Self-Attention

- Goal: Aggregate arbitary-length input to fixed-size vector.
  Allow data-driven, trainable aggregation.

Given the sequence of inputs $x_1, \ldots, x_n$:

- Create three "views" of them: queries, keys, values.
- Using trained matrices W^Q, W^K, W^V.