**Data explanation**

**Dataset**

The dataset consists of three sparse matrices describing a variable in the galaxy formation. The three variables are mass of the halo progenitors, distance to the main branch and a categorical variable that describes if the progenitor is a main halo or a subhalo. The value 0.0 depicts that a progenitor is not present, while any other value describes the value of the variable of the present progenitor. If a progenitor is present, the value is positive, which means there are no negative values [2].

**3 variables**

* **Mass**
  + The mass variable describes the mass of the progenitor. Mass, or the mass assembly history of a halo is arguably the most influential and important variable in a merger tree [3]. In galaxy formation models, mass growth is the main input (Springel et al. ([2001](javascript:;))), which emphasizes the importance of the mass variable in merger trees, galaxy formation models, and in these experiments.
* **Dist**
  + The distance variable describes the “the physical distance between merging haloes” [3], or “distance between the centers of a progenitor in a given branch with respect to the progenitor in the main branch at the same step in time” [2]. In contrast to the mass variable, the distance variable is usually not required in SAMs of galaxy formations to evolve galaxies (e.g. Somerville et al. [2008](javascript:;)), which means the distance variable is less crucial for our experiments. However, managing to include the distance variable in a model and predicting it would be beneficial for the field to advance further. That is seen in GALFORM (COLE et al. 2000) which managed to use distance, to redegine the progenitor type (subhalo or main halo) and mass.
  + “Note that, together, mass and distance are a measure of the gravitational pull between merging progenitors” [2]
* **Subhalo**
  + The subhalo/type variable describes the halo-type of the progenitor. This is a categorical variable that can take the values 0.0, which implies there is no progenitor, 0.5 which implies the progenitor is a subhalo (or satellite halo), or 1.0 which implies that the progenitor is a main halo. The progenitor type variable has become an important, almost necessary, part in galaxy formation SAMs such as SAG (Cora et al. 2018), GALACTICUS (Benson 2012), SAGE (Croton et al. 2016) and GALFORM (Cole et al. 2000).
  + Whether a progenitor is a main or a sub-halo have great influence on the other two variables behavior.
    - In terms of distance are subhaloes expected to be closer to the progenitor in the main branch they are merging with. [2]
    - That means that in terms of mass, the subhalo masses can decrease over time as long as they approach the merging event [2].

**Picture – the data will be treated as a picture**

A blue and green pixelated logo

Description automatically generatedIt is possible to represent the data as an image since we have three variables, to experiment with single variables, a mix of two variables or all three variables is possible. Paper [2] manages to get great results, especially using all three variables.Representing the data as an image is done by storing the variables in separate matrices that illustrate the merger tree structure. [2] The columns in the matrix format represents branches in the merger trees, and the rows represent snapshots, that is, time until present day. The last row represents the present day. The leftmost column (branch), represent the main branch, which is the longest branch in the merger tree. In the last row (present day), the main branch only contains the halo which the merger tree mass assembly history informs about, which is called the last descendant.

Using this matrix representation of each variable, then “merging” the matrices together, can be seen as an image with three channels, where the three channels are the variables.

In this paper, the merger trees are mostly going to be represented as 29 x 10 pixels images, in a format like the one to the right. Here we can see a consistent merger tree with all its 7 branches, where they start, where they stop, and the main branch (left).

**Consistancy & properties**

In terms of generating new galaxy formations, it is important that the generated merger trees have the same traits as the original galaxy formations merger trees, that means being consistent in terms of structure and the different variables. Here are five main consistency point for the structure of the generated merger trees:

* **Reappearing merger consistency:** There should be no reappearing haloes after a merger. [2] That means there should be no gaps between haloes in a branch (column) of the image. ***Example image***
* **Branch gap consistency:** There should be no gaps between branches within a galaxy formation. That means that if there is an empty branch (column), that means all branches in that galaxy formation has been represented. There should not be an empty branch (column), followed by a branch with any progenitors. In other words, all x branches should be in the x leftmost columns. ***Example image***
* **Main branch consistency:** The main branch is in the first column, all the way to the left. It can start whenever, but when it start, it should have progenitors until the last time step (row) ***Example image***
* **Last row consistency:** The last row should only contain the last descendant, which means only one halo. This halo should be in the main branch. [2] ***Example image***
* **Progenitor absence 0 consistency:** the absence of a progenitor in a branch should be represented by a zero value.[2] And if there is an absence of a progenitor, the value should be 0.0 for all variables.

In terms of the three variables, there are patterns that need to be consistent or half-consistent for a successfully generated merger tree. Here are descriptions of the consistency requirements for each variable in generated trees based on patterns, structure, and consistency of the original merger trees:

* **Mass:** 
  + The generated mass should be within the mass range of the original merger trees
  + Usually, progenitor masses in a branch are expected to increase monotonically as time (number of row) elapses [3], if not, it is expected to preserve its mass. [2, 3] This is an assumption for SAMs Lacey & Silk, 1991; White & Frenk, 1991; Cole et al., 1994; 2000; Gonzalez-Perez et al., 2014; Lacey et al., 2016; Lagos et al., 2018).
    - It is worth noting that the behavior of other variables like progenitor type and distance to main branch affects the expected behavior of the mass assembly history structure and patterns. [2, 3]
* **Distance:** 
  + The distance variable should decrease the closer the branch gets to a merge in time. However, there are no set rule for the exact time of occurrence. [2, 3].
  + There should be no dramatic jumps in location. [2]
    - This would be a great way to evaluate if the distance variable is well reproduced in the generated merger trees. This evaluation could be done by comparing the original and generated distance distributions with the Kolmogorov-Smirnov (KS) test. [2]
  + In contrast with the mass variable, there are no set range for the distance variable. The distance between two merging progenitors has no set range. [2]
* **Subhalo:**
  + In the main branch, progenitor types are expected to be main haloes. There are a few exceptions early in time. [2, 3]
  + In other branches, the progenitor types can be both main halo or satellite/subhalo. Usually, progenitors become satellites as they approach the other merging halo, as a result of gravitational infall [2, 3]. The exact time of occurrence vary from tree to tree and depends on masses and distance of merging progenitors (Diemand, Kuhlen & Madau [2006](javascript:;); Elahi, Thacker & Widrow [2011](javascript:;); Muldrew et al. [2011](javascript:;); Han et al. [2012](javascript:;); Onions et al. [2012](javascript:;), [2013](javascript:;)) [3].

**Dataset**

One last thing to note; the dataset is relatively small and limited. The main dataset utilized in this paper contains less than 40 000 training images with ranging from five to ten branches. However, given as seen in the following data analysis, separating the merger trees by the number of branches they contain, all of them have over 1 000 training examples, which was mentioned as a requirement for generating well-constructed images with a given number of branches. Thus, even though the dataset is limited, it should be large enough to generate merger trees according to paper 3.