The scatter plot displays the distribution of the questions in lower dimensional space. By using cosine distance and neural network distribution we try to compare the two approaches. Our intuition is that cosine distances perform well for short and simple sentences, whereas the neural networks perform well when there is a comparison between long and complex meaning sentences. When the points in the scatter plot show an overlap for question 1 and question 2 we can infer that the questions are duplicates and vice versa

The reasoning for using two different scatterplot for each method is so that we can compare and analyse which questions are selected as similar questions to the inputs by each method and where they lie in the lower dimensions. . We incorporate the **Shneiderman mantra** by displaying all points in default state and then enable the user to filter the questions using on click events and view details as per his demand.

Our heatmap visualization is basically validating whether the questions that fall in a particular distribution bucket are similar to each other. The idea being that if a question is similar to the input it should also be similar to an extent to other questions that are similar to it. (transitivity relation) The heatmap has a max element count of 10. These are the top 10 questions lying in every bucket of the probability distribution. This visualization serves the **task of analyzing our model performance** using the generated outputs(in this case the similar questions of the distribution bucket).

The Sankey diagram is used to list the questions that are top 10 in every bucket to a particular input question. The width of the links depends on the value of the probability/normalized cosine distance. The user now basically has a ranked list of similar questions to the input questions that he may want to look at. Also, by displaying the list of questions in parallel plots we enable **connectivity and proximity principles**.