For large N, we need to avoid the partition function

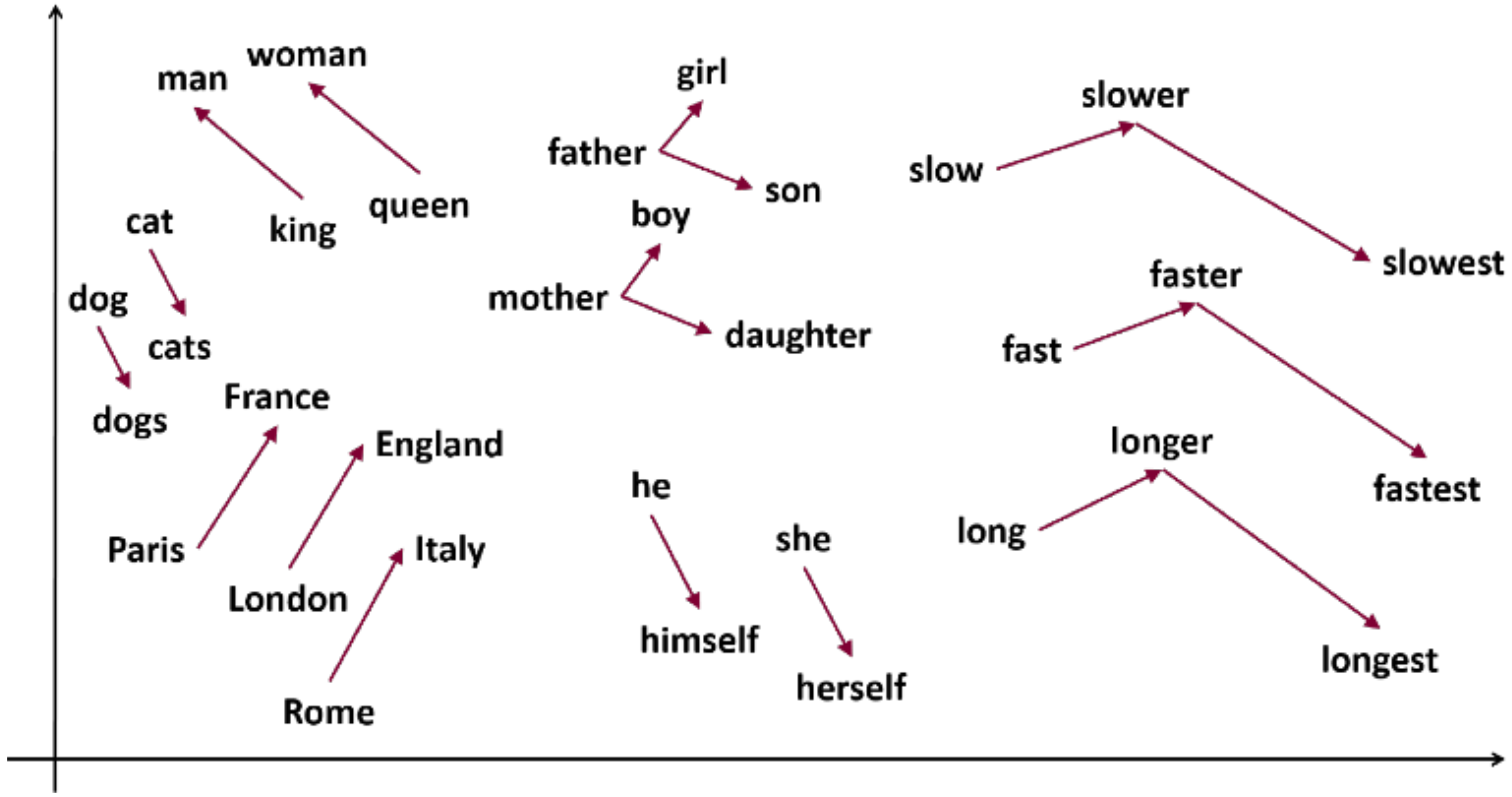
- The approach on the previous slide is the preferred method when N is a manageable size, say $N < 10^5$.

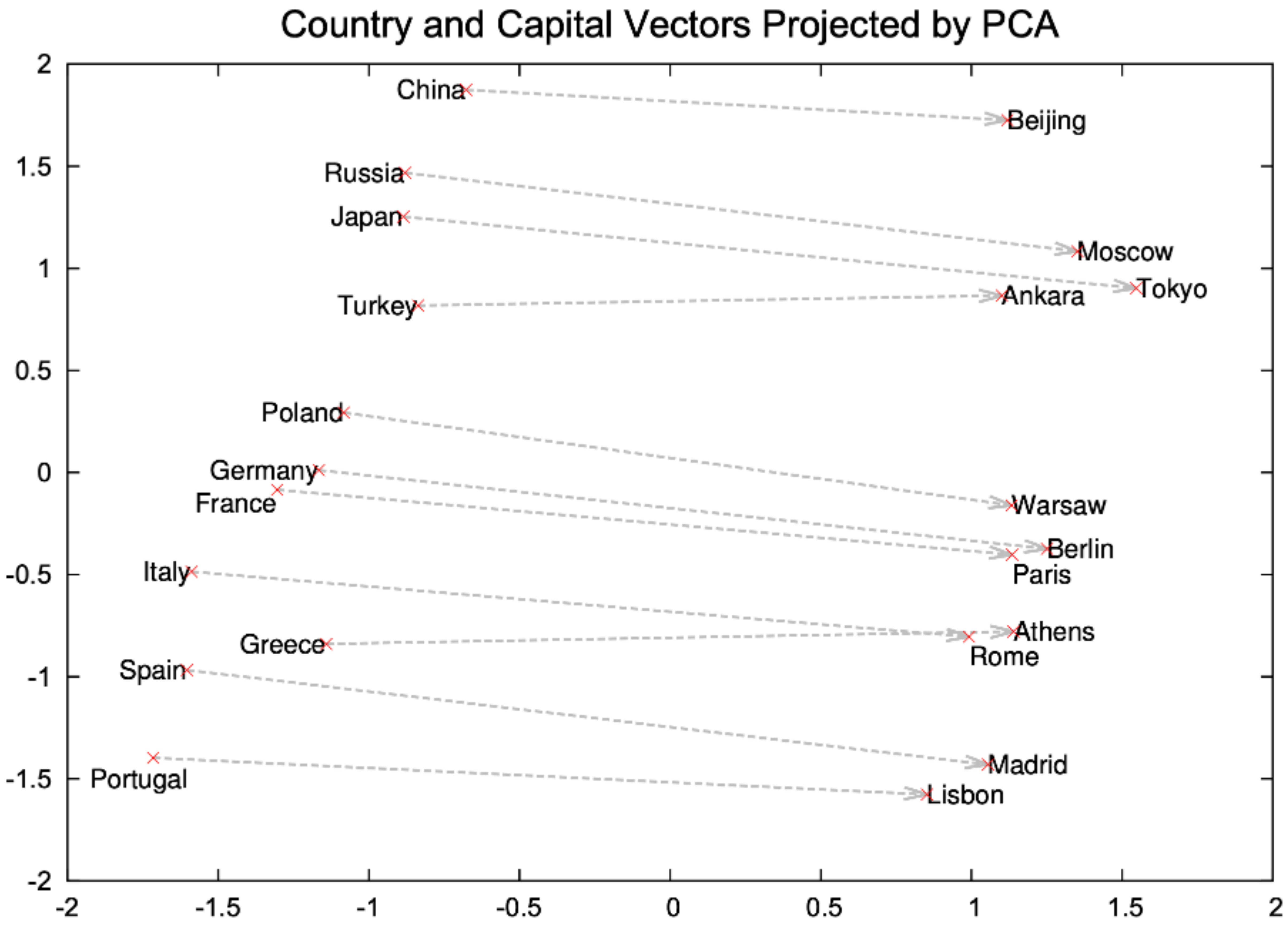
- When $N > 10^6$, the partition function (denominator of the softmax function) becomes prohibitively expensive to compute due the O(NK) scaling.

- There are several approaches to get around having to compute the (full) partition function:
    - Hierarchical softmax
    - Importance sampling (IS)
    - Adaptive IS
    - Target sampling
    - Noise contrastive estimation (NCE)
    - Negative sampling

$\sim 10 - 100$x speedup

Word2vec: structured semantic relationships





Mikolov et al., 2013