Artificial Intelligence for Observational Cosmology: Machine Learning and Deep Learning with Astronomical Data

by
Professor John Suárez-Pérez

Workshop Outline

- **Lecture:** Introduction to DESI Data

- **Hands on DESI Data: Learning-Based Approaches**
  - DESI spectroscopic classification
  - Predicting the Photo-redshift of Bright Galaxies

Introduction to DESI data

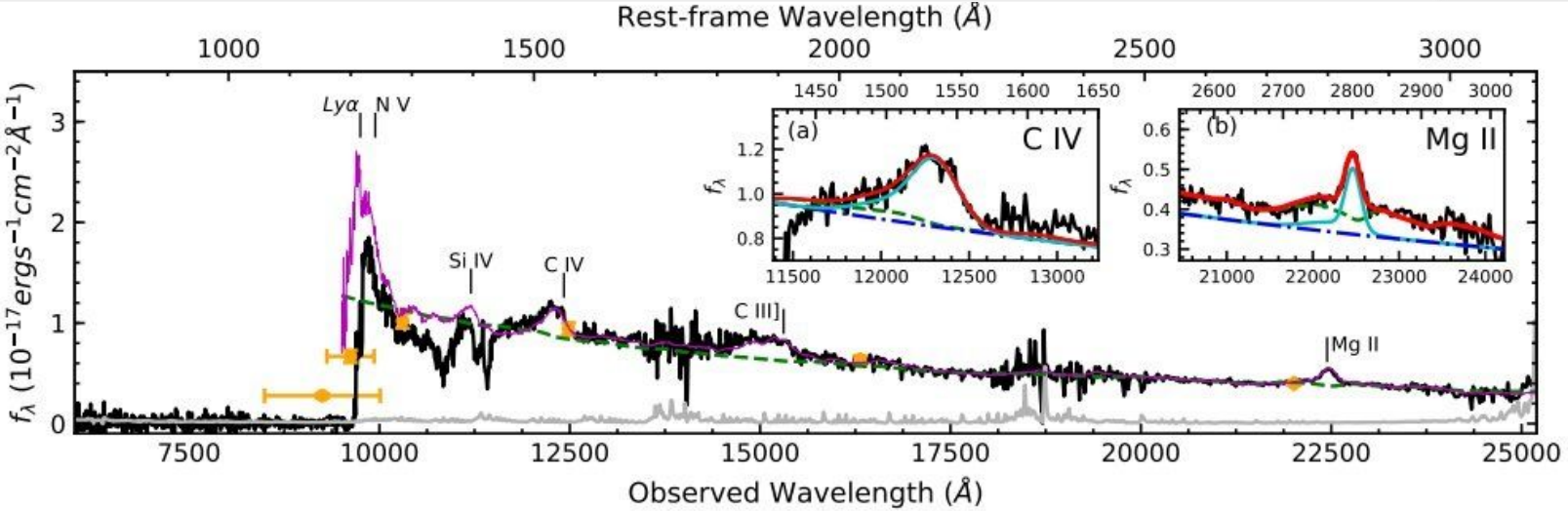
Why and How Galaxies are Observed?

- **Understanding Fundamental Physics Laws.**
  - Observations provide insights into how galaxies form and evolve over time.
  - Possibility to understand the physics behind their formation and evolution.

- **Origins and Evolution.**
  - To learn about its origins, such as the Big Bang.
  - Subsequent evolution (expansion) over billions of years.

- **Galaxies emit light that could be captured with CCD cameras attached to the telescopes.**
  - **Photometric** :
    - Light captured using different colored filters.
    - Magnitude (brightness), distance (redshift, z_photo), and other properties.
  - **Spectroscopic** :
    - Light captured using spectrographs.
    - What galaxies are made of and how far away they are (z_spec, more accurate than z_phot).

Redshift Measurements



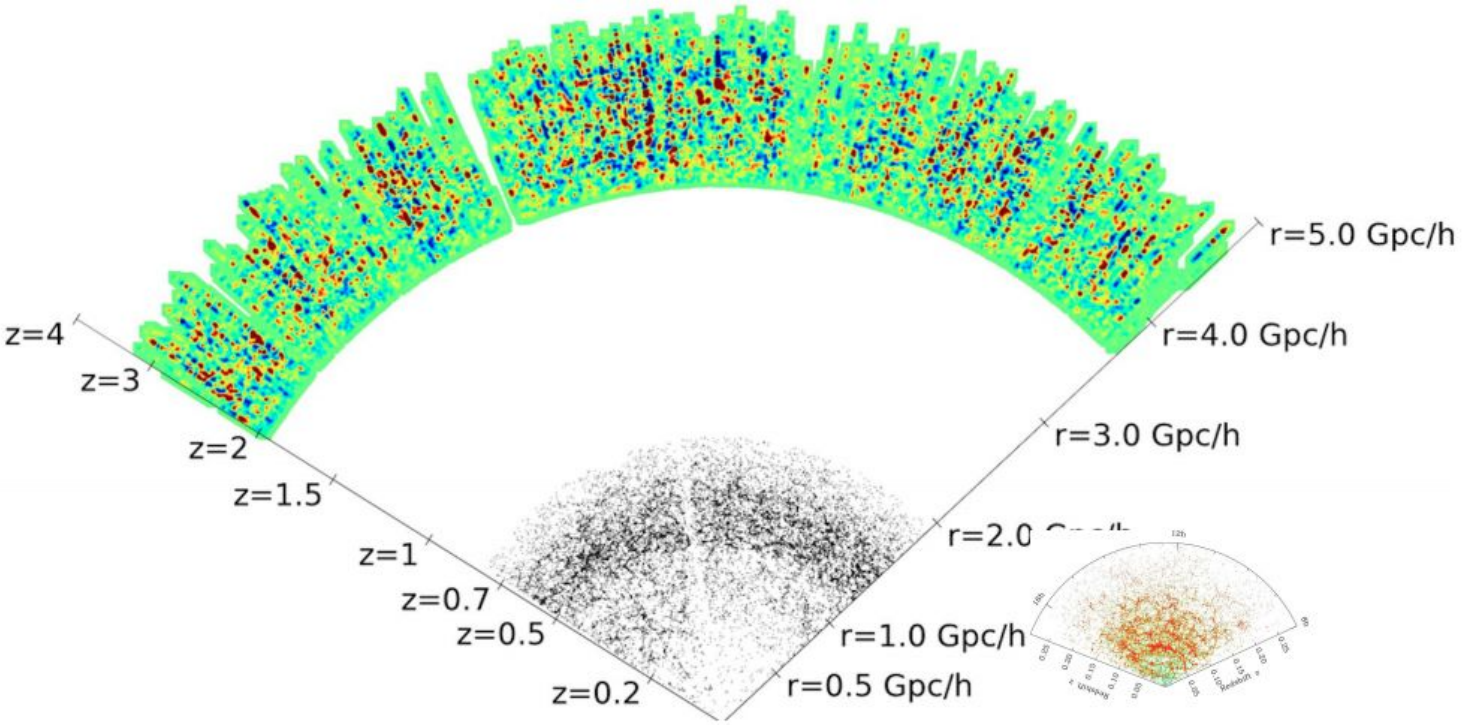
- **Observed** when the Universe **was only ~770 million years old (< 6% of its current age)**
- Provides a direct window into the Cosmic Dawn / Epoch of Reionization
- Era when the first galaxies and black holes began shaping the intergalactic medium.

$$1 + z = \frac{\lambda_{obsv}}{\lambda_{emit}} \qquad 1 + z \approx \frac{12200Å}{1520Å}$$

$$z \approx 7$$

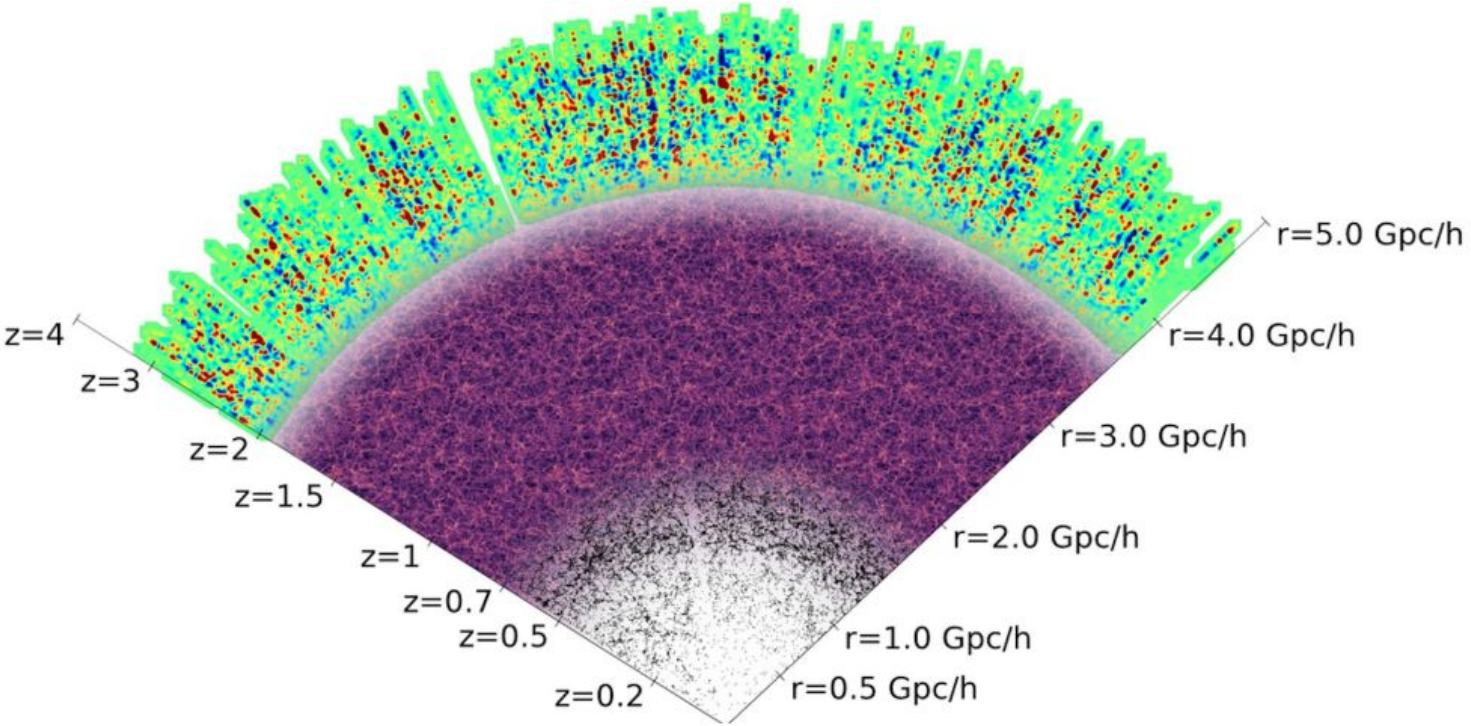
The final calibrated spectrum of DELS J003836.10–152723.6. Credit: Wang et al., 2018.

The evolution of the experiments



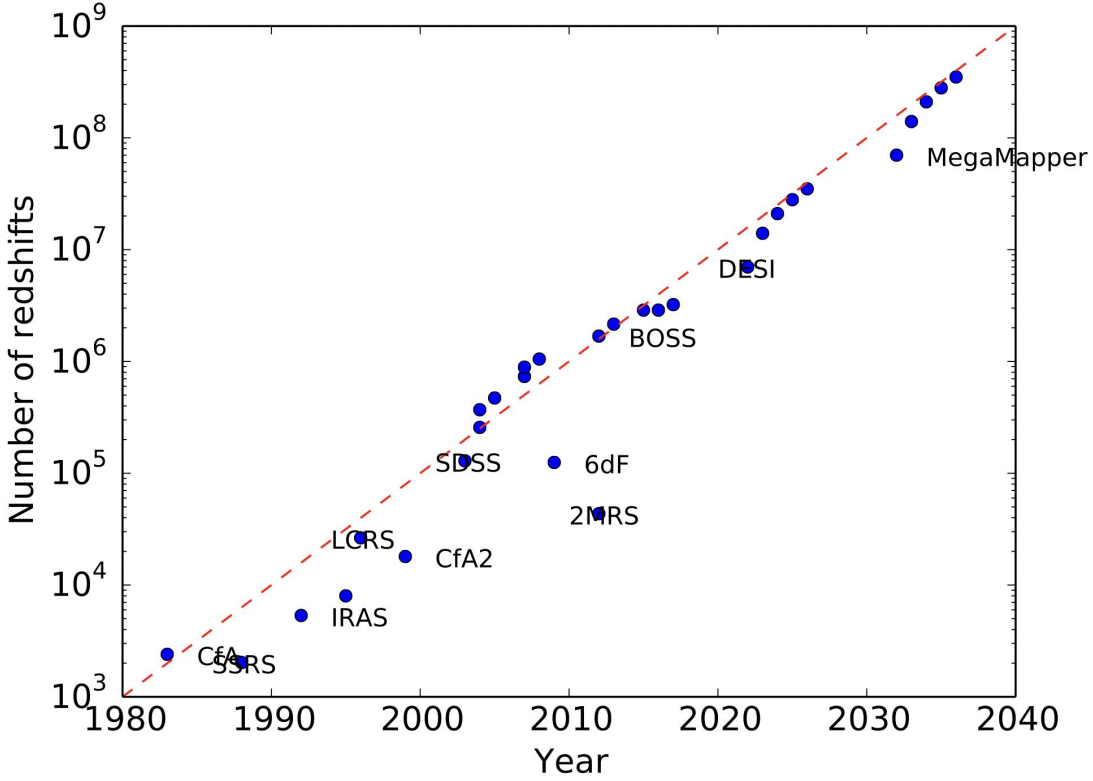
Source: http://dstn.astrometry.net/talks/2017-02-14-desi-northwestern.pdf

The experimental challenge of DESI



Source: http://dstn.astrometry.net/talks/2017-02-14-desi-northwestern.pdf

The experimental challenge of DESI

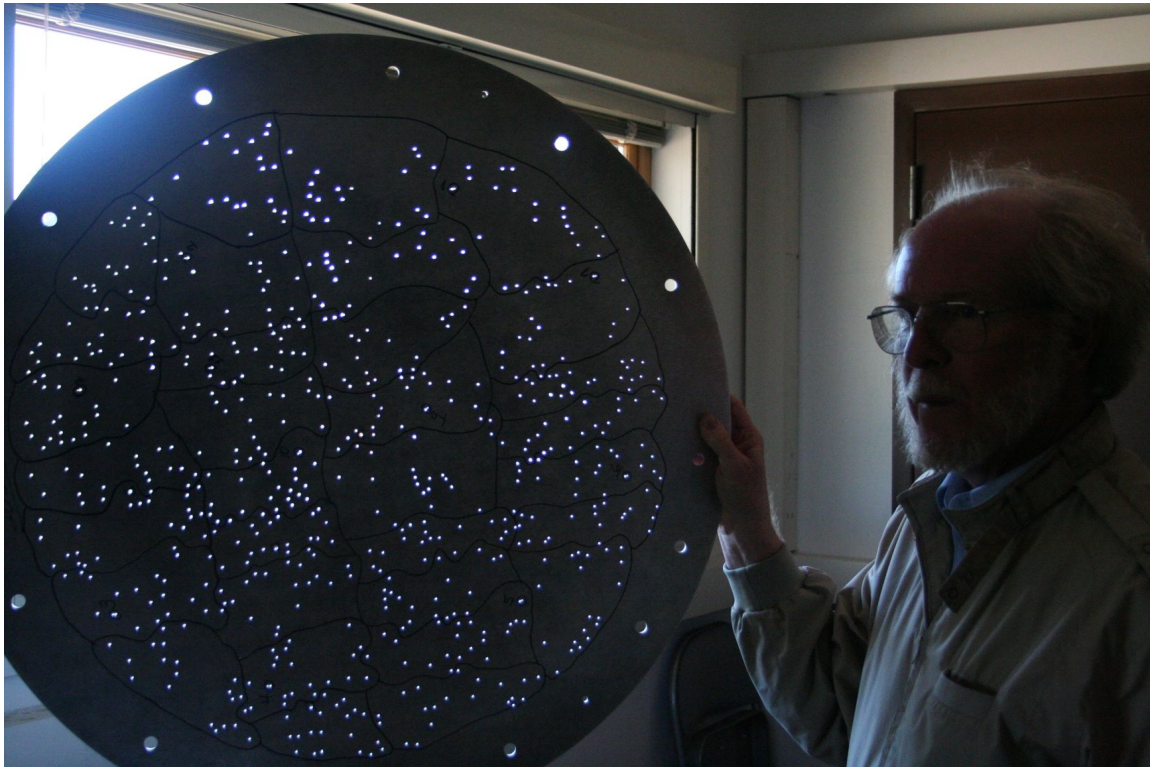


Schlegel et al. (2022), 2209.04322

The SDSS - BOSS technology

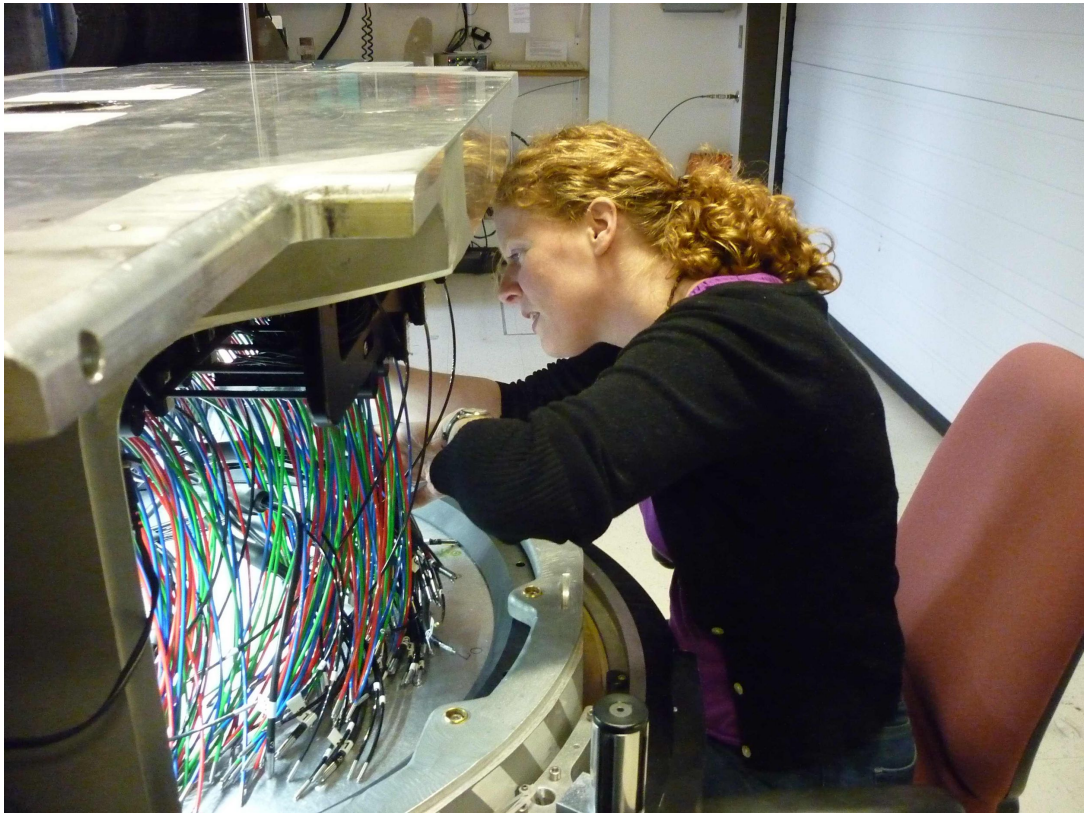
Source: https://pages.astronomy.ua.edu/keel/telescopes/apo.html

The SDSS - BOSS technology



Sclegel et al. (2022), 2209.04322

The SDSS - BOSS technology



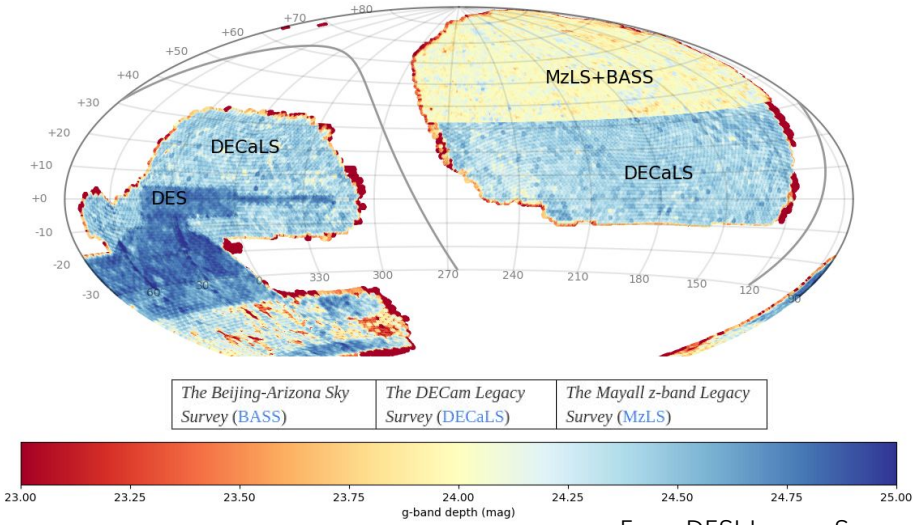
Source: https://blog.sdss.org/image-gallery/

The new DESI technology

Source: https://blog.sdss.org/image-gallery/

Legacy Imaging Survey (Photometric)

- BASS+DeCaLS+MzLZ(WISE)
- ⅓ of the observable sky
- Four Classified Galaxy Classes
  - Luminous Red Galaxies up to z = 1.0
  - Emission Line Galaxies up to z = 1.6
  - Quasars with z > 2.1
  - Bright Galaxies out to $z \approx 0.6$



From DESI Legacy Survey Webpage

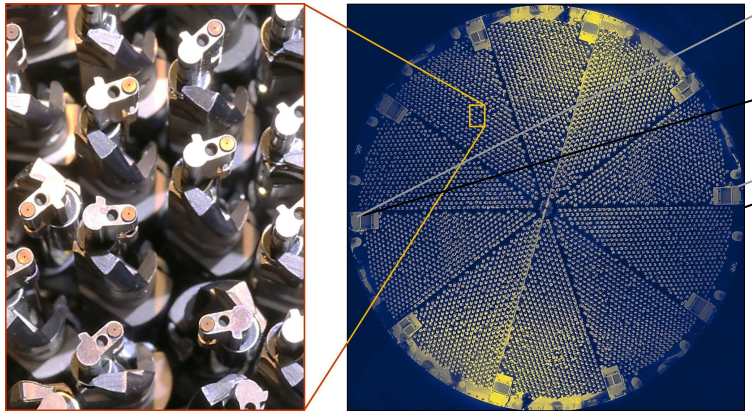
- Images of 256px * 256px (67.07 arcsec * 67.07 arcsec)
- Photometry of g,r,z and from four mid-infrared bands (at 3.4, 4.6, 12, and 22 μm)
- Magnitudes, photo-redshift and other physical features
- RA-Dec Coordinates

Legacy Imaging Survey (Photometric)

http://legacysurvey.org/viewer

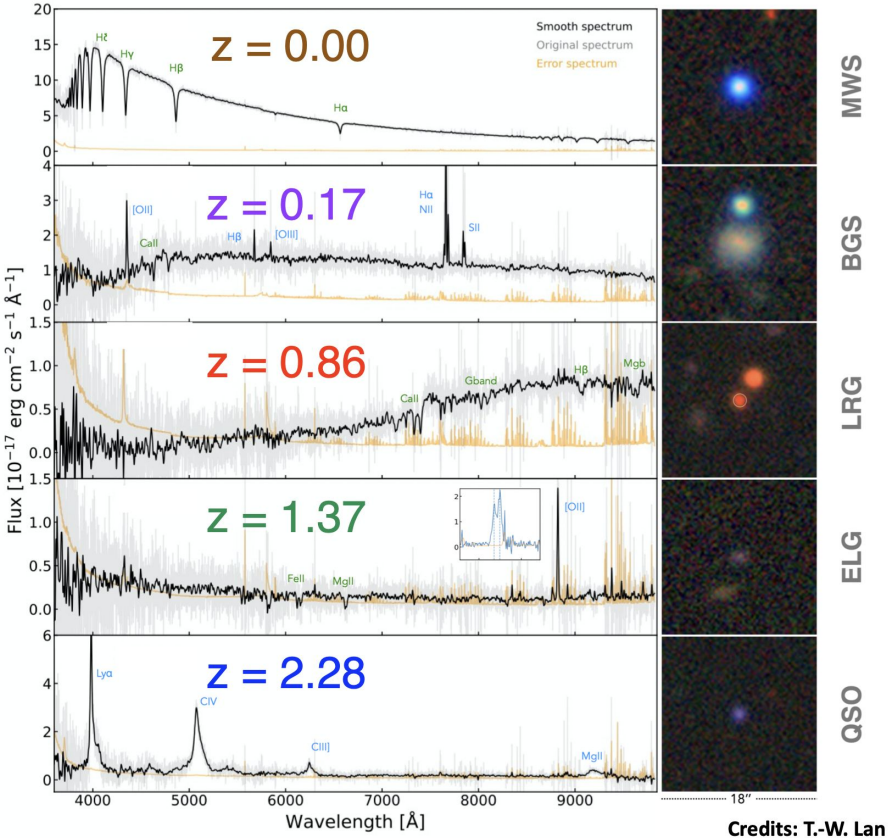
Dark Energy Spectroscopic Instrument (DESI)

- Mayall Telescope
- 5000 automatic positioners
- Fibers with 40m of length
- 3 bands B,R,Z
- 3600 Å - 9600 Å
- DESI goal is 40M of objects
- Spectroscopic redshift z_spec (redrock)



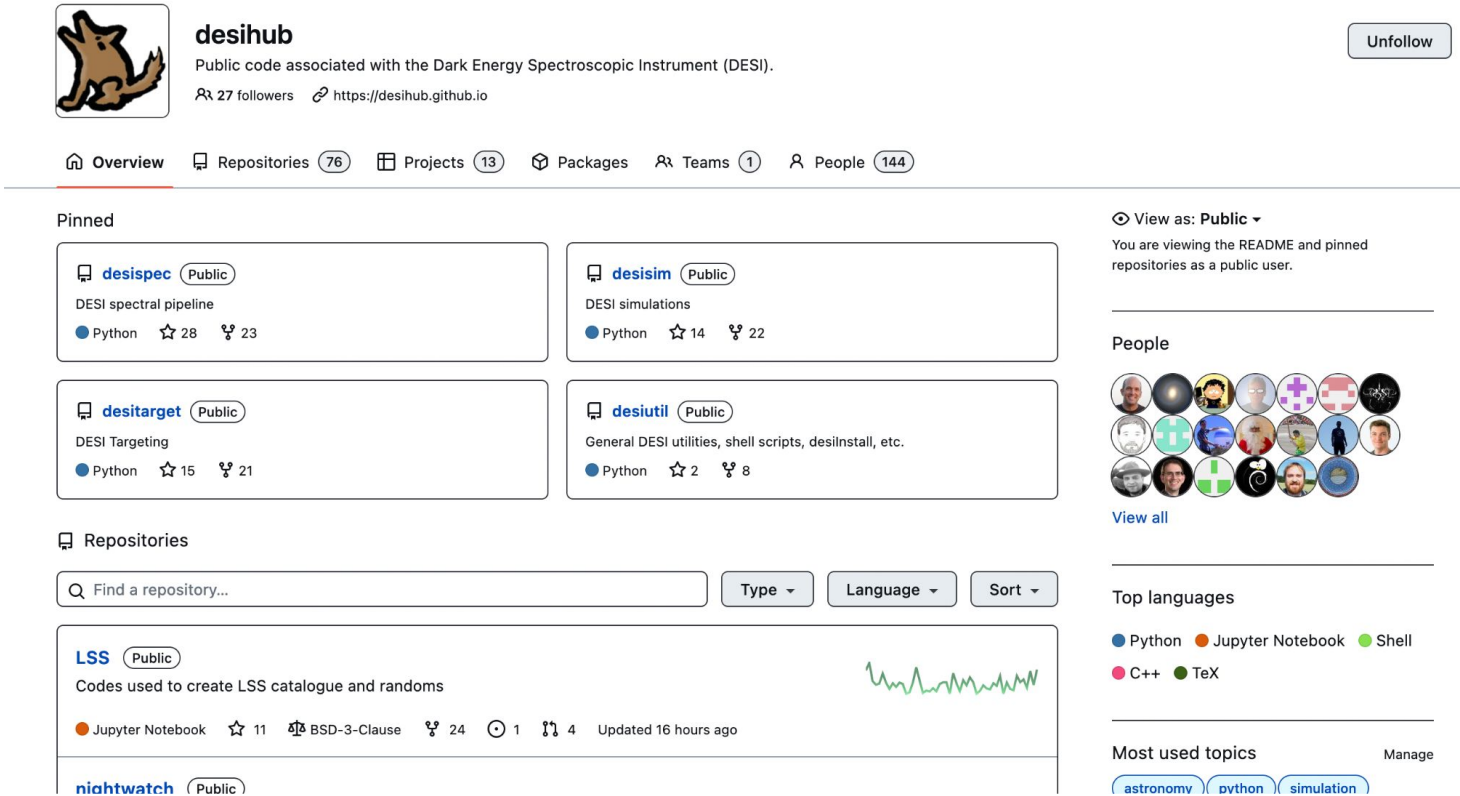
Credit to David Kirkby. From DESI Blog.

DESI objects sample

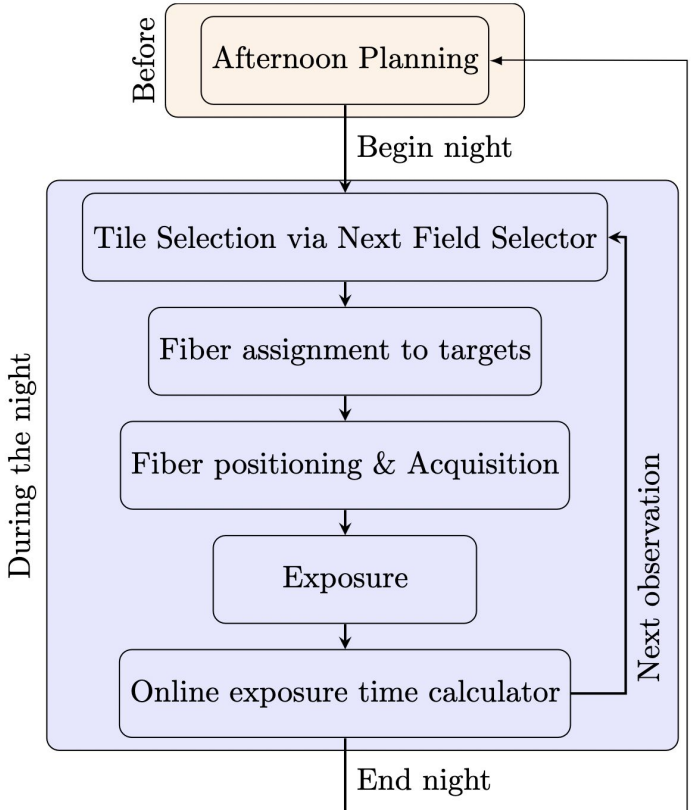


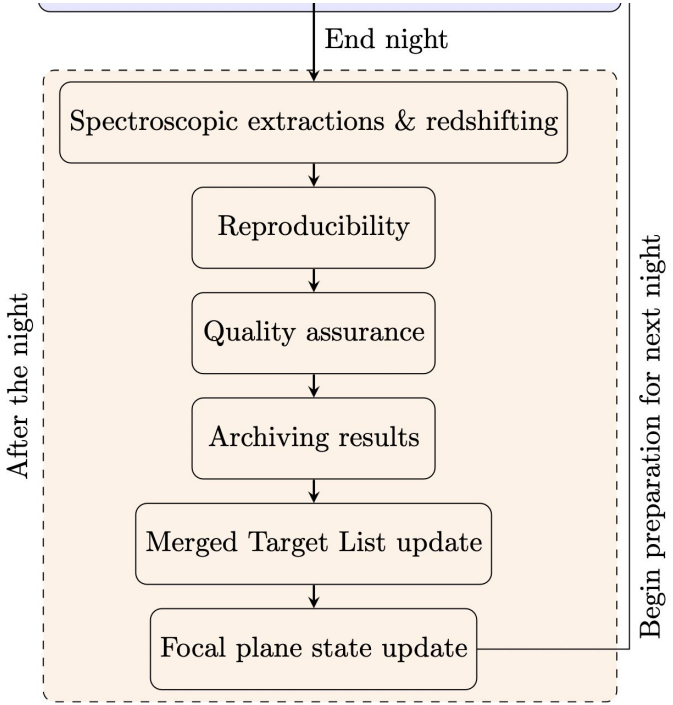
Credit to T-W Lan

The DESI data framework `desihub`



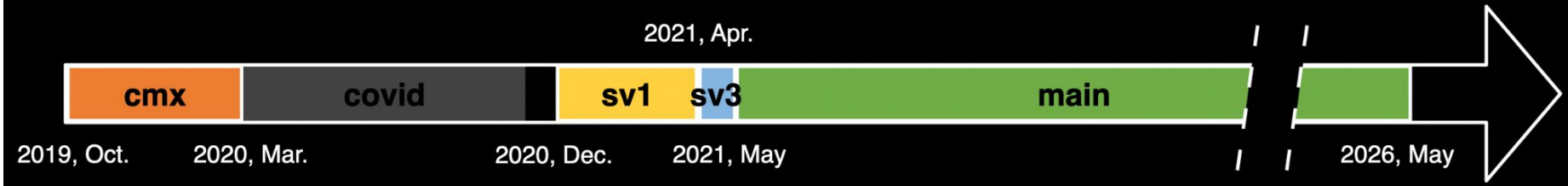
Daily Operations



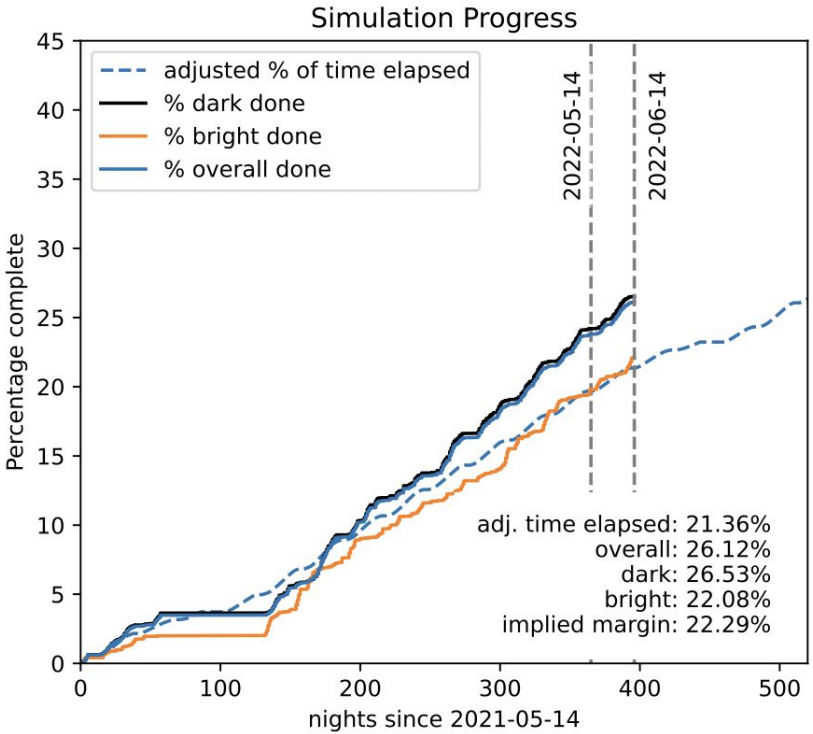


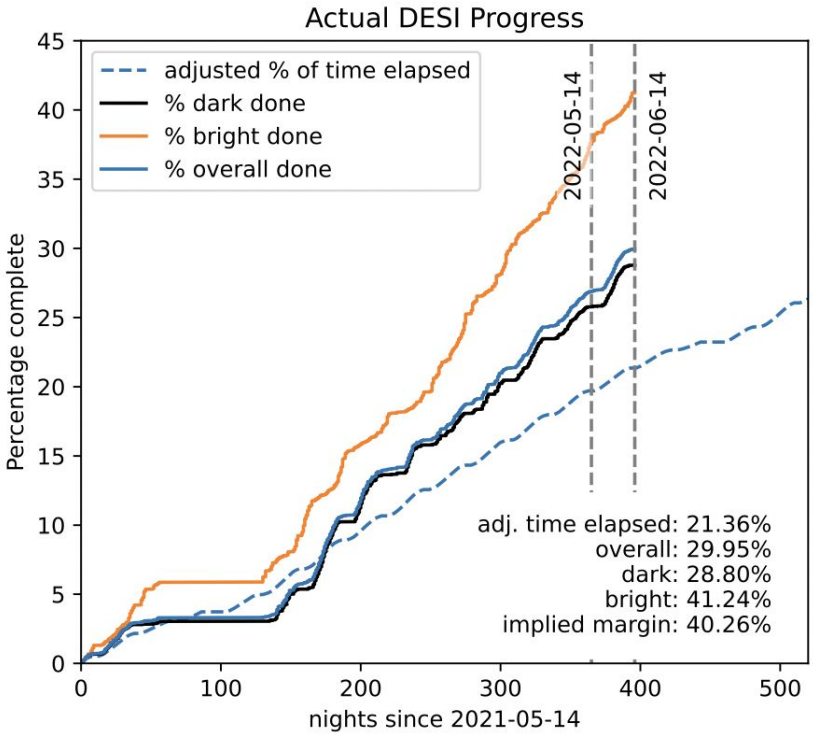
First Measurements (2021)

Phase		Short name	Goal
Commissioning		cmx	Validate Instrument
Survey Validation	Target Selection Validation	sv1	Validate Target Selections
	One Percent Survey	sv3	Final rehearsal + clustering for mocks
Main Survey		main	Constrain Dark Energy

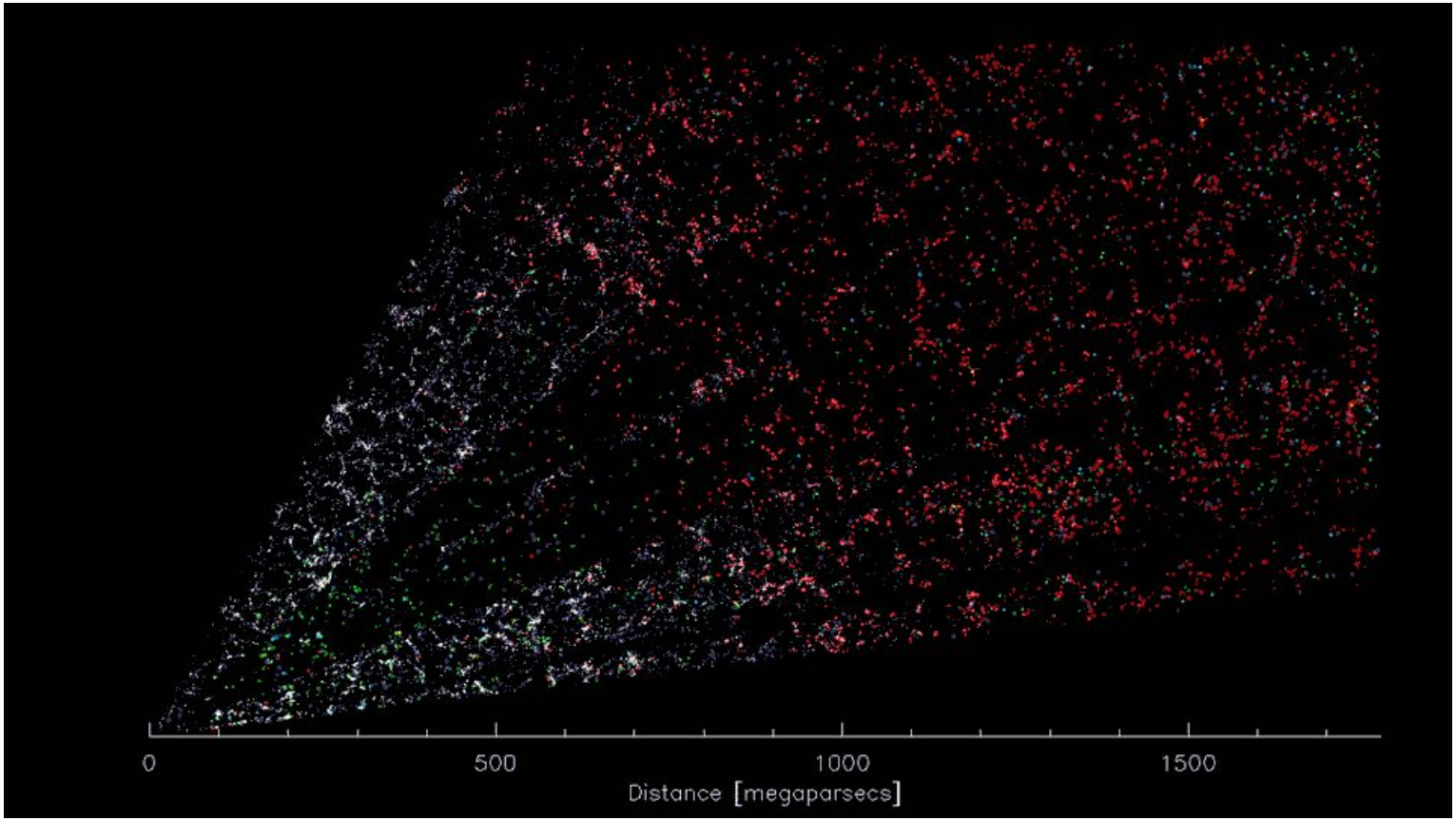


The Y1 progress

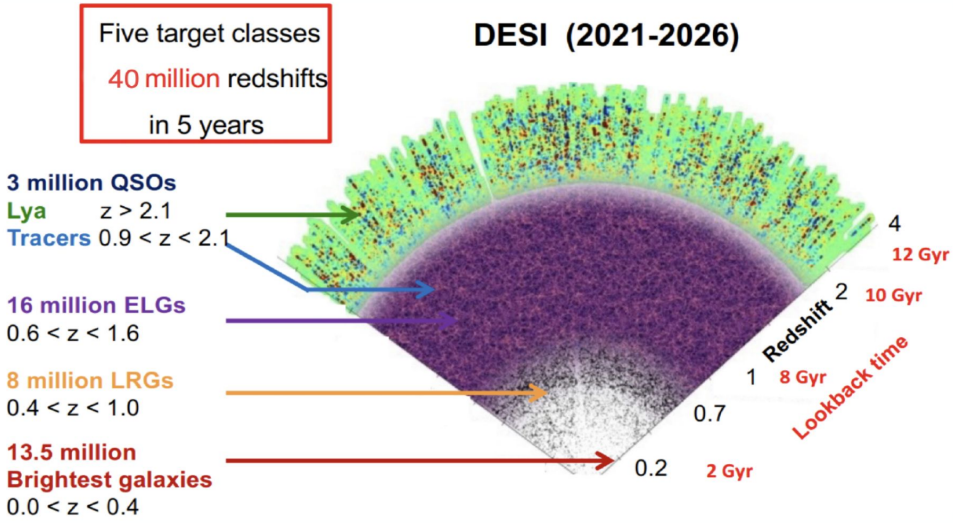


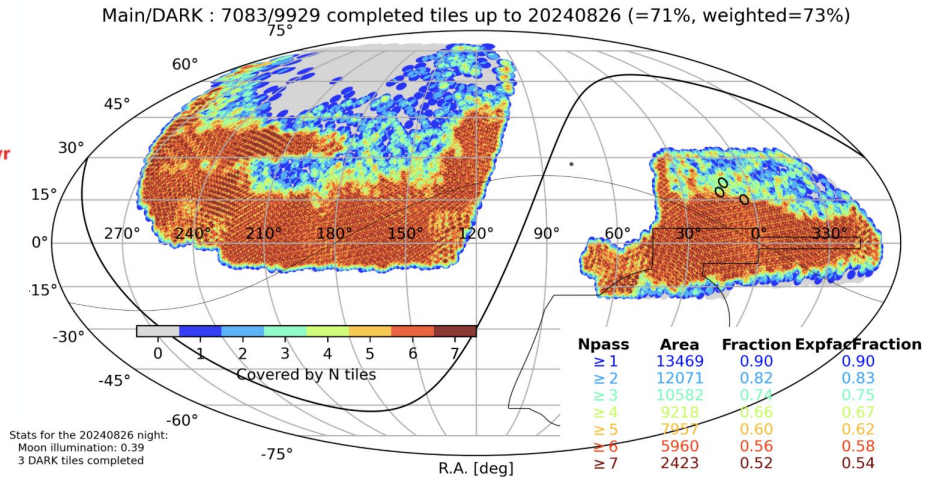


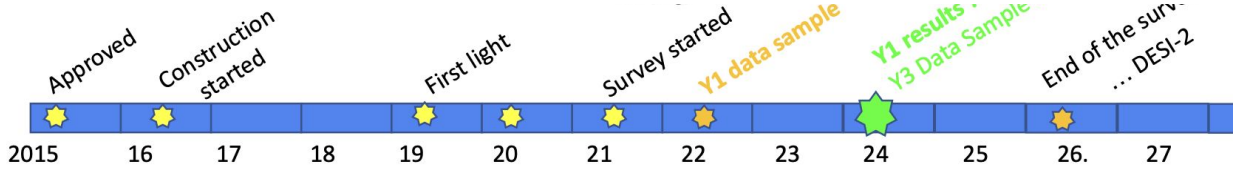
First results



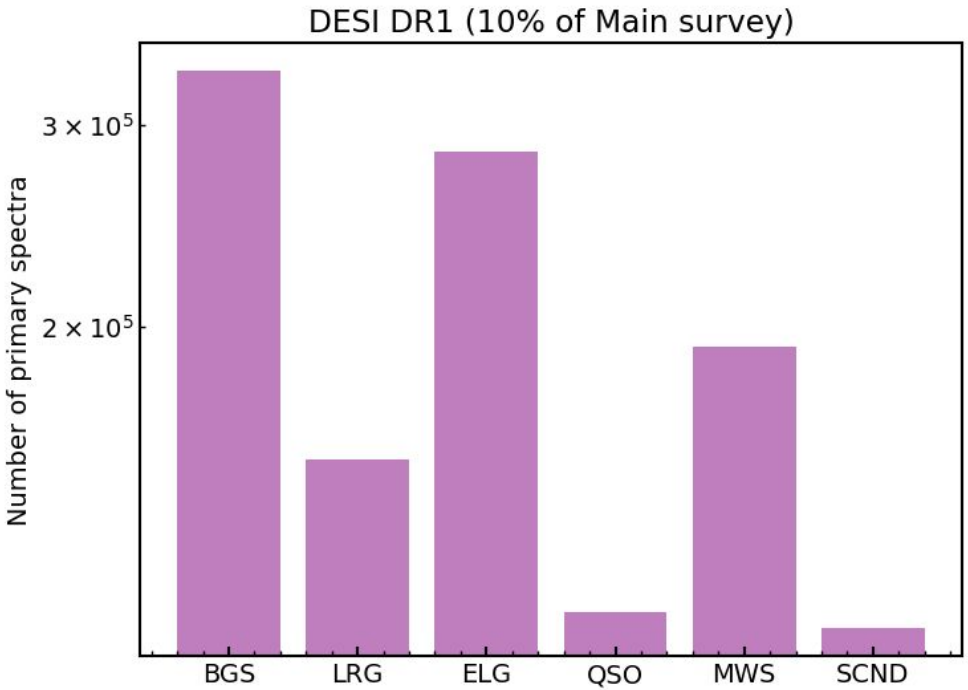
Current Progress

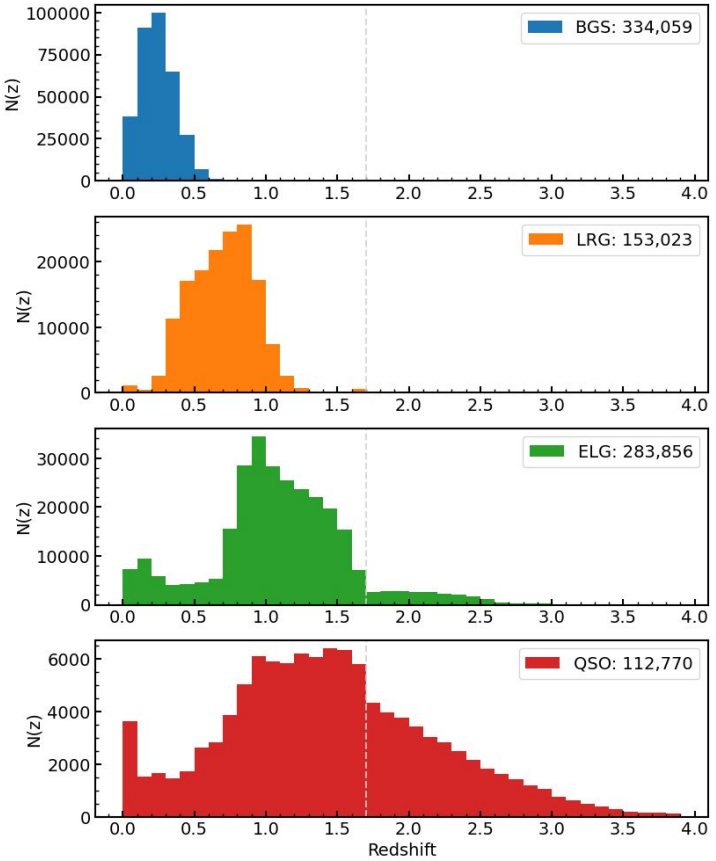






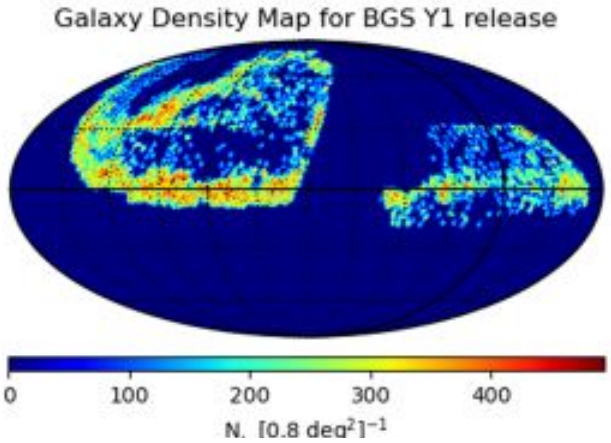
The Data Release 1 (DR1)

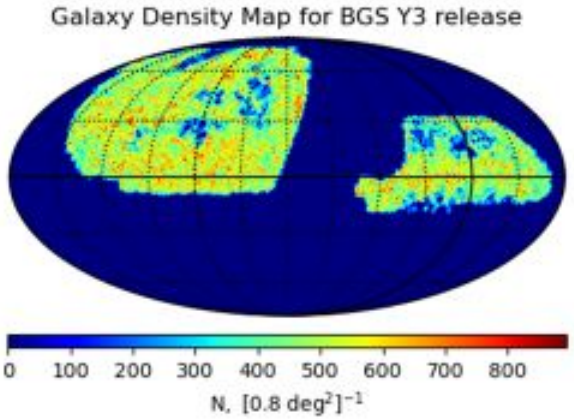




Bright Galaxies Survey

BGS: Bright Galaxies → z<0.5





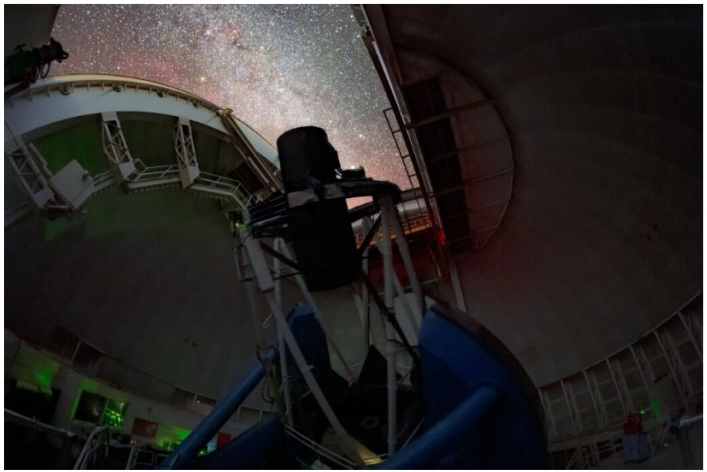
DESI's Observational Capacity

- The spectral collection rate of DESI
  - ~5000 spectra per exposure (~20 minutes)
  - In good conditions: 100,000 - 150,00 spectra per night.
  - Record-setting nights: more than 40 reconfigurations in a single session.

A Record-Breaking Night

Joan Najita (NOIRLab)

12 February 2024 was a spectacular night for DESI: it broke its own record and acquired nearly 200,000 redshifts in a single night. The figure is remarkable, especially in the context of history.



The Dark Energy Spectroscopic Instrument (DESI) observing the night sky on the Nicholas U. Mayall 4-meter Telescope at Kitt Peak National Observatory in Arizona. Credit: KPNO/ NOIRLab/NSF/AURA/T. Slovinský

Credit to DESI website "A record-breaking night" post

DESI's 50M Milestone

- Quanta Magazine 2024: DESI named among the Biggest Breakthroughs in Physics.
- **Highlight:** Possible evidence that dark energy may be weakening.
- First-year data (Y1) → largest 3D map of the Universe to date.
- More results expected in 2025 with 3-year dataset.

> In April, physicists detected a hint of a signal suggesting that dark energy, the mysterious energy of space itself, may be weakening. "Hint" is the preferred term because the sign in the heavens isn't quite robust enough to be called "evidence," to say nothing of "discovery." Astrophysicists used the Dark Energy Spectroscopic Instrument (DESI) to map millions of galaxies at different distances in space and time, and from this map they inferred how the universe has expanded over its history. The data confirmed — as we've known since 1998 — that the cosmos's expansion is accelerating, driven by what we call dark energy. But DESI's data hints that the rate of acceleration has been dropping.

Credit to DESI website "A record-breaking night" post.

DESI's Current Progress

- 50 million spectra measured (as of Dec 2024) (36.3 M galaxies & quasars. 13.7 M stars.)
- Data Release 1 (DR1): 18 million unique objects.
- Data Release 2 (DR2, March 2025): includes the first 3 years of observations.
- Backup program: ~7 M Milky Way stars observed (1.2 M in DR1).

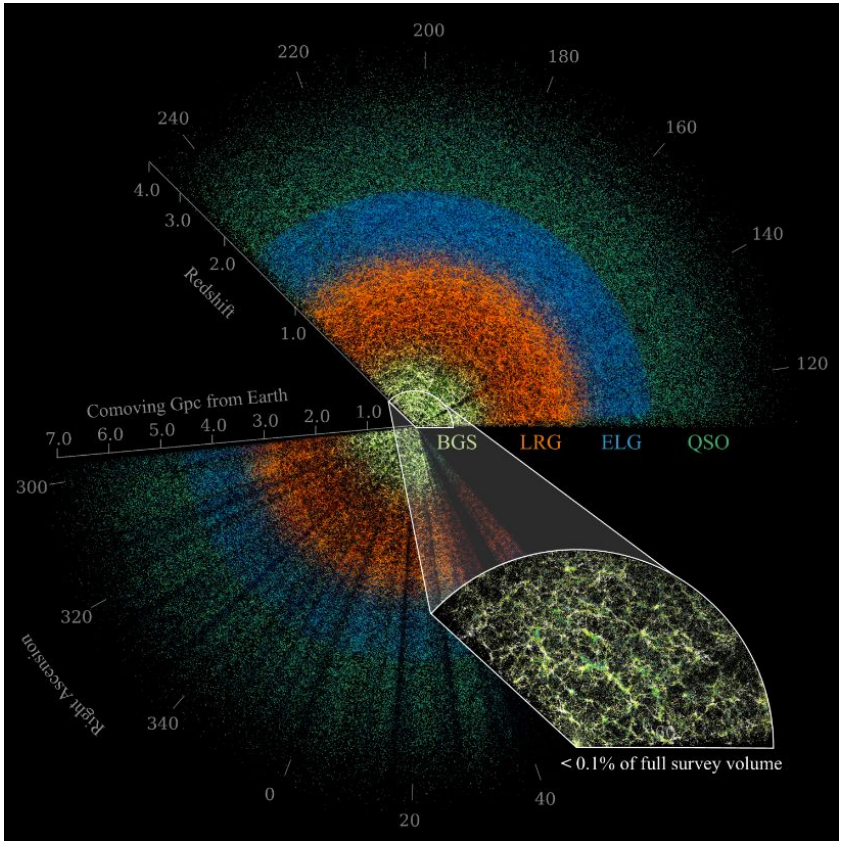


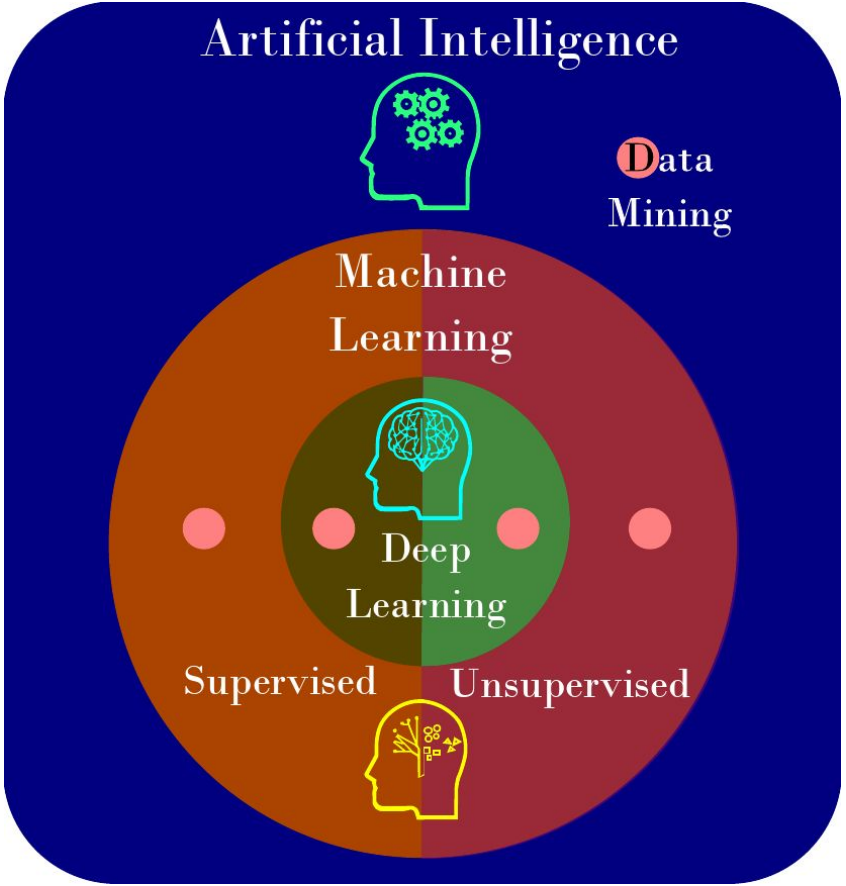
Image from [Data Release 1 of the Dark Energy Spectroscopic Instrument] paper.

How to analyze all of these data?

AI-based
Approaches to DESI data

What is AI?

- Simulating of human intelligence in machines (visual perception, speech recognition, decision-making, language processing).
- Algorithms and models that can process large amounts of data, trained with techniques as **Supervised** or **Unsupervised** learning.
- Main subfields of AI are Machine & Deep learning. Data Mining cut across both.



Supervised & Unsupervised Learning

Supervised

- Used for classification or regression tasks.
- Requires to use labels to make predictions.

ML Algorithms

* Support Vector Machines
* K-nearest neighbors
* Decision Trees
* Random Forest.

DL Algorithms

* Multi-Layer Perceptron
* Convolutional Neural Network
* Recurrent Neural Network
* Transformers

Unsupervised

- Used for clustering or dimensionality reduction tasks.
- Doesn't require to use labels. Used to find patterns.

DL Algorithms

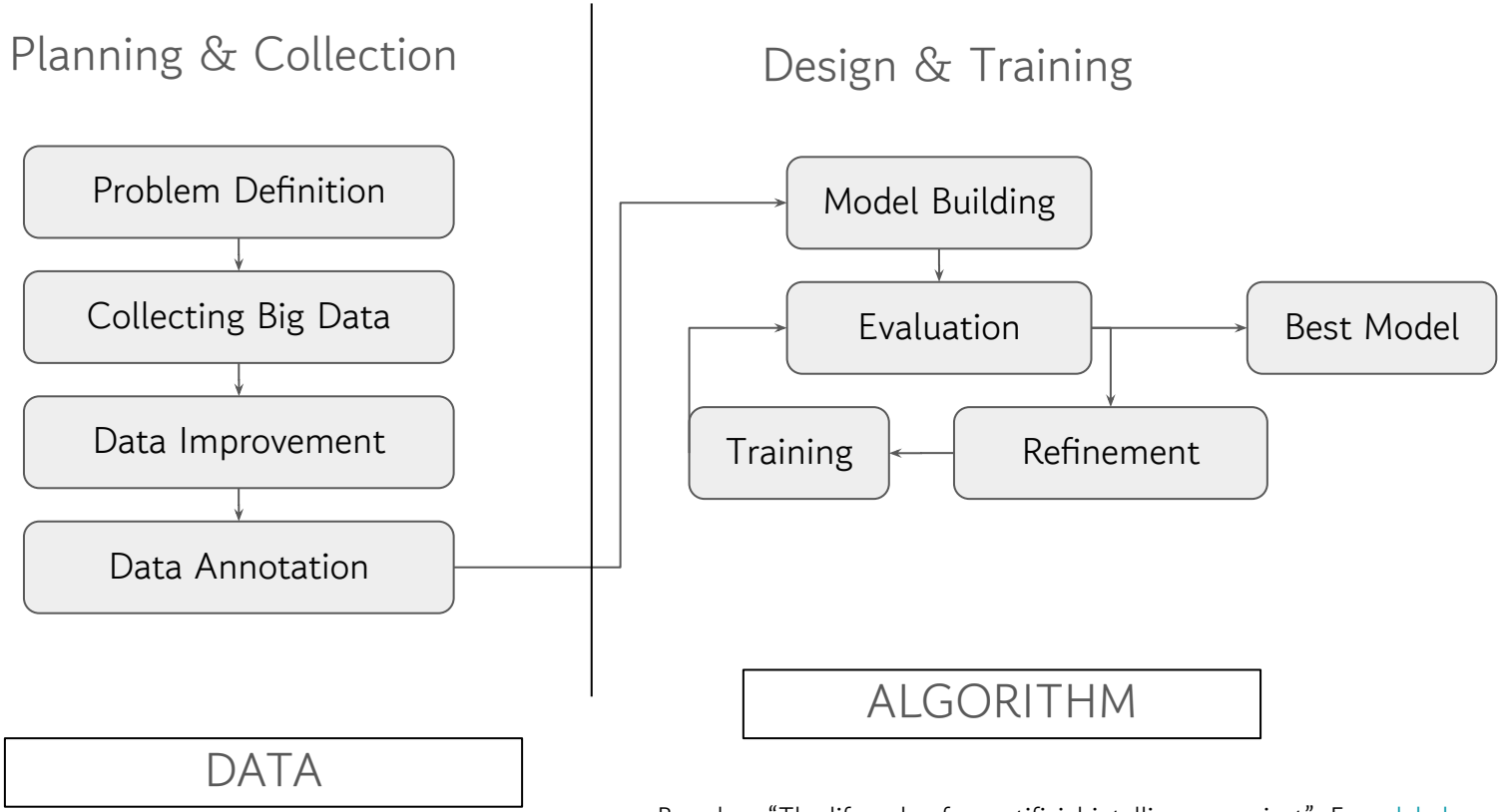
* Autoencoders

ML Algorithms

Clustering:
* K-means clustering
* DBScan
* Gaussian Mixture Models

Dimensionality Reduction:
* PCA
* Isometric Map
* T-SNE
* UMAP

AI project stages



Based on “The lifecycle of an artificial intelligence project”. From labelyourdata.

Learning-based Approaches

I. Assessing the quality of DESI Spectroscopic Survey

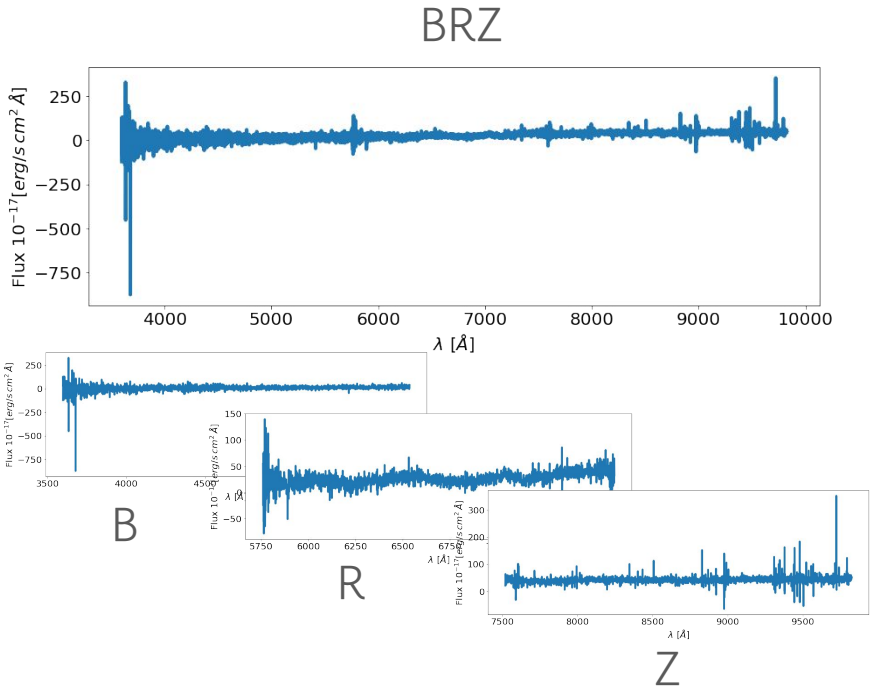
The Anomaly Detection Challenge

- Detecting instrumental errors in (near) real-time.
- On massive surveys an automated algorithm is necessary (learning-based).
- On DESI, the *Quality Assurance step* assess the quality of DESI observations.
- I proposed a learning-based method to identify outliers.

- Examination of the QA plots for each observed region to identifies any notable aspects (regions that present abnormal behaviors and potential issues) labeled as **unsure** .
- Most common are extremely bright stars, minor air leaks within the spectrograph, turbulence, albeit occurring rarely.
- Imperfect sky subtraction under intense brightness conditions can yield not optimal redshift measurements.
- Only good observed tiles are added to the MLT.

The Outlier Detection Method (Data)

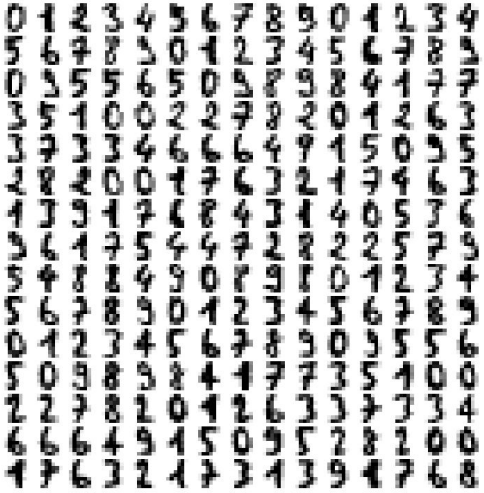
- I used observations from the first observation year to test and validate the outlier detection method.
- Each spectra has a length of ~2.7K points (by band) an ~8K in BRZ.
- Applied by observation night (~150K observation by night).

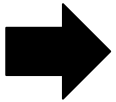


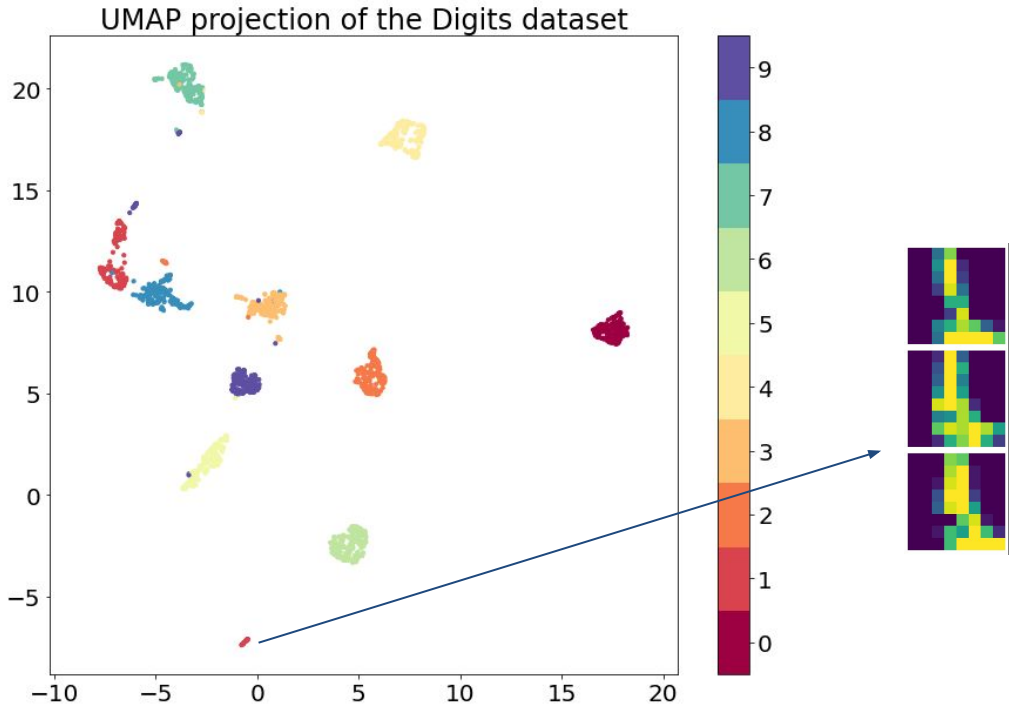
The Outlier Detection Method

- Uniform Manifold Approximation and Projection (UMAP) algorithm.
  - Unsupervised ML algorithm.
  - No linear dimensionality reduction for visualization, and outliers detection.

64-Dimension (8x8)







[1802.03426] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

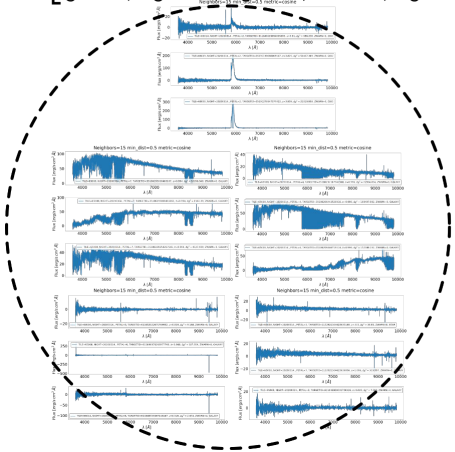
The Outlier Detection Method

1) Dimensionality Reduction Algorithm (UMAP)
2) Group finder (FoF)

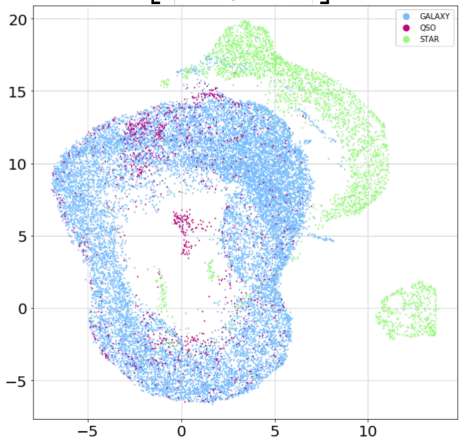
UMAP
>Number of neighbors (Nn): [0 - 50] steps 5
>Minimal Distance (Md): [0.0 - 0.5] steps 0.1
>Metrics (Me): [‘euclidean’, ‘cosine’, ‘braycurtis’]
>Dimension (D): 2

FoF (Friend of Friends)
>Linking Length (Ll): [0.1 - 0.5] steps 0.1

$[f_1, f_2, f_3, \ldots, f_n]$



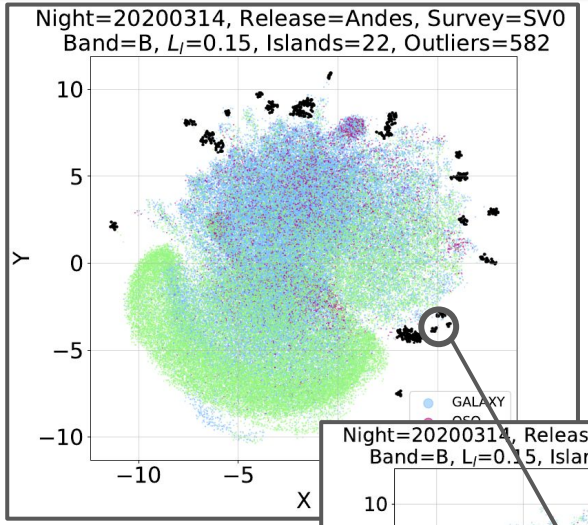
$[x_1, x_2]$



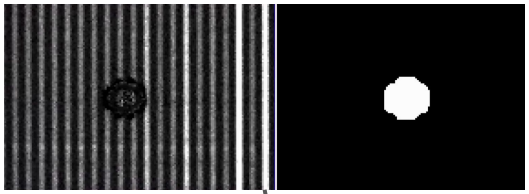
- I used observations from the first observation year to test and validate the outlier detection method.
- Applied by observation night (~150K observation by night).
- Each spectra has a length of ~2.7K points (by band) an ~8K in BRZ.

jsuarez314/assessing_DESI_survey

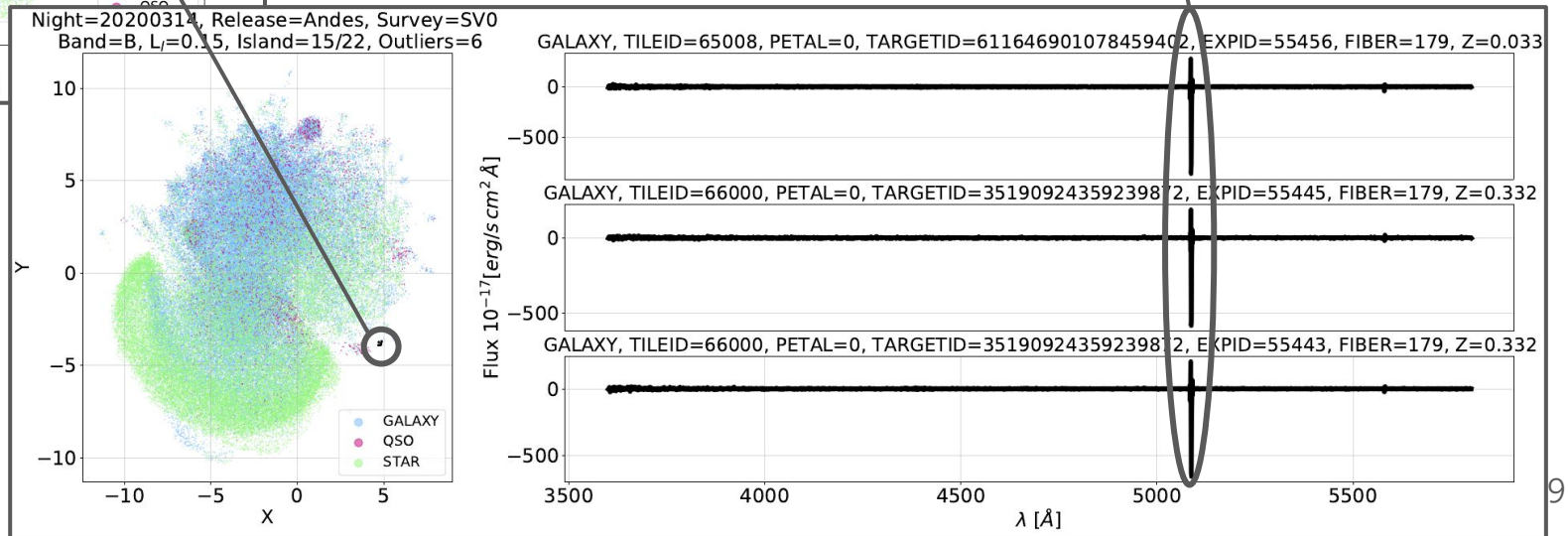
Quality Assessment



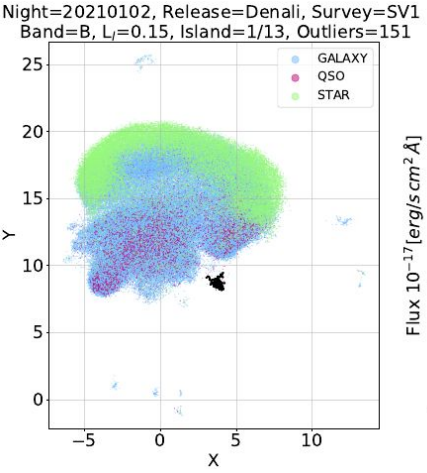
Desispec Issue Reported on Github as **defect in the CCD** , #983

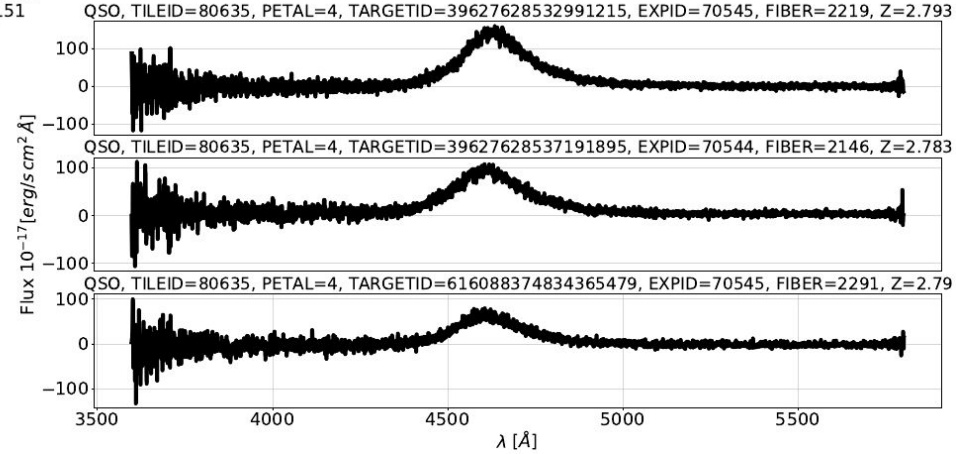


Me = ‘cosine’
Nn = 45
Md = 1.0
Ll = 0.15

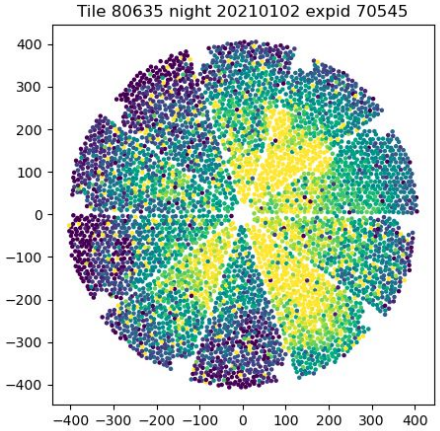


Quality Assessment



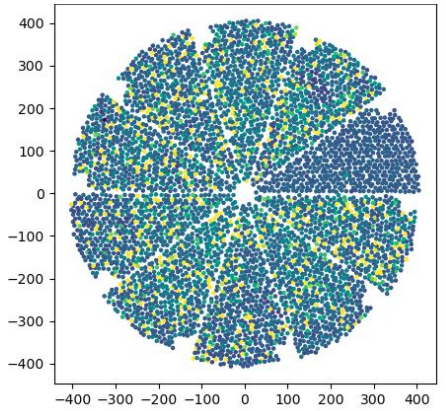


Fibers contaminated



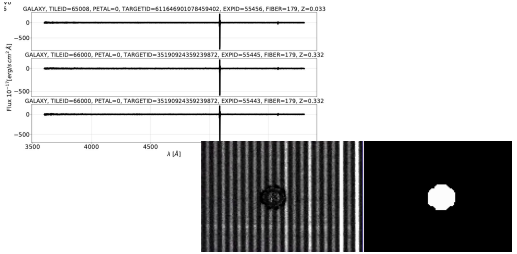
DesiSpec Issue Reported on Github as **contamination** , #1262

Fibers not contaminated

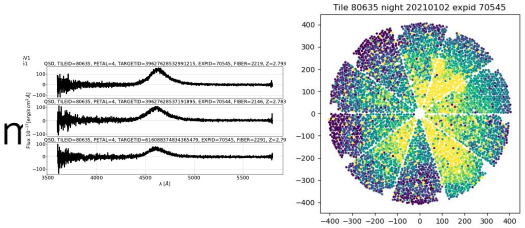


Results

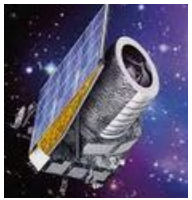
- This learning-based method using UMAP and FoF is a good strategy for assessing the quality of massive spectroscopic surveys. Identification of instrumental errors show the power of our method.



- This method is promissory to detect outliers and could be added to the pipeline analysis as an early alert system to identify instrumental errors.



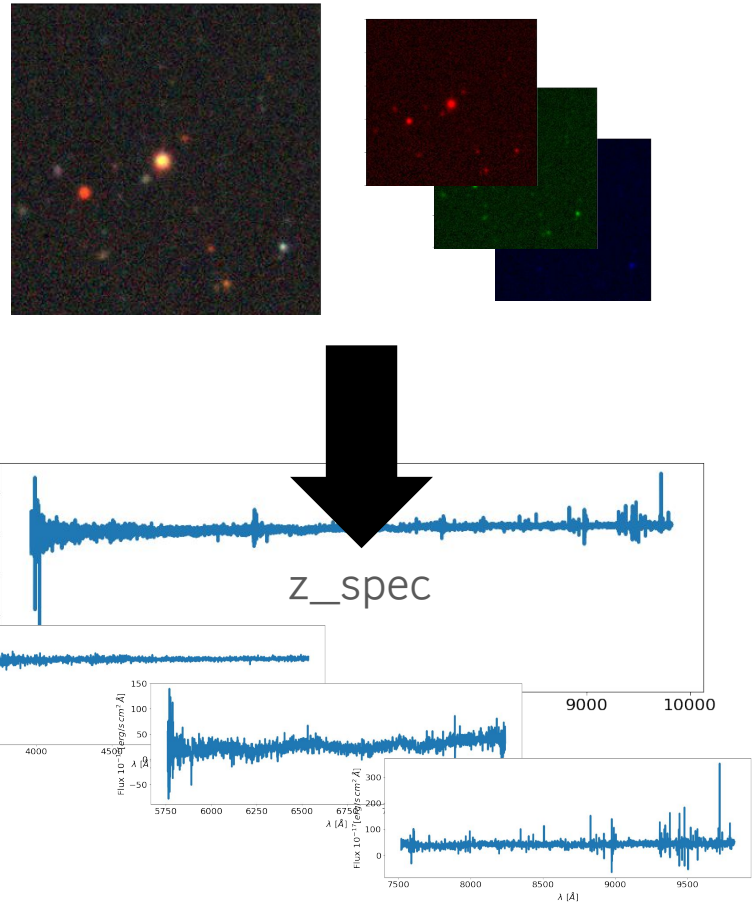
- This mechanism could be implemented in the QA DESI validation data process.
  In future surveys like the WEAVE, Subaru/PSF, 4MOST, Euclid and others.



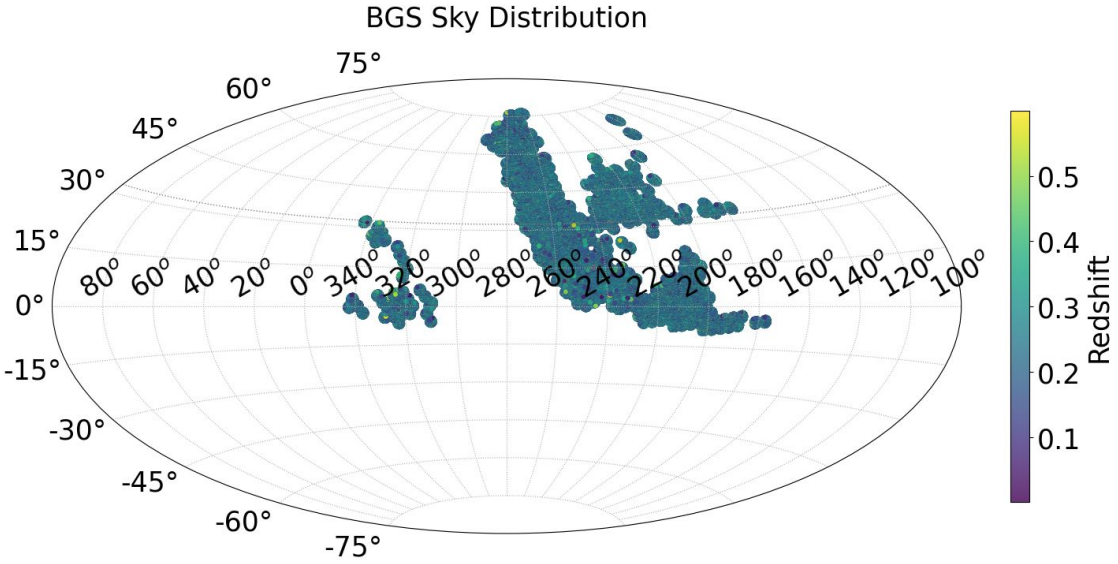
II. Deep-PhotoZ: Predicting the Photo-redshift of Bright Galaxies (Friday)

The Photometric-Redshift Challenge

- Redshifts are really important in astrophysics.
- Train an algorithm that learns from photometry to predict z_spec.
- In contrast with spectroscopic surveys, photometric are cheaper and faster.
- Spectroscopic just for pre-selected galaxies due limitation approach.
- Future massive photometric surveys & photo-z estimation requirements.



DESI - Bright Galaxy



- First study focused on Bright Galaxies (Nearest)
- 820,455 Bright Galaxies
- $z<0.6$ - Mean: 0.23

Learning-based Strategies

Using photometric features
+ ML algorithm
(Current State of Art)

Using pixel-level (imaging)
+ DL algorithm
(Proposed)

Photo-z estimation metrics

- *Prediction bias* defined as the average value of the prediction error.

$$bias = \left\langle \frac{\Delta z}{1 + z_{spec}} \right\rangle$$

$$\Delta z = z_{phot} - z_{spec}$$

- Normalized Median Absolute Deviation (NMAD).

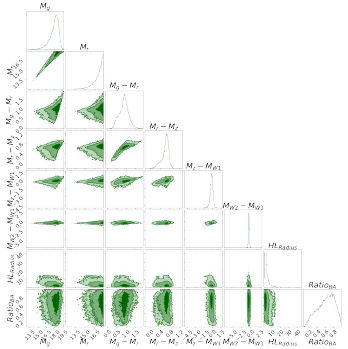
$$\sigma_{NMAD} = 1.4826 \times \text{Median}\left(\left| \frac{\Delta z}{1 + z_{spec}} - \text{Median}\left(\frac{\Delta z}{1 + z_{spec}} \right) \right| \right)$$

- Fraction of Outliers (fout) defined as the fraction of z predictions with error greater than ϵ.

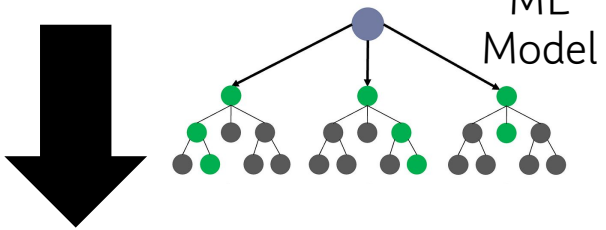
$$f_{out} = \left| \frac{\Delta z}{1 + z_{spec}} \right| > \epsilon$$

Photometric Feature-Based Strategy

- Extract some photometric-features to train a classical ML model (Catalogs from DESI LEgacy Survey DR9).
- Random Forest & Gradient Boosting Regressors:
  - >Number of Estimators (NE): From 20 to 400 in steps of 5
  - >Max Depth (MD): From 10 to 30 in steps of 5.
- Training: 70%, Validation: 20% & Testing: 10%.



Photometric Extraction



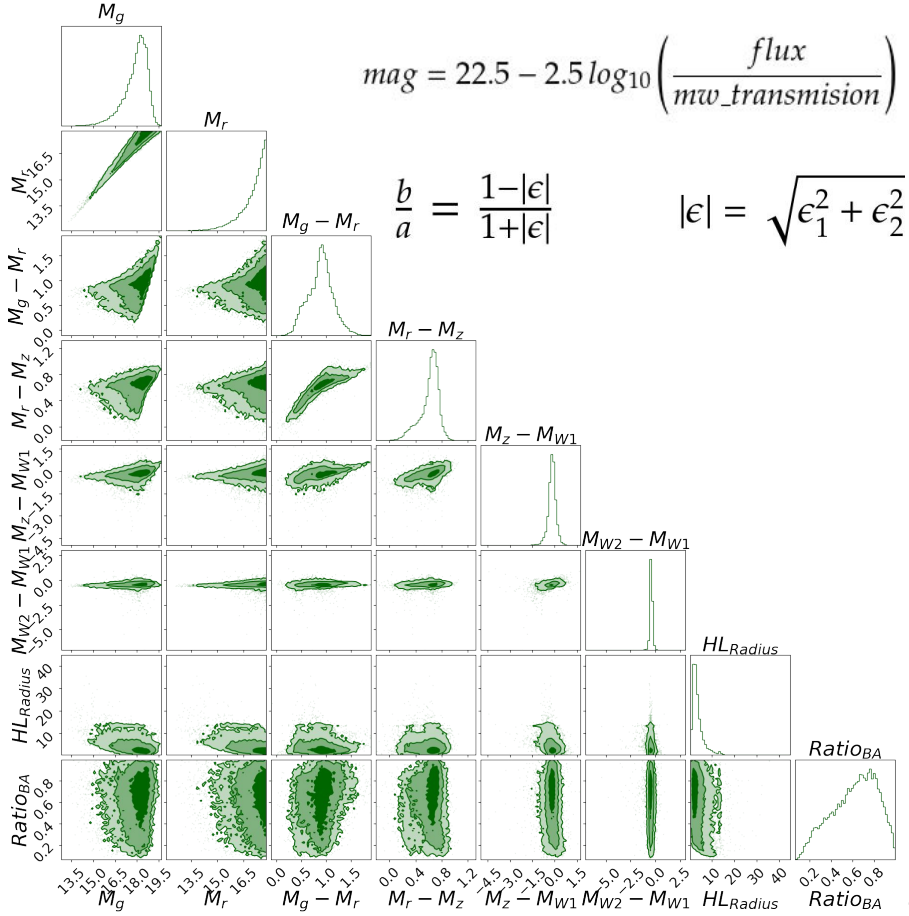
Redshift (spectroscopic)

Photometric Feature-Based Strategy

- Mg : The magnitude in the g-band.
- Mr: The magnitude in the r-band.
- Mg – Mr : The difference between the magnitude in the g-band and the magnitude in the r-band.
- Mr – Mz: The difference between the magnitude in the r-band and the magnitude in the z-band.
- Mz – MW1 : The difference between the magnitude in the z-band and the magnitude in the W1-band.
- MW1 – MW2 : The difference between the magnitude in the W1 -band and the magnitude in the W2-band.
- HLradius: radius from within which half of the galaxy light is contained.
- RatioBA: The ratio between the main axis of the galaxy.

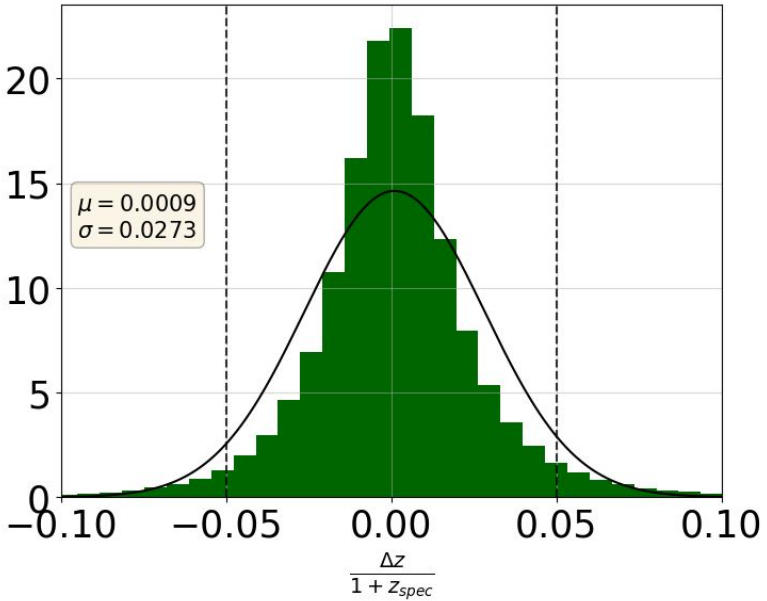
$$mag = 22.5 - 2.5\log_{10}\left(\frac{flux}{mw\_transmision}\right)$$

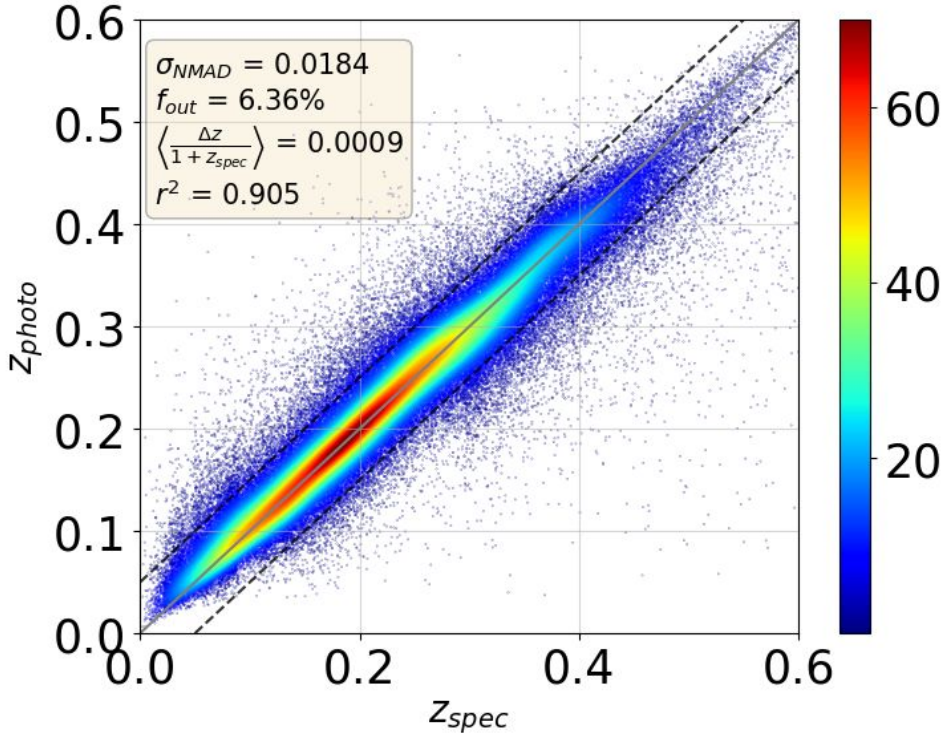
$$\frac{b}{a} = \frac{1-|\epsilon|}{1+|\epsilon|} \qquad |\epsilon| = \sqrt{\epsilon_1^2 + \epsilon_2^2}$$



Photometric Feature-Based Strategy

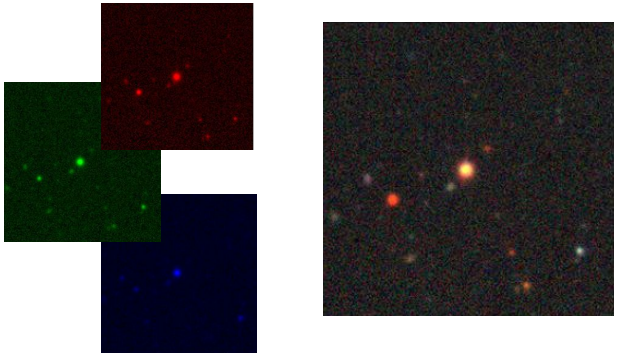
- Best result for the Random Forest model with NE=400 and MD=30.



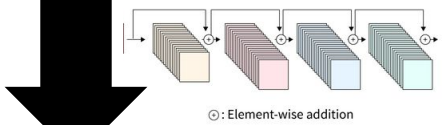


Deep-Photo Z model

- 3 color channels
- Using known NN structures, as CNN, ResNet & Transformers.
  - ResNet 18, ResNet 50, ResNet 152.
  - batch_size: 8, 16, 32 & 64.
  - learning_rate: 0.1 & 0.01
  - drop_out: 0.1, 0.3 & 0.5
- Transfer Learning:
  - layers_frozen: 0, 3 & 6 (Pretrained model on Imagenet)
- Training: 70%, Validation: 20% & Testing: 10%.



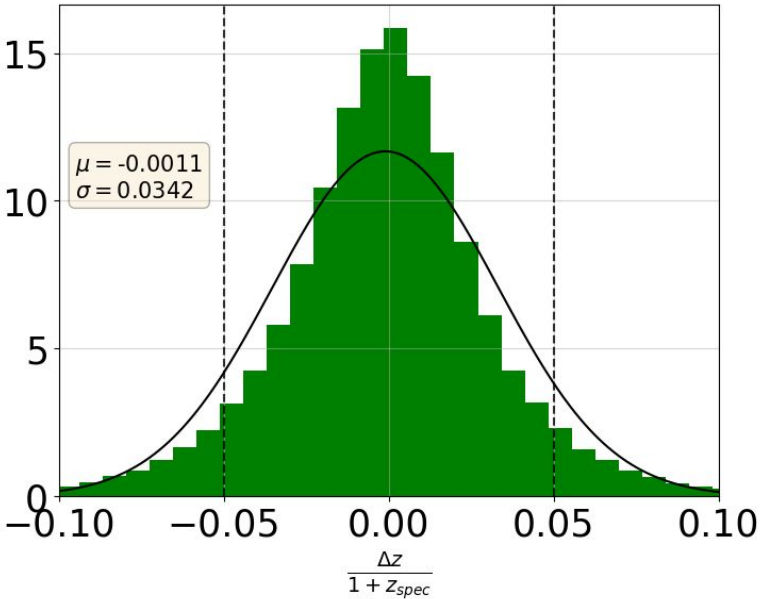
DL Model

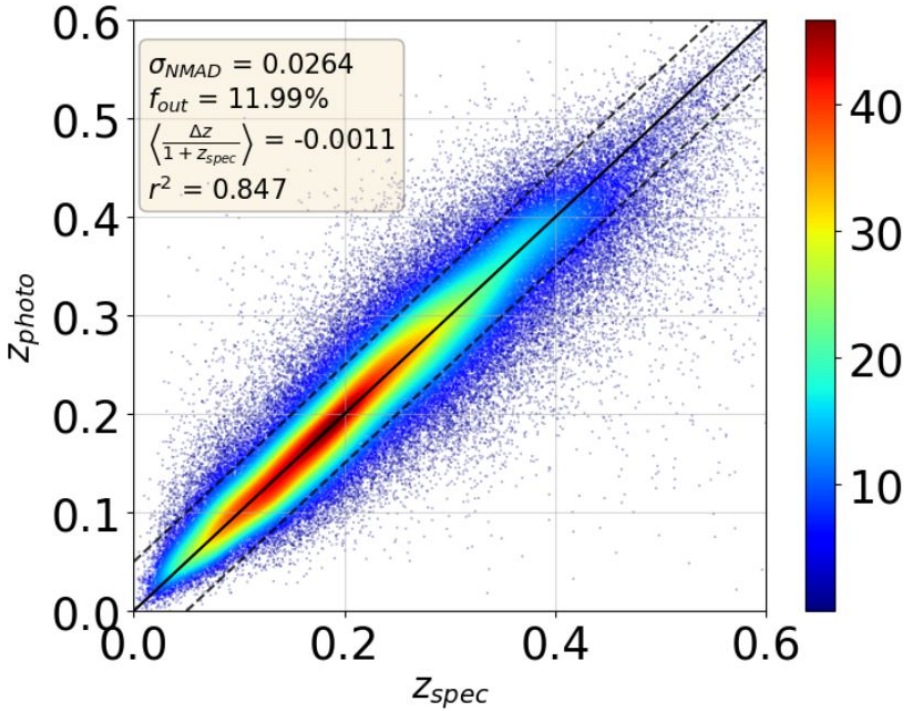


Redshift (spectroscopic)

Deep-Photo Z model

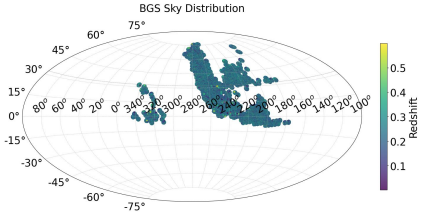
- Best result
  - ResNet 18, ResNet 50, **ResNet 152** .
  - batch_size: 8, 16, 32 & 64.
  - learning_rate: 0.1 & 0.01
  - drop_out: 0.1, 0.3 & 0.5
  - layers_frozen: 0, 3 & 6.



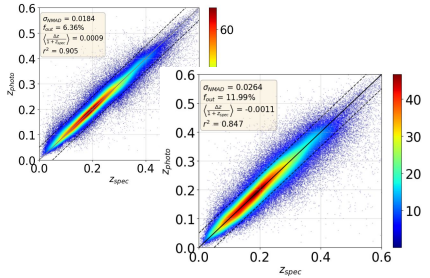


Results

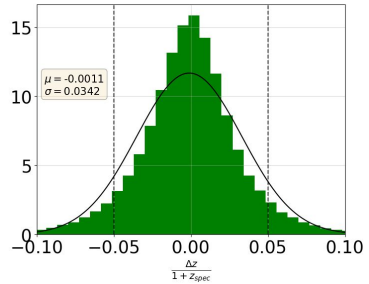
- This first approximation in the prediction of BG photo-z is directed to perform the redshift estimation of galaxies in the local Universe.



- Our method are comparable to those familiar works. We achieve comparable results in terms of the proposed metrics in both, based-features (state-of-the-art) and pixel-level strategy.



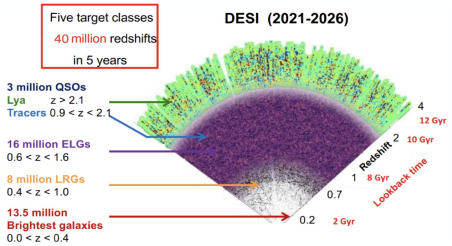
- Pointing in the direction of future photometric surveys, our results are a good step in the use of DL model for the redshift estimation.



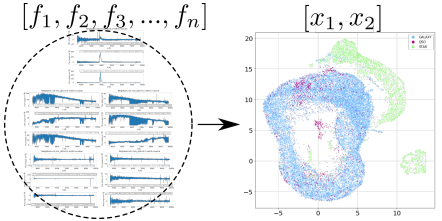
Conclusions

Conclusions

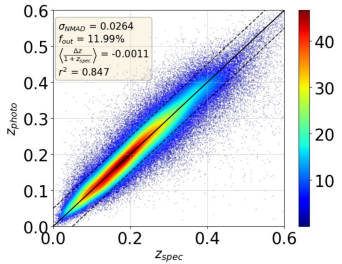
- The large amount of data observed by DESI presents a great opportunity not only to study the evolution of the Universe, but also to explore modern learning-based algorithms that enable us to analyze this data and extract important features.



- Automatic identification of outliers in massive spectroscopic surveys (in this case, DESI), improve the quality of observed spectra & the full survey. The proposed unsupervised learning-method show the power to identify instrumental and environmental errors with efficiency.



- In the era of massive surveys, where a new set of photometric experiments are ready to start (Euclid), a method capable to predict the photo-z with high accuracy is necessary. This approach DL method-based is a first approximation to solve this problem.



Thanks!

John Suárez-Pérez
https://jsuarez314.gitlab.io