Finding structures in astronomical data with machine learning

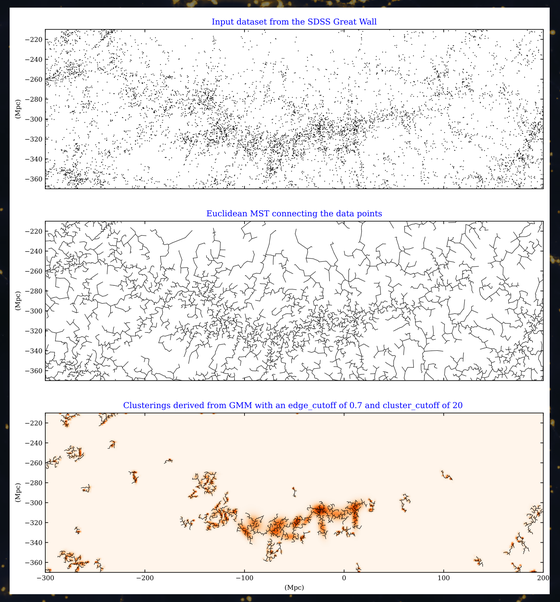
Wanying Zhao

Supervisor: Dr. Siri Chongchitnan

WARWICK
UNDERGRADUATE RESEARCH SUPPORT SCHEME

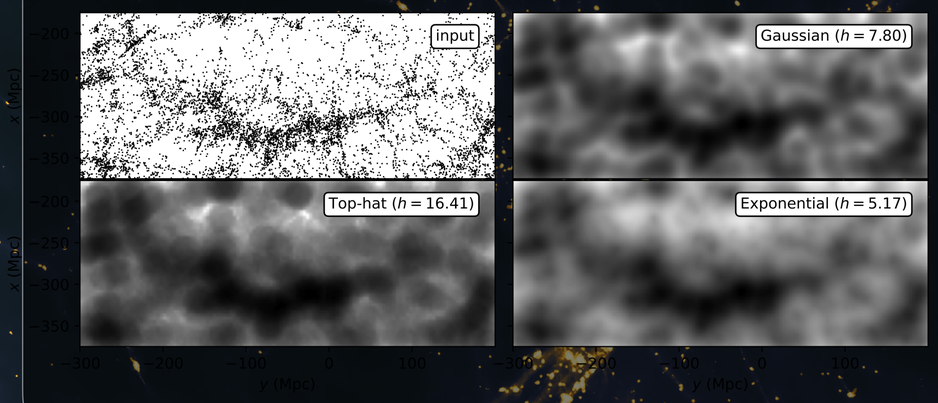
My project demonstrates how we can use **Python** to perform machine learning algorithms on large-scale astronomical datasets.

Astronomy is experiencing a rapid growth in complexity and data size. Just so you get the picture, in my project I dealt with a dataset called the SDSS Great Wall, which consists of more than 8000 stars! This is only considered a medium-sized dataset, so clearly we have to do something smart to extract some structure and trend out of it.



We also wish to be able to divide the stars into categories. Usually, this requires some prior knowledge about what they should look like, but with the sheer amount of data this gets tricky very quickly. Fortunately, we have a class of methods called **Hierarchical clustering.** Loosely speaking, this allows us to find all clusters at all scales by letting the points organise themselves in some optimal way, and we can just choose the right scale at which the desired features live. Top: all the data points, Middle: what happens when the points reorganise themselves, Bottom: the remaining features after we chose the right scale

It is often useful to estimate the density of our data — for instance, roughly how many stars do we expect to see when we look at a certain part of the sky? We can do this via a procedure called **Kernel density estimation.** Loosely speaking, this involves applying a filter to our data points to get a smoothened picture. The effect looks like this when we do this to our SDSS data set.



To pick the right clustering scheme for our dataset, we would need to evaluate the performance of our models quantitatively. There are two methods to do this — the **Akaike information criterion (AIC)** and **the Bayesian information criterion (BIC).** These two schemes account for both model performance and complexity, thus gives us a model that achieves a fine balance between power and efficiency.