# Machine Learning for Natural Language Processing

Arieda Muço

Central European University

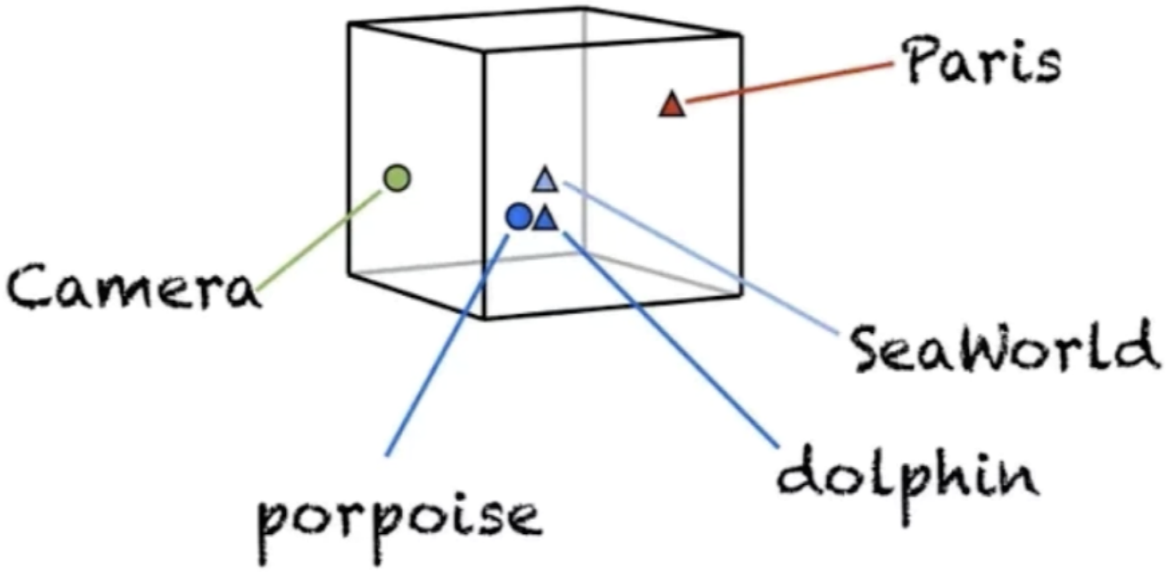
# Word Embeddings

- Fancy word, old concept
- Vector representation of a word (we have already seen count-vectorizer, tf-idf)
- What we mean by word embedding is that we are embedding a categorical entity into a vector spacee

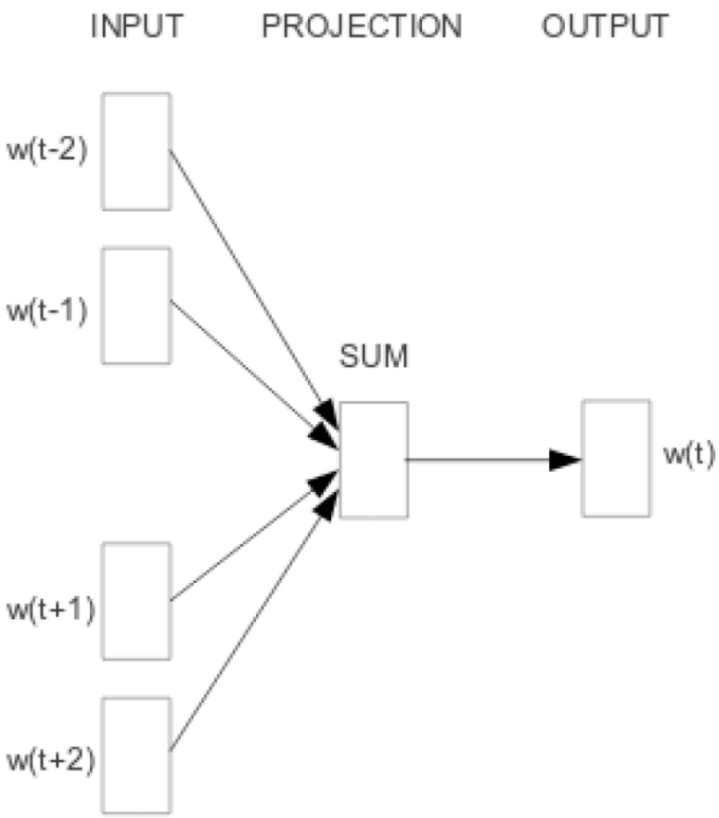
# Idea

- Unsupervised extraction of semantics using large corpus (Wikipedia etc)
- Input: one-hot representation of word (as in BoW)
- Use auxiliary task to learn continuous representation

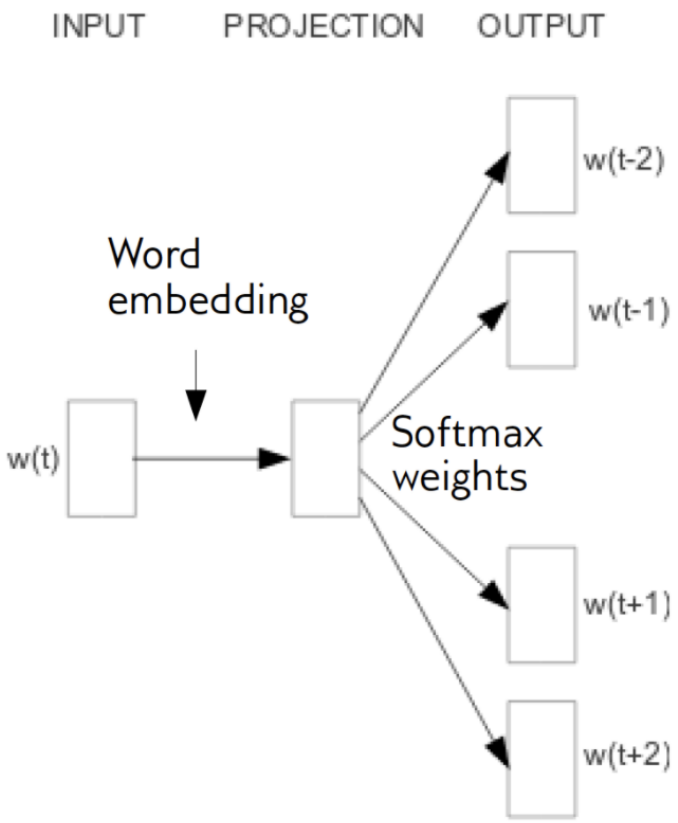
# Word Embeddings



# Continious Bow vs SkipGram




CBOW

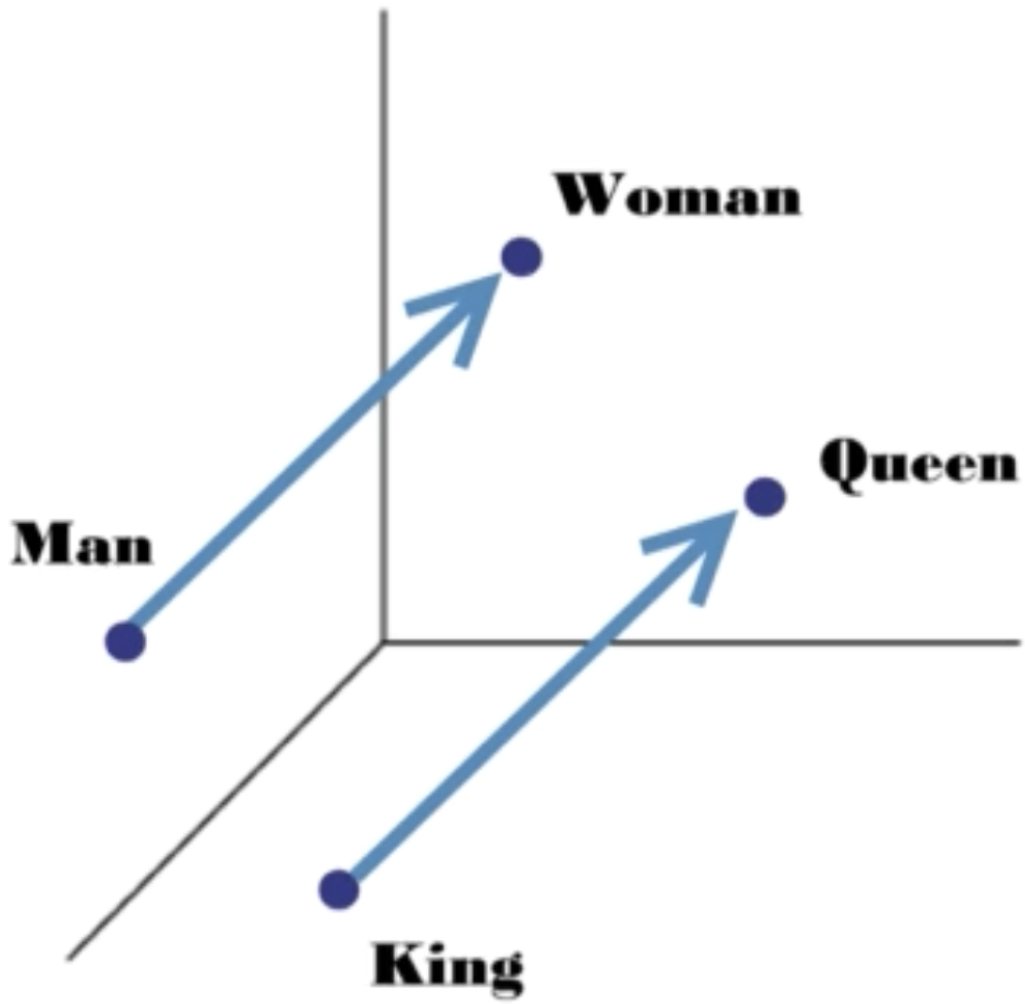



Skip-gram

# Examples

- King - Queen $\sim=$ Prince - Princess
- France - Paris $\sim=$ Germany - Berlin
- Japan - Japanese $\sim=$ China - Chinese
- Brother - Sister $\sim=$ Uncle - Aunt
- Walk - Walking $\sim=$ Swim - Swiming

# Visualizing Analogies



# Code

```
closest_distance = infinity
best_word = None
test_vector = king - man + woman
for word, vector in vocabulary:
        distance = get_distance(test_vector, vector):
        if distance < closest_distance:
                closest_distance = distance
                best_word = word
```

- Use Numpy to do this
- Use Cosine Distance (or 1- Cosine Similarity) as a distance measure. Alternatively, use Euclidian Distance