Making astronomical data AI-ready across surveys

Lessons-learned from the Multimodal Universe Project

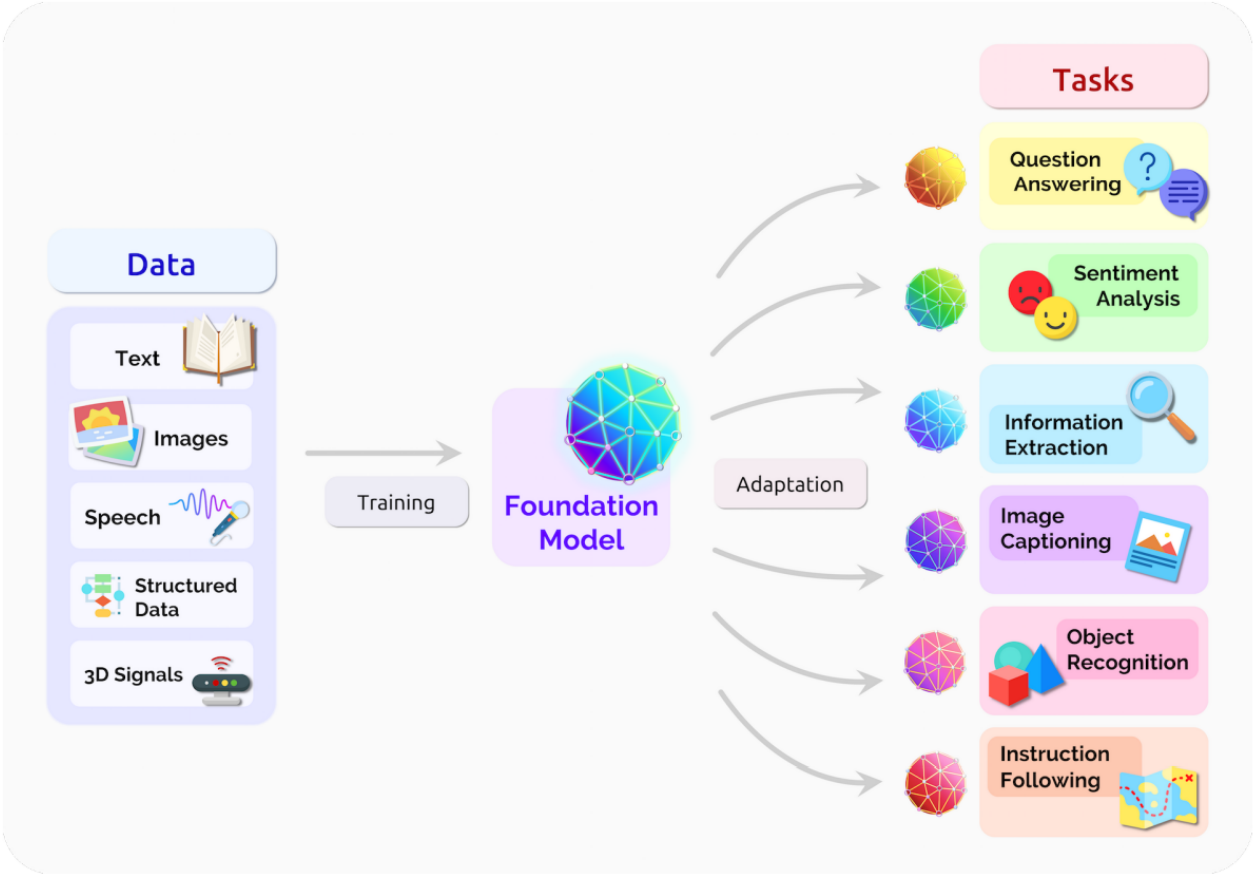
François Lanusse

National Center for Scientific Research (CNRS)

Polymathic AI

The Rise of The Foundation Model Paradigm

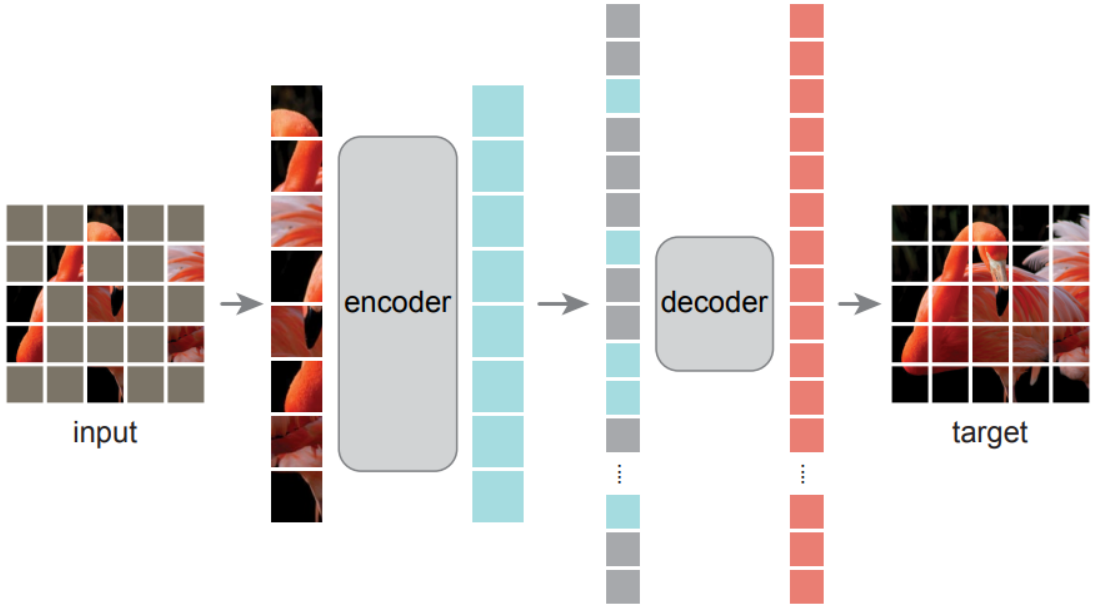
- **Foundation Model approach**
  - **Pretrain** models on pretext tasks, without supervision, on very large scale datasets.



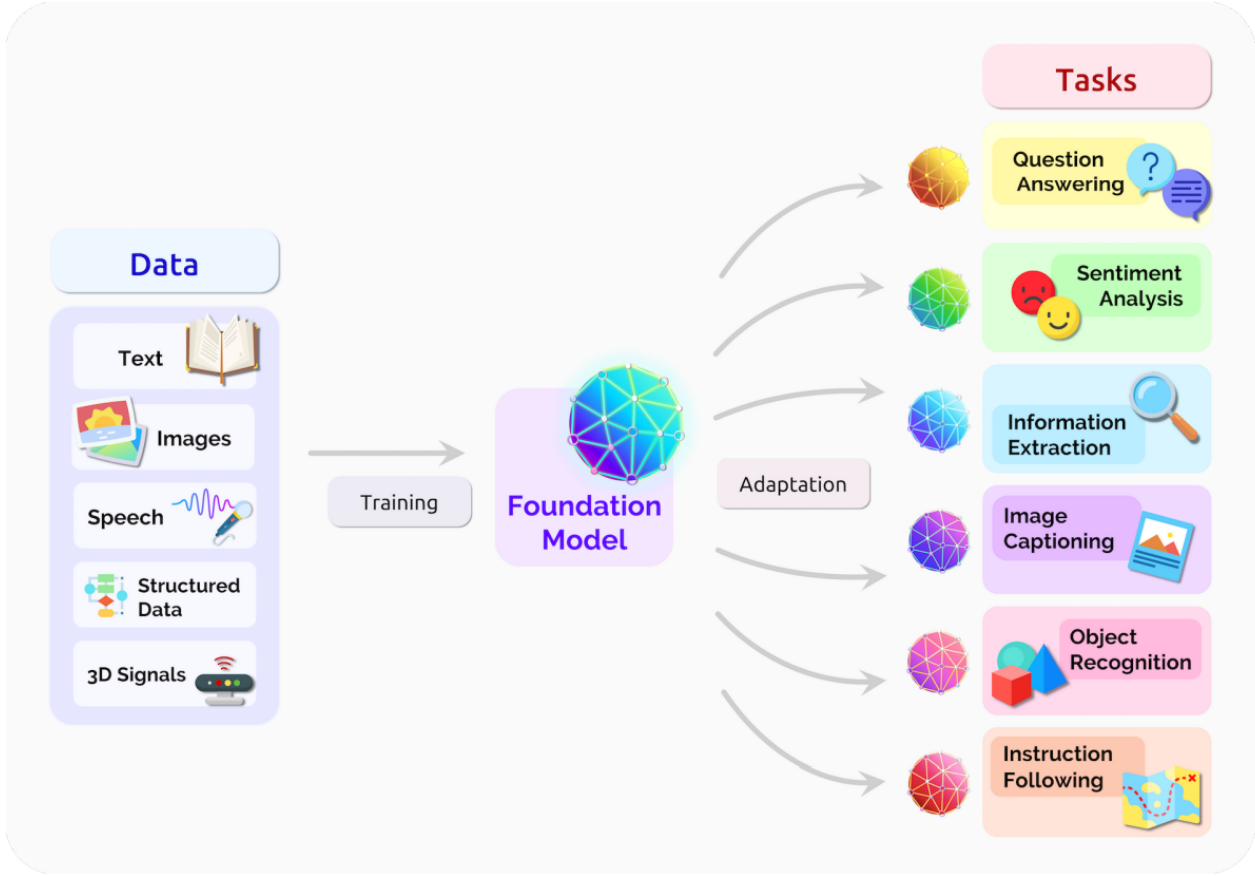
Bommasani et al. 2021

The Rise of The Foundation Model Paradigm

- **Foundation Model approach**
  - **Pretrain** models on pretext tasks, without supervision, on very large scale datasets.



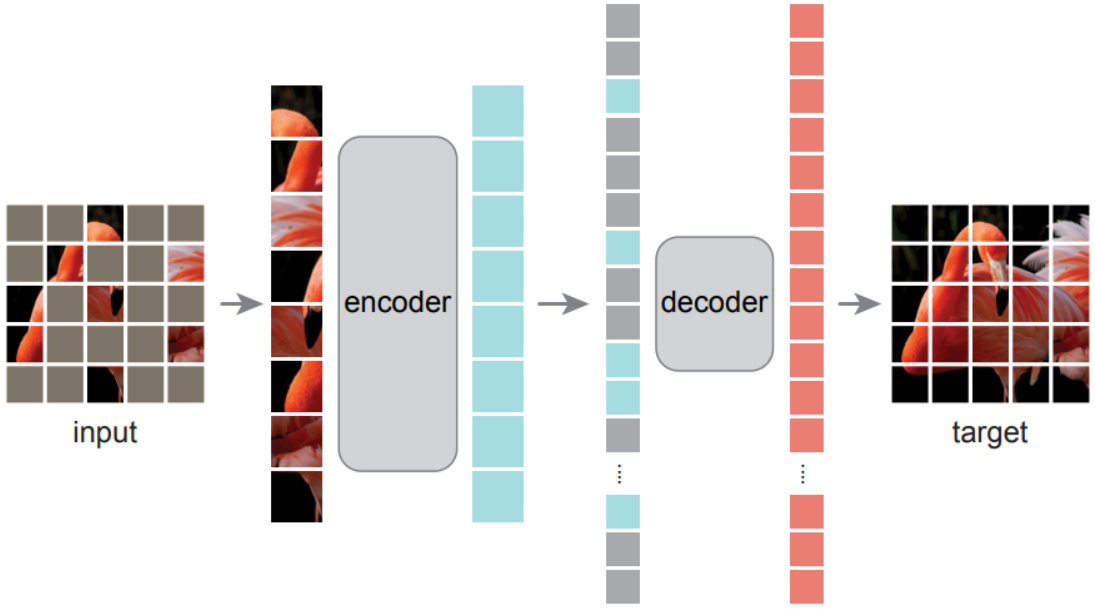
He et al. 2021



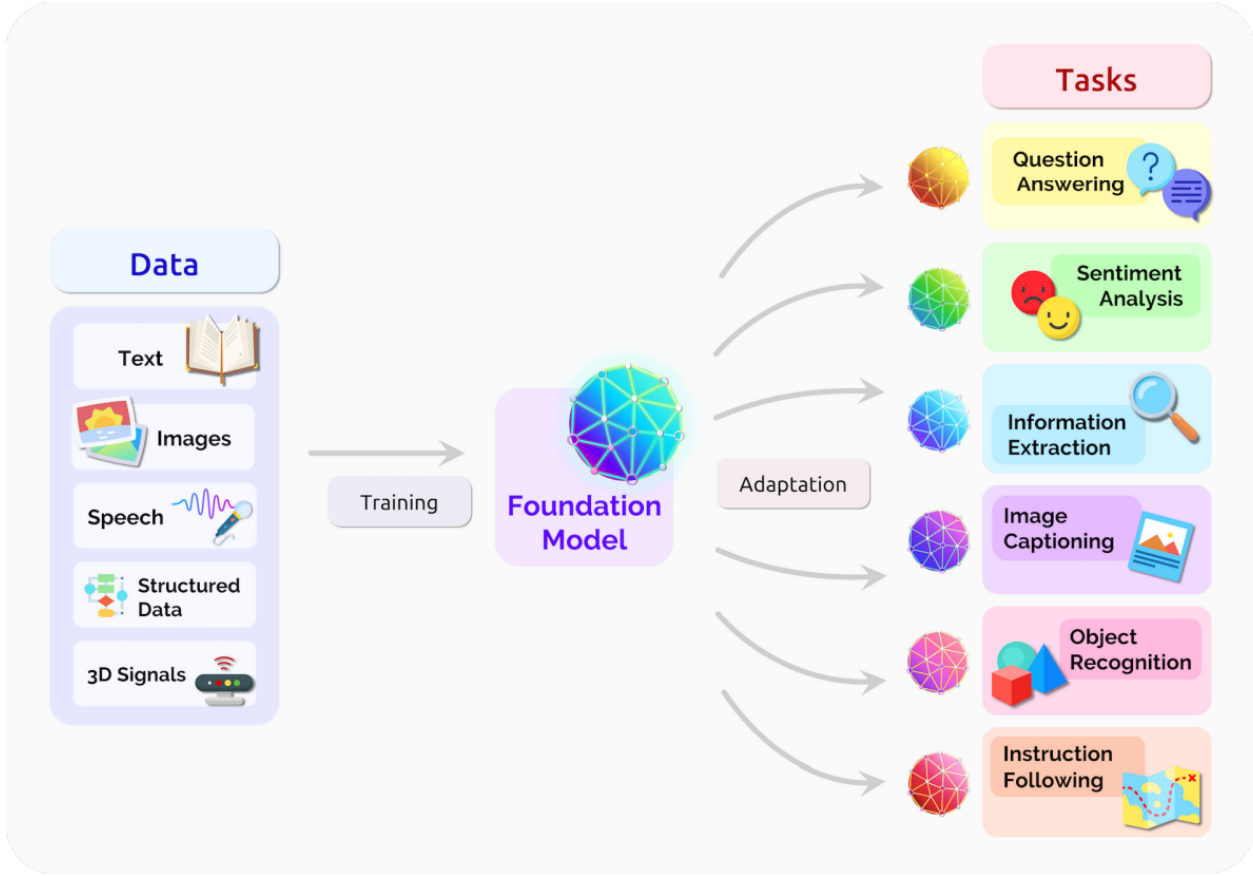
Bommasani et al. 2021

The Rise of The Foundation Model Paradigm

- **Foundation Model approach**
  - **Pretrain** models on pretext tasks, without supervision, on very large scale datasets.
  - **Adapt** pretrained models to downstream tasks.

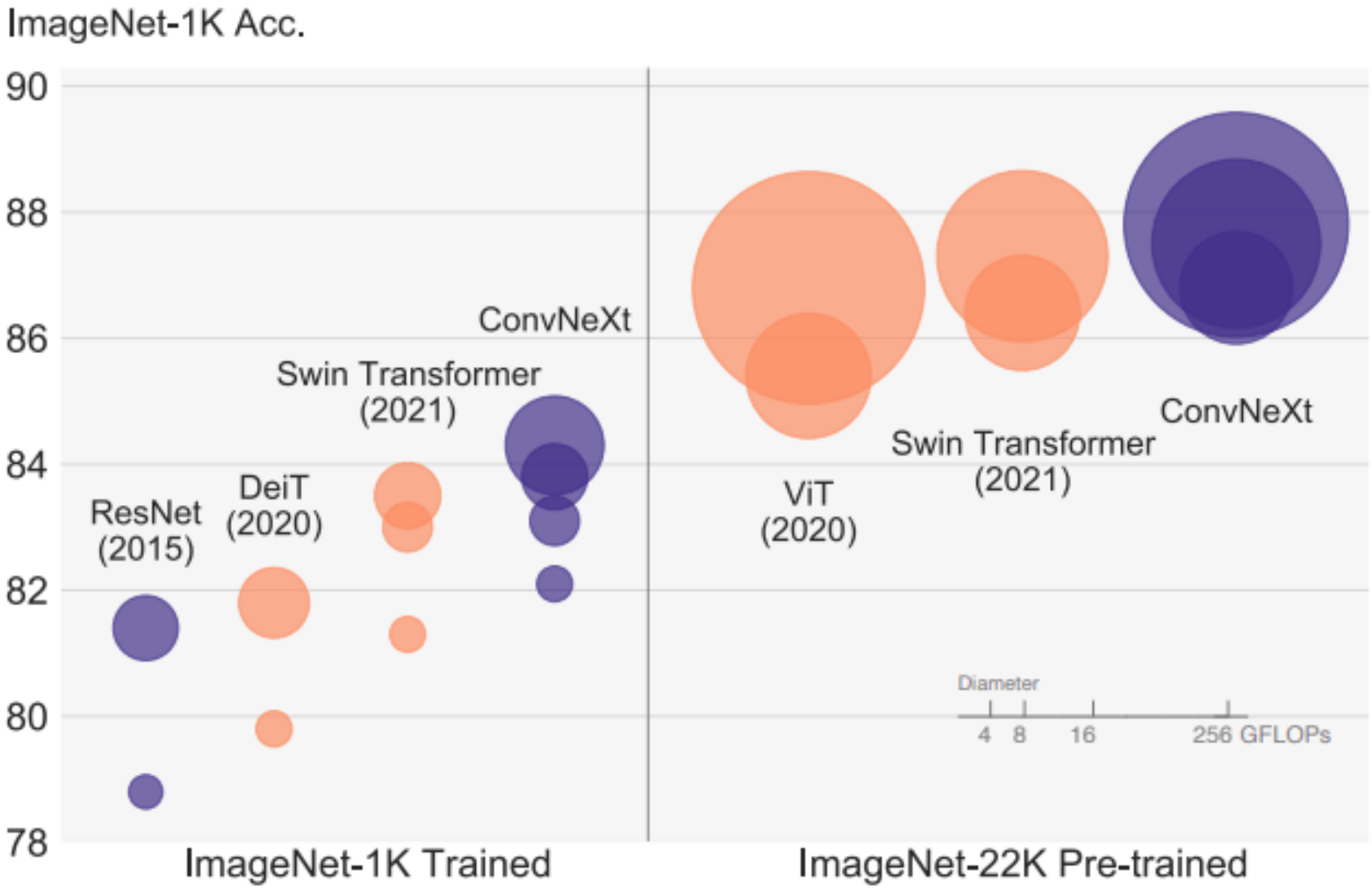


He et al. 2021



Bommasani et al. 2021

The Advantage of Scale of Data and Compute



Liu et al. 2022

What This New Paradigm Could Mean for Us

Polymathic

- **Never have to retrain my own neural networks** from scratch
  - Existing pre-trained models would already be near optimal, no matter the task at hand

What This New Paradigm Could Mean for Us

- **Never have to retrain my own neural networks** from scratch
  - Existing pre-trained models would already be near optimal, no matter the task at hand
- Practical large scale Deep Learning even in **very few example regime**
  - Searching for very rare objects in large surveys like Euclid or LSST becomes possible

What This New Paradigm Could Mean for Us

- **Never have to retrain my own neural networks** from scratch
  - Existing pre-trained models would already be near optimal, no matter the task at hand
- Practical large scale Deep Learning even in **very few example regime**
  - Searching for very rare objects in large surveys like Euclid or LSST becomes possible
- If the information is embedded in a space where it becomes linearly accessible, **very simple analysis tools are enough** for downstream analysis
  - In the future, survey pipelines may add vector embedding of detected objects into catalogs, these would be enough for most tasks, without the need to go back to pixels

AstroCLIP

Cross-Modal Pre-Training for Astronomical Foundation Models

astro-ph.IM arXiv:2310.03024





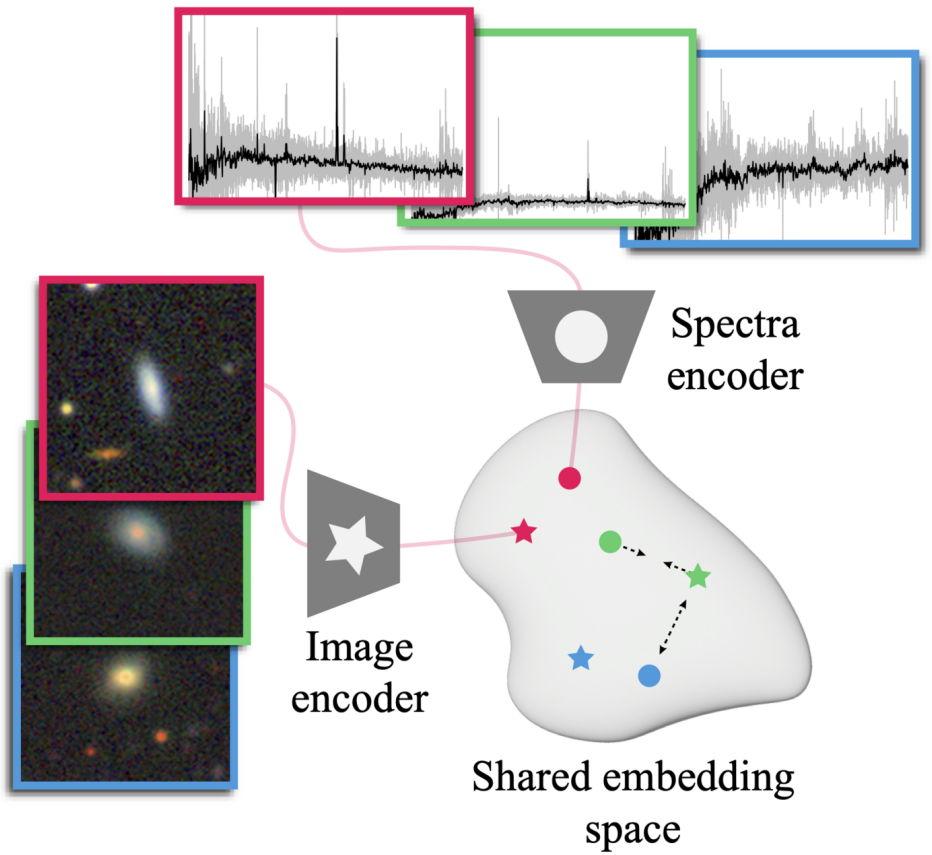
Project led by Liam Parker, Francois Lanusse, Leopoldo Sarra, Siavash Golkar, Miles Cranmer

Polymathic

Accepted contribution at the NeurIPS 2023 AI4Science Workshop

Published in the Monthly Notices of Royal Astronomical Society

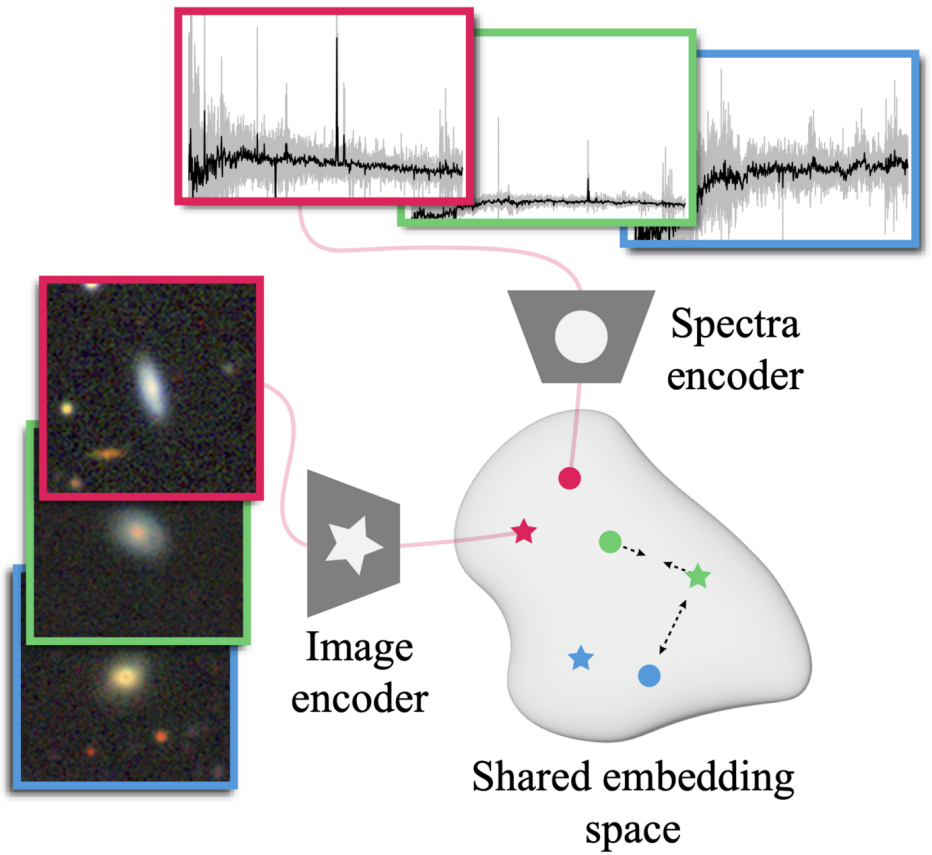
The AstroCLIP approach



$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

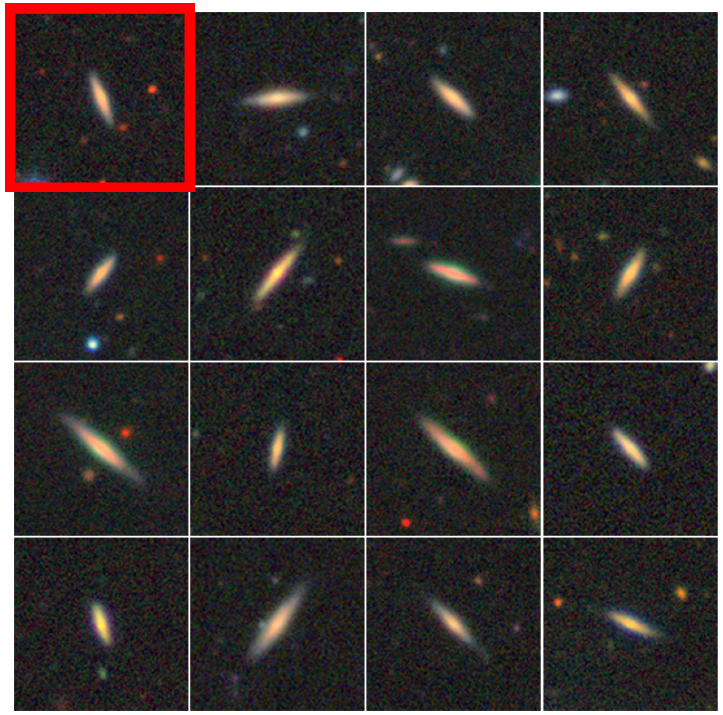
- We use **spectra** and multi-band **images** as our two different views for the same underlying object.
- DESI Legacy Surveys *(g,r,z)* images, and DESI EDR galaxy spectra.

The AstroCLIP approach



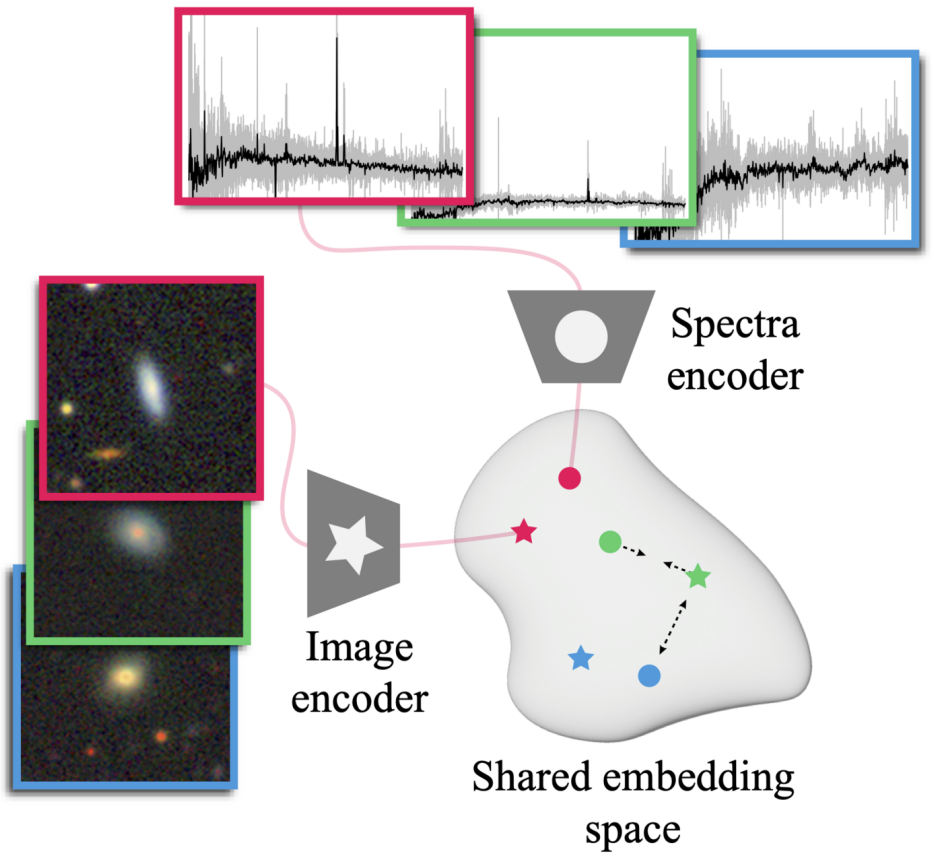
$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

- We use **spectra** and multi-band **images** as our two different views for the same underlying object.
- DESI Legacy Surveys *(g,r,z)* images, and DESI EDR galaxy spectra.



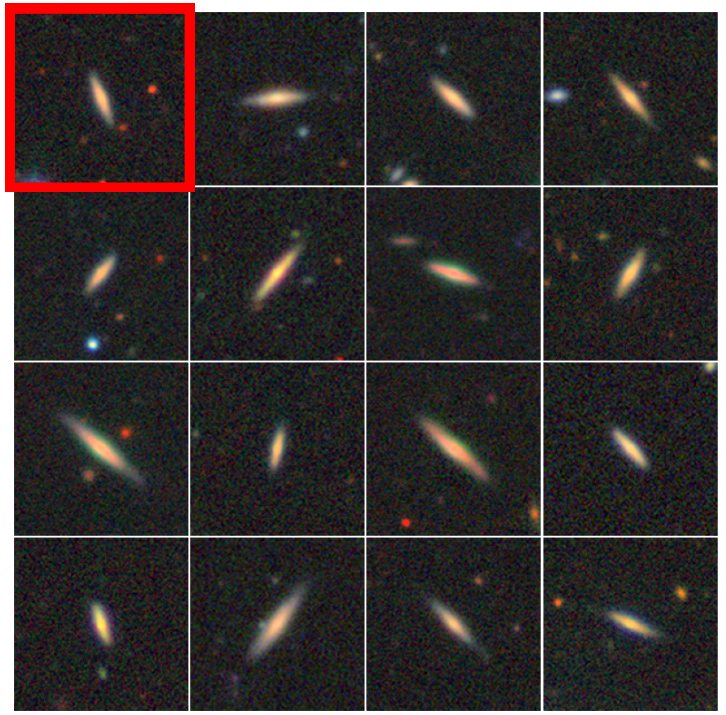
Cosine similarity search

The AstroCLIP approach

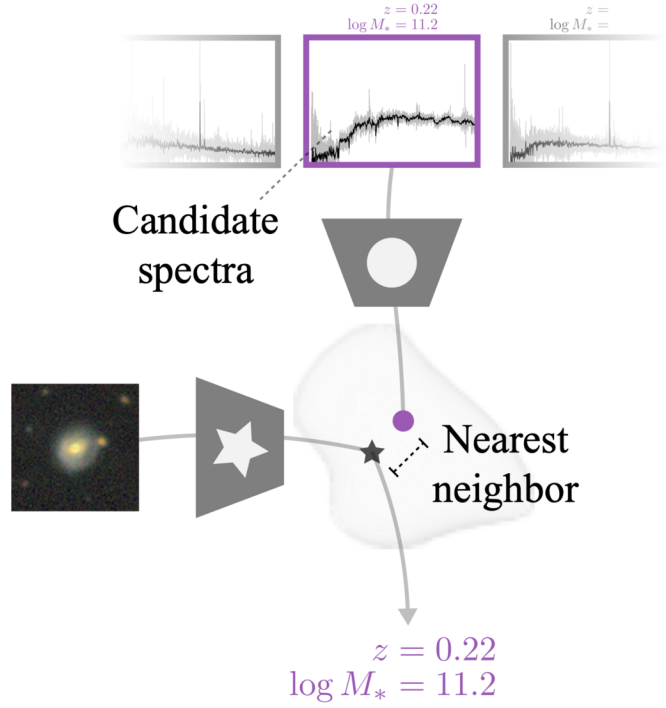


$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

- We use **spectra** and multi-band **images** as our two different views for the same underlying object.
- DESI Legacy Surveys *(g,r,z)* images, and DESI EDR galaxy spectra.



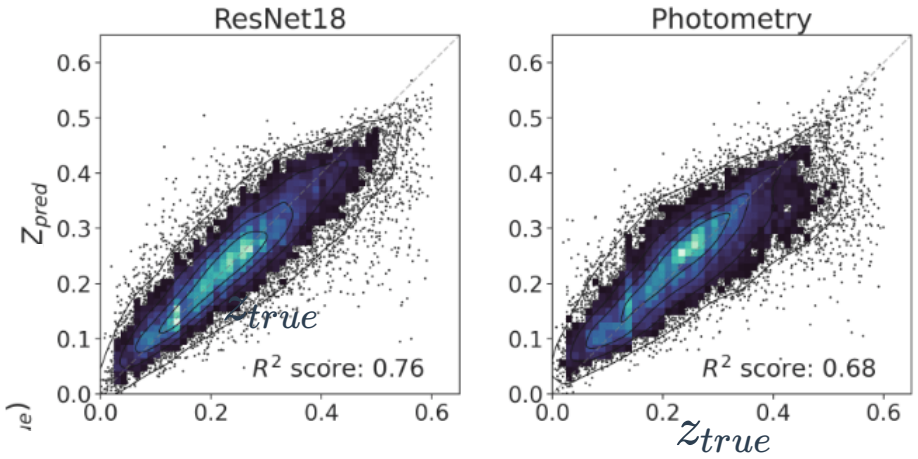
Cosine similarity search



Zero-shot prediction

Evaluation of the model: Parameter Inference

- Redshift Estimation From Images

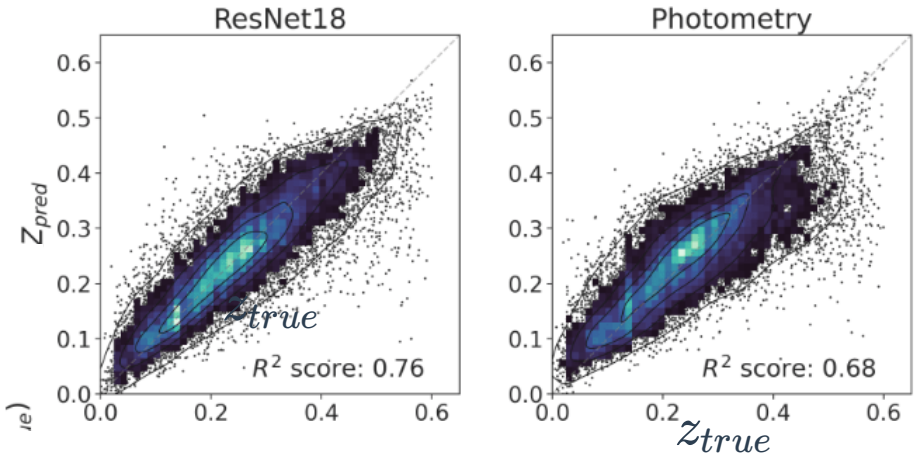


Supervised baseline

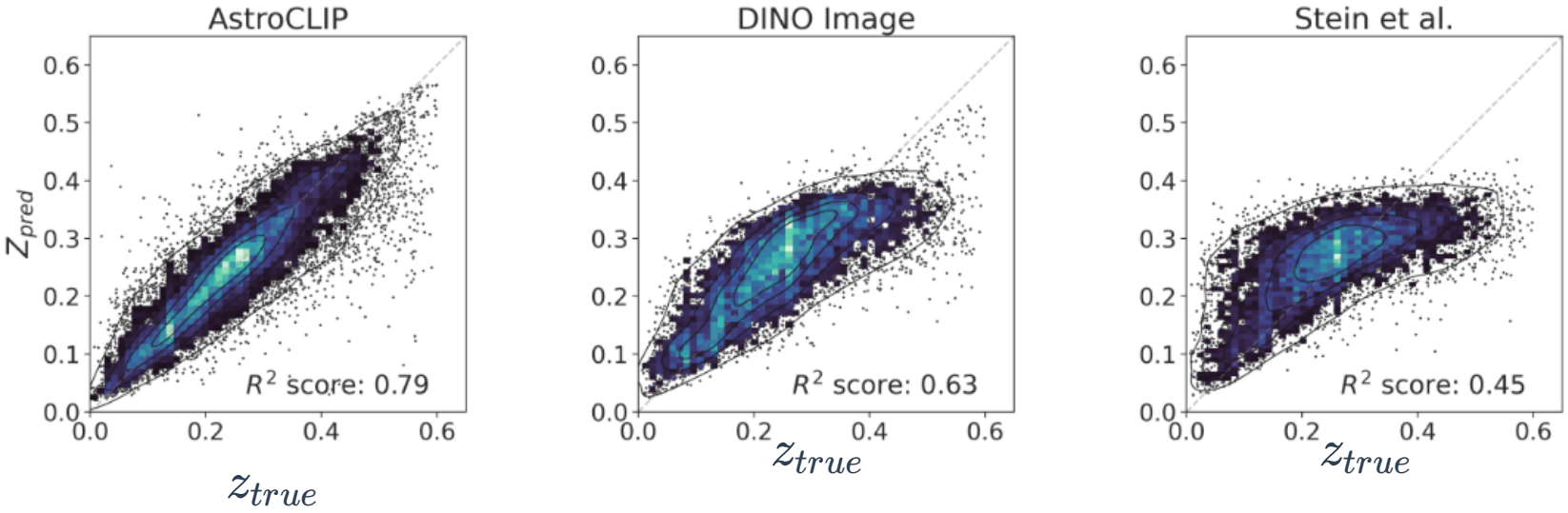
Evaluation of the model: Parameter Inference

Polymathic

- Redshift Estimation From Images



Supervised baseline



- Zero-shot prediction
  - k-NN regression

AION-1
Omnimodal Foundation Model for Astronomical Surveys

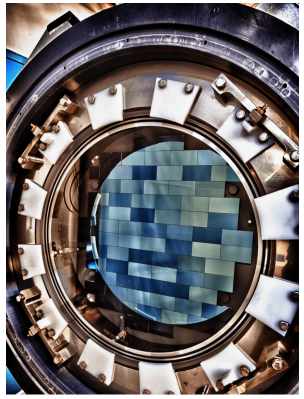
NEURAL INFORMATION PROCESSING SYSTEMS

Project led by:

Francois Lanusse, **Liam Parker**, **Jeff Shen**, **Tom Hehir**, **Ollie Liu**, Lucas Meyer, Sebastian Wagner-Carena, Helen Qu, Micah Bowles

Accepted at NeurIPS 2025, spotlight presentation at NeurIPS 2025 AI4Science Workshop

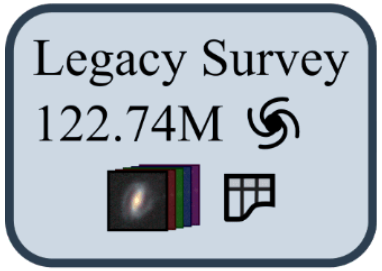
The AION-1 Data Pile

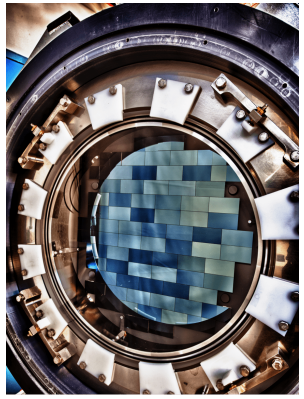

(Blanco Telescope and Dark Energy Camera.

Credit: Reidar Hahn/Fermi National Accelerator Laboratory)

Cuts: extended, full color *griz*, $z < 21$



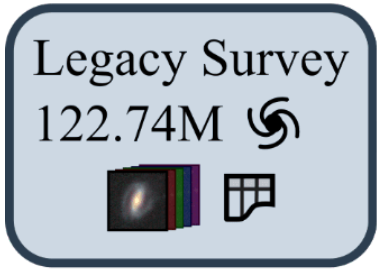
The AION-1 Data Pile


(Blanco Telescope and Dark Energy Camera.

Credit: Reidar Hahn/Fermi National Accelerator Laboratory)

Cuts: extended, full color *griz*, $z < 21$

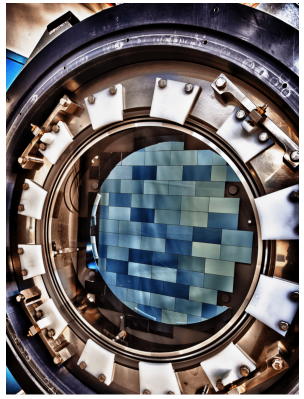


Cuts: extended, full color *grizy*, $z < 21$



(Subaru Telescope and Hyper Suprime Cam. Credit: NAOJ)

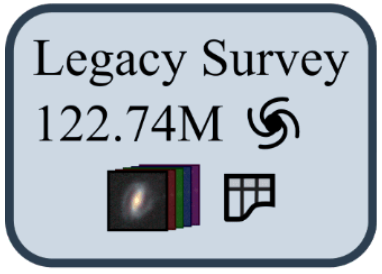
The AION-1 Data Pile


(Blanco Telescope and Dark Energy Camera.

Credit: Reidar Hahn/Fermi National Accelerator Laboratory)

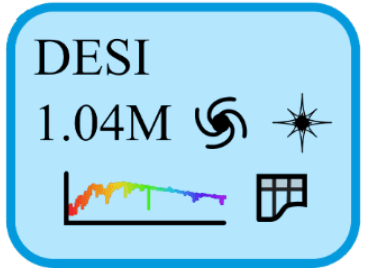
Cuts: extended, full color *griz*, $z < 21$

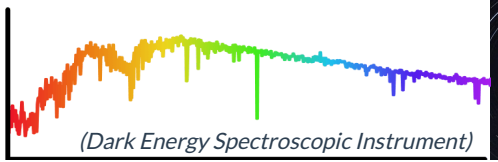


Cuts: extended, full color *grizy*, $z < 21$

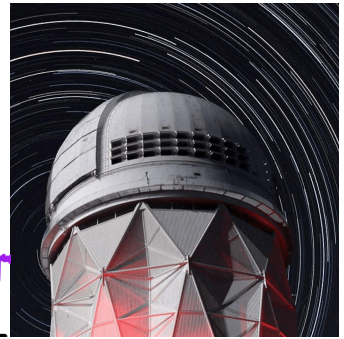
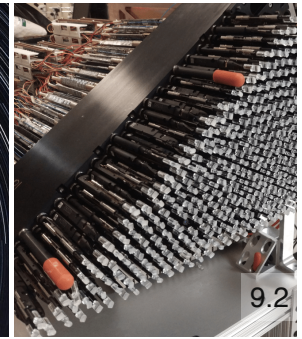


(Subaru Telescope and Hyper Suprime Cam. Credit: NAOJ)

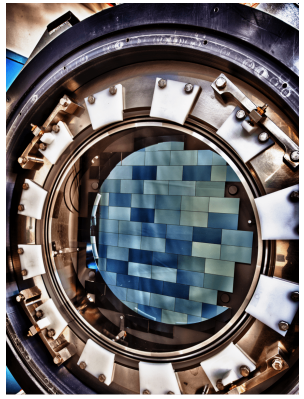




(Dark Energy Spectroscopic Instrument)

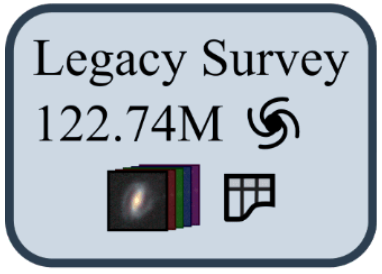
The AION-1 Data Pile

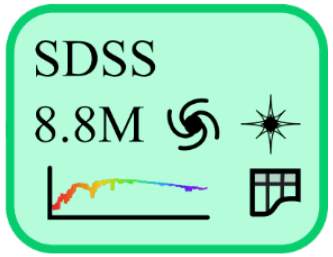



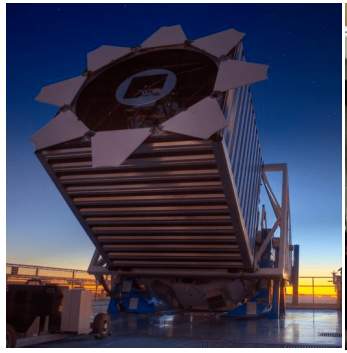

(Blanco Telescope and Dark Energy Camera.

Credit: Reidar Hahn/Fermi National Accelerator Laboratory)

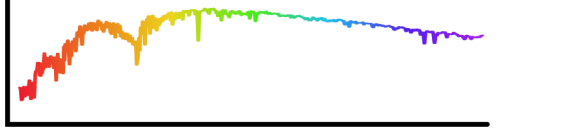
Cuts: extended, full color *griz*, $z < 21$







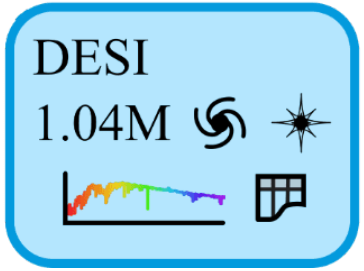

(Sloan Digital Sky Survey. Credit: SDSS)

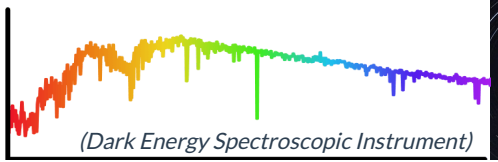


Cuts: extended, full color *grizy*, $z < 21$

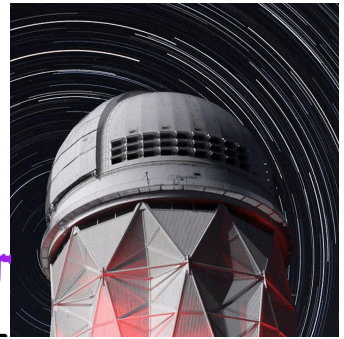
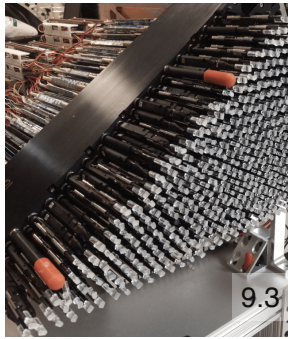


(Subaru Telescope and Hyper Suprime Cam. Credit: NAOJ)

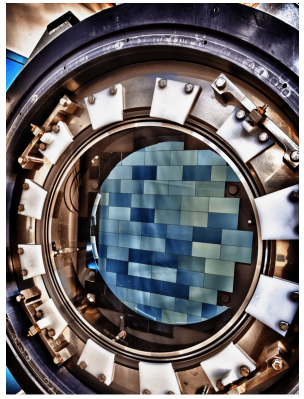




(Dark Energy Spectroscopic Instrument)

The AION-1 Data Pile


(Blanco Telescope and Dark Energy Camera.

Credit: Reidar Hahn/Fermi National Accelerator Laboratory)

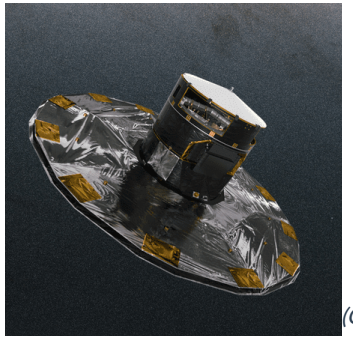
Cuts: extended, full color *griz*, $z < 21$



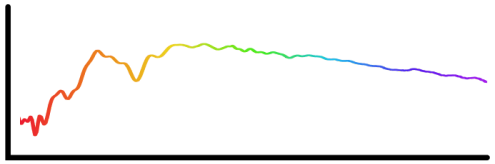
(Subaru Telescope and Hyper Suprime Cam. Credit: NAOJ)

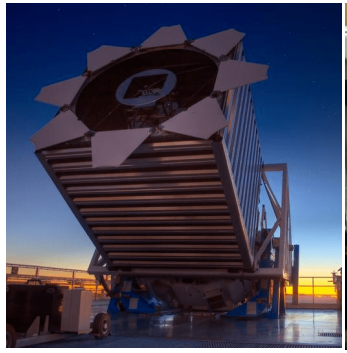
Cuts: extended, full color *grizy*, $z < 21$

Cuts: parallax / parallax_error > 10

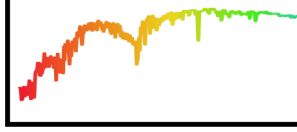


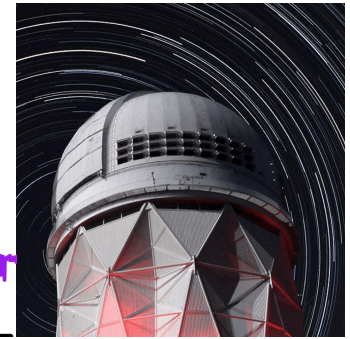
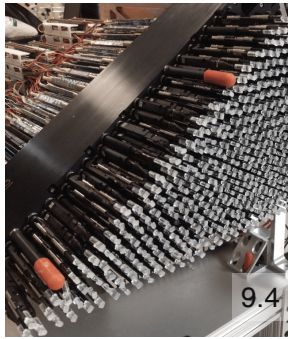
(Gaia Satellite. Credit: ESA/ATG*)*



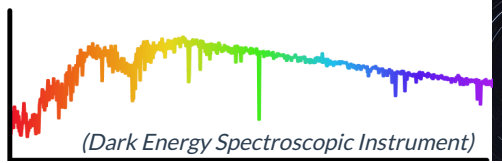



(Sloan Digital Sky Survey. Credit: SDSS)

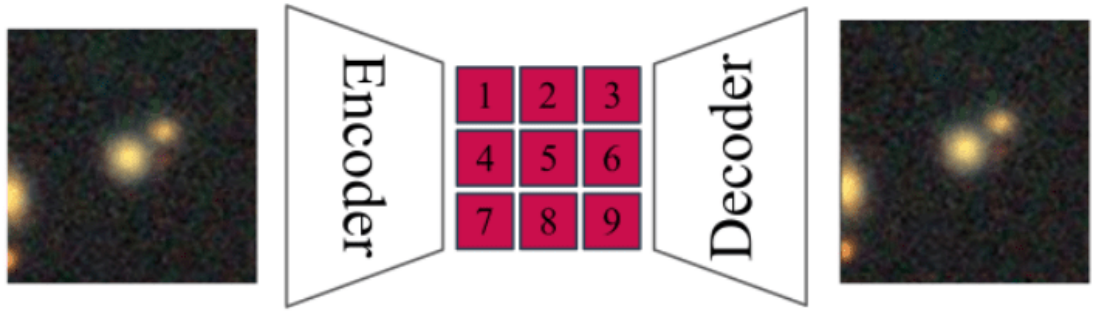


(Dark Energy Spectroscopic Instrument)

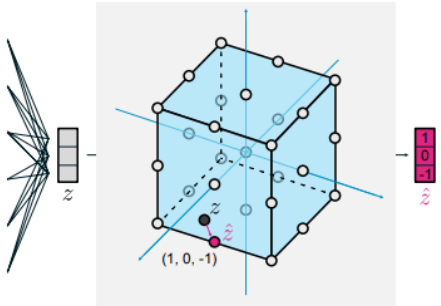


Standardizing all modalities through tokenization



$$\mathcal{L} = \| \Sigma^{-\frac{1}{2}} \left(x - d_\theta(\lfloor e_\theta(x) \rfloor_{\text{FSQ}}) \right) \|_2^2$$

FSQ

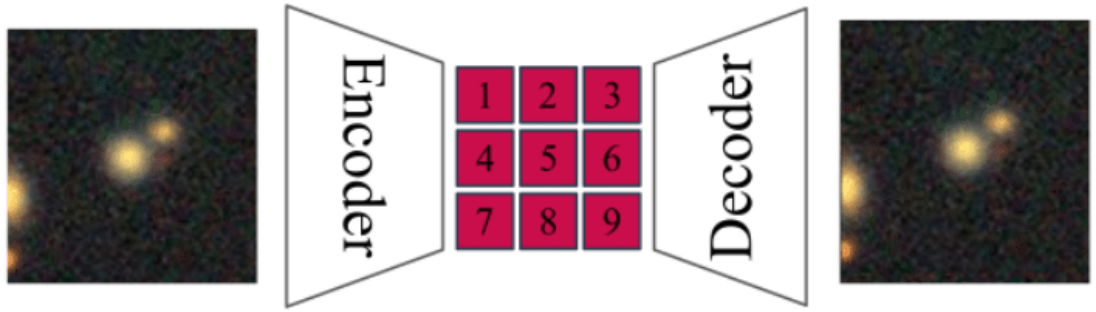


(Mentzer et al. 2023)

- For each modality category (e.g. image, spectrum) we build dedicated tokenizers
  => Convert from any data to **discrete tokens**

	VQ	FSQ
Quantization	argmin_c || z-c ||	round(f(z))
Gradients	Straight Through Estimation (STE)	STE
Auxiliary Losses	Commitment, codebook, entropy loss, ...	N/A
Tricks	EMA on codebook, codebook splitting, projections, ...	N/A
Parameters	Codebook	N/A

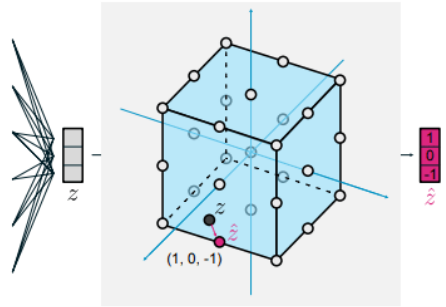
Standardizing all modalities through tokenization



- For each modality category (e.g. image, spectrum) we build dedicated tokenizers
  => Convert from any data to **discrete tokens**

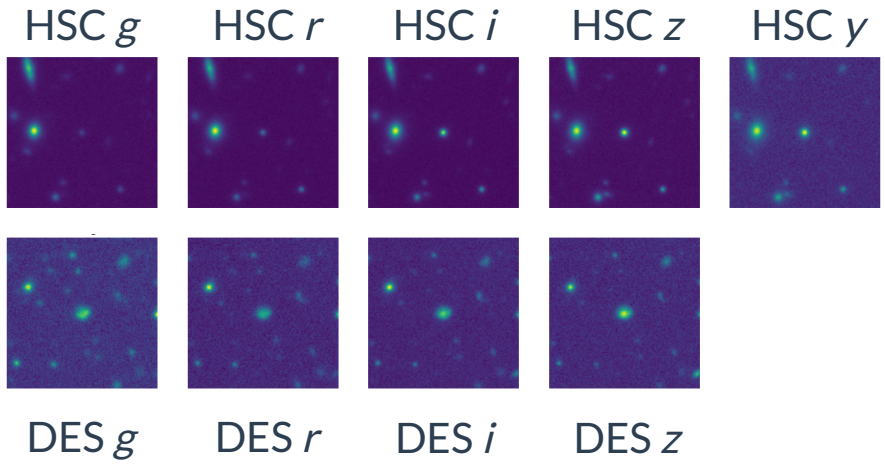
$$\mathcal{L} = \| \Sigma^{-\frac{1}{2}} \left(x - d_\theta(\lfloor e_\theta(x) \rfloor_{\text{FSQ}}) \right) \|_2^2$$

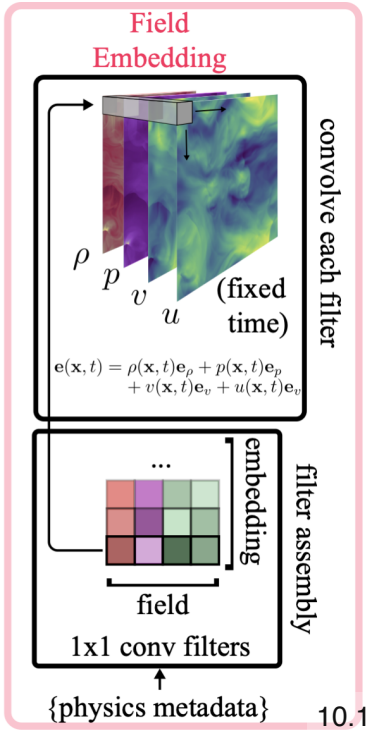
FSQ



(Mentzer et al. 2023)

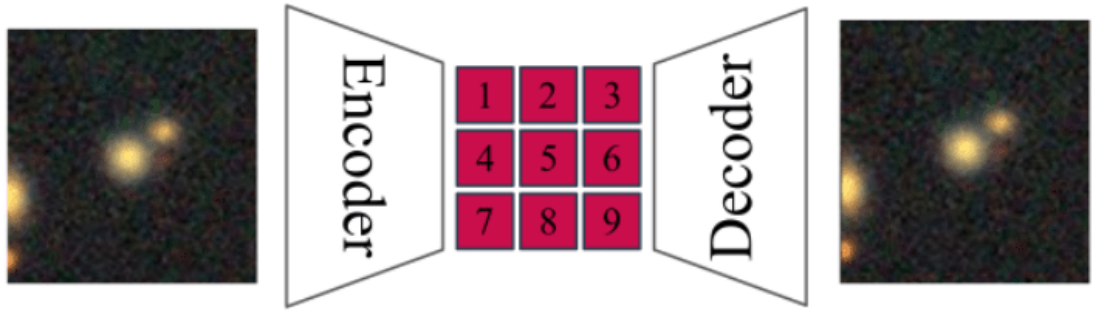
	VQ	FSQ
Quantization	argmin_c || z-c ||	round(f(z))
Gradients	Straight Through Estimation (STE)	STE
Auxiliary Losses	Commitment, codebook, entropy loss, ...	N/A
Tricks	EMA on codebook, codebook splitting, projections, ...	N/A
Parameters	Codebook	N/A

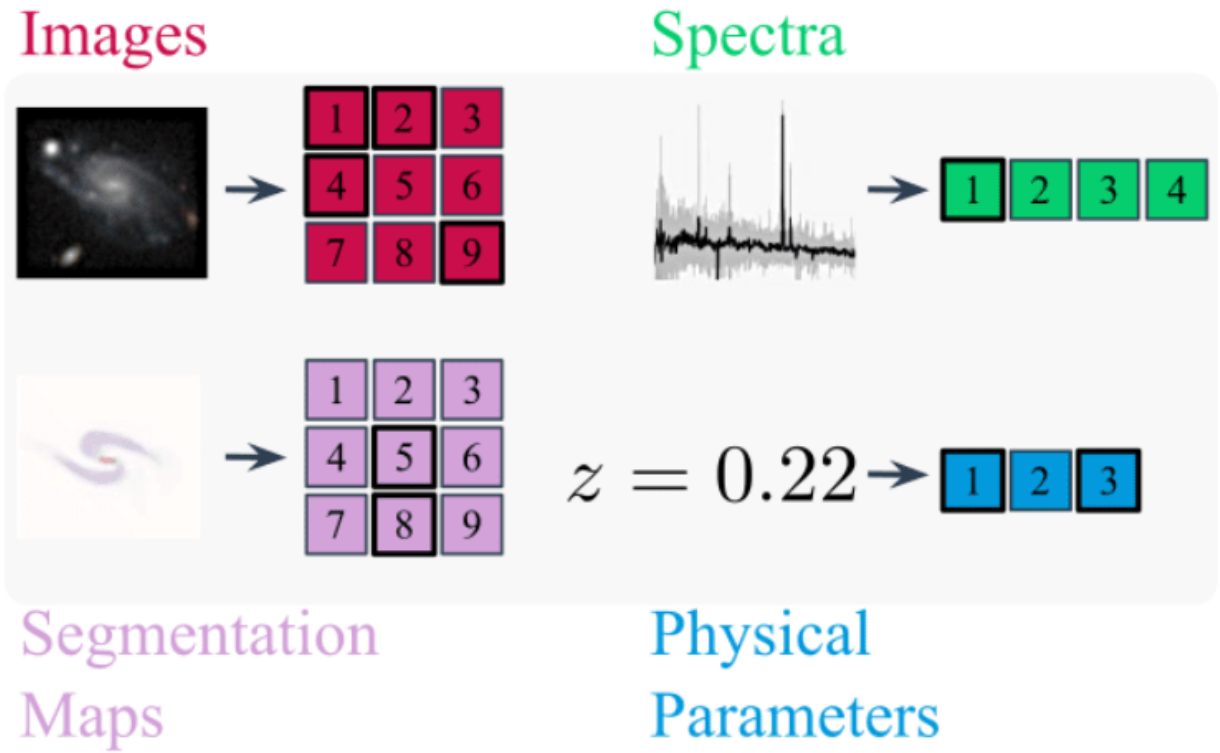




Field Embedding Strategy Developed for Multiple Physics Pretraining (McCabe et al. 2023)

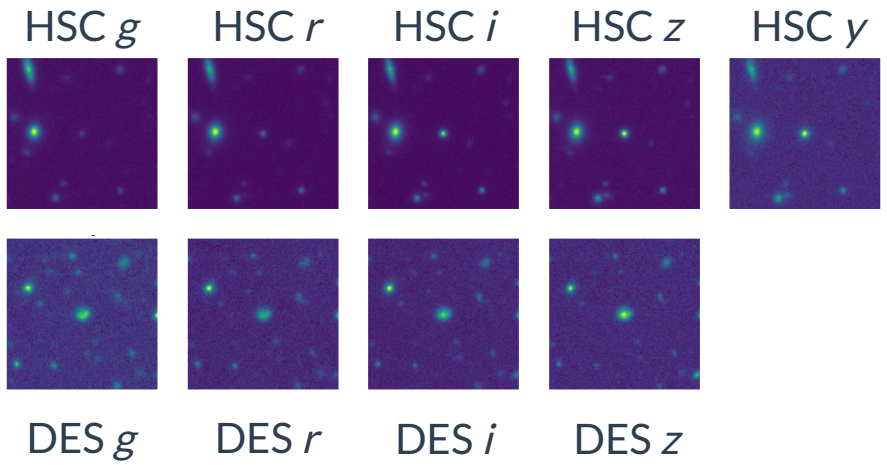
Standardizing all modalities through tokenization

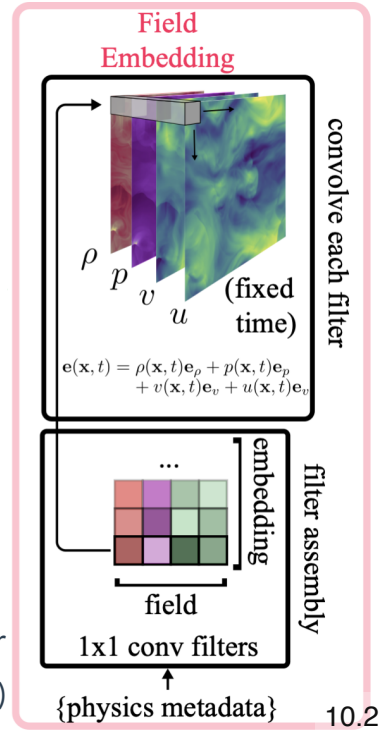




- For each modality category (e.g. image, spectrum) we build dedicated tokenizers => Convert from any data to **discrete tokens**

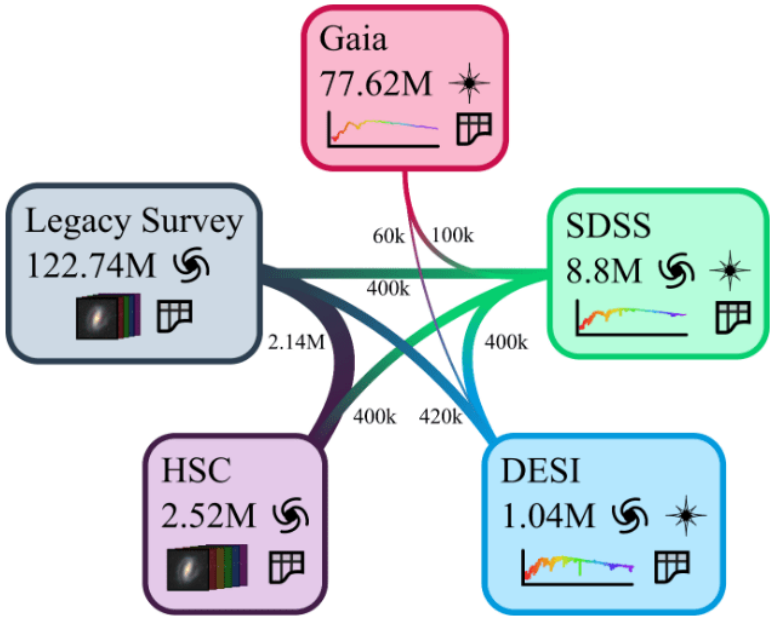
- For Aion-1, we integrate **39 different modalities** (different instruments, different measurements, etc.)

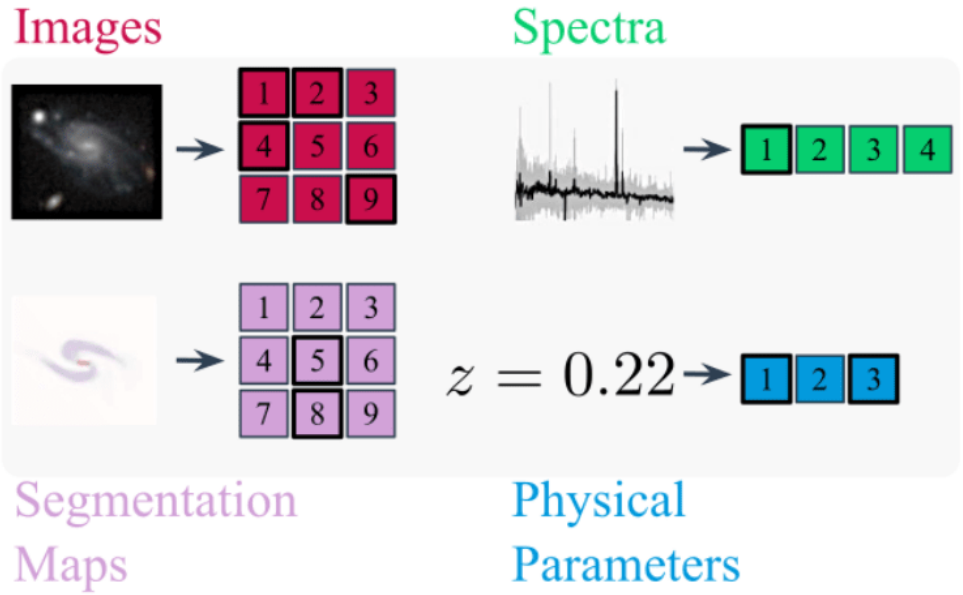


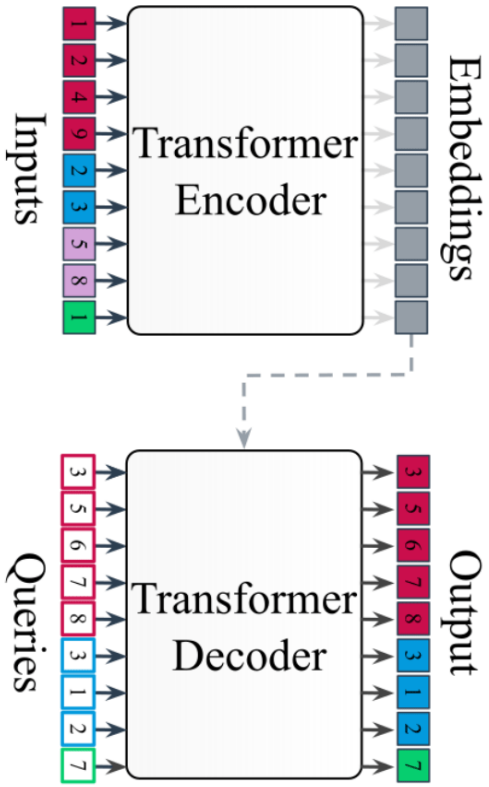


Field Embedding Strategy Developed for Multiple Physics Pretraining (McCabe et al. 2023)

Any-to-Any Modeling with Generative Masked Modeling



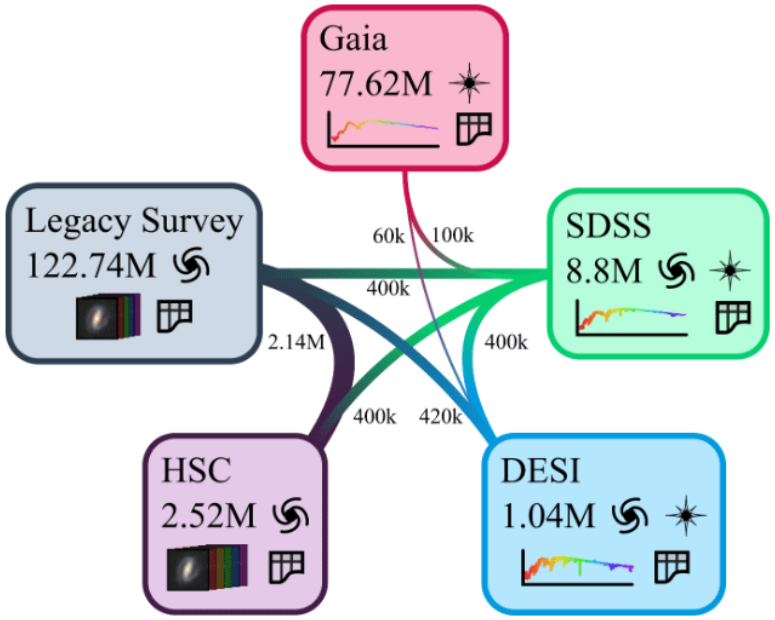


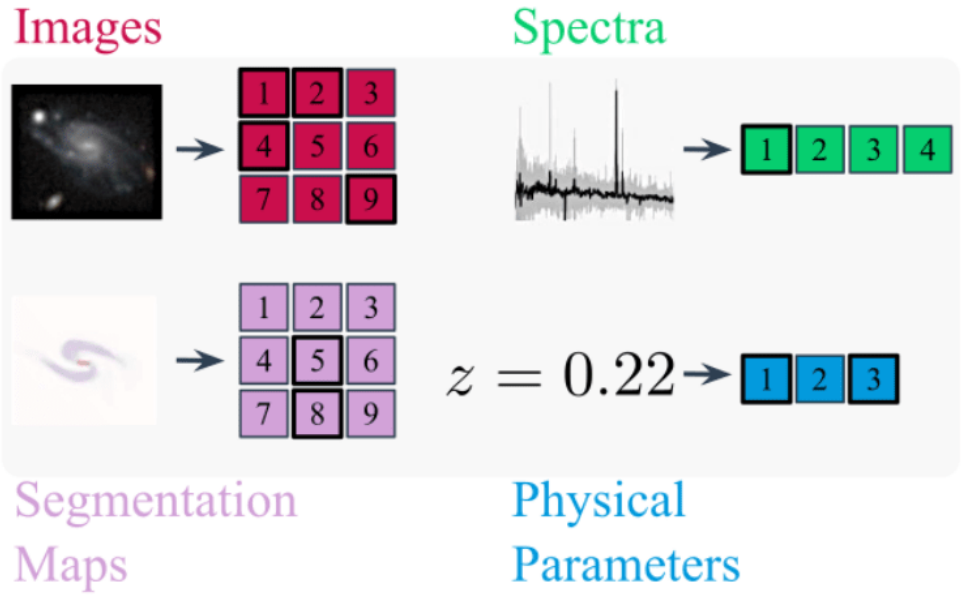


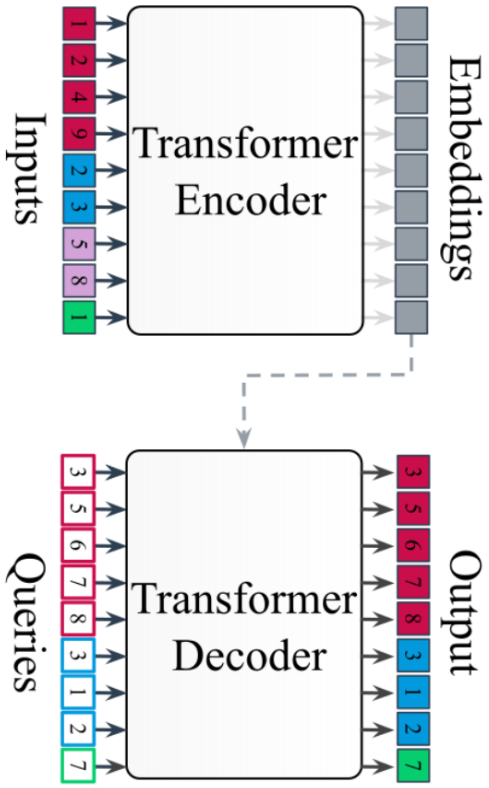
- Given standardized and cross-matched dataset, we can feed the data to a large **Transformer Encoder Decoder**
  - Flexible to **any combination of input data**, can be prompted to **generate any output**.

Any-to-Any Modeling with Generative Masked Modeling

Polymathic



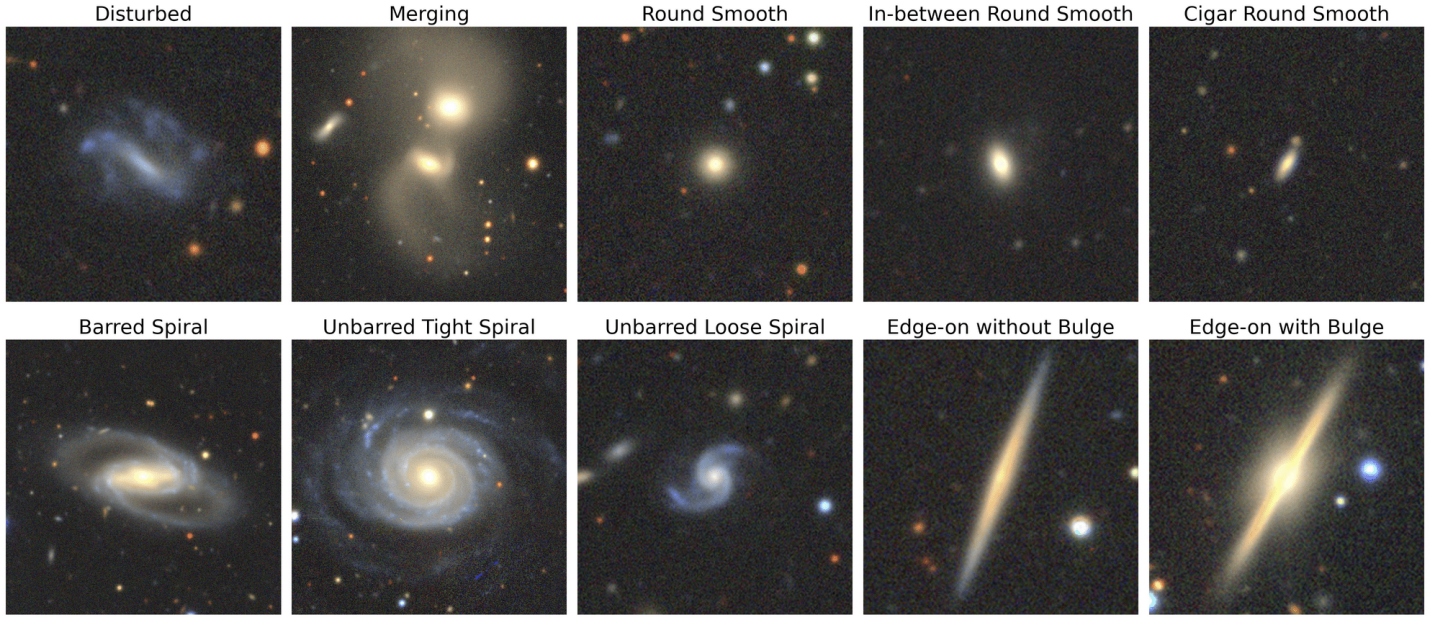




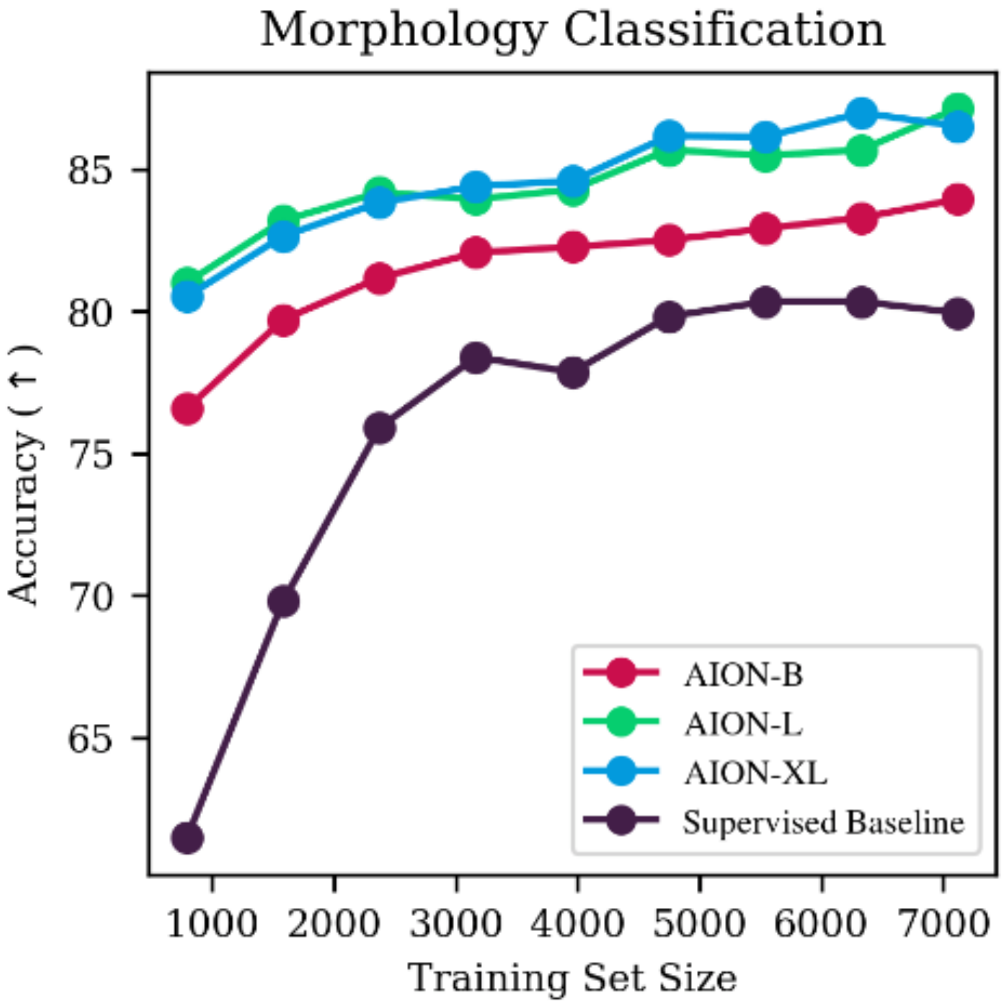
- Given standardized and cross-matched dataset, we can feed the data to a large **Transformer Encoder Decoder**
  - Flexible to **any combination of input data**, can be prompted to **generate any output**.
- Model is trained by cross-modal generative masked modeling
  => Learns the ***joint and all conditional distributions of provided modalities***: $\forall m, n \quad p_\theta(x_m | x_n)$

Morphology classification by Linear Probing

Example images of each class from Galaxy10 DECals



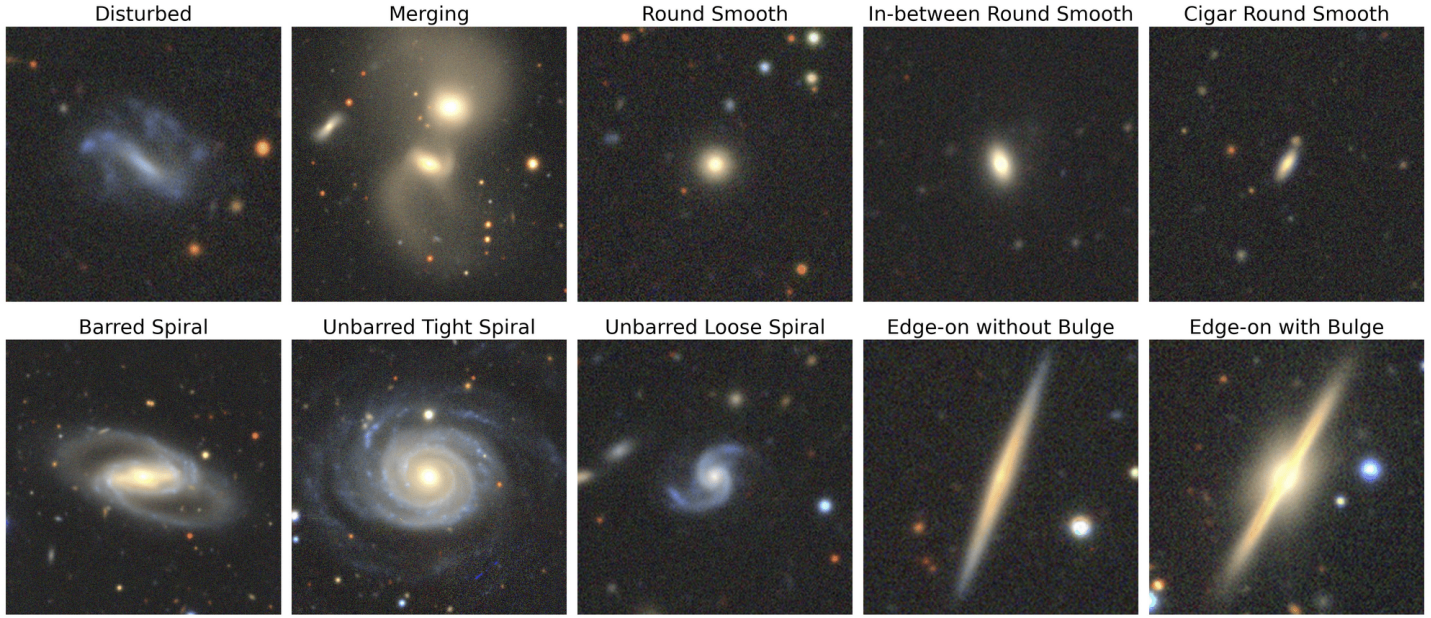
Galaxy10 DECals: Henry Leung/Jo Bovy 2021, Data: DECals/Galaxy Zoo



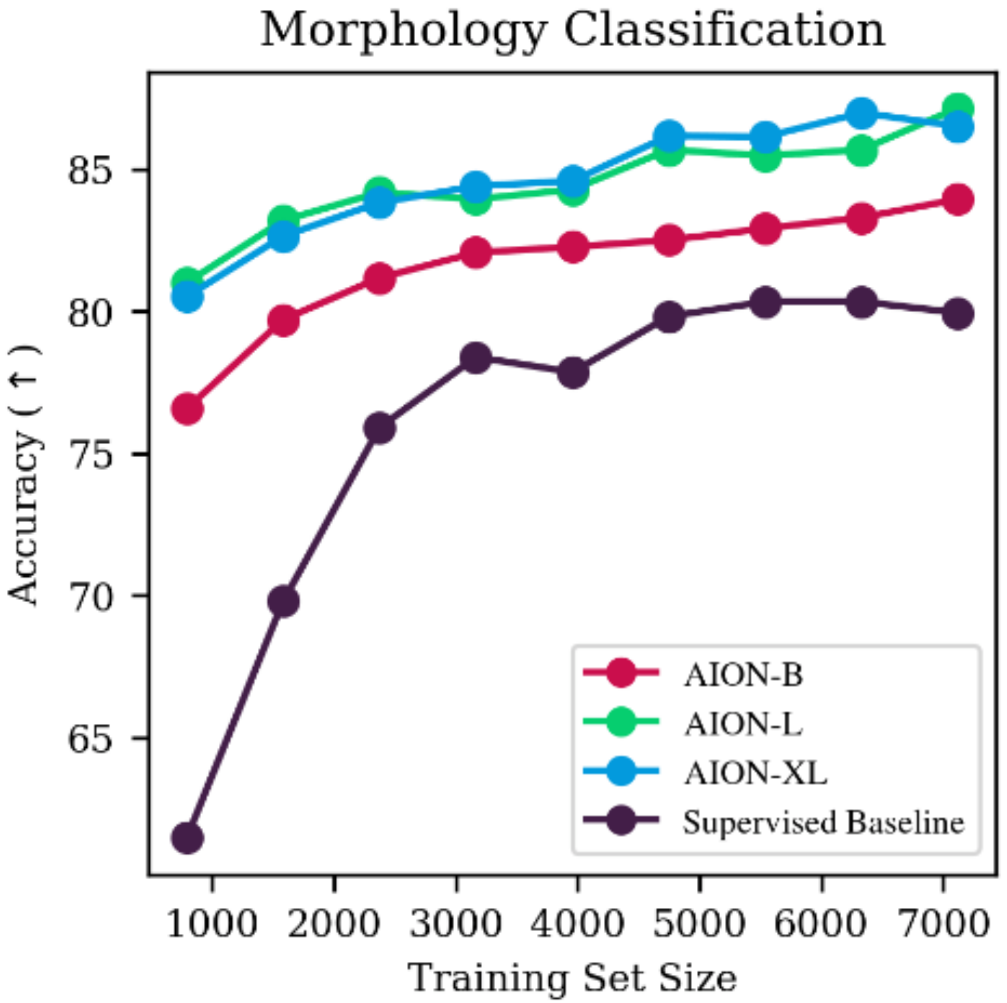
Morphology classification by Linear Probing

Polymathic

Example images of each class from Galaxy10 DECals



Galaxy10 DECals: Henry Leung/Jo Bovy 2021, Data: DECals/Galaxy Zoo



	AION-1-B	AION-1-L	**AION-1-XL**
Trained on -> Legacy Survey	83.95	87.16	86.99
Eval on -> HSC	84.15	85.66	85.91

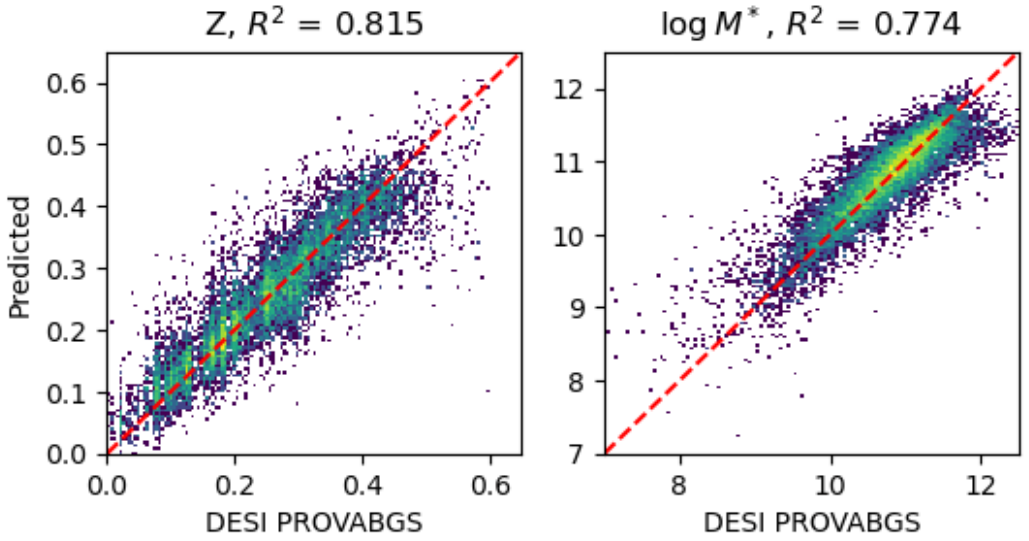
Table 5: **Transfer accuracy (↑) of AION-1 models on morphology classification.**

Physical parameter estimation and data fusion

Polymathic

Inputs:

measured fluxes



Physical parameter estimation and data fusion

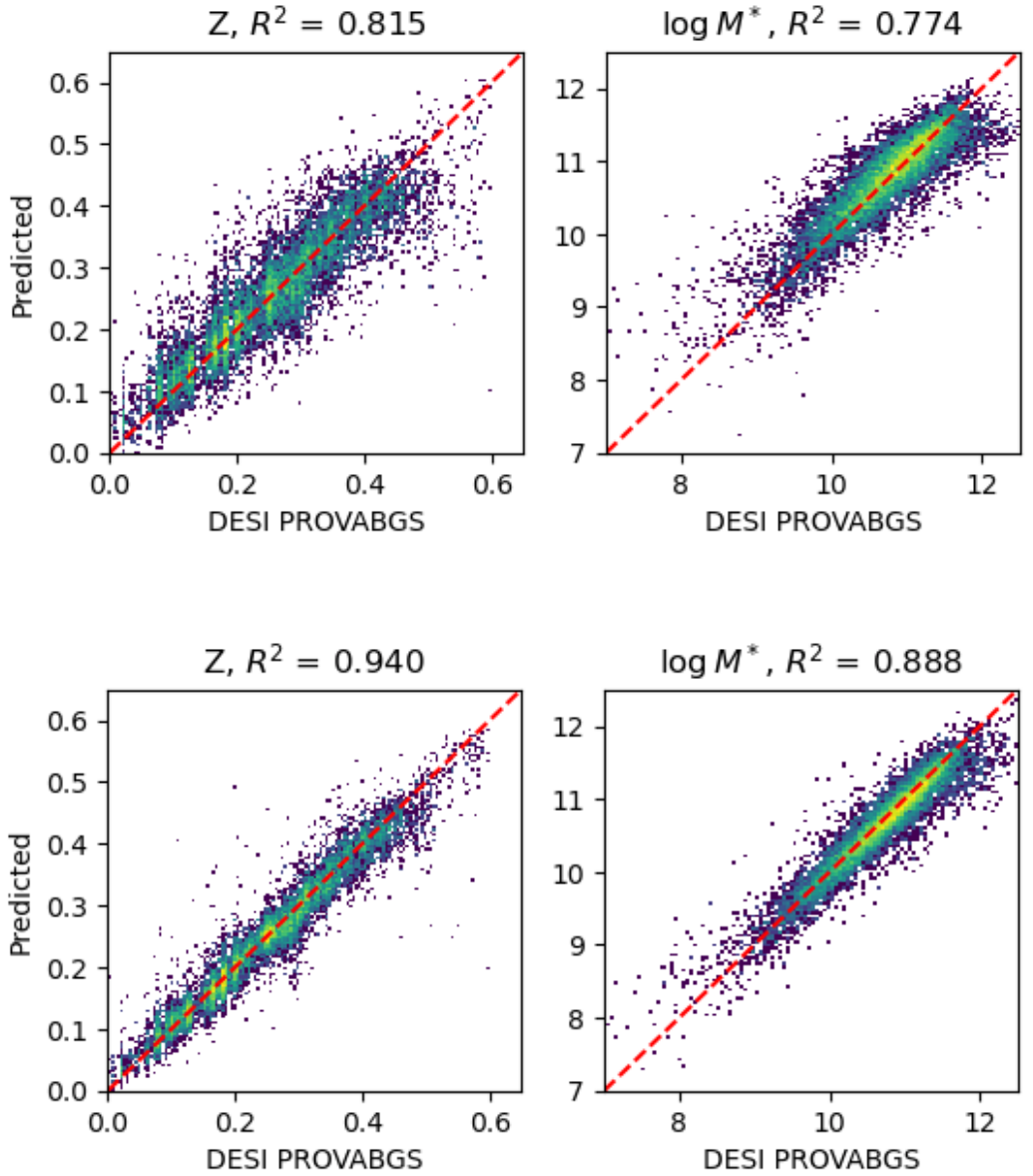
Polymathic

Inputs:

measured fluxes

Inputs:

measured fluxes
+ image



Physical parameter estimation and data fusion

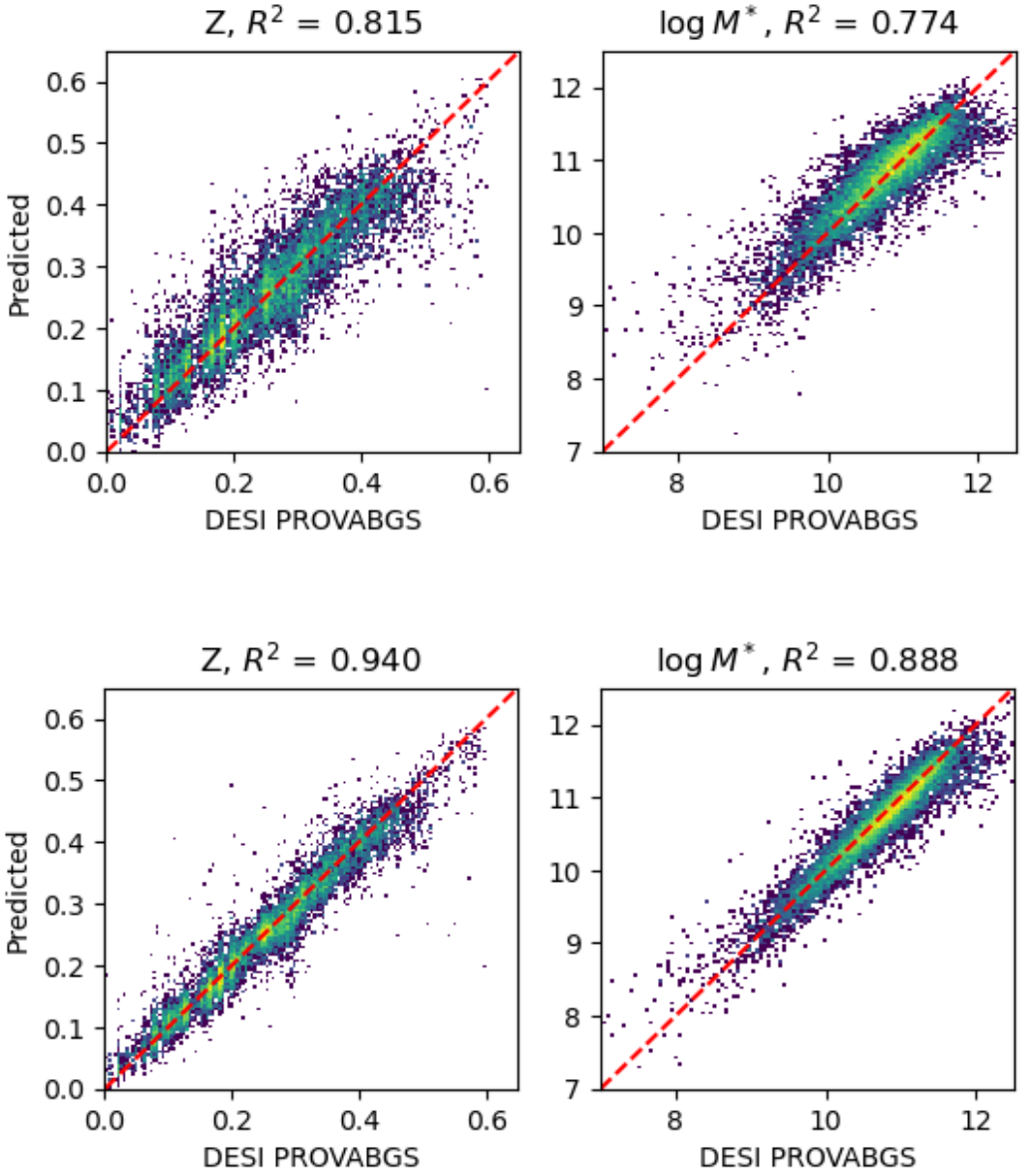
Polymathic

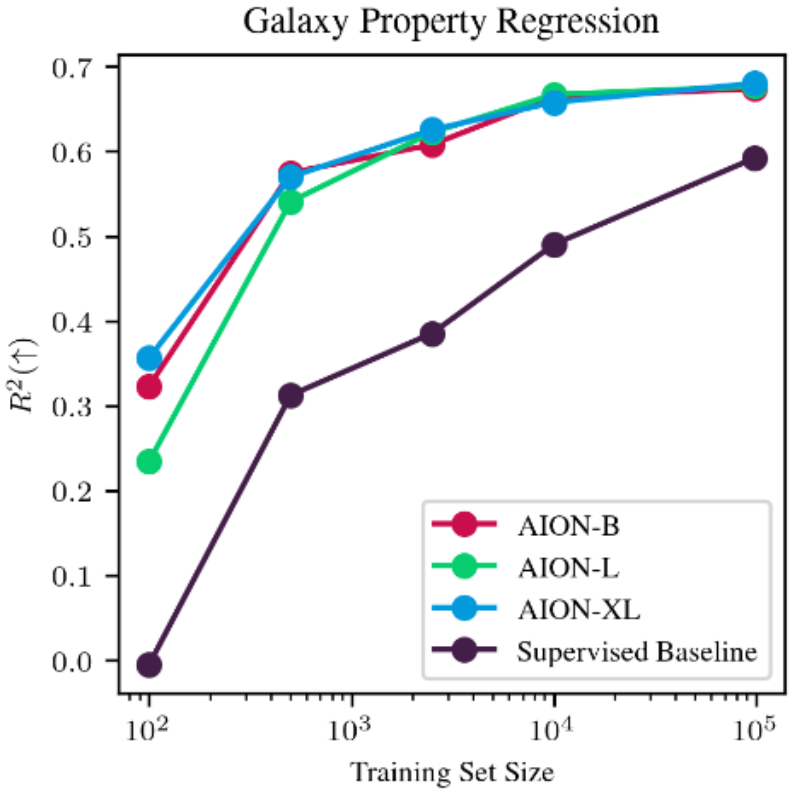
Inputs:

measured fluxes

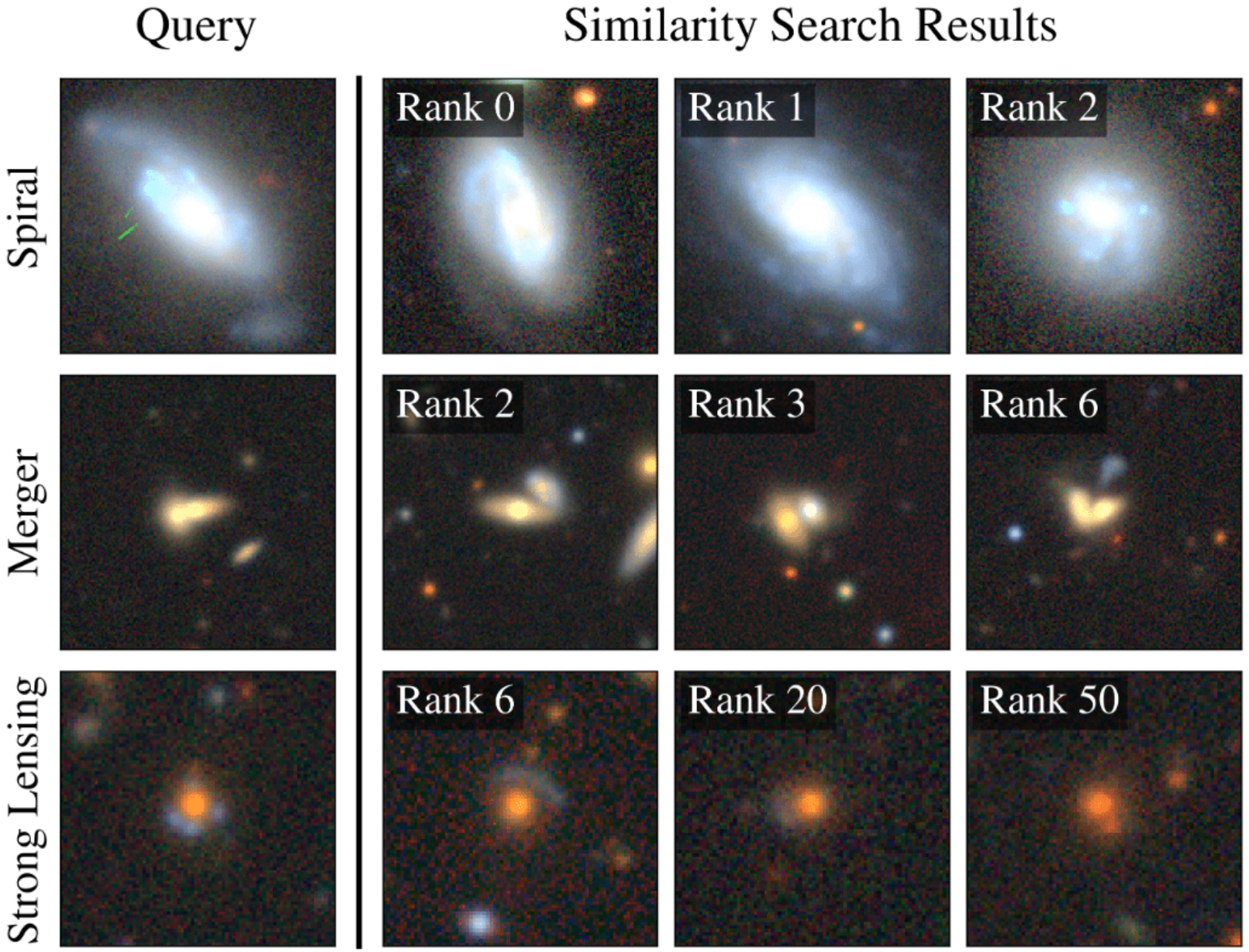
Inputs:

measured fluxes
+ image





Example-based retrieval from mean pooling



Where do we get the data to train these models?

The Scientific Data Curation Challenge

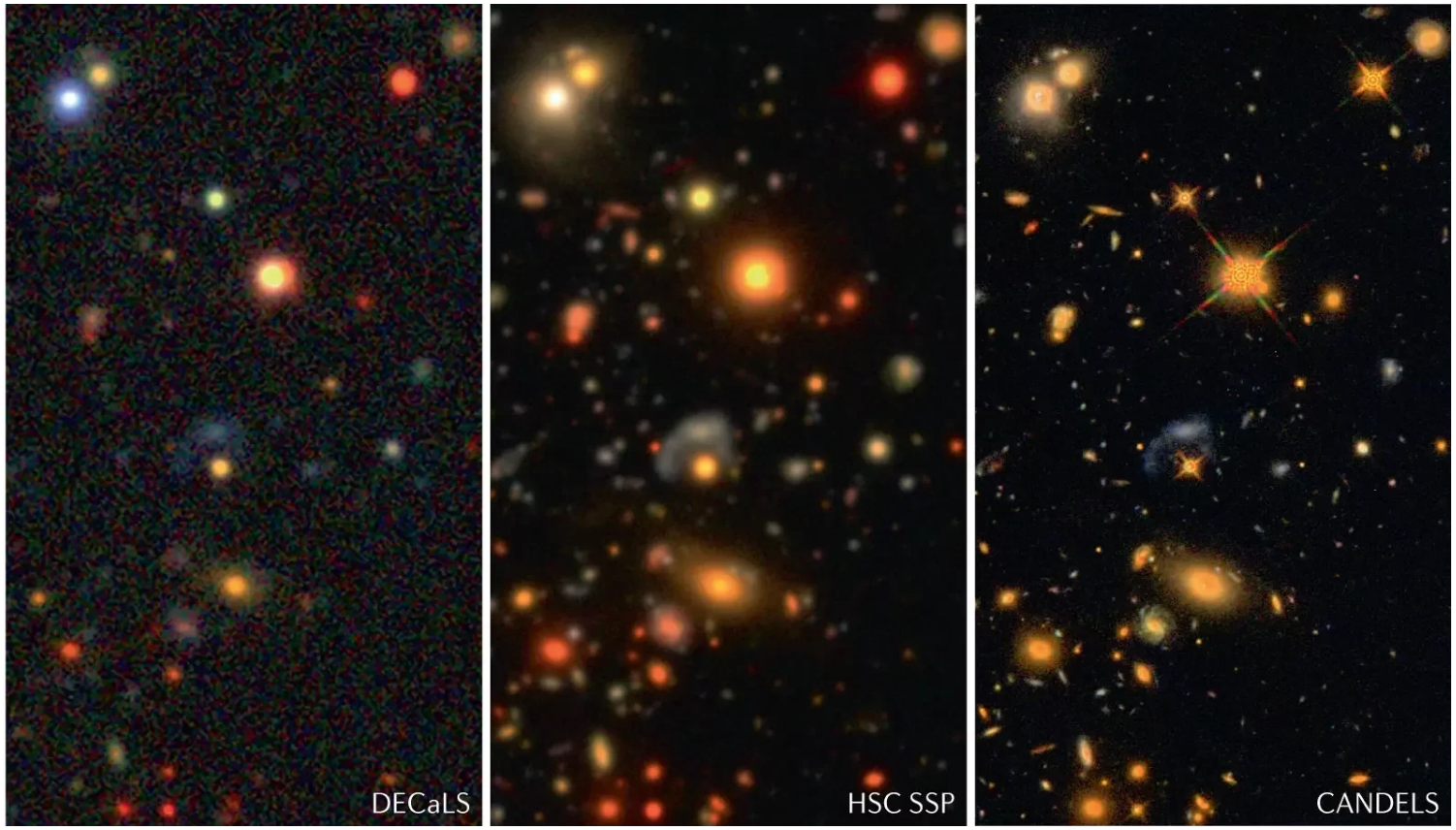
Polymathic

Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Schuhmann et al. (2022)

- Success of foundation models is driven by **large corpora of uniform data** (e.g LAION 5B).

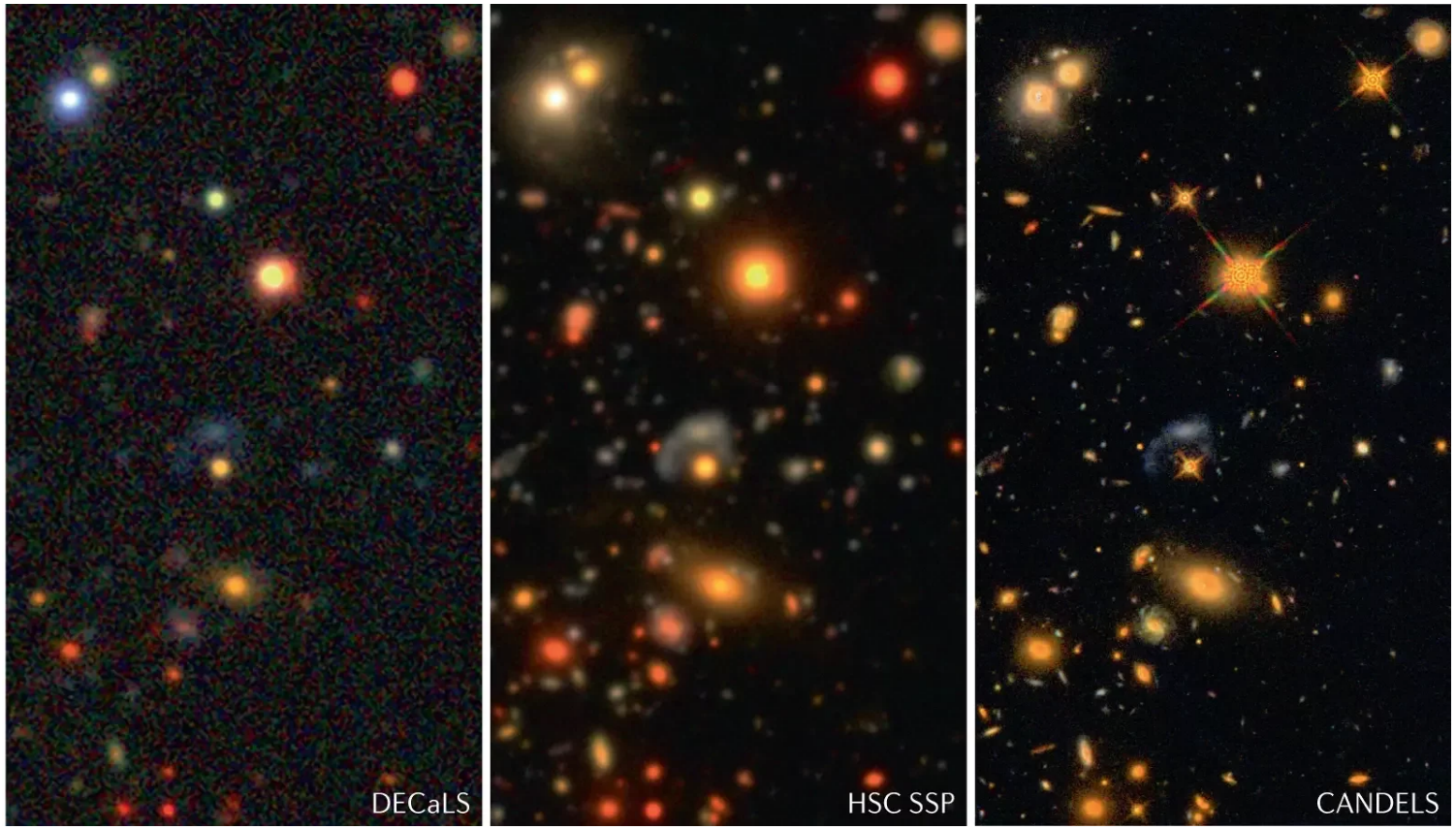
The Scientific Data Curation Challenge



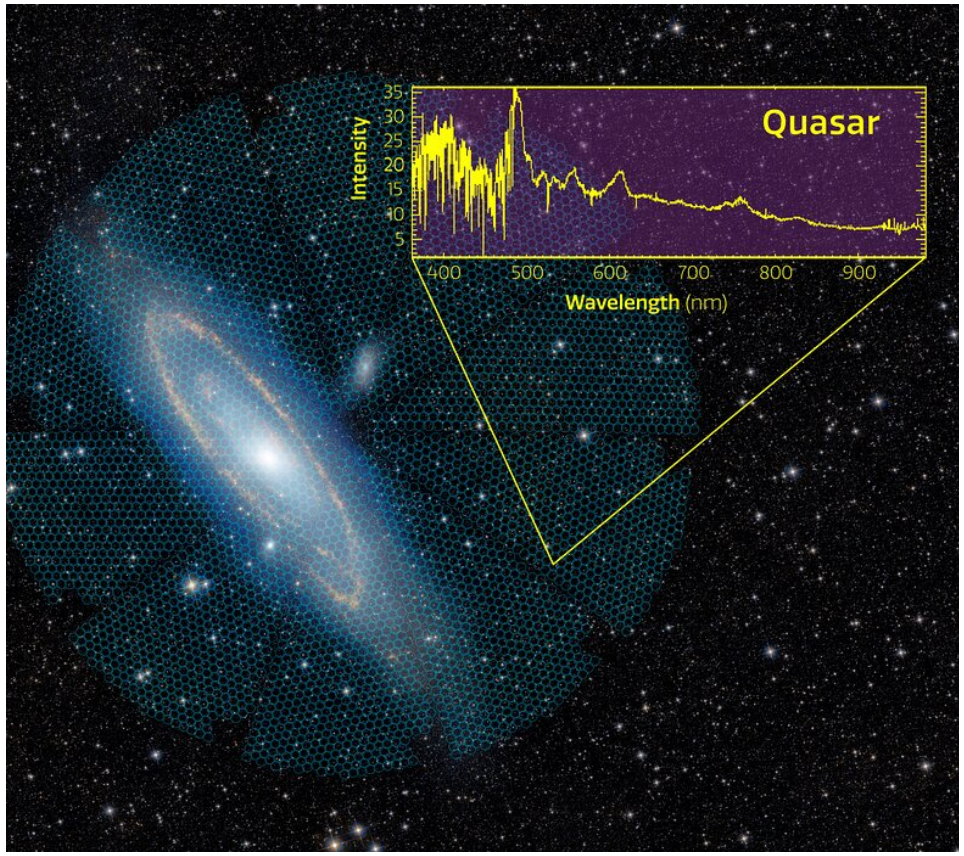
Credit: Melchior et al. 2021

- Success of foundation models is driven by **large corpora of uniform data** (e.g LAION 5B).
- Scientific data comes with many additional challenges:
  - Metadata matters

The Scientific Data Curation Challenge



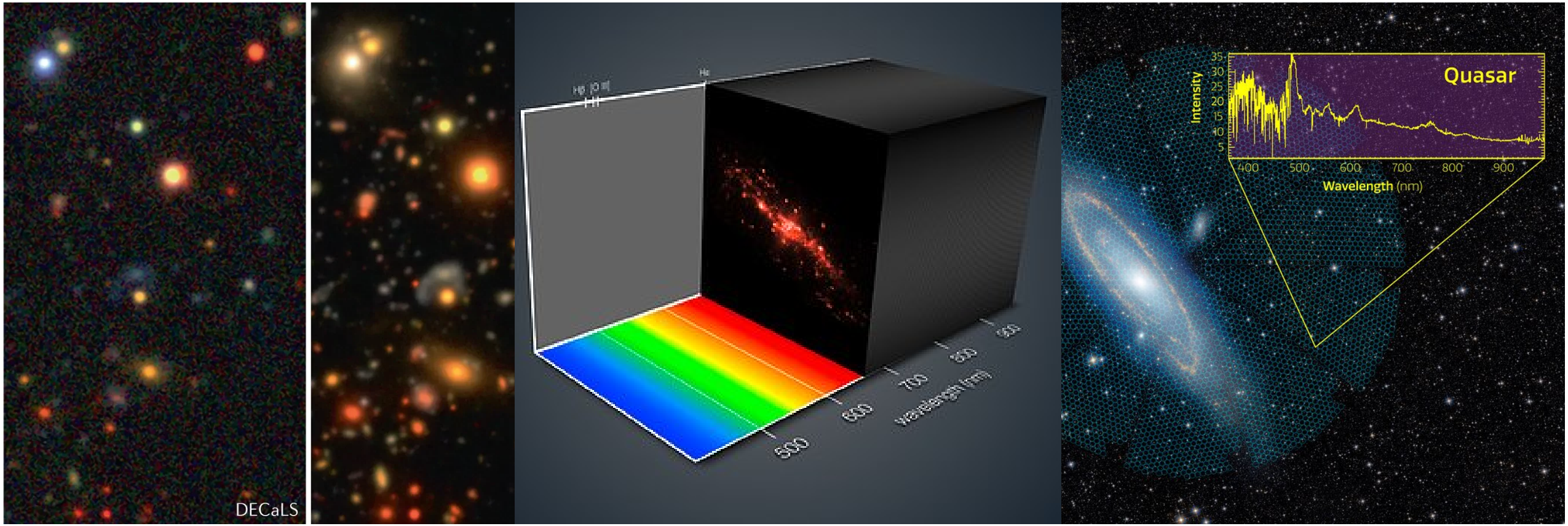
Credit: Melchior et al. 2021



Credit:DESI collaboration/DESI Legacy Imaging Surveys/LBNL/DOE & KPNO/CTIO/NOIRLab/NSF/AURA/unWISE

- Success of foundation models is driven by **large corpora of uniform data** (e.g LAION 5B).
- Scientific data comes with many additional challenges:
  - Metadata matters
  - Wide variety of measurements/observations

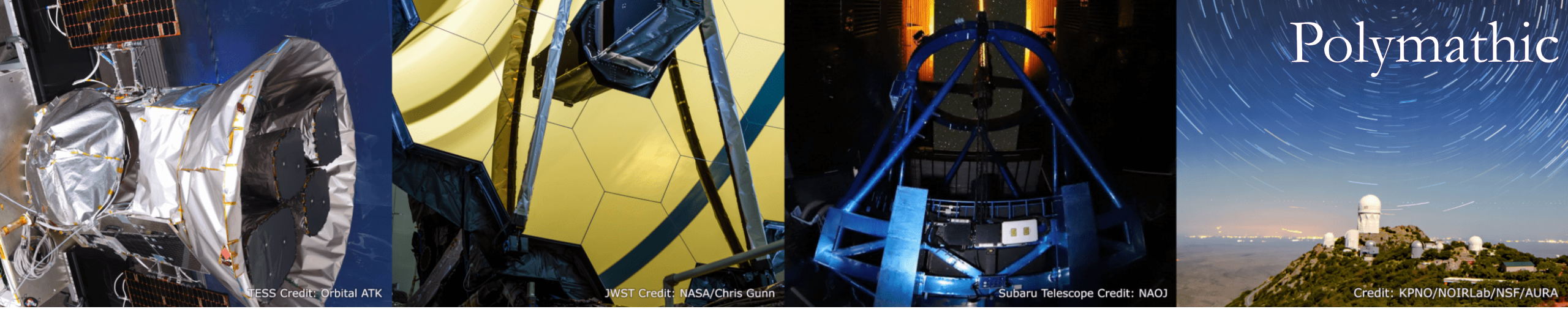
The Scientific Data Curation Challenge



Credit: Melchior et al. 2021

Credit:DESI collaboration/DESI Legacy Imaging Surveys/LBNL/DOE & KPNO/CTIO/NOIRLab/NSF/AURA/unWISE

- Success of foundation models is driven by **large corpora of uniform data** (e.g LAION 5B).
- Scientific data comes with many additional challenges:
  - Metadata matters
  - Wide variety of measurements/observations



The Multimodal Universe

Enabling Large-Scale Machine Learning with 100TBs of Astronomical Scientific Data

arXiv 2412.02527 NeurIPS 2024 Stars 455

Collaborative project with about 30 contributors

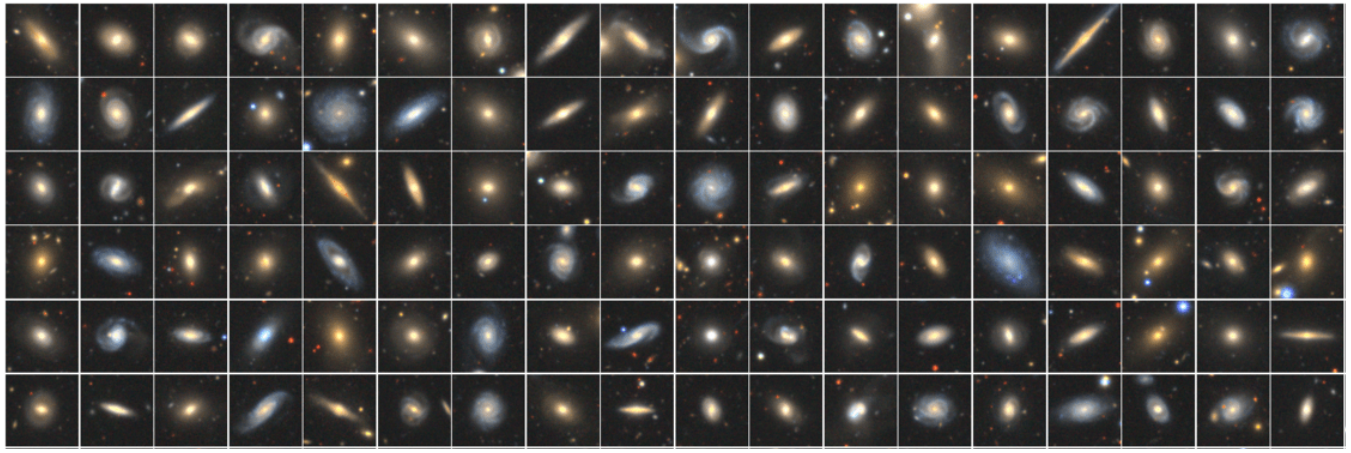
Presented at **NeurIPS 2024 Datasets & Benchmark track**

MULTIMODAL UNIVERSE

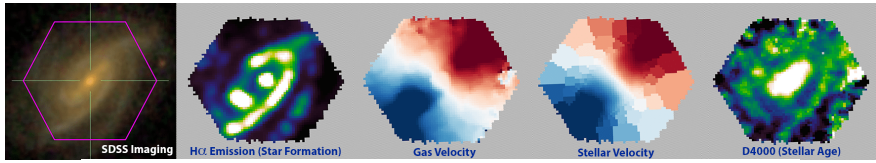
NEURAL INFORMATION PROCESSING SYSTEMS

The MultiModal Universe Project

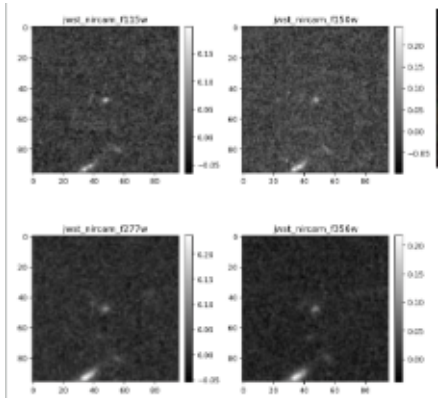
- **Goal**: Assemble the first large-scale multi-modal dataset for machine learning in astrophysics.



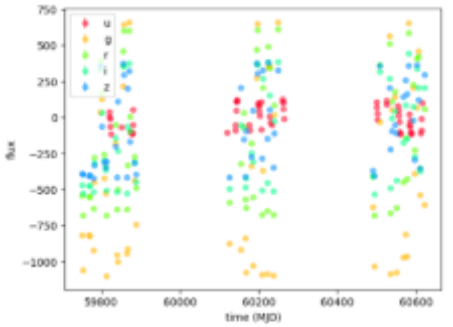
Polymathic



Hyperspectral Images from MaNGA



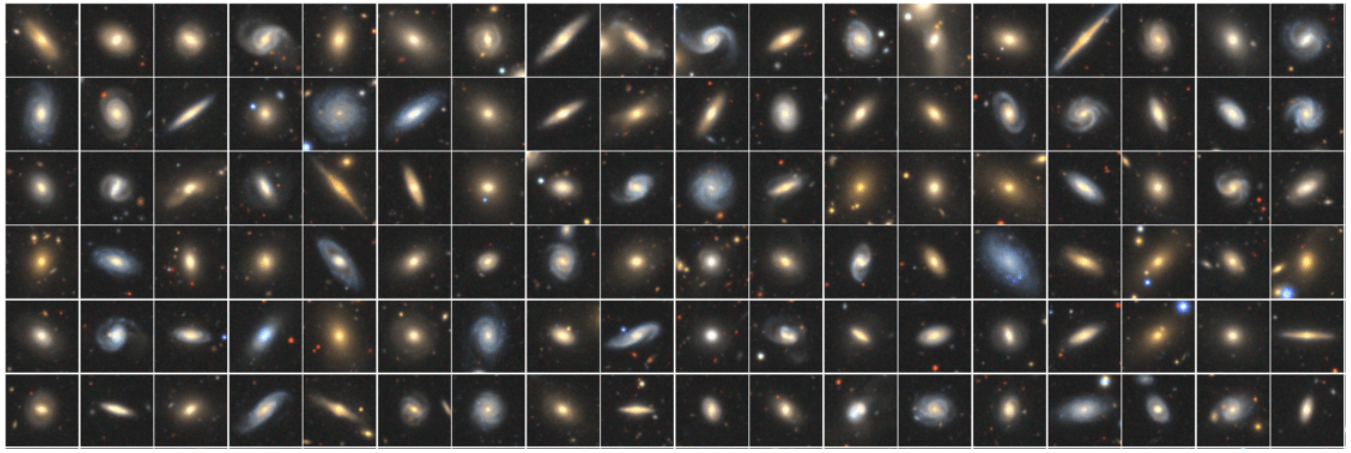
JWST Images



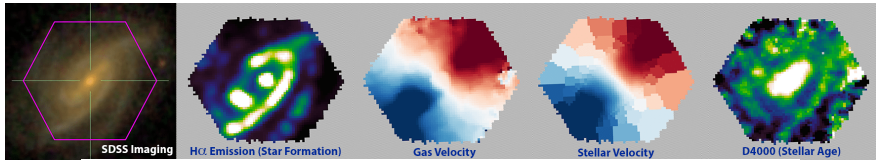
Time series of supernovae and stellar variability

The MultiModal Universe Project

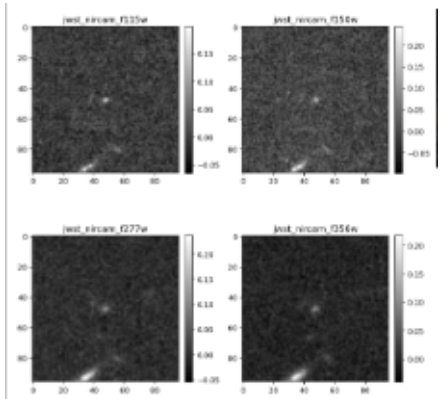
- **Goal**: Assemble the first large-scale multi-modal dataset for machine learning in astrophysics.
- **Main pillars**:
  - Engage with a **broad community of AI+Astro experts**.
  - Adopt **standardized conventions for storing and accessing data** and metadata through mainstream tools (e.g. **Hugging Face Datasets**).
  - Target large astronomical surveys, varied types of instruments, many **different astrophysics sub-fields**.



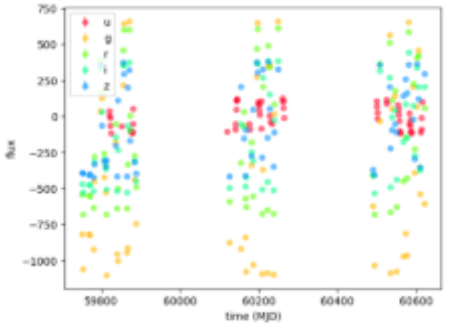
Polymathic



Hyperspectral Images from MaNGA



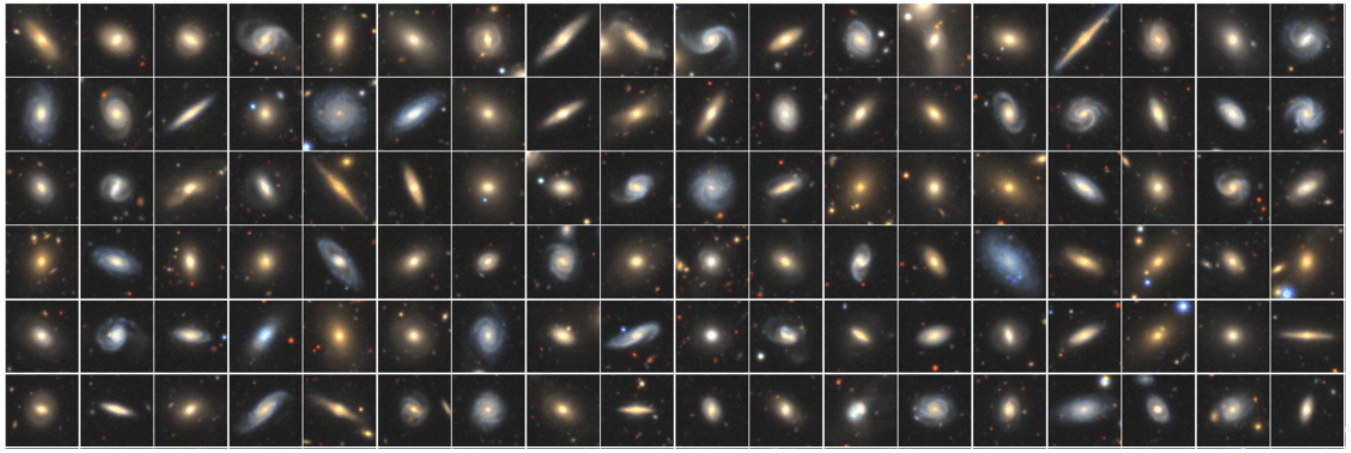
JWST Images



Time series of supernovae and stellar variability

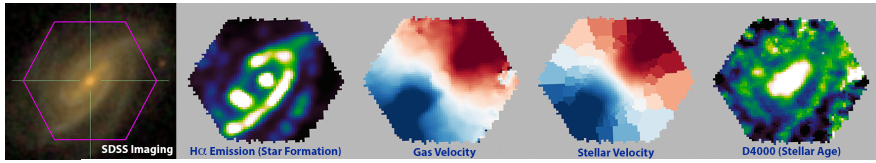
The MultiModal Universe Project

- **Goal**: Assemble the first large-scale multi-modal dataset for machine learning in astrophysics.
- **Main pillars**:
  - Engage with a **broad community of AI+Astro experts**.
  - Adopt **standardized conventions for storing and accessing data** and metadata through mainstream tools (e.g. **Hugging Face Datasets**).
  - Target large astronomical surveys, varied types of instruments, many **different astrophysics sub-fields**.

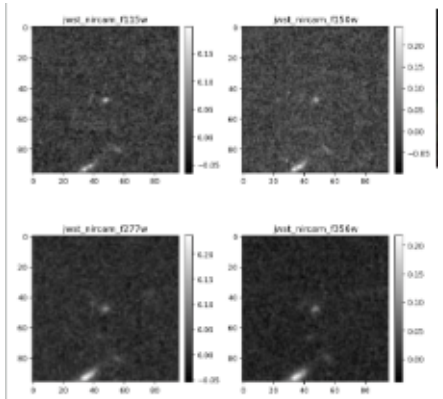


Multiband images from Legacy Survey

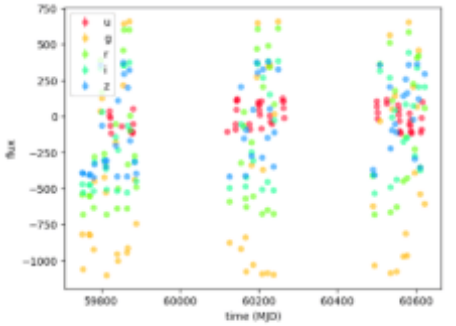
Polymathic



Hyperspectral Images from MaNGA



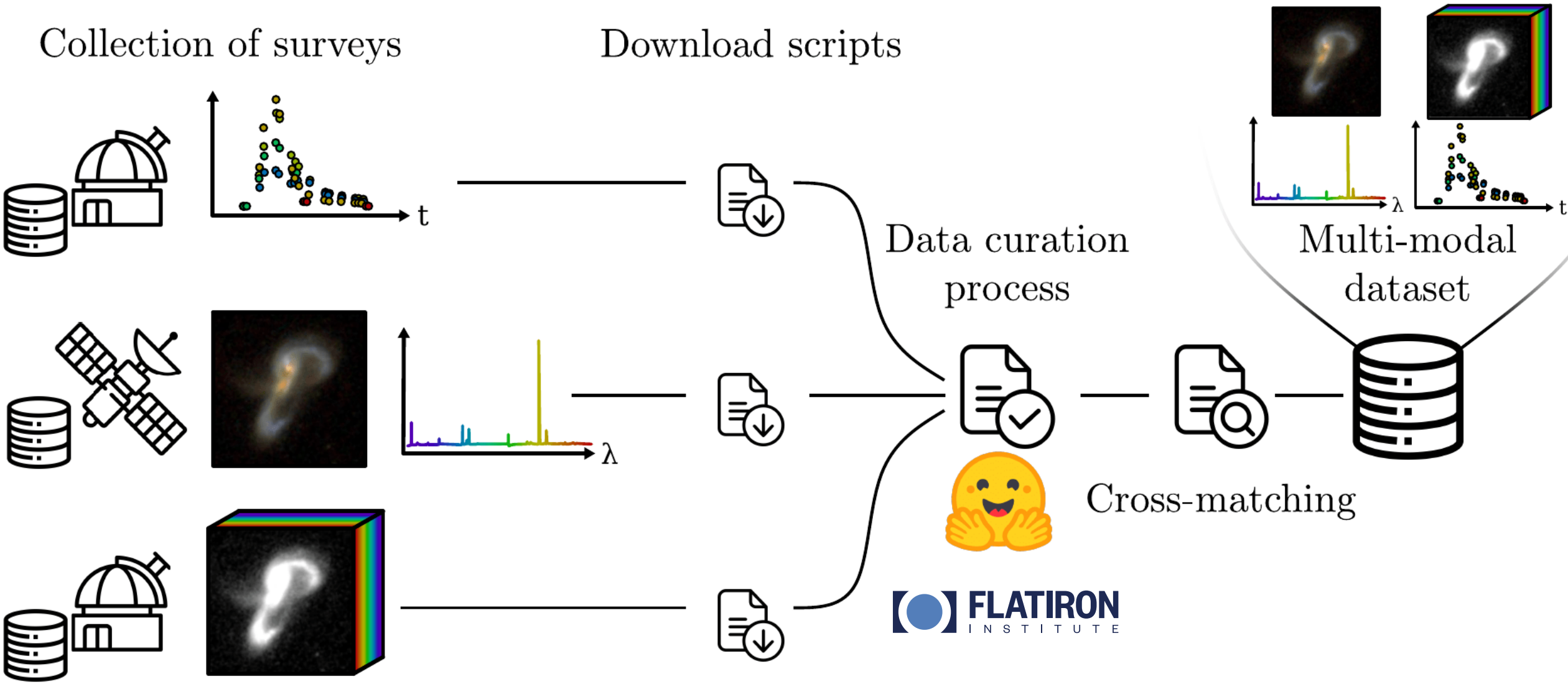
JWST Images



Time series of supernovae and stellar variability

MMU Infrastructure

Polymathic



Content of v1

Polymathic

Modality	Source Survey	N_c	Shape	Number of samples	Main science
Images	Legacy Surveys DR10 [43]	4	160×160	124M	Galaxies
	Legacy Surveys North [43, 134]	3	152×152	15M	Galaxies
	HSC [5, 3]	5	160×160	477K	Galaxies
	BTS [56, 114, 120]	3	63×63	400K	Supernovae
	JWST[13, 14, 50]	6-7	96×96	300K	Galaxies
Spectra	Gaia BP/RP [59]	-	110[1]	220M	Stars
	SDSS-II [1]	-	Variable	4M	Galaxies, Stars
	DESI [41]	-	7081	1M	Galaxies
	APOGEE SDSS-III [6]	-	7514	716k	Stars
	GALAH [28]	-	Variable	325k	Stars
	Chandra [51]	-	Variable	129K	Galaxies, Stars
	VIPERS [126]	-	557	91K	Galaxies
Hyperspectral Image	MaNGA SDSS-IV [2]	4563	96×96	12k	Galaxies
Time Series	PLAsTiCC[2][138]	6	Variable	3.5M	Time-varying objects
	TESS [121, 33]	1	Variable	1M	Exoplanets, Stars
	CfA Sample [68, 69, 18, 70]	5-11	Variable	1K	Supernovae
	YSE [7]	6	Variable	2K	Supernovae
	PS1 SNe Ia [127]	4	Variable	369	Supernovae
	DES Y3 SNe Ia [24]	4	Variable	248	Supernovae
	SNLS [63]	4	Variable	239	Supernovae
	Foundation [53, 81]	4	Variable	180	Supernovae
	CSP SNe Ia [36, 135, 86]	9	Variable	134	Supernovae
	Swift SNe Ia[26]	6	Variable	117	Supernovae
Tabular	Gaia [59]	-	-	220M	Stars
	PROVABGS [65]	-	-	221K	Galaxy
	Galaxy10 DECaLS [147, 92]	-	-	15K	Galaxy

Data schema and storage

Polymathic

- For each example MMU expects a few **mandatory fields**:
  - object_id, ra, dec

Data schema and storage

- For each example MMU expects a few **mandatory fields**:
  - object_id, ra, dec

- For each modality, MMU expects the **data** to be **formatted according to a fixed schema which only contains strictly necessary metadata.**

Table 2: Description of standardized fields and metadata provided for the main modalities. These fields represent necessary and near-sufficient information to allow for the consistent interpretation of observations from multiple surveys or instruments.

Modality	Field	Description
Images	flux	Array of flux measurements of the image
	ivar	Inverse variance of noise in the image
	band	Key indicating the wavelength range of the image
	psf_fwhm	Size of the instrumental response (Point Spread Function)
	scale	Scale of pixels on the sky
Spectra	flux	Measured flux as a function of wavelength
	ivar	Inverse variance of noise on measured flux
	lsf_sigma	Size of the instrumental response (Line Spread Function)
	lambda	Wavelength of each flux measurement
Time Series	flux	Measurements of flux as a function of time
	flux_err	Uncertainty on flux measurement
	band	Key indicating the wavelength range of the measurement
	time	Time of observation

Data schema and storage

- For each example MMU expects a few **mandatory fields**:
  - object_id, ra, dec

- For each modality, MMU expects the **data** to be **formatted according to a fixed schema which only contains strictly necessary metadata.**

- Data is stored in **HDF5 files**, split according to **HEALPix regions** for efficient cross-matching and easy access

Table 2: Description of standardized fields and metadata provided for the main modalities. These fields represent necessary and near-sufficient information to allow for the consistent interpretation of observations from multiple surveys or instruments.

Modality	Field	Description
Images	flux	Array of flux measurements of the image
	ivar	Inverse variance of noise in the image
	band	Key indicating the wavelength range of the image
	psf_fwhm	Size of the instrumental response (Point Spread Function)
	scale	Scale of pixels on the sky
Spectra	flux	Measured flux as a function of wavelength
	ivar	Inverse variance of noise on measured flux
	lsf_sigma	Size of the instrumental response (Line Spread Function)
	lambda	Wavelength of each flux measurement
Time Series	flux	Measurements of flux as a function of time
	flux_err	Uncertainty on flux measurement
	band	Key indicating the wavelength range of the measurement
	time	Time of observation

```
hsc
├── hsc.py
├── pdr3_dud_22.5
│   ├── healpix=1104
│   │   └── 001-of-001.hdf5
│   ├── healpix=1105
│   │   └── 001-of-001.hdf5
│   ├── healpix=1106
│   │   └── 001-of-001.hdf5
│   ├── healpix=1107
│   │   └── 001-of-001.hdf5
│   ├── healpix=1171
```

Datasets: MultimodalUniverse / legacysurvey

like 3 | Following Multimodal Universe 78

Modalities: Image | Text | Formats: parquet | Size: 10K - 100K | ArXiv: arxiv:1804.08657 | Libraries: Datasets | Dask | Croissant | +1

Dataset Viewer

Auto-converted to Parquet | API | Embed | Data Studio

Split (1)
train · 90.6k rows

Search is not available for this dataset

image sequence	blobmodel image · width (px) 160 160	rgb image · width (px) 160 160	object_mask image · width (px) 160 160	catalog sequence
{"band":["des-g","des-r","des-i","des-z"],"flux":[[[0.0007380799506790936,-0.0001605062279850244 5,0.…				{"FLUX_G":[1.3640868663787842,0
{"band":["des-g","des-r","des-i","des-z"],"flux":[[[0.00069914769846946,0.00088390929158777,-0.00082…				{"FLUX_G":[0.5193391442298889,0
{"band":["des-g","des-r","des-i","des-z"],"flux":[[[0.0026036326307803392,0.002823472488671541,-0.00…				{"FLUX_G":[0.5193391442298889,0
{"band":["des-g","des-r","des-i","des-z"],"flux":[[[0.029560092836618423,0.02978307381272316,0.02261…				{"FLUX_G":[0.9414768218994141,0

Previous | 1 | 2 | 3 | ... | 906 | Next

Downloads last month 530

View full history

Use this dataset | Edit dataset card

Size of downloaded dataset files: 64.5 GB

Size of the auto-converted Parquet files: 64.5 GB

Number of rows: 90,580

Models trained or fine-tuned on MultimodalUniv…

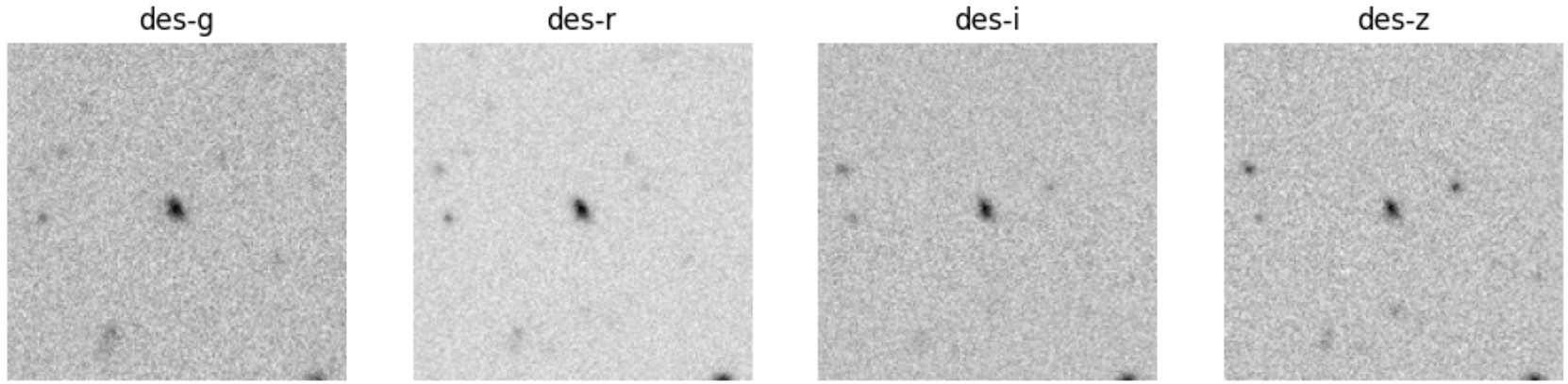
polymathic-ai/aion-base
0.3B · Updated 24 days ago · 893 · 13

Usage example


```python
from datasets import load_dataset

# Open Hugging Face dataset
dset_ls = load_dataset("MultimodalUniverse/legacysurvey",
                       streaming=True,
                       split='train')
dset_ls = dset_ls.with_format("numpy")
dset_iterator = iter(dset_ls)

# Draw one example from the dataset iterator
example = next(dset_iterator)

# Let's inspect what is contained in an example
print(example.keys())

figure(figsize=(12,5))
for i,b in enumerate(example['image']['band']):
  subplot(1,4,i+1)
  title(f'{b}')
  imshow(example['image']['flux'][i], cmap='gray_r')
  axis('off')
```

Open in Colab

```
dict_keys(['image', 'blobmodel', 'rgb',
'object_mask', 'catalog', 'EBV', 'FLUX_G',
'FLUX_R', 'FLUX_I', 'FLUX_Z', 'FLUX_W1',
'FLUX_W2', 'FLUX_W3', 'FLUX_W4', 'SHAPE_R',
'SHAPE_E1', 'SHAPE_E2', 'object_id'])
```



Some considerations for AI pre-training

- Design should facilitate streaming tensors to GPUs in conventional frameworks (e.g. Hugging Face Datasets or similar)

Some considerations for AI pre-training

- Design should facilitate streaming tensors to GPUs in conventional frameworks (e.g. Hugging Face Datasets or similar)

- Homogeneity is good but does not have to be imposed across data sources
  - i.e. images can have different pixel scale, normalizations, etc... as long as provenance is *somehow* captured alongside the data.
  - Neural networks will automatically learn to interpret the relative difference between datasets

Some considerations for AI pre-training

- Design should facilitate streaming tensors to GPUs in conventional frameworks (e.g. Hugging Face Datasets or similar)

- Homogeneity is good but does not have to be imposed across data sources
  - i.e. images can have different pixel scale, normalizations, etc... as long as provenance is *somehow* captured alongside the data.
  - Neural networks will automatically learn to interpret the relative difference between datasets

- Multimodal training requires cross-matching across surveys
  - Implies a strategy to obtain O(1000) postage stamps per seconds between 2 surveys
  - With MMU and training AION-1 this cross-matching is done offline: slow, not very flexible

Some considerations for AI pre-training

- Design should facilitate streaming tensors to GPUs in conventional frameworks (e.g. Hugging Face Datasets or similar)

- Homogeneity is good but does not have to be imposed across data sources
  - i.e. images can have different pixel scale, normalizations, etc... as long as provenance is *somehow* captured alongside the data.
  - Neural networks will automatically learn to interpret the relative difference between datasets

- Multimodal training requires cross-matching across surveys
  - Implies a strategy to obtain O(1000) postage stamps per seconds between 2 surveys
  - With MMU and training AION-1 this cross-matching is done offline: slow, not very flexible

- Versionning and production of static curated datasets is challenging
  - Requires a local copy of all datasets (postage stamps APIs of most surveys are not fast enough to support generating millions of stamps), non trivial storage, and non trivial compute.
  - MMU contains 120TB of data, versionning that data and incremental additions is not trivial to manage.

Polymathic

Thank you for listening!

Follow us online