To prepare the data for our experiments, we employed basic tactics for preprocessing. First, we split the log into a training and a test set. The test set will act as our primary source for evaluating factuals, that are completely unknown to the model. We further split the training set into a training set and validation set. This procedure is a common tactic to employ model seletection techniques. In other words, Each dataset is split into 25% Test and 75 remaining and from the remaining we take 25 val and 75 train.

First, we filter out every case, whose' sequence length exceeds 25. We keep this maximum threshold for most of the experiments that focus on the evolutionary algorithm. The reason is the polinomial computation time of the viability measure. The similarity and sparsity components of the proposed viability measure have a run time complexity of at least N^2. Hence, limiting the sequence length saves a substantial amount of temporal resources.

Next, we extract time variables if they are provided in the log. Then, we normalise the values. For a proper time-format, we encode all information from seconds to a year. If the full log occurs during one time-unit only, e.g. every event happened within a year, drop the column that was extracted. Afterwards, we standard scale all remaining time features.

Each categorical variable is converted using binary encoding. Binary encoding is very similar to onehot encoding. However, it is still distinct. Binary encoding uses a binary representation for each class encoded. This representation saves a lot of space as binary encoded variables are less sparse, than one-hot encoded variables.

We also add an offset of 1 to binary and categorical columns to introduce a symbol which represents padding in the sequence. All numerical columns are standardized to have a zero mean and a standard deviation of 1.

We omit the case id, the activity and label column from this preprocessing procedure, for reasons explained in **??**. The activity is label-encoded. Hence, every category is assigned to a unique integer. The label column is binary encoded, as we focus on outcome prediction.