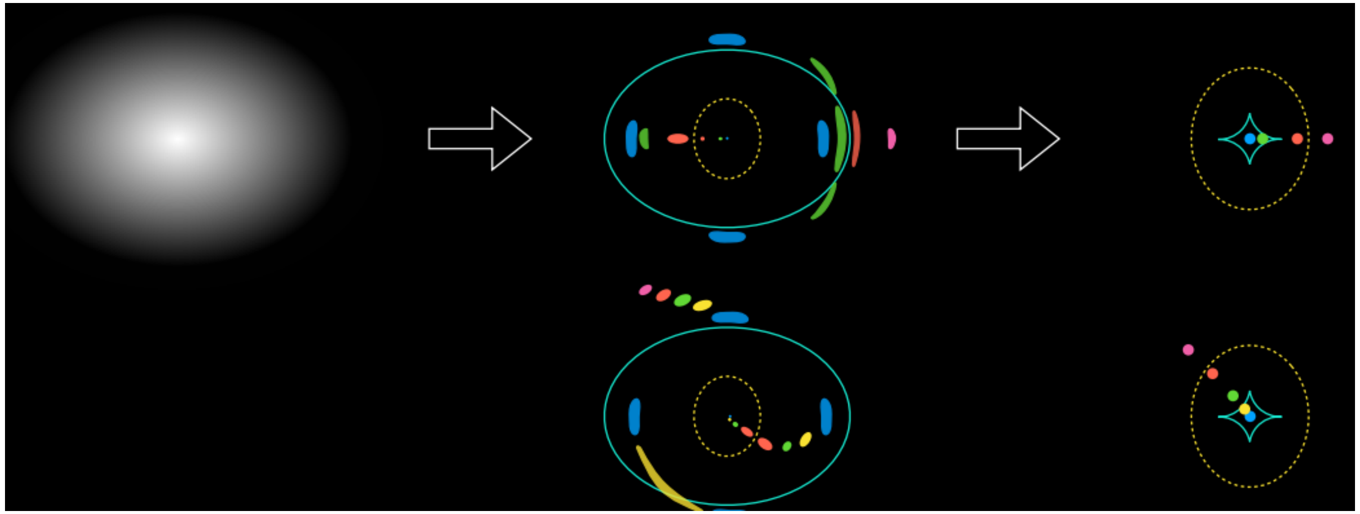
z ≈ **0,0036** light comming from 55million year

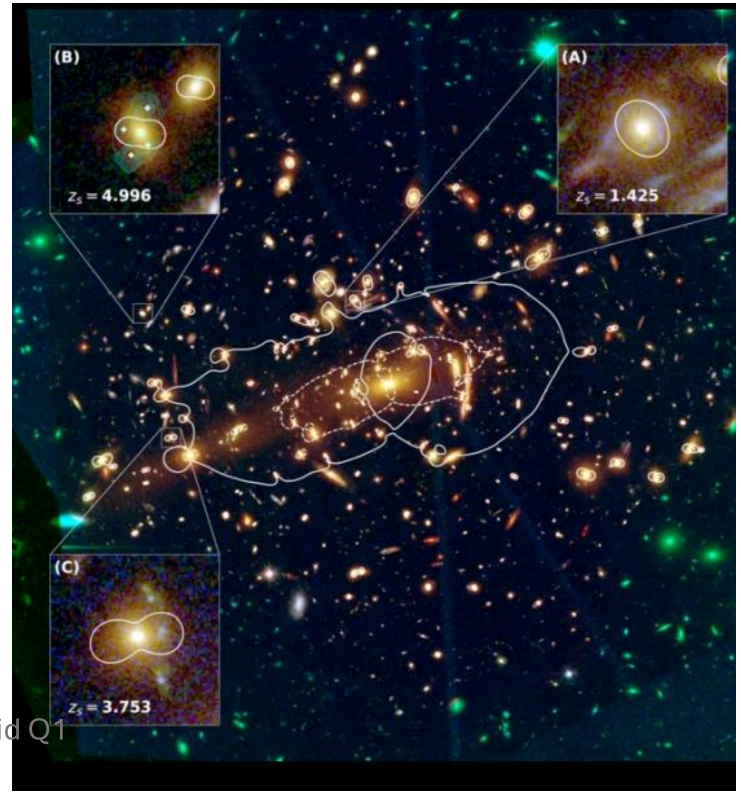

Expansion history o
the Universe
Deceleration from dark matter
Acceleration from dark energy
Big bang
10 billion years ago
5 billion years ago
Today

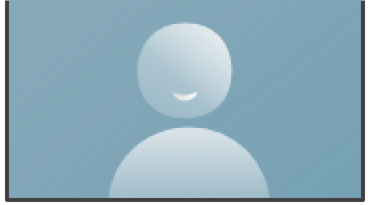

DESI+CMB+Pantheon+
DESI+CMB+Union3
DESI+CMB+DESY5
DESI+CMB
w_a
w_0
0
−1
−2
−3
−1.0
−0.8
−0.6
−0.4
−0.2
0.0


distance (left axis)
time (right axis)
comoving distance (giga light years)
lookback time (giga years)
redshift (z)
0.0
5.0
10.0
15.0
20.0
25.0
30.0
35.0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0


No lensing
Weak lensing
Flexion
Strong lensing
Large-scale structure
Substructure, outskirts of halos
Cluster and galaxy cores

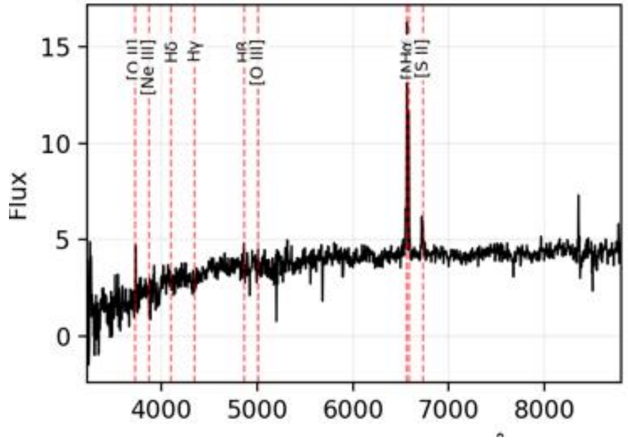


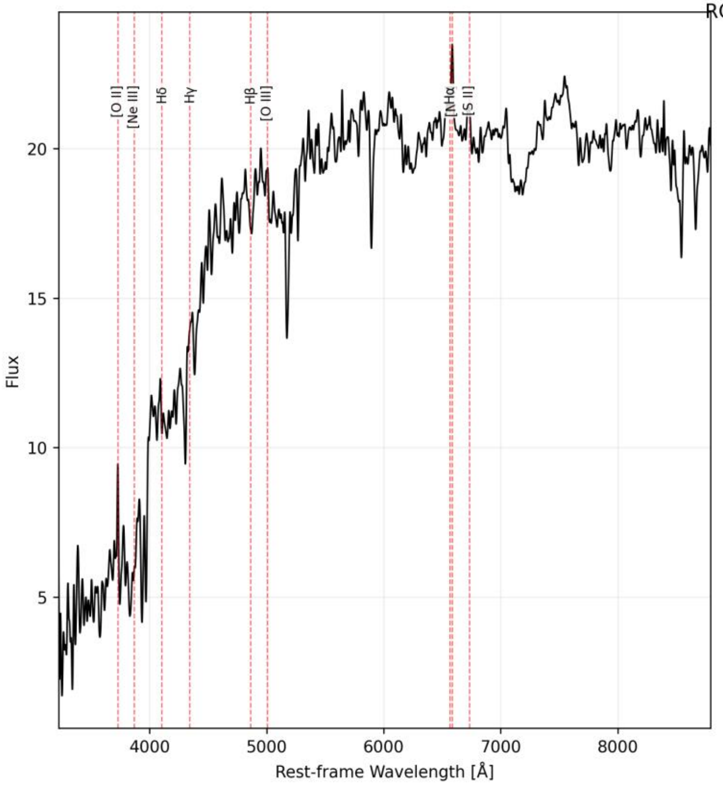

(B)
$z_s = 4.996$
(A)
$z_s = 1.425$
(C)
$z_s = 3.753$

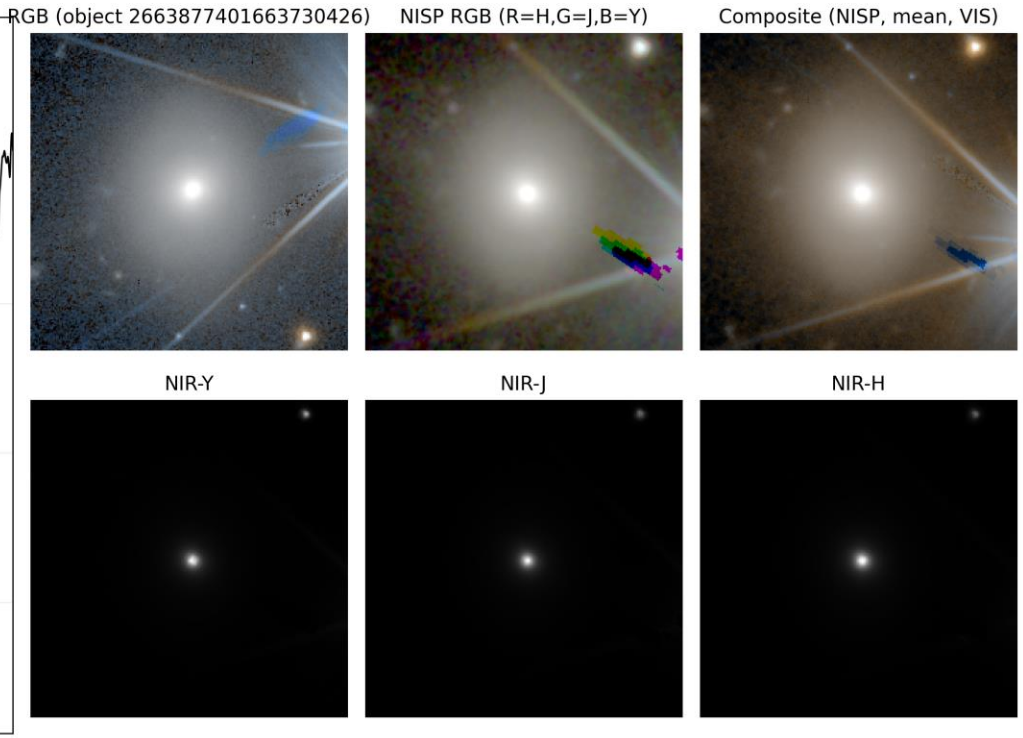


First results

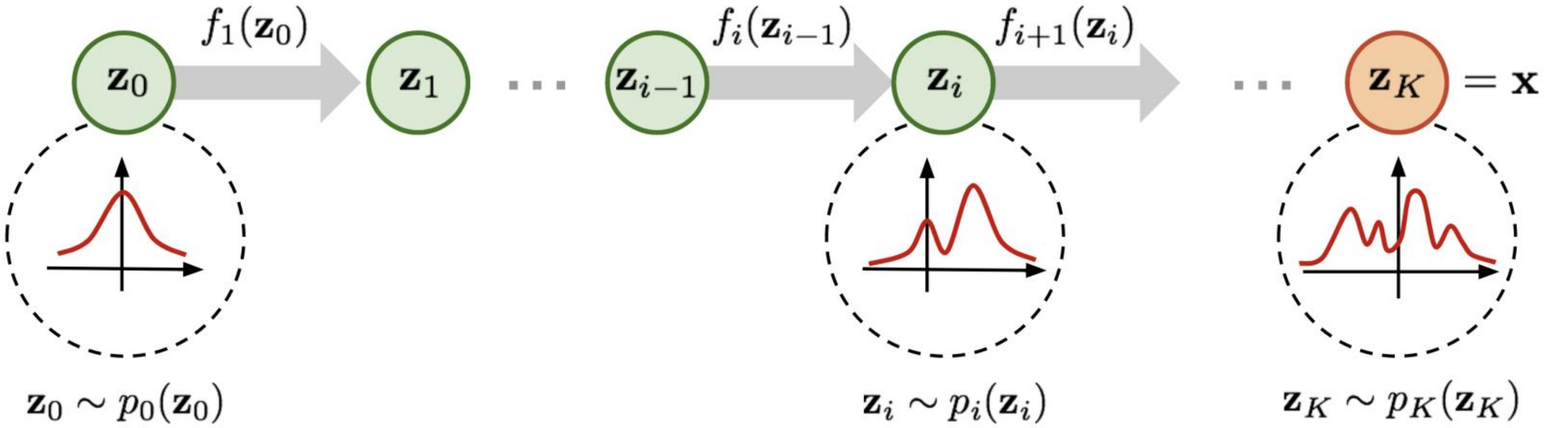
2667128432657198680 (z=0.116)







Normalizing Flows



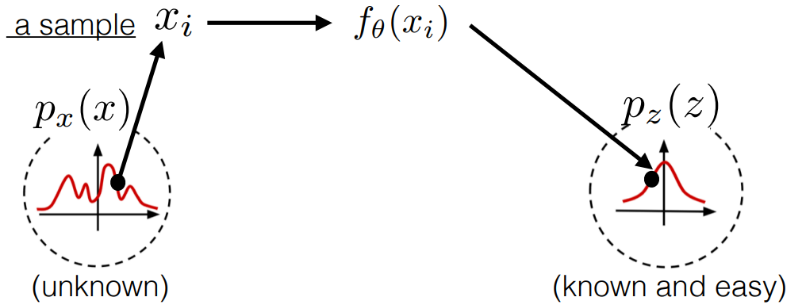
$$p_X(x) = p_Z(f(x))\ |\det J_f(x)|, \qquad J_f(x) = \frac{\partial f(x)}{\partial x^\top}$$

unknown pdf

tractable pdf
(typically Gaussian)

Invertible differentiable
function $z \sim \mathcal{N}(0,1)$

Determinant of the Jacobian
of f(x)



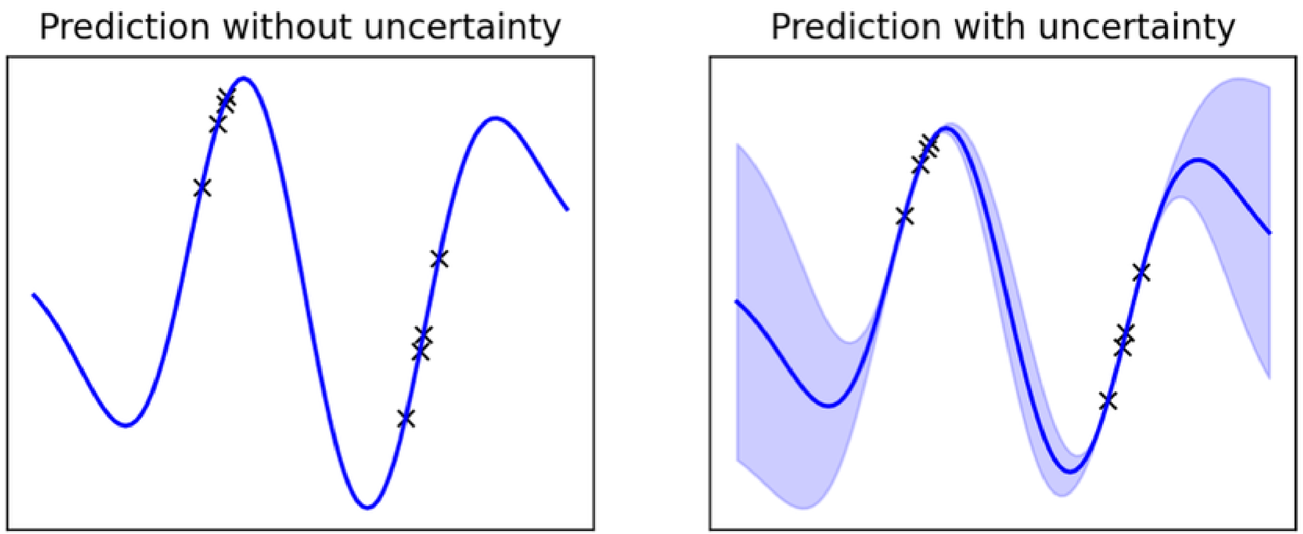
$z_i = f_i(z_{i-1})$ are **invertible** and **differentiable** transformations

$f = f_1 \circ f_2 \ldots \circ f_{k-1} \circ f_k$ is also invertible and differentiable

Bayesian Deep learning

$$posterior = \frac{likelihood * prior}{evidence}, \; or \; p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

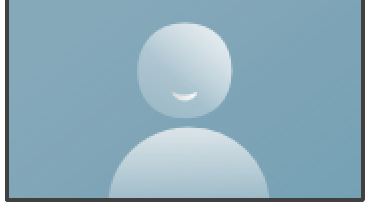
$$evidence = p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw$$

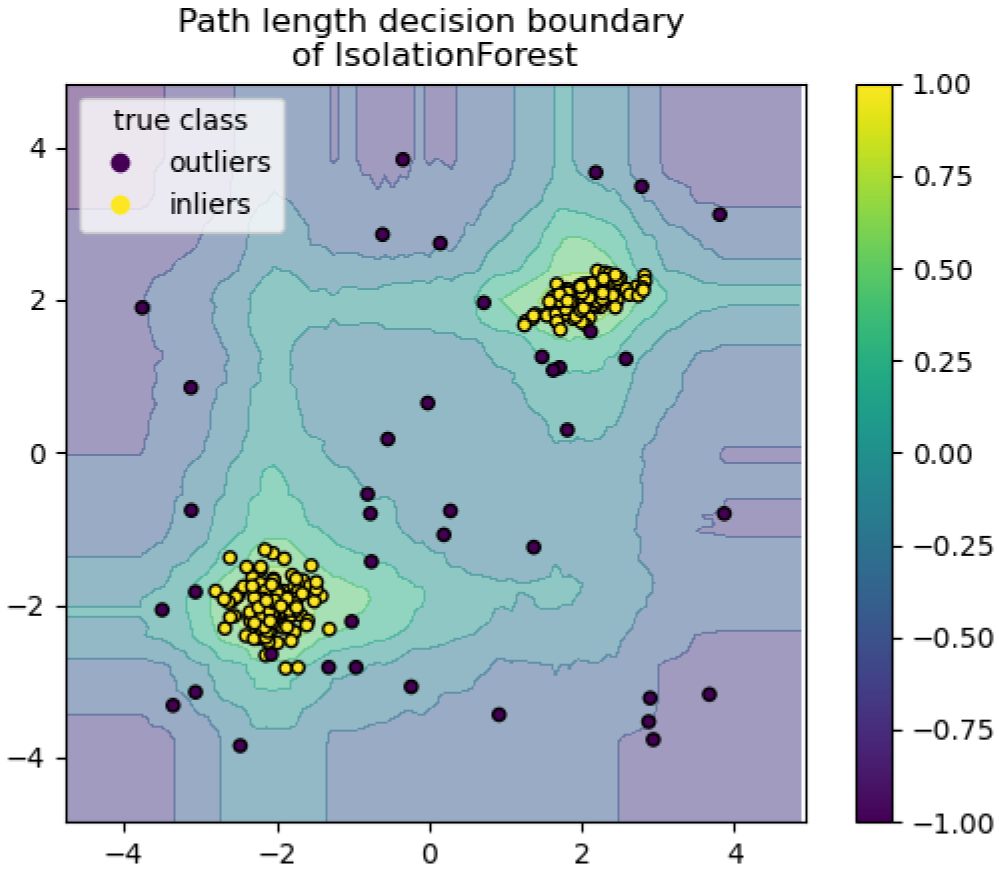


1 - Probabilize and sample weight Space

Or

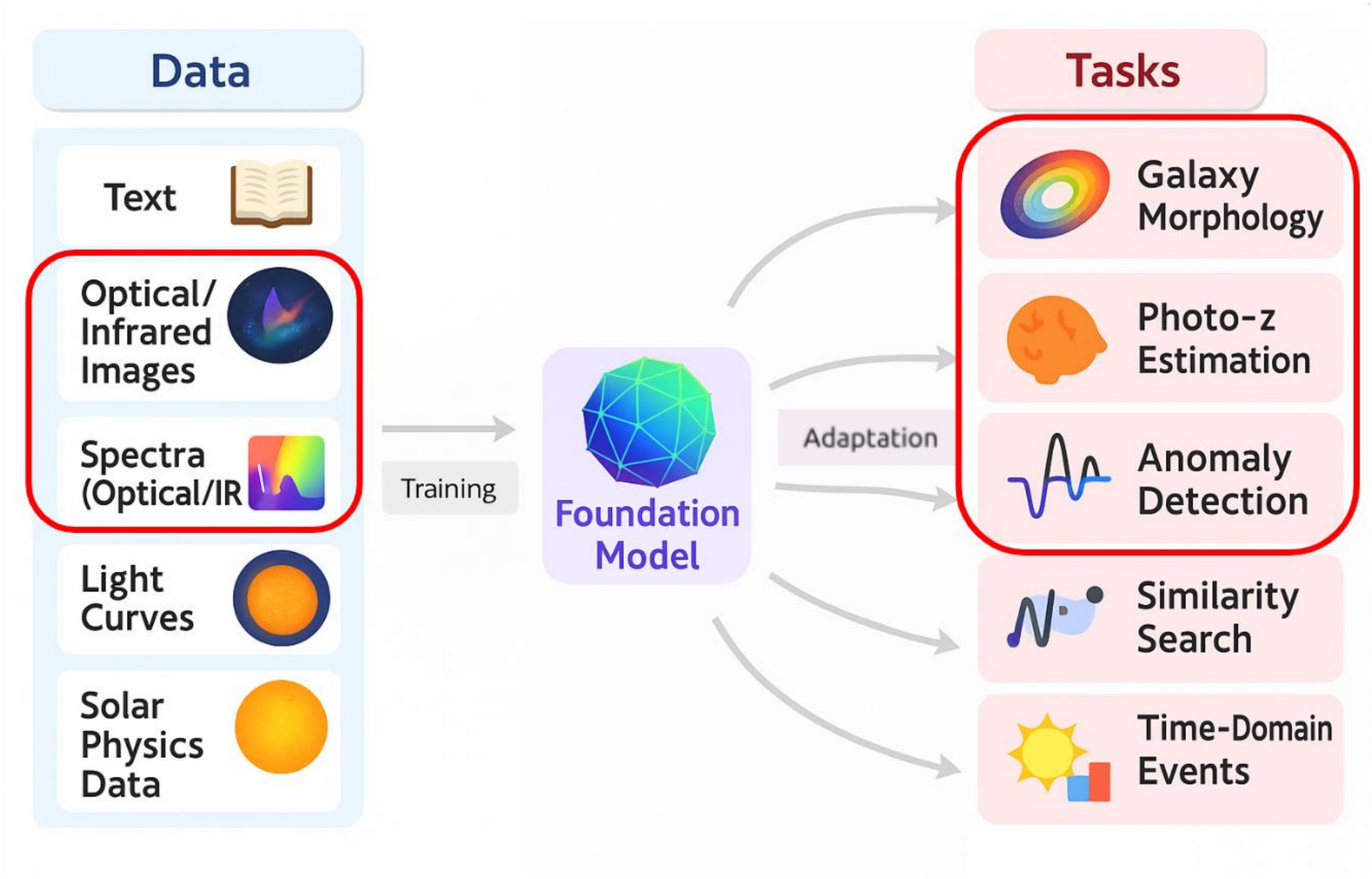
2 – Predict mu and sigma for each points and compute full gaussian likelihood

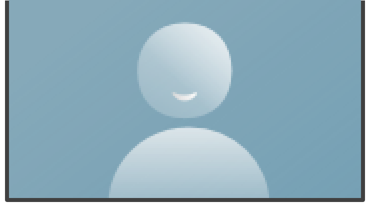


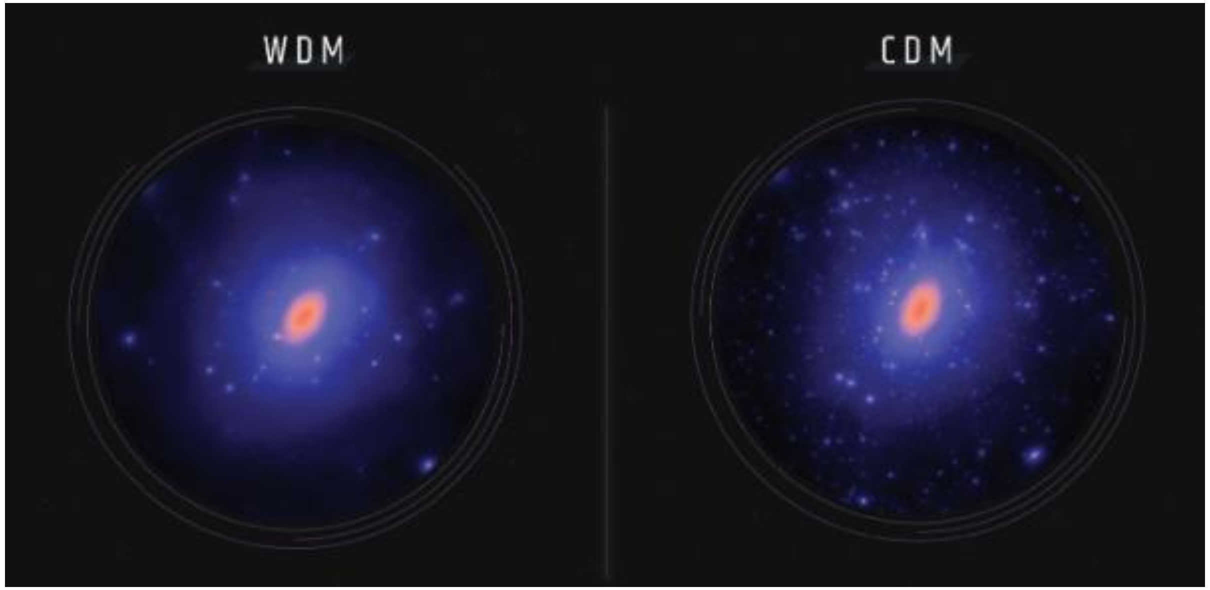


- It isolates observations by recursively splitting the feature space with random thresholds, creating random "isolation trees."
- Outliers tend to be isolated in fewer splits because they lie in sparse, atypical regions of the data space.
- Inliers require more splits to isolate, as they are embedded in dense, common regions.
- The average path length across many trees becomes an anomaly score, with shorter paths indicating stronger anomalies.

Astronomical Foundation Models







$$D_{\Delta t} = (1 + z_l)\,\frac{D_l D_s}{D_{ls}}$$

where

- z_l is the lens redshift,
- D_l the angular diameter distance to the lens,
- D_s the distance to the source, and
- D_{ls} the distance between lens and source.
- $D_{\Delta t}$ is the **time-delay distance**, which depends on the cosmological parameters H_0, Ω_m, and Ω_Λ.

Sersic index