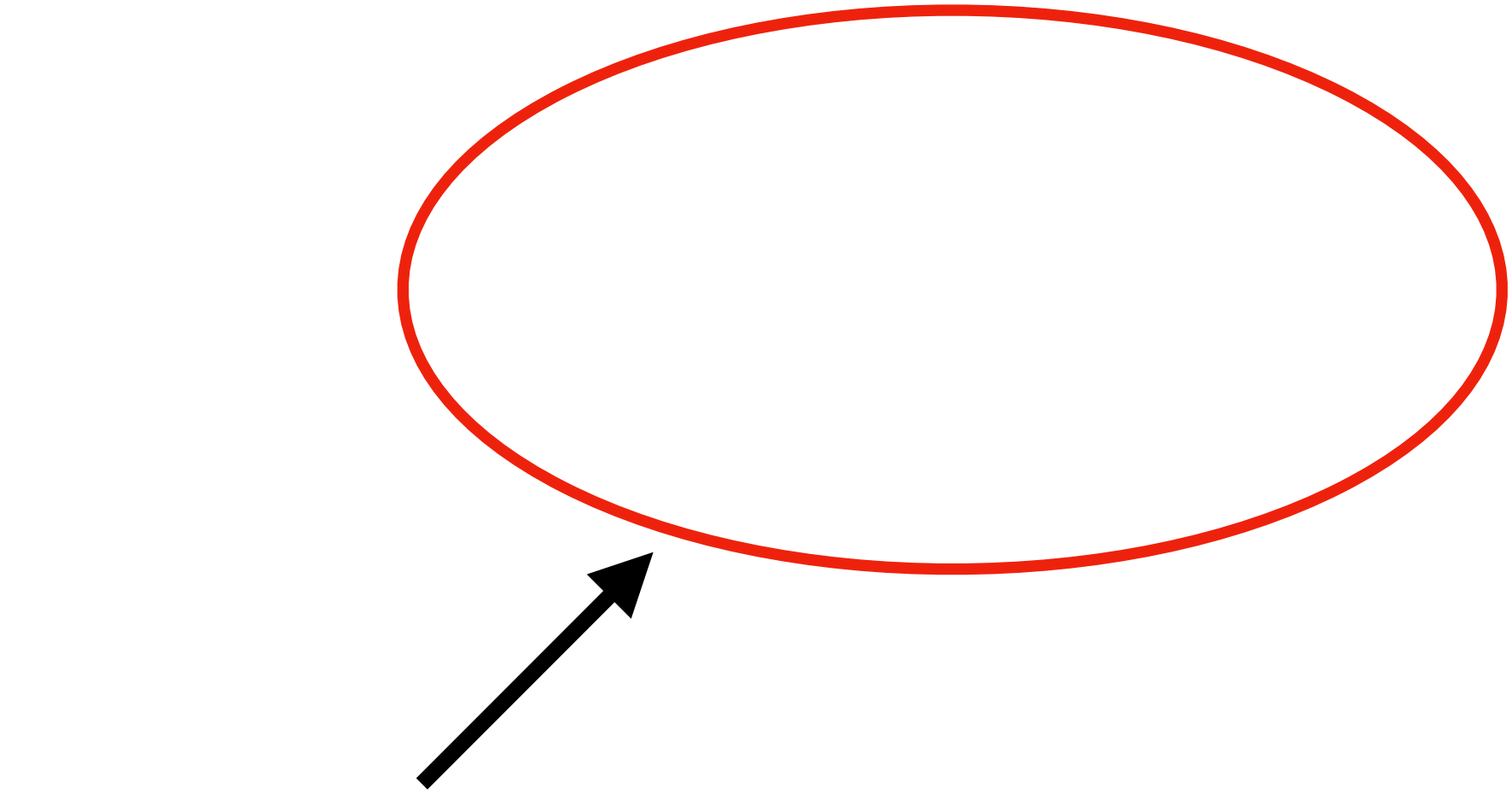
Attention Mechanism

- Representation of the context:

$$\sum_{i \leq 7} \frac{\exp(\langle q_7, k_i \rangle)}{\sum_{i' \leq i} \exp(\langle q_7, k_{i'} \rangle)} v_i$$


Attention weight p_i

- $p_1, \ldots, p_7$ is a probability distribution and the representation is an "expectation"

- Assign context-dependent probabilities to the tokens

Attention Mechanism

- Representation of the context:

$$\sum_{i \leq 7} \frac{\exp(\langle q_7, k_i \rangle)}{\sum_{i' \leq i} \exp(\langle q_7, k_{i'} \rangle)} v_i$$

Attention weight p_i

- $p_1, \ldots, p_7$ is a probability distribution and the representation is an "expectation"
- Assign context-dependent probabilities to the tokens

Maximal Data Use