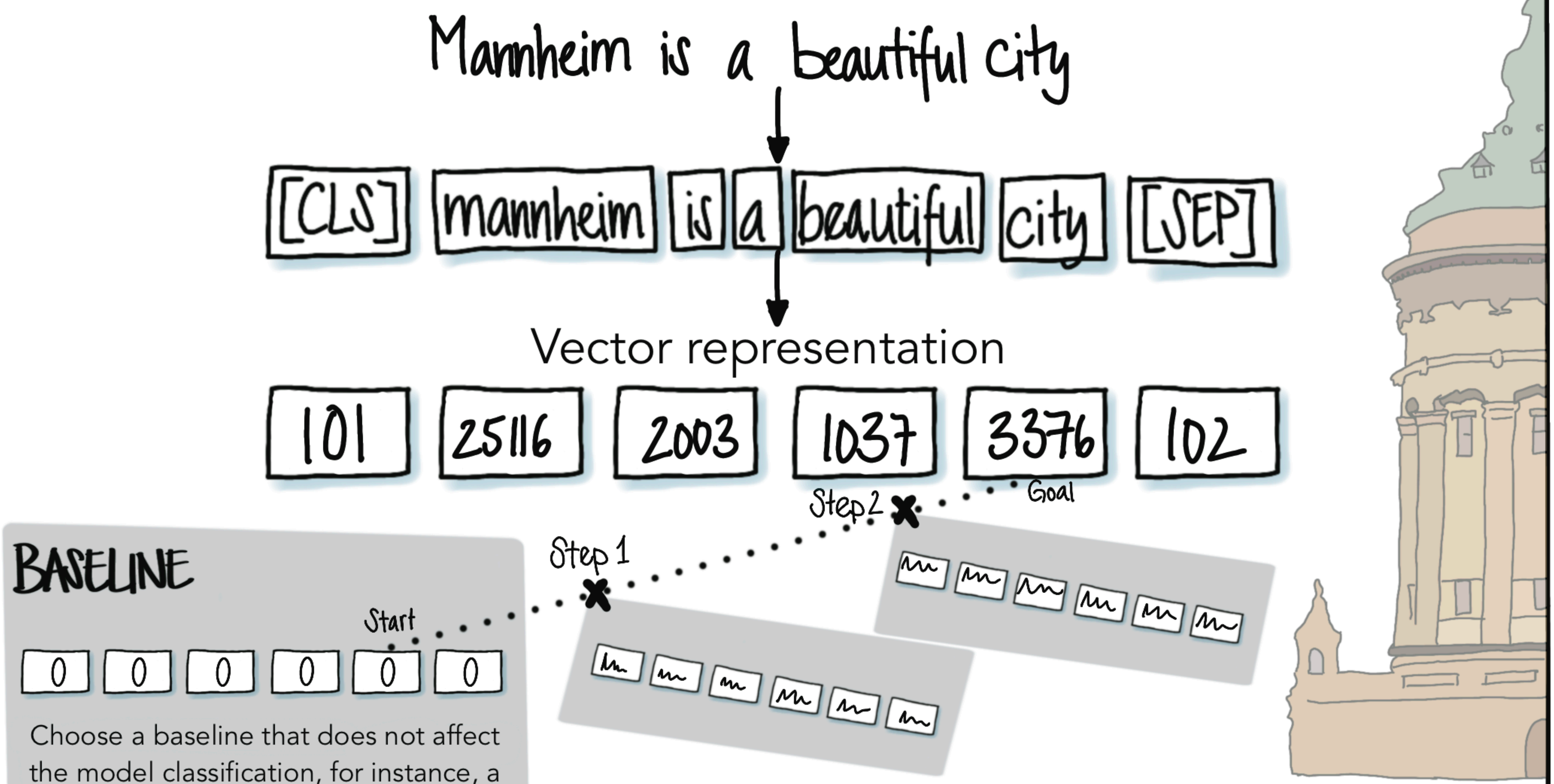
UNDERSTANDING INTEGRATED GRADIENTS



Vector representation

Choose a baseline that does not affect the model classification, for instance, a sequence of zeros

The algorithm calculates the model prediction for each input sequence at each step and compares it to the baseline. The difference in the model's prediction at each step is then multiplied by the corresponding step size. We can then derive the integrated gradients for each input feature by summing up the results of these calculations to identify how important each feature is for the prediction.

cosima_meyer