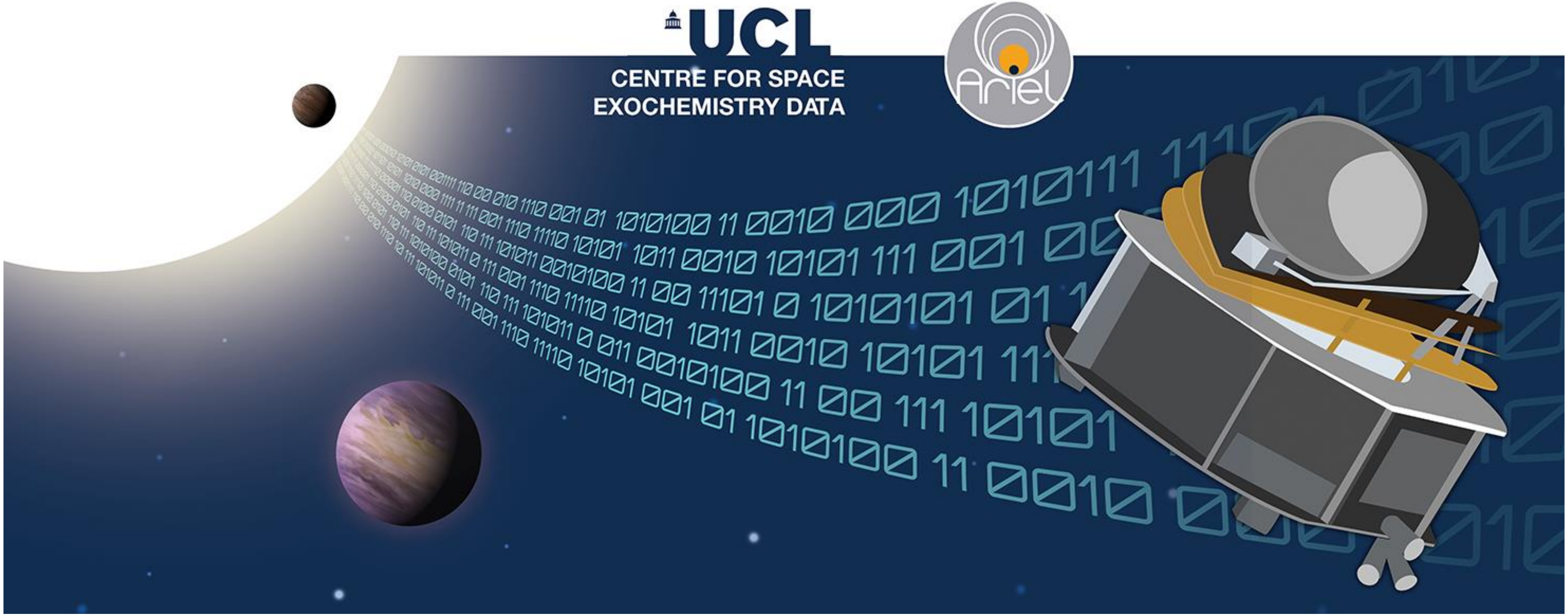


Lecture 1: Introduction to Ariel Data Challenge 2024

Tutorials: [NeurIPS - Ariel Data Challenge 2024]

Presenter: kaggle君-sakura（bili_sakura@zju.edu.cn）

Date: October 8, 2024

Image Credit: NeurIPS - Ariel Data Challenge 2024 | Kaggle

Outline

- Background
- Data Overview
- Research Objective & Evaluation Metrics
- Introduction to Regression/Prediction Task
- Quick Start on Kaggle for Submission

Background

- European Space Agency's Ariel Mission: **Atmospheres of ~1,000 exoplanets**

[European Space Agency](#)

The mission

ESA's mission Ariel, Atmospheric Remote-sensing Infrared Exoplanet Large-survey, will inspect the atmospheres of a thousand planets in our galaxy orbiting stars other than the Sun. Ariel will reveal the ingredients of their atmospheres and the presence of clouds, and monitor how weather conditions change over time. From rocky to gas-giant exoplanets, Ariel will deepen our understanding of these distant worlds.

The launch

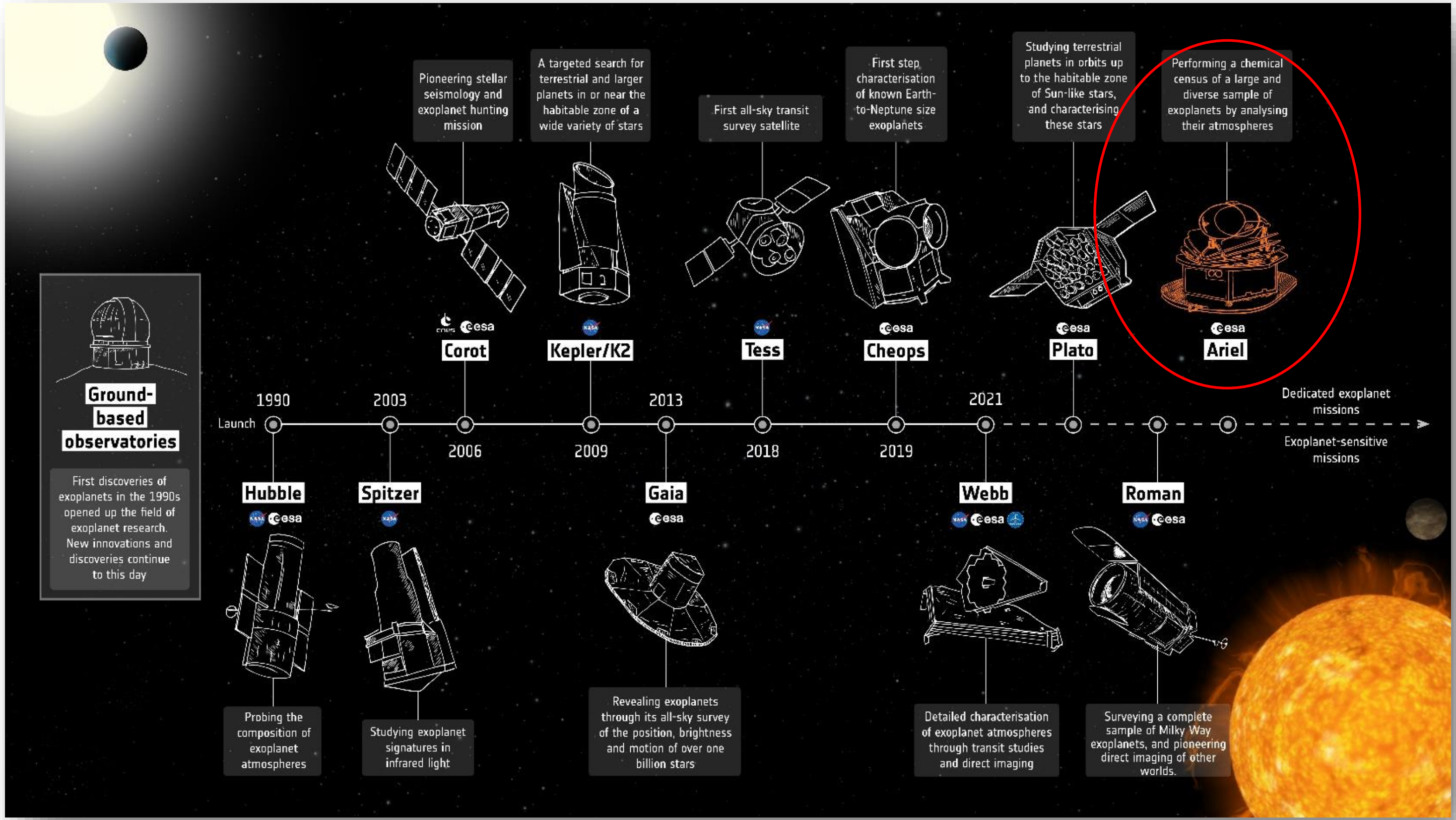
Launch: planned for 2029

Launch location: Europe's Spaceport in French Guiana

Launch vehicle: Ariane 6

Orbit: halo orbit around Sun-Earth Lagrange point L2

ESA - [Ariel Mission](#)



[Exoplanet mission timeline - Ariel](#)

Background

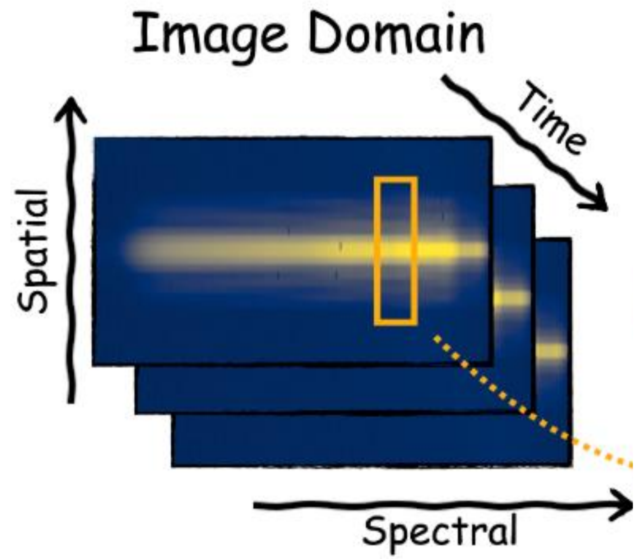
The background of the Ariel Data Challenge 2024 is rooted in the European Space Agency's Ariel Mission, which aims to study the **atmospheres** of roughly 1,000 **exoplanets**. The mission will analyze the chemical composition of these distant worlds by observing them as they transit their host stars. These observations will be done using powerful instruments that collect limited amounts of light (photons), which are further complicated by significant noise due to spacecraft vibrations and other factors.

The primary challenge for the project is to extract faint exoplanetary signals from the noisy data, particularly focusing on the **spectral data** that reveals atmospheric composition. Participants are tasked with designing machine learning models that can process **time series data** from two instruments onboard the Ariel satellite (FGS1 and AIRS-CH0), apply **noise reduction techniques**, and accurately extract the atmospheric spectra along with their uncertainties.

This project is significant because it directly contributes to the growing field of exoplanet research, helping scientists understand the chemical makeup of planets outside our solar system. Ariel is expected to launch in 2029, but this simulated data provides an opportunity to develop tools and methods well in advance.

这个项目非常重要，因为它直接推动了系外行星研究的发展，帮助科学家了解太阳系外行星的化学组成。虽然Ariel计划在2029年发射，但这些模拟数据提供了提前开发工具和方法的机会。

Data



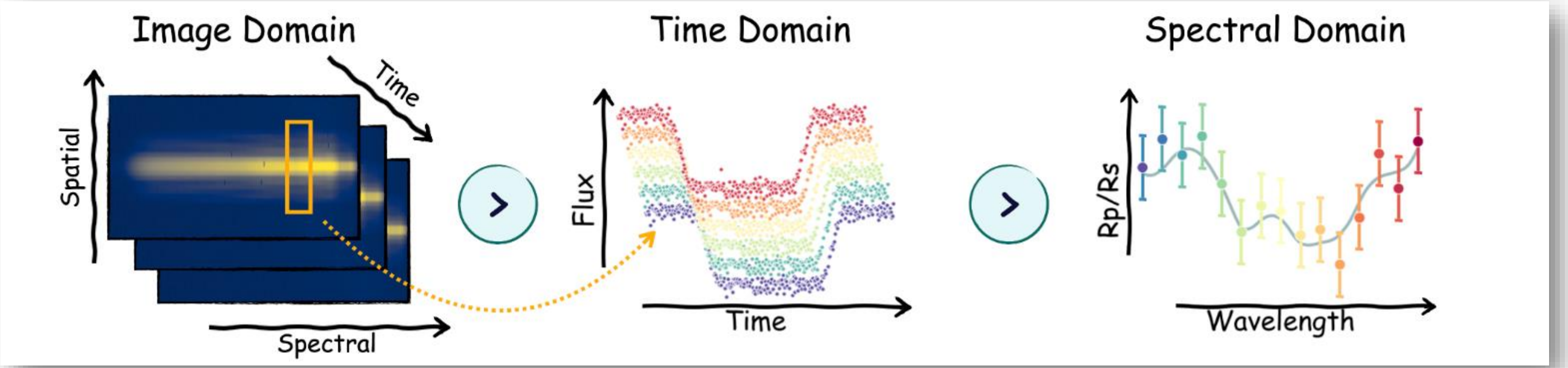
Let’s just ignore processing it into other format currently.

This is a multimodal supervised learning task. Participants can choose to detrend this jitter noise in either modality (i.e. the image, time or spectral domains) or combinations thereof. Each modality bears different advantages. Here we outline two common training strategies.

Signal Data Visualization

Data – Signal Files

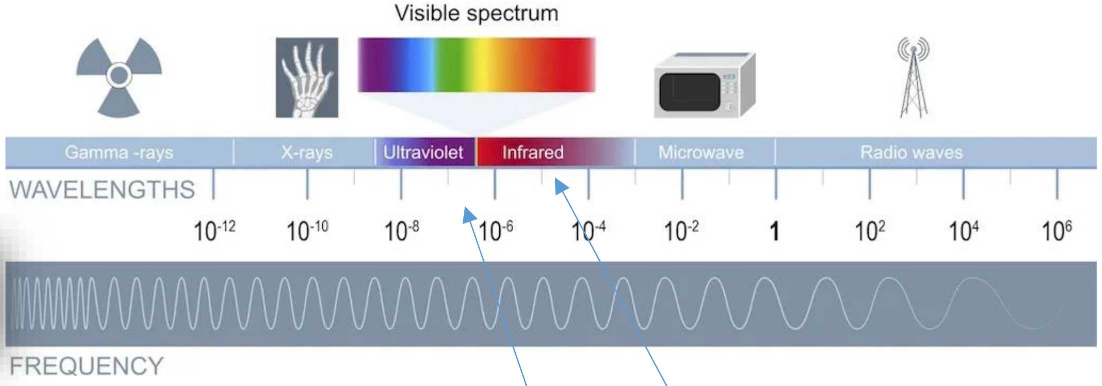
AIRS-CH0_signal.parquet: **11,250 frames (time), 32 × 356 (spatial × spectral)**
FGS1_signal.parquet: **135,000 frames (time), 32 × 32 (spatial × spectral)**
Note: *.parquet file is a compressed file which can be easily convert into numpy array/ tensor. Also, you should reshape it into 3-d data in this case.



For simplicity, suggested by official documentation, you can convert 3D data into 2D/1D by averaging the spatial dimension.

Data – Metadata Files

	A	B	C	D	E	...	JS	JT	JU	JV	JW
1	wl_1	wl_2	wl_3	wl_4	wl_5		wl_279	wl_280	wl_281	wl_282	wl_283
2	0.705	1.95176	1.96061	1.96945	1.97827		3.87503	3.88006	3.88506	3.89006	3.89504



Left: wavelengths.csv, the value denote the wavelength in μm for this column.
Right: The electromagnetic spectrum, from highest to lowest frequency wavelengths. (Image credit: Shutterstock)

data > axis_info.parquet

Data Query Schema Metadata

AIRS-CH0-axis0-h	AIRS-CH0-axis2-um	AIRS-CH0-integration_time	FGS1-axis0-h
0.000027777777777778	4.078463140898873	0.1	0.000027777777777778
0.0000555555555555556	4.074022593360798	4.5	0.0000555555555555556
0.0013611111111111111	4.069568218878434	0.1	0.000138888888888889
0.001388888888888889	4.065100017451781	4.5	0.000166666666666666
0.0026944444444444446	4.060617989080837	0.1	0.00025
0.0027222222222222222	4.056122133765605	4.5	0.000277777777777778
0.004027777777777778	4.051612451506082	0.1	0.00036111111111111115
0.004055555555555555	4.0470889423022705	4.5	0.000388888888888889
0.005361111111111112	4.042551606154168	0.1	0.000472222222222222
0.00			
0.00			
0.00			
0.008027777777777778	4.024263992118866	0.1	0.0006944444444444445
0.008055555555555555	4.019657521249316	4.5	0.0007222222222222223
0.009361111111111112	4.015037223435477	0.1	0.0008055555555555556
0.009388888888888888	4.010403098677347	4.5	0.0008333333333333333
0.010694444444444444	4.005755146974929	0.1	0.0009166666666666668
0.010722222222222223	4.001093368328221	4.5	0.0009444444444444444
0.012027777777777778	3.996417762737223	0.1	0.0010277777777777778
0.012055555555555555	3.991728330201936	4.5	0.0010555555555555557

Showing 1 of 6750 pages   Page Size 20   First Prev Next Last

axis_info.parquet: Axis information for both instruments.

Axis information for both instruments (i.e. AIRS-CH0 and FGS).

FGS1 is the first channel of Ariel's Fine Guidance System (FGS). The main task of the Fine Guidance System is to enable centering, focusing, and guiding of the satellite but it will also provide high-precision photometry of the target star in the **visible spectrum**. It has a sensitivity between 0.60 and 0.80 μm. AIRS-CH0 is the first channel (CH0) of the Ariel InfraRed Spectrometer (AIRS). It is an **infrared spectrometer** with a sensitivity between 1.95 and 3.90 μm, and has a resolving power of approximately R=100. For more information about Ariel please visit the Ariel red book.

Data – Metadata Files

[train/test]_adc_info.csv: Contains analog-to-digital (ADC) **conversion parameters** (gain and offset) for **restoring** the original dynamic range of the data. Also includes a star column identifying which star was used for that planet's simulation.包含模数转换(ADC)参数（增益和偏移），用于恢复原始数据的动态范围。此外，还包含每个行星模拟所用恒星的信息。

data > train_adc_info.csv > data

```
planet_id,FGS1_adc_offset,FGS1_adc_gain,AIRS-CH0_adc_offset,AIRS-CH0_adc_gain,star
785834,-343.33593795992203,0.8372436618056899,-778.9165332747325,0.9247461003693224,1
14485303,-366.38199233481527,0.8429826098421321,-740.3232124446863,0.9317273406892734,1
17002355,-386.1070370237384,1.0417005186749702,-808.6906591858508,1.5135411221718669,0
24135240,-339.7374904920832,0.8402386328699937,-776.1241671086959,0.9312772364580252,1
```

A

data > test_adc_info.csv > data

```
planet_id,FGS1_adc_offset,FGS1_adc_gain,AIRS-CH0_adc_offset,AIRS-CH0_adc[illegible]n,star
499191466,-331.0330086611678,0.8234419796809459,-537.6920996424188,0.937[illegible]3551582,1
```

B

```python
def ADC_convert(signal, gain, offset):
    signal = signal.astype(np.float64)
    signal /= gain
    signal += offset
    return signal
```

I think “/=“ should be “*=” actually.

C

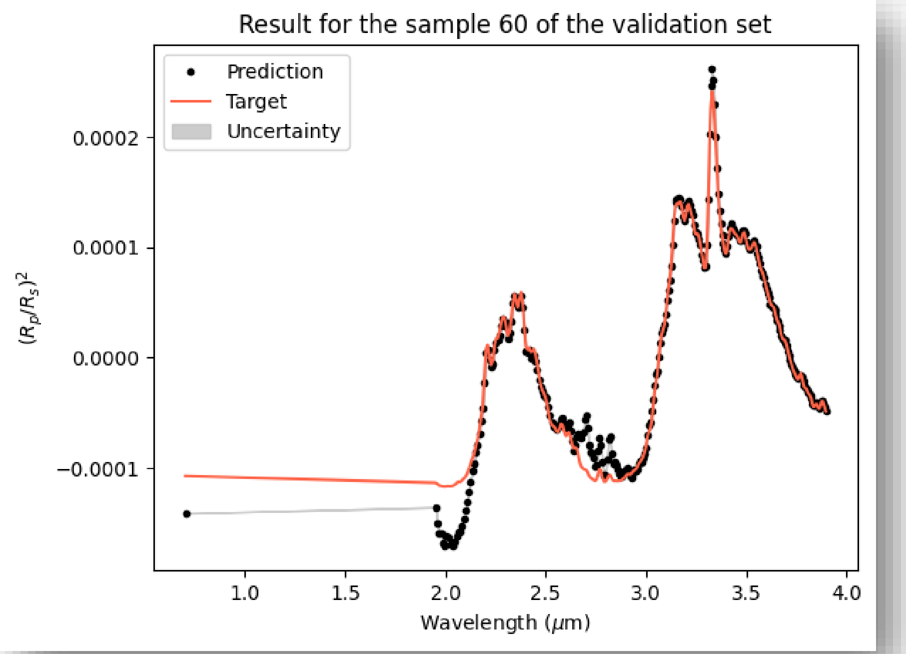
A: train_adc_info.csv. B: test_adc_info.csv. C: Conversion with gain and offset given from adc_info file.

Note: For detailed calibration steps, refer to [official notebook](#).

Objective & Metrics

	A	B	JX	JY	UU
1	planet_id	wl_1	wl_283	sigma_1	sigma_283
2	499191466	0.123	0.123	0.123	0.123
3	…				

This is a Regression task!



Visualization of prediction results from official notebook.

$$GLL = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma_{user}^2) + \frac{(y-\mu_{user})^2}{\sigma_{user}^2}\right)$$

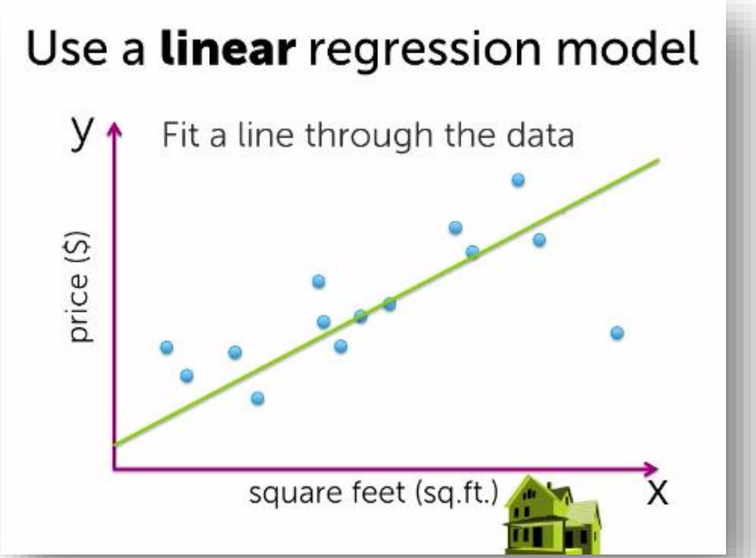
Evaluated by Gaussian Log-likelihood (GLL) function

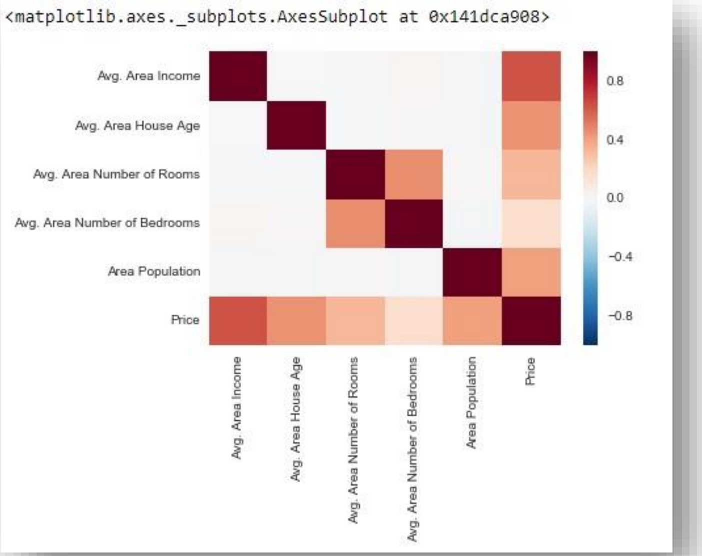
$$score = \frac{L - L_{ref}}{L_{ideal} - L_{ref}}$$

We define L_{ideal} as the case where the submission perfectly matches the ground truth values, with an uncertainty of 10 parts per million (ppm). This ideal case is defined based on Ariel's Stability Requirement. For L_{ref} is defined using the mean and variance of the training dataset as its prediction for all instances. The score will return a float in the interval [0, 1], with higher scores corresponding to better performing models. Any score below 0 will be treated as 0.

Regression

Consider a housing price estimation task.





	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

https://medium.com/@kaushiksimran827/house-price-prediction-a-simple-guide-with-scikit-learn-and-linear-regression-f91a27b9d650
https://medium.com/@yennhi95zz/2-introduction-to-linear-regression-predicting-house-prices-a36f6172030

How to do with Kaggle

- Register your Kaggle Account. (recommend to use your edu.cn email)
- Join the competition. (Click *I accept* and feel free.)
- Try your first submission. (You can copy others' work freely from Code panel to give a first try. As they are publicly available, there is no copyright issue, feel free again. ☺)

Let's go further.

- If you want to get a medium score, using simple machine learning algorithm is plausible, so that training is fast. You can do it easily in your kaggle notebook.
- If you want to be competitive, there is no way out except considering it as a multimodal task. You are going to training a deep neural network much similar as a traditional convolutional neural network. BUT, as the data is not a normal 2D image with timestamp rather a (time, spectral, y-axis) tuple data, pre-trained vision foundation models would not work! You need train a model based on the given data from scratch with your own designed model architecture (I think 3D convolution module would work.). This leads to a large amount of training compute, it is highly recommended to do your training on your own GPU clusters because kaggle notebook GPU (T4) is not quite good and it is painful to manage your project on cloud.

Thanks for listening!