Abstract

Hello abstraxt!!

Introduction

Hello introduction!!

1 SDSS and MaNGA data

Small intro related to galaxies:
Galaxies in a nutshell:

- "ΛCDM tells us that dark matter halos grow from the 'bottom up', hierarchically assembling into increasingly massive structures with time" (Bundy et al., 2015)
- "The fundamental components of present-day galaxies formed as a result of various complex processes that act on the baryons residing in these evolving halos" (Bundy et al., 2015)

SDSS and MaNGA have and will help us understand the physics of galaxies as well as the mechanisms that drive their evolution.

1.1 Sloan Digital Sky Survey (SDSS)

SDSS Scientific Motivation

SDSS Survey Design

SDSS Data Components and Metadata

SDSS Data Curation

1.2 Mapping Nearby Galaxies at APO: MaNGA

MaNGA is a survey component from the Sloan Digital Sky Survey (SDSS) (Bundy et al., 2015; Ahumada et al., 2020). Its mission is to observe with Integral Field Units (IFUs) approximately 10.000 nearby galaxies. Thanks to the IFUs, MaNGA is able to measure the spectra of different regions of the same galaxy, depending on the number of fibers in the IFU.

MaNGA Scientific Motivation

MaNGA is helping us to understand the physics of the mechanisms that drive the evolution of galaxies by:

- Trying to understand the present growing process of a galactic disk:
  - What fuels this process?
- Trying to understand how stellar accretion, mergers, and secular evolution influence the growth of bulges and ellipticals.
- Trying to understand the influence of internal and external processes in the shutdown of star formation.
- Trying to probe the distribution of mass and angular momentum among different components.

REFRAME IT to fii natural in this section: As mentioned previously, the MaNGA sample size is $\sim$ 10.000 galaxies. The natural question to ask is: why is this a good number for the science intended to do? A simple estimation can help us understand that:

The approximate number of bins needed per axis in the parameter space that defines a galaxy population is ~ 6. These axes are Stellar-mass, SFR (or morphology), and environment. The number of galaxies required per bin, satisfying the condition that the precision of a measurement is approximately equal to the expected variation of it among the bins, is ~ 50. This gives a rough approximation of $50 \times 63 = 10,800$ galaxies

With this sample size, MaNGA is expected to explore new regions of the parameter's space of galaxy populations. **Meaning there is an open window for new discoveries!**

MaNGA Survey Design

MaNGA uses the two Baryon Oscillation Spectroscopic Survey (BOSS) spectrographs, covering a wavelength range from 360 nm to 1040 nm with a resolving power of ~ 2000. MaNGA has a configuration that allows for the observation of 17 galaxies at a time. Each fiber used in MaNGA has a diameter of 120 microns, equivalent to 2" and they are packed in ferrules forming a hexagonally shaped IFU, as shown in Fig. 1 below [put figure obviously].

The number of IFUs used for galaxies are 2x19, 4x37, 4x61, 2x91, and 5x127. For instance, 4x61 means that there are 4 IFUs packed in a bundle of 61 fibers. Besides 12x7 IFUs are used for calibration and 92 fibers are used to sample the sky near each target galaxy, this for a total of 1423 fibers.

1. The fiber diameter is of $\sim 2''$, matched to a typical ground-based seeing.
2. S/N requirement in the outskirts: $5 - 10$ Å^{-1}
3. The outskirts of the target are assumed to be located at $1 - 2\ R_e$
4. Required exposure time: $\sim 3\ hr$
5. To reach a physical scale of $\delta = 1\ kpc$ with a $2''$ fiber diameter, it demands a $z \sim 0.03$
6. Volumetric density of galaxies with $M > 109 M_\odot$ + observing area of $\sim 3500\ deg^2 \rightarrow$ $z > 0.03$, in order to obtain ~ 10.000 galaxies.
7. The number of fibers used is 1423
8. Fiber budget + IFUs size + dynamic range of $R_e \rightarrow 0.03 < z < 0.10$
9. It translates in targeting larger galaxies at greater distances.
10. ~ 20 IFUs (3 FoV) + 3 hr exposure + ~ 500 plates $\rightarrow$ 190 nights.
11. Accounting for weather and dark time limitations, plus time shared with eBOSS, the survey has a duration of 6 yr

In regard to the fibers and the fiber bundles, we have:

1. The number of fibers used is 1423
2. Each fiber has a $2''$ diameter on-sky.
3. The IFUs are 19 to 127 fibers in size, packed in an hexagonal shape.
4. The diameters range is from $12''$ to $32''$ on-sky.
5. Fibers are inserted by hand in metal ferrules.
6. The IFUs used for galaxies: 2×19; 4×37; 4×61; 2×91 ; 5×127, (1247)
7. 12 IFUs packed in bundles of 7 are used for calibration.
8. The remaining 92 fibers are used to sample the sky near each target
9. The IFUs have a fill factor of $\sim 56\%$.
10. The cross-talk between spectral traces of adjacent fibers is of $\sim 10\%$

Data Components and Metadata

The MaNGA Data Reduction Pipeline (DRP) (Law et al., 2016) takes the raw MaNGA data and generates a data-cube, a three-dimensional data array containing spectra for a single galaxy. The array has two spatial dimensions and a wavelength dimension representing the spectra of a galaxy at each of its spatial measured points. The DRP generates the DRPall summary fits file containing the metadata for all the observations, galaxies, and non-galaxies. In the case of galaxies, it includes information such as the redshift, among others drawn from the NASA-Sloan Atlas. The DAP processes only the log-linearly binned data cubes (LOGCUBE files) from the DRP to produce stellar and ionized gas kinematics, among others, stored in the MAPS files. Each MaNGA observation is uniquely identified with the plate-IFUdesing, which allows us to find a specific galaxy observation. There is also the MaNGA-ID which uniquely identifies a galaxy, this is something to have in mind since a galaxy might have multiple observations, having, therefore, several plate-IFUdesigns

MaNGA data processing needs are not a trivial task, for instance, we can plot some numbers. The sample size is of $\sim 10,000$ low redshift galaxies, out of wich there are approximately 100 million raw-frame spectra which after reduction, go down to about 10 million reduced spectra. The processing of the raw data has to take into account many variables related to the exposition time, too.

The MaNGA Data Reduction Pipeline (DRP) is aimed at producing a flux calibrated spectral cube of galaxies in a FITS format. The DRP is written mainly in IDL, aided by C and python scripts. The DRP consists of two stages:

1. 2d stage: flux-calibrated fiber spectra

2. 3d stage: products from 2D stage + astrometric information $\rightarrow$ data cubes

For the 2d stage the product consists of flux-calibrated individual exposures of an entire plate, that is, the science and calibration exposures.All the calibrations exposures are processed to determine the spatial trace of the fiber on the detector in order to extract fiber flat-field and wavelength calibration vectors. The products above are used to process science frames. From the sky fibers a super-sampled model of the sky background is created to then be subtracted from each fiber spectra. The mini-bundles used for standard stars are used to determine the flux calibration vector for the exposure. The final product is a 2D FITS file with Row-Stacked Spectra (RSS), ie, each row is the spectra of each of the 1423 fibers.

For the 3d stage The products consist of data cubes and RSS for a galaxy from all the exposures. Once the needed number of exposures is completed per plate, the 3d stage is triggered using all the RSS FITS files generated in the 2d stage. The RSS FITS files of

a plate are combined in this stage. The relevant spectra of a target for the exposures are identified and sorted using astrometric information. A data cube is generated for each target, having additional products: griz image, mask cube, reference information and more. The final products of the 3d stage are used later with the Data Analysis Pipeline (DAP) built for MaNGA to produce science data products.

Given the complexity of all the data processing stage, it is good to keep in mind that in addition to what was mentioned above different operational variables change and in different time scales, therefore the need for metadata track, eg: fiber bundle metrology, cartridge layout, fiber plugin locations and more. Metadata is essential if we want to rerun the DRP to obtain data products.

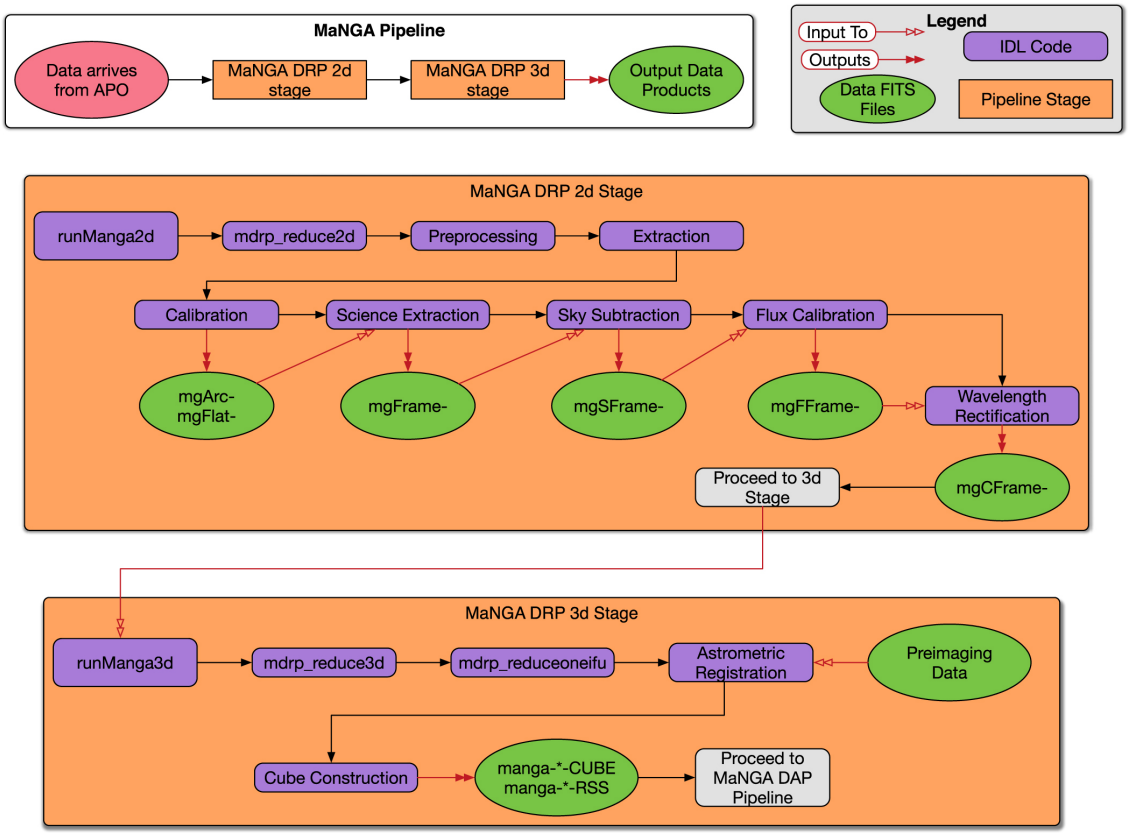

Figure 1: MaNGA DRP diagram (Law et al., 2016)

MaNGA Data Curation

ML algorithms are agnostic in the sense that they don't know what is the phenomenology behind the data used for their training. For instance, let's assume we have the same spectrum of a galaxy but with different redshifts. If we feed these spectra to the ML algorithm, it is going to think that we have two different objects. Therefore before developing an ML model it

is necessary to correct for these kinds of situations in order to make comparisons meaningful. Additionally, we can not just feed raw data to the algorithm because there are scenarios where we have missing values, artifacts. MaNGA data provides several flags signaling scenarios of these kinds for the data. Table 1 provides these values:

Key	Bit	Description
FORESTAR	0	There is a FORESTAR region within the data cube
BADZ	1	NSA redshift does not match derived redshift (placeholder)
LINELESS	2	No emission lines in data cube (placeholder)
PPXFFAIL	3	pPXF failed to fit this object (placeholder)
SINGLEBIN	4	Voronoi binning resulted in all spectra in a single bin
BADGEOM	5	Invalid input geometry; elliptical coordinates and effective radius are meaningless.
DRPCRIT	28	A critical failure in DRP
DAPCRIT	29	A critical failure in DAP
CRITICAL	30	A critical failure in DRP or DAP

Table 1: MaNGA Data Quality Mask Bits

During the data curation process, all MaNGA data with any of the mask-bits listed in Table 1 were ignored.

Some data products are two-dimensional maps of:

1. stellar velocity and velocity dispersion,
2. mean stellar age and star formation history,
3. stellar metallicity,
4. element abundance ratio,
5. stellar mass surface density,
6. ionized gas velocity,
7. ionized gas metallicity,
8. star formation rate and
9. dust extinction

2 The Outlier Problem

Hello outlier problem!!

Algorithm 1 ODA based on AE

Require: Observed spectrum O
Require: Trained AE AE
Require: Outlier score function s
$R \leftarrow AE.predict(O)$ ▷ R : reconstructed spectrum
$score \leftarrow s(O, R)$ ▷ $score$: outlier score of O
return $score$

3 Auto-Encoders (AEs)

Hello AEs!!

4 Random Forest (RF)

Hello RF!!

5 Manifold Learning

Hello manifold learning!!

6 Explainable Artificial Intelligence (xAI)

1. LIME
2. SAHP

6.1 xAI for Unsupervised Learning

LIME and SHAP are designed to interpret the predictions made by a given classifier or regressor model f. Therefore, the f comes from a SL setting. For the case of the URF, this is not much of a problem because it is a classifier that distinguishes whether a spectrum is real or synthetic. However, the ODA based on the URF is a regressor since it outputs a continuous variable, the outlier score. On the other hand, the AE is an UL algorithm, i.e, it is neither a regressor nor a classifier. The ODA based on the AE has the workflow illustrated in algorithm (1), meaning this ODA can be framed as a regressor. Having this in mind, we can conclude that for any ODA that computes an outlier score, we can frame it as a regressor model f for LIME and compute an explanation model ξ for a given prediction.

6.2 Local Interpretable Model-agnostic Explanations (LIME)

Important, an explanation is a model, this are posthoc models. In LIME's paper (Ribeiro et al., 2016), the authors concider the following desiderata for an xAI algorithms:

1. **Interpretability**: the idea of explainig a prediction is for it to be easily understandable for humans. Therefore the explanation of a prediction made by a ML model must be human-friendly. Let us consider the case of a ML model that takes a 64×64 image of a galaxy with three chanels and outpust its star formation rate (SFR). In that case, the number of features used by the model to predict the SFR, corresponds to $64 \times 64 = 12288$. Having a table that indicates the relevance of each of these pixel values for the prediction constitues an overwelhmingly high amount of information to be handled by an expert astronomer. In that sense, human-friendly explanations must be constrained by our human limitations. That being said, Ribeiro et al. (2016) proposed for such cases the use of **interpretable features** for the explanation model, that can be constructed from the original features used for the prediction. Remaining with the same example of the image of a galaxy, an interpretable feature could be a the collection of neighboring pixels into a super-pixels. The python implementation of LIME by Ribeiro et al. (2016) has a class to work with models that use as input image data. In that case, the super-pixels are created using estandar segmentation algorithms. In the figure fig: andromeda segmented, we can see a couple of examples for this case:
2. **Local fidelity**:
3. **Model-Agnostic**:
4. **Global perspective**:

To address the desireata exposed above, Ribeiro et al. (2016)...

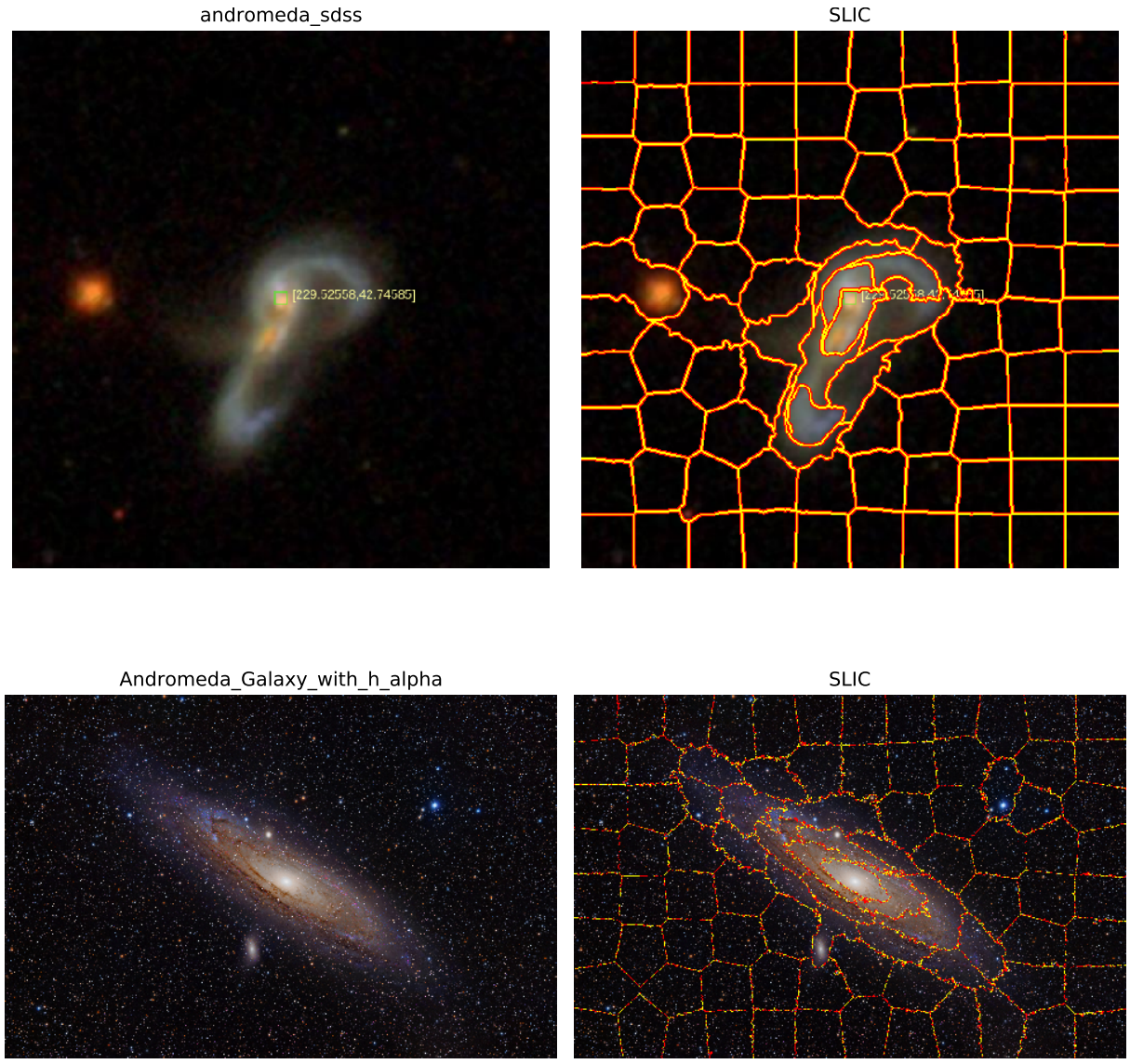

Figure 2: Two examples of segmented images for the Andromeda galaxy. Up: SDSS, Down: HST (review)

References

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3

Bundy, K., Bershady, M. A., Law, D. R., et al. 2015, ApJ, 798, 7

Law, D. R., Cherinka, B., Yan, R., et al. 2016, AJ, 152, 83

Ribeiro, M. T., Singh, S., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: Association for Computing Machinery), 1135–1144