Python and Machine learning: An Introduction

ISSAA 2022

Day #2

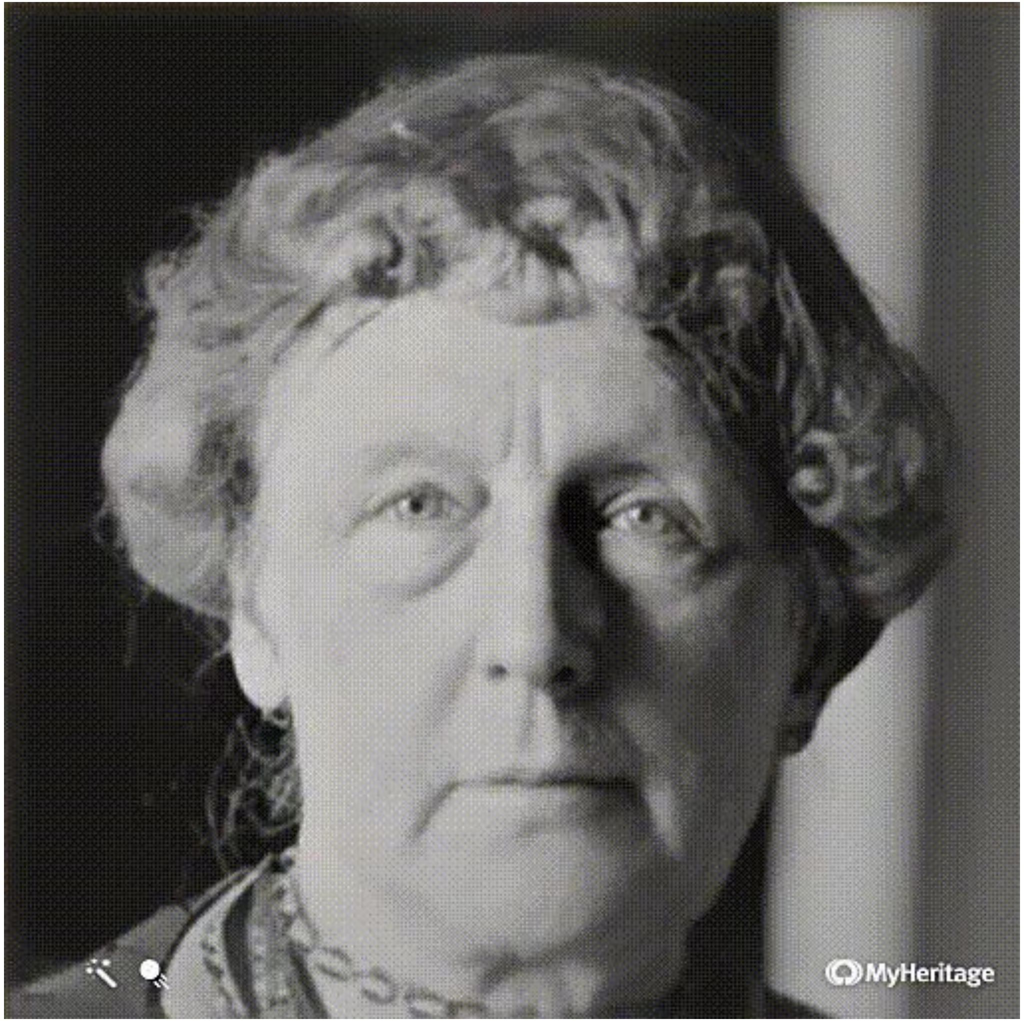
Vishal Upendran
IUCAA

Yesterday's question

How is + translated to __add__ ?

Answer here:

https://stackoverflow.com/questions/13334218/where-are-operators-mapped-to-magic-methods-in-python



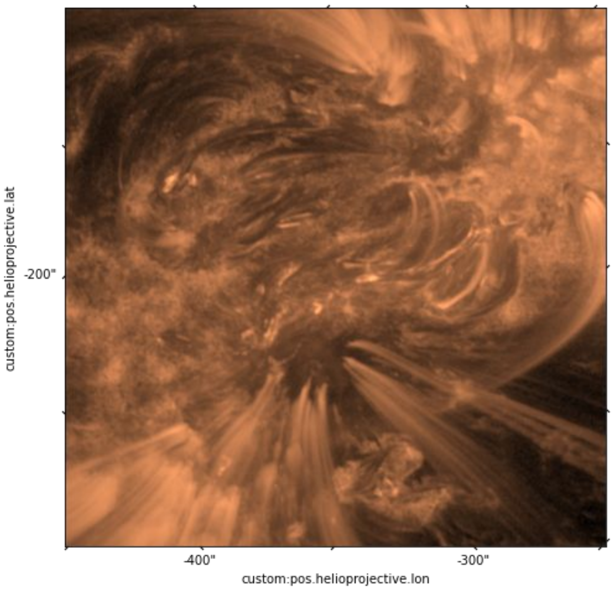
She was the first to classify stars based on their spectral signatures. Who is this?

Menu for today

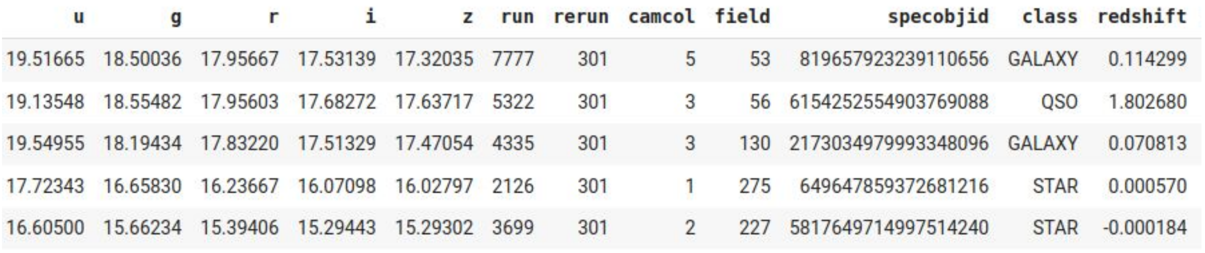
1. Recap of data.
2. What is Machine learning?
3. Machine learning algorithms.
4. Classification: exercise #1
    a. Using Scikit-learn
5. Regression: exercise #2
    a. Using Pytorch

The data:

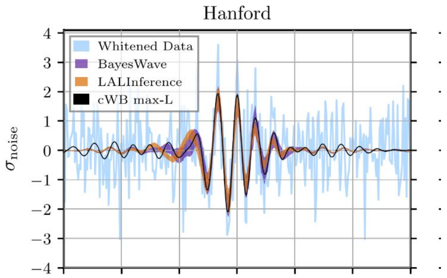
Looking at data

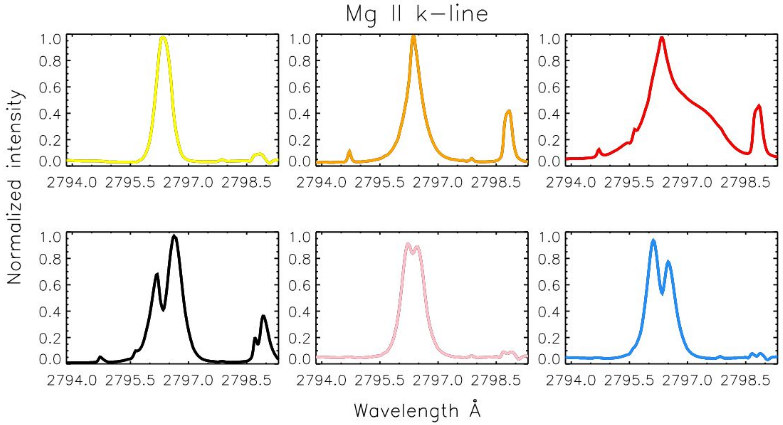


custom:pos.helioprojective.lat
-200"
-400"
-300"
custom:pos.helioprojective.lon



u	g	r	i	z	run	rerun	camcol	field	specobjid	class	redshift
19.51665	18.50036	17.95667	17.53139	17.32035	7777	301	5	53	8196579232391106560	GALAXY	0.114299
19.13548	18.55482	17.95603	17.68272	17.63717	5322	301	3	56	6154252554903769088	QSO	1.802680
19.54955	18.19434	17.83220	17.51329	17.47054	4335	301	3	130	2173034979993348096	GALAXY	0.070813
17.72343	16.65830	16.23667	16.07098	16.02797	2126	301	1	275	649647859372681216	STAR	0.000570
16.60500	15.66234	15.39406	15.29443	15.29302	3699	301	2	227	5817649714997514240	STAR	-0.000184





Keyword: Features

What is Machine learning?

> Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data.
>
> - *Wikipedia*

- Computer algorithms.
- Improve automatically.
- Use data.

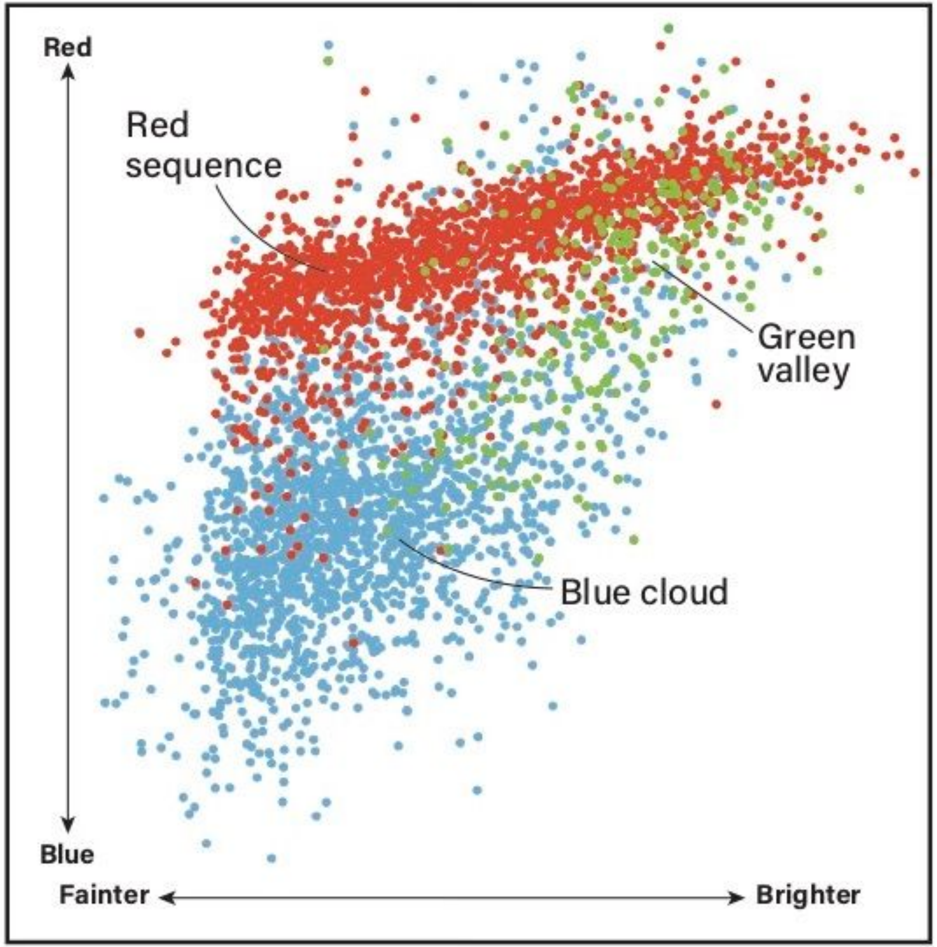
Why Machine learning?

GPUs:

One task, many times.

Typical ML problems #1

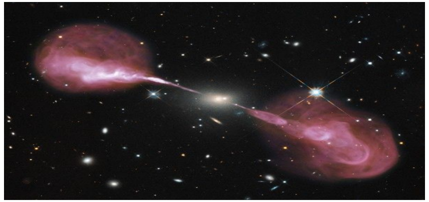
Easy question: How many clusters are present in this pic?

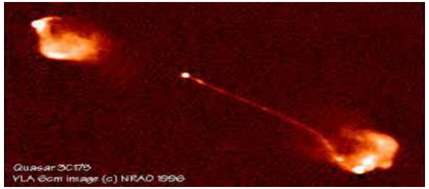


Typical ML problems #2

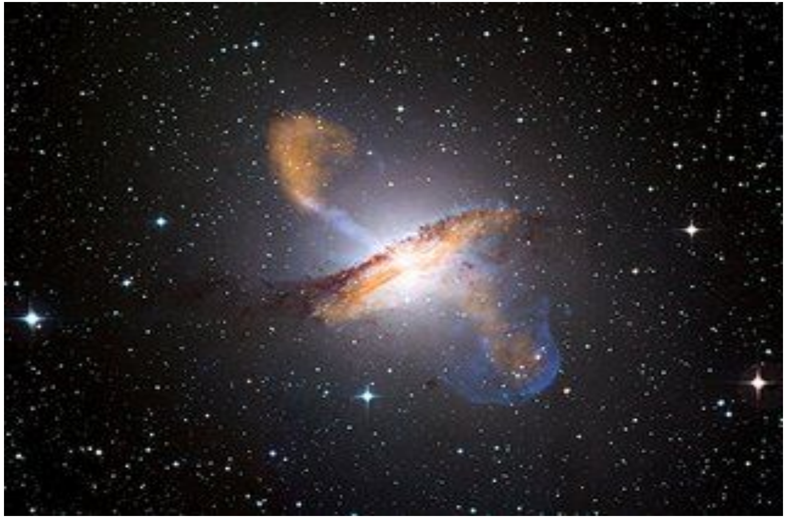
Galaxies

Jets

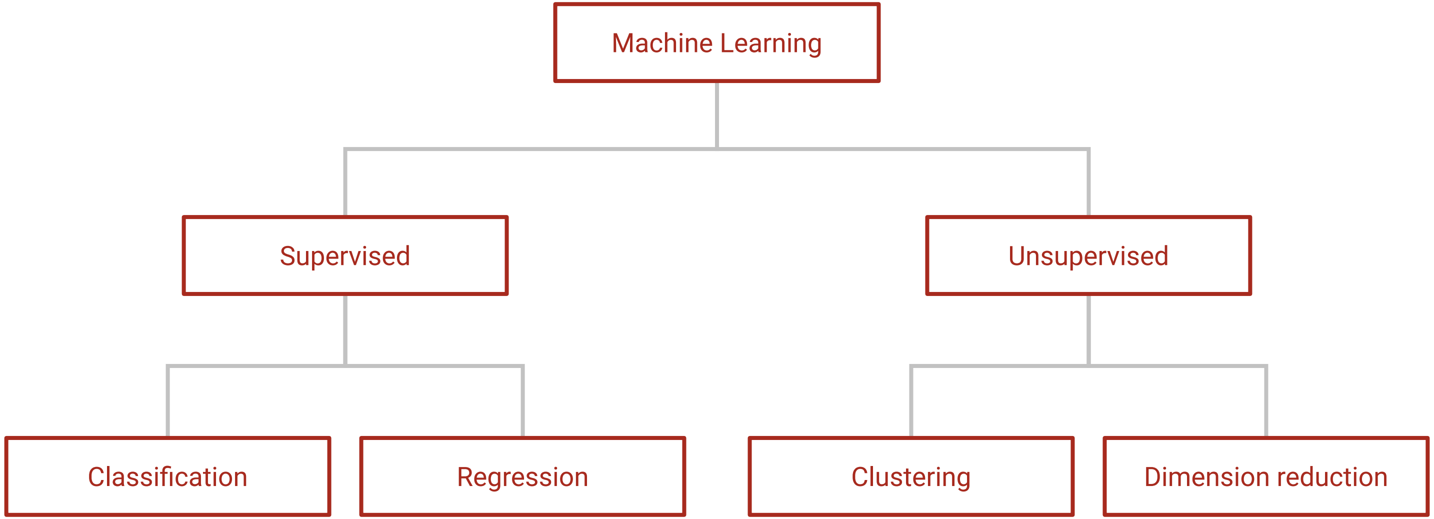




Galaxy or Jet?



Typical ML problems



Machine Learning

Supervised

Unsupervised

Classification

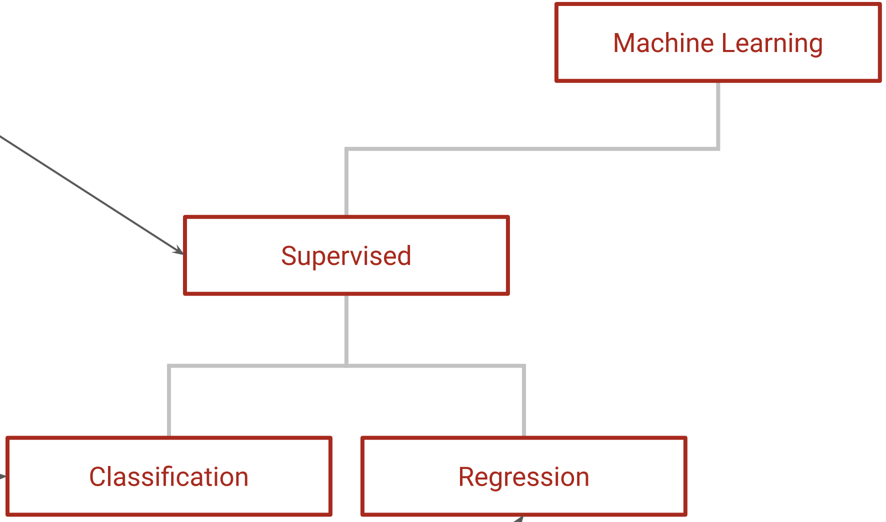
Regression

Clustering

Dimension reduction

Supervised learning

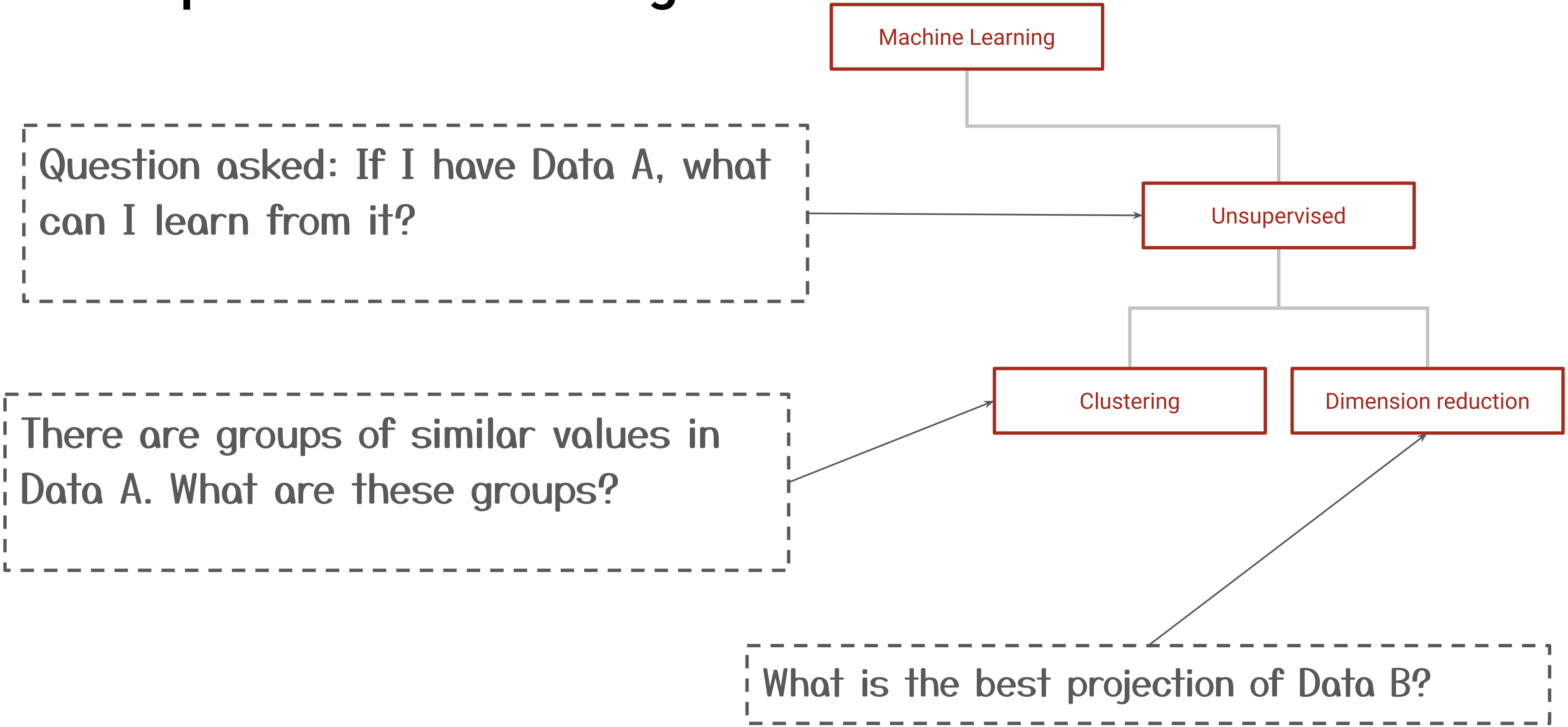
Question asked: If I have Data A, what is the value of Data B?



Data B is a bunch of discrete variables: “Star”, “Galaxy”, “Quasar”

Data B is continuous valued: “Redshift”

Unsupervised learning



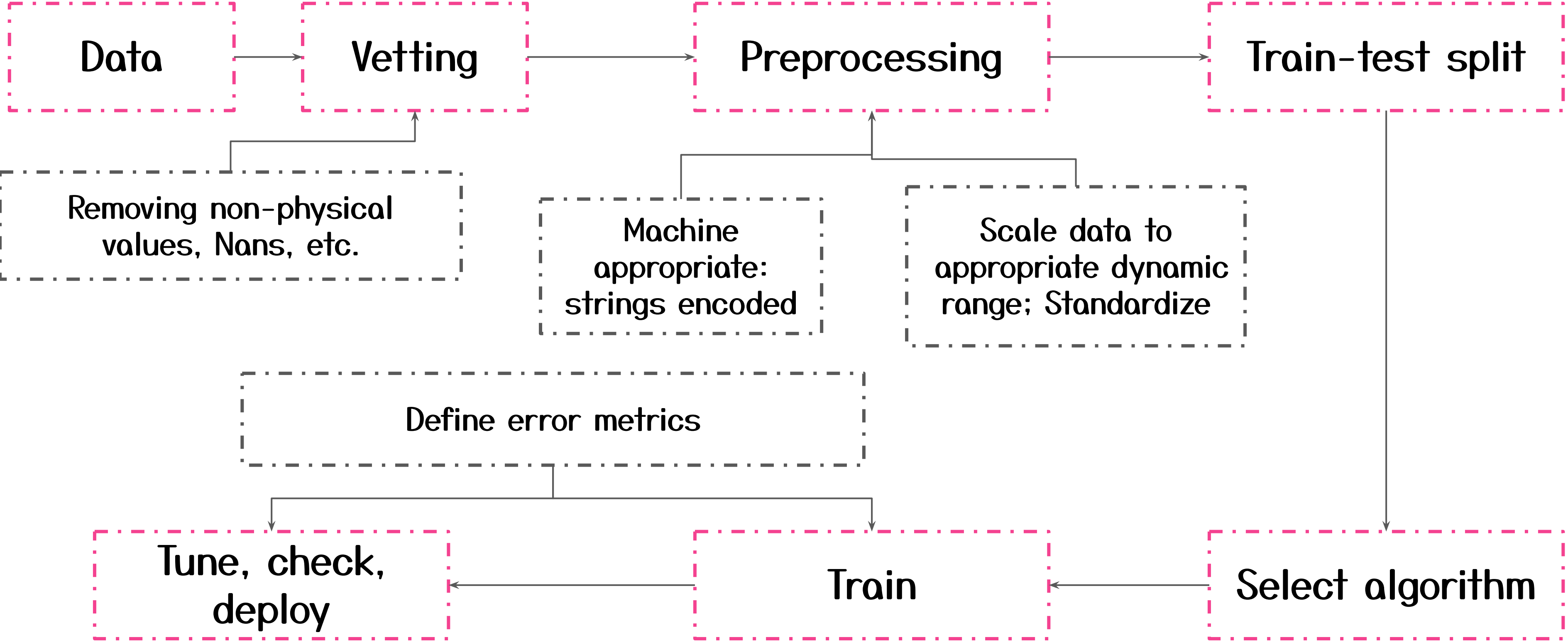
Question asked: If I have Data A, what can I learn from it?

There are groups of similar values in Data A. What are these groups?

What is the best projection of Data B?

What is the procedure to ML?

Frame the correct science question.



Data

Vetting

Preprocessing

Train-test split

Removing non-physical values, Nans, etc.

Machine appropriate: strings encoded

Scale data to appropriate dynamic range; Standardize

Define error metrics

Tune, check, deploy

Train

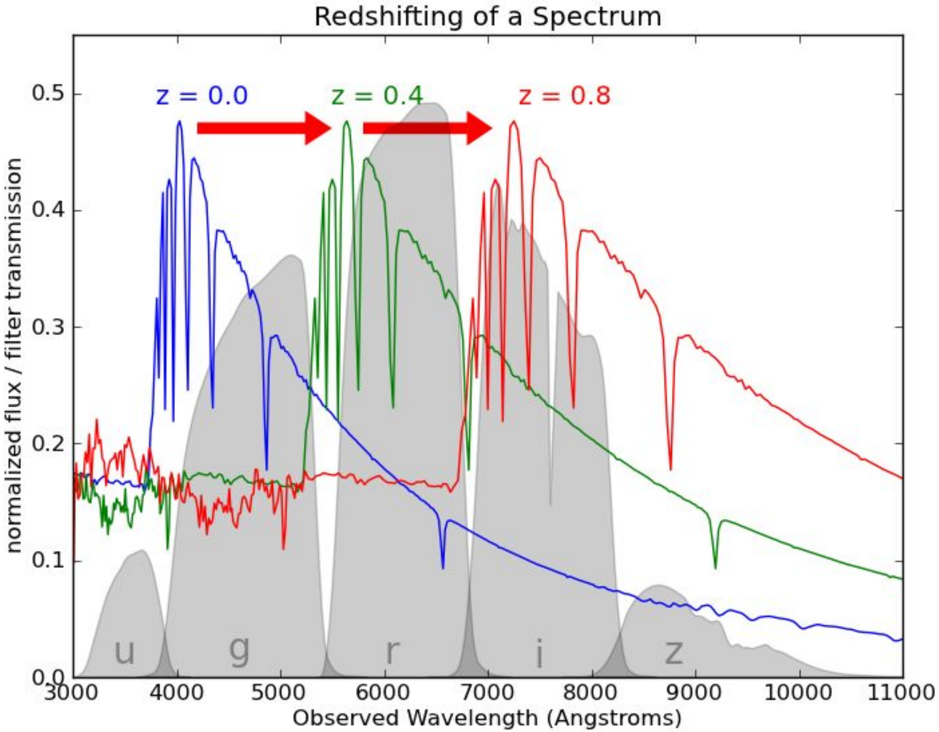
Select algorithm

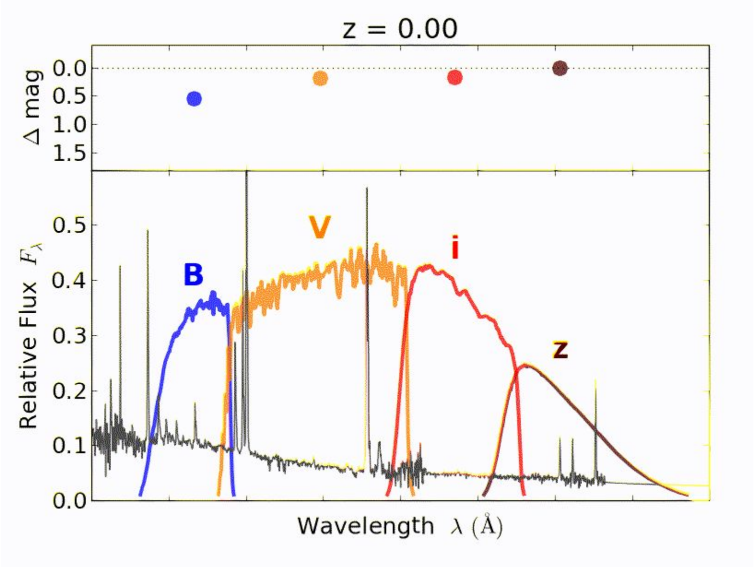
Task for the day

Estimate spectroscopic redshift from photometric colors ⇒

We will use a simple Linear Regression and a Deep neural network.

Why should it work?





From https://www.kaggle.com/c/photometric-redshift-estimation-2019

Metrics

$$\frac{1}{N}\Sigma(z_{pred} - z_{known})^2$$

Mean square error

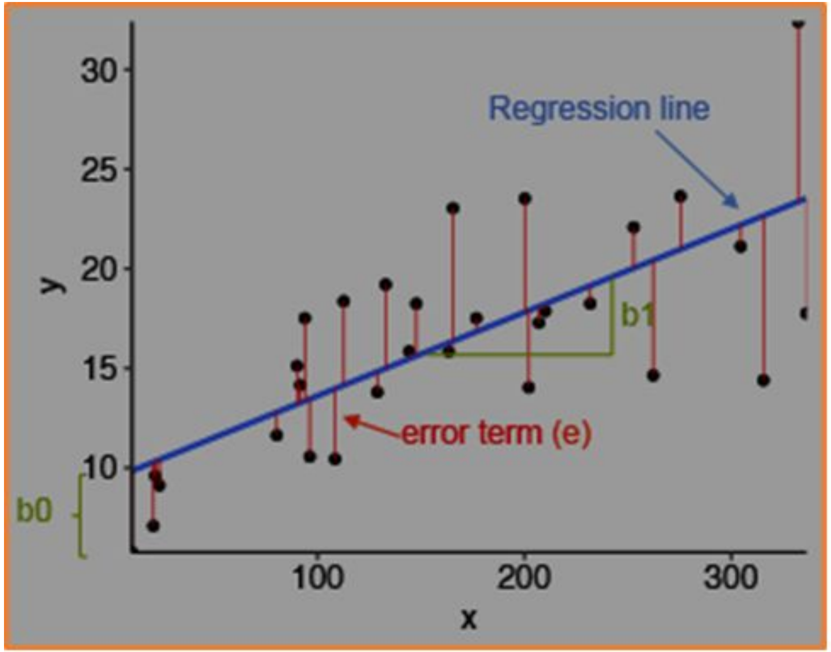
$$\frac{1}{N}\sum_{i=0}^{N}\frac{|z_{known,i} - z_{pred,i}|}{max(\epsilon, |z_{known,i}|)}$$

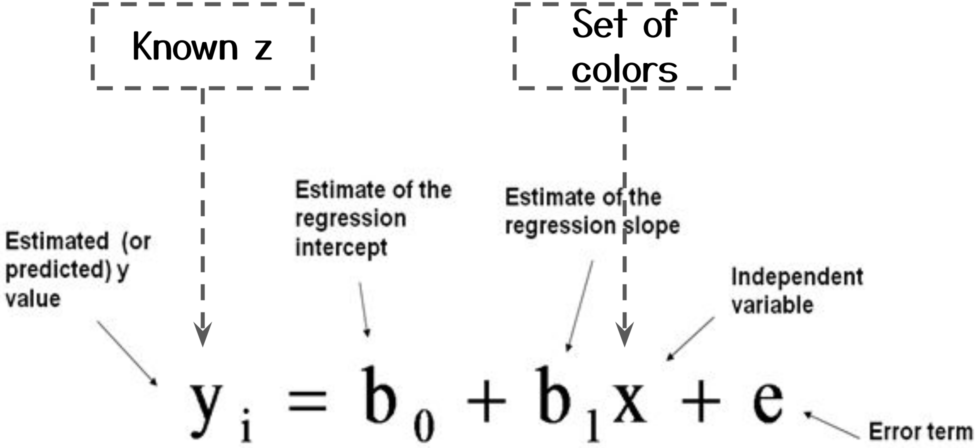
Mean absolute % error

$$1 - \frac{\Sigma(z_{pred} - z_{known})^2}{\Sigma(z_{known} - \mu(z_{known}))^2}$$

Coefficient of determination

Linear regression





Known z

Set of colors

Estimated (or predicted) y value

Estimate of the regression intercept

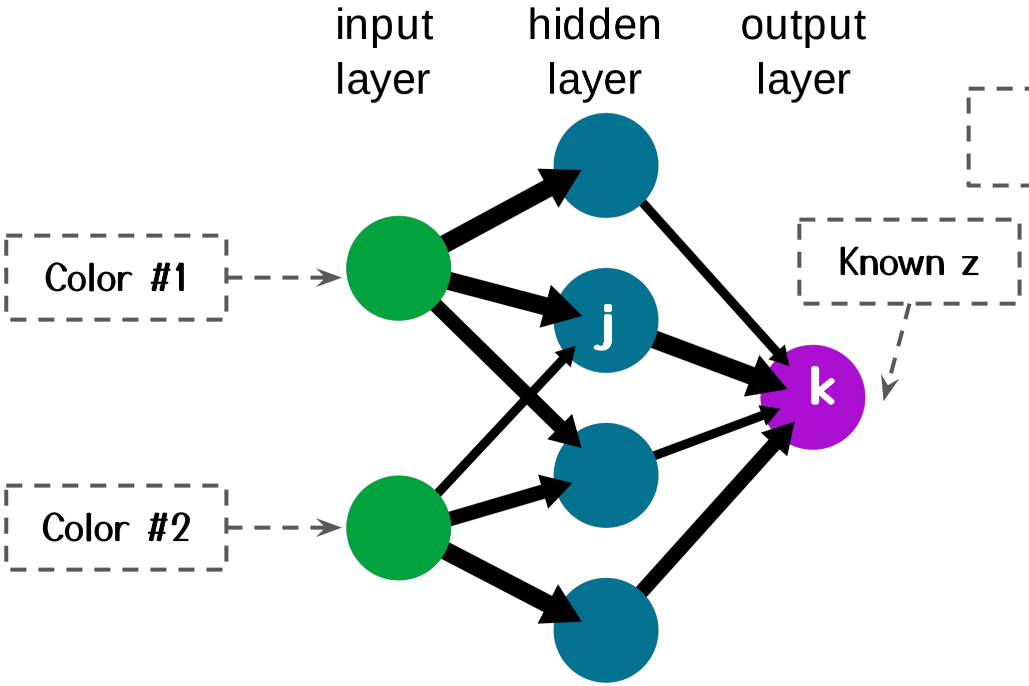
Estimate of the regression slope

Independent variable

$$y_i = b_0 + b_1 x + e$$

Error term

Neural network



Weight variable of neuron j

Bias variable of neuron j

Output from neuron j

$$y_j = f(\Sigma_i w_{ij} x_i + b_j)$$

Input to neuron j from all neurons i

Non linear activation function

$$y_k = f(\Sigma_j w_{jk} y_j + b_k)$$

Neural network: Training

Step 1: Initialize w and b for all layers with some non-zero values.

Step 2: Calculate output from NN for input set:

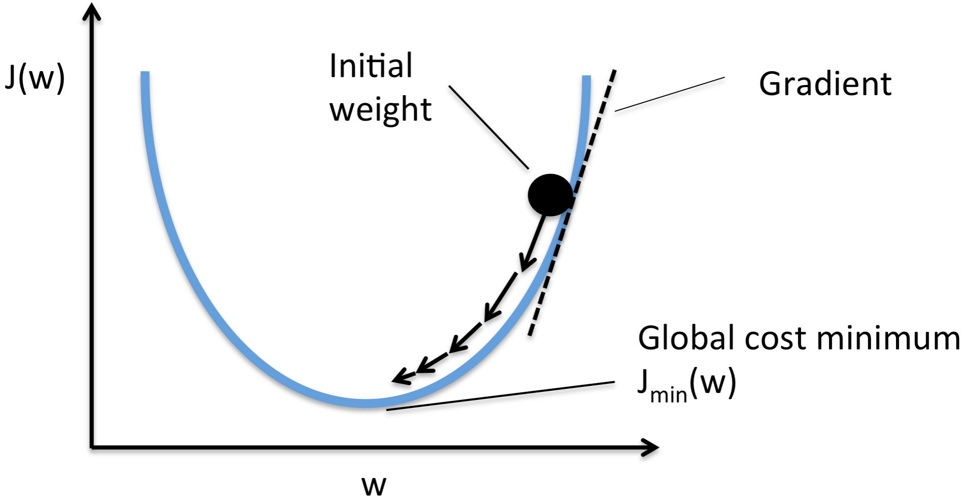
$$z_{pred} = f(..f(\Sigma_k f(\Sigma_j w_{jk} x_j + b_k) + b_k)...$$

Step 3: z_{pred} will not match z_{known}. So get the error between these two values.

Step 4: Now you update weights and biases using this error:

$$w \to w - \alpha \frac{\partial loss}{\partial w}$$

Step 5: Repeat till convergence!



Let us move on to Jupyter →

References for further reading

1. Andrew Ng's course on Machine learning in coursera: https://www.coursera.org/learn/machine-learning
2. Fast AI deep learning course: https://www.fast.ai/
3. Analytics vidhya and Towards Data Science are good blogs too: https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/, https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916 .
4. Advanced: Bishop's book on Pattern recognition and Machine learning; Ian Goodfellow's book on Machine learning.