Word2Vec

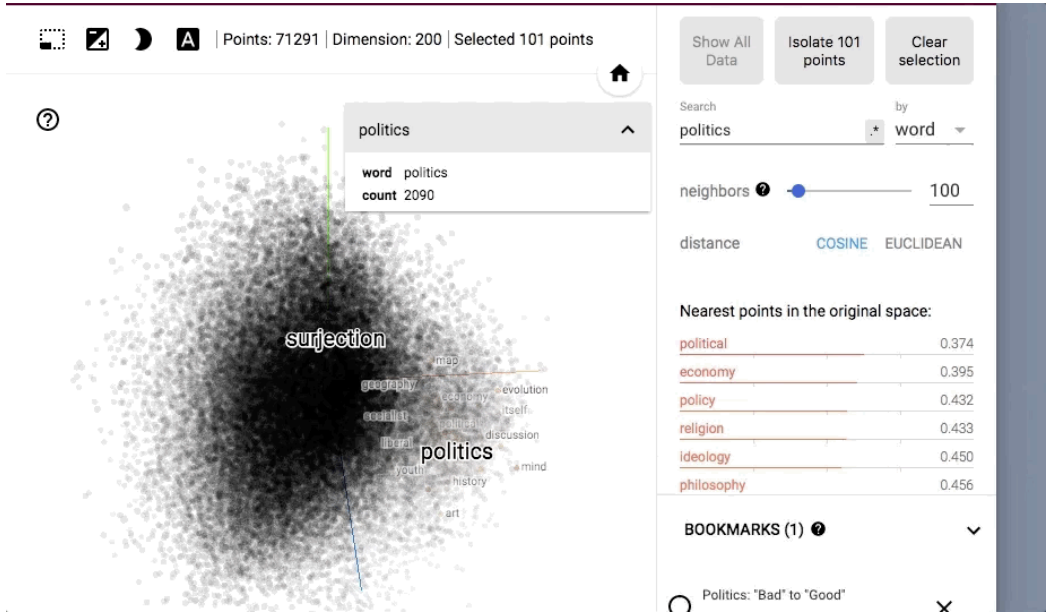
Outline

- Motivation
- Word2Vec from Scratch
- The 20 Newsgroup Text Dataset

Motivation

Word2Vec

- **Input:** Wikipedia corpus
- **Dimension:** 200
- **Output (CBoW):** The model predicts the current word from a window of surrounding context words
- **Output (CSG):** The model uses the current word to predict the surrounding window of context words

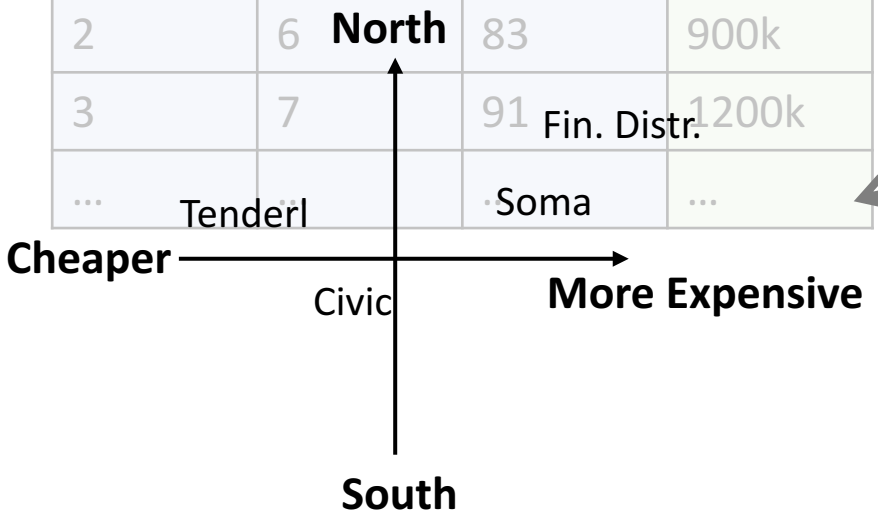


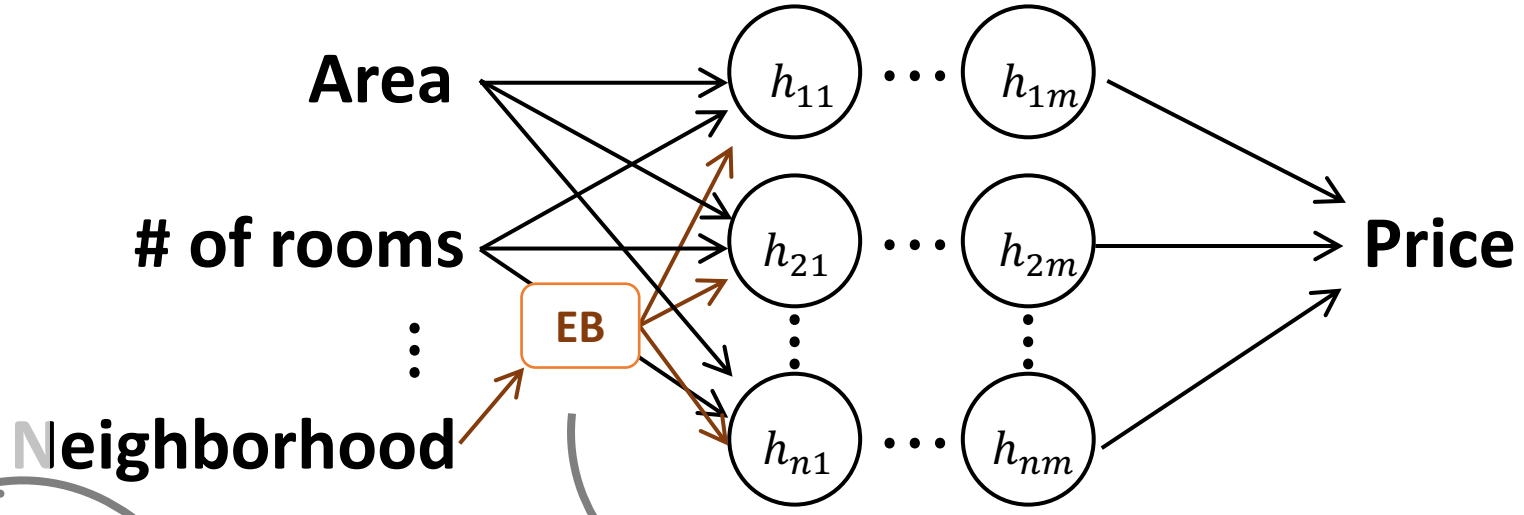
Examples from: http://projector.tensorflow.org/

Embeddings

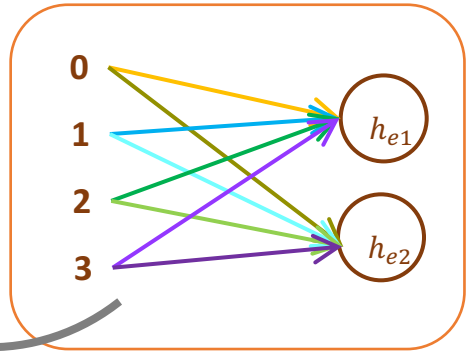
- Recap from last Class

Nbhd	rooms	Area(m2)	Price
0	4	61	500k
1	5	72	600k
2	6	83	900k
3	7	91	1200k
…	…	…	…





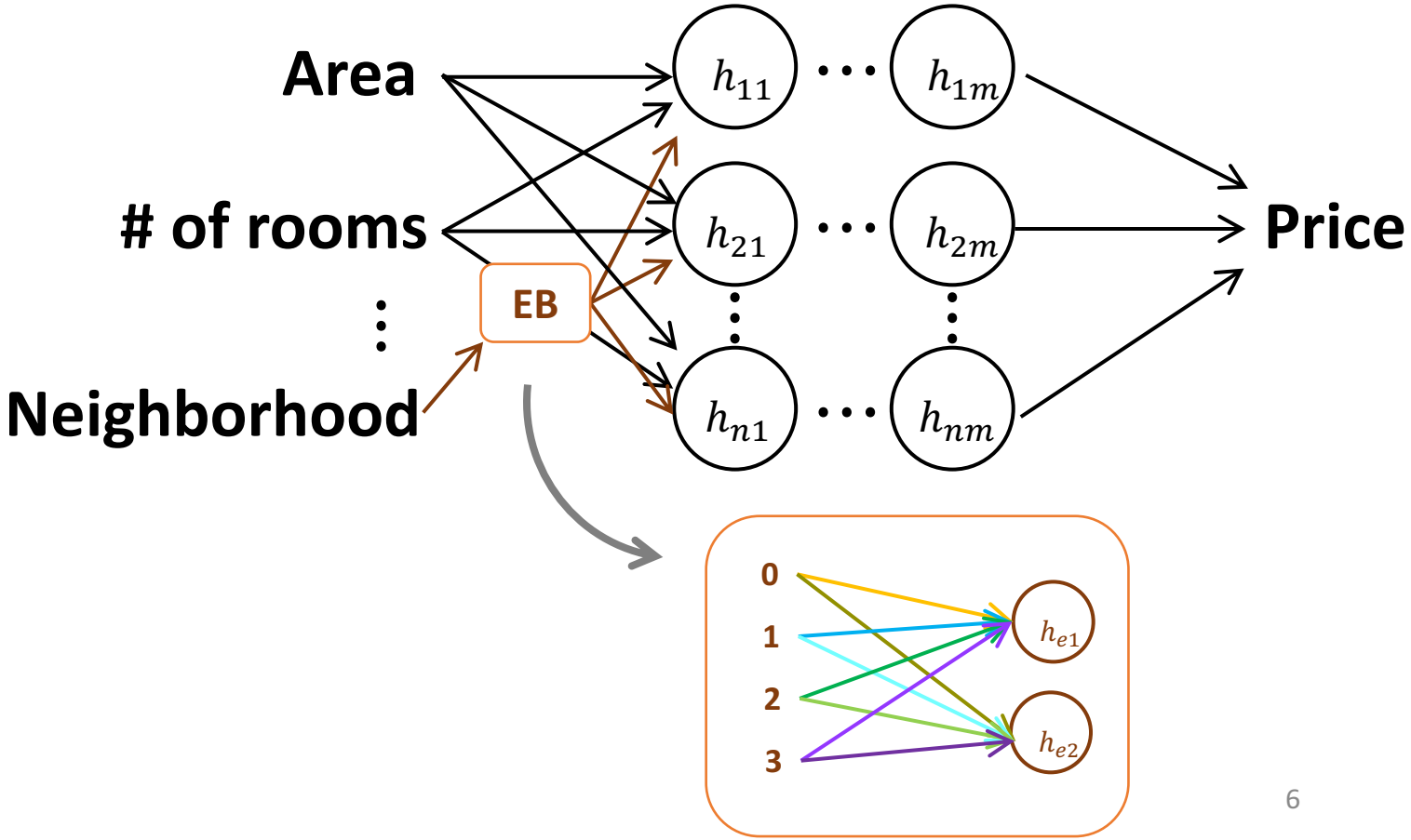
	h_{e1}	h_{e2}
0	w1	w2
1	w3	w4
2	w5	w6
3	w7	w8




```
neibhd_to_idx={0:'Tender',1:'Civic',2:'FDist',3:'Soma'}
```

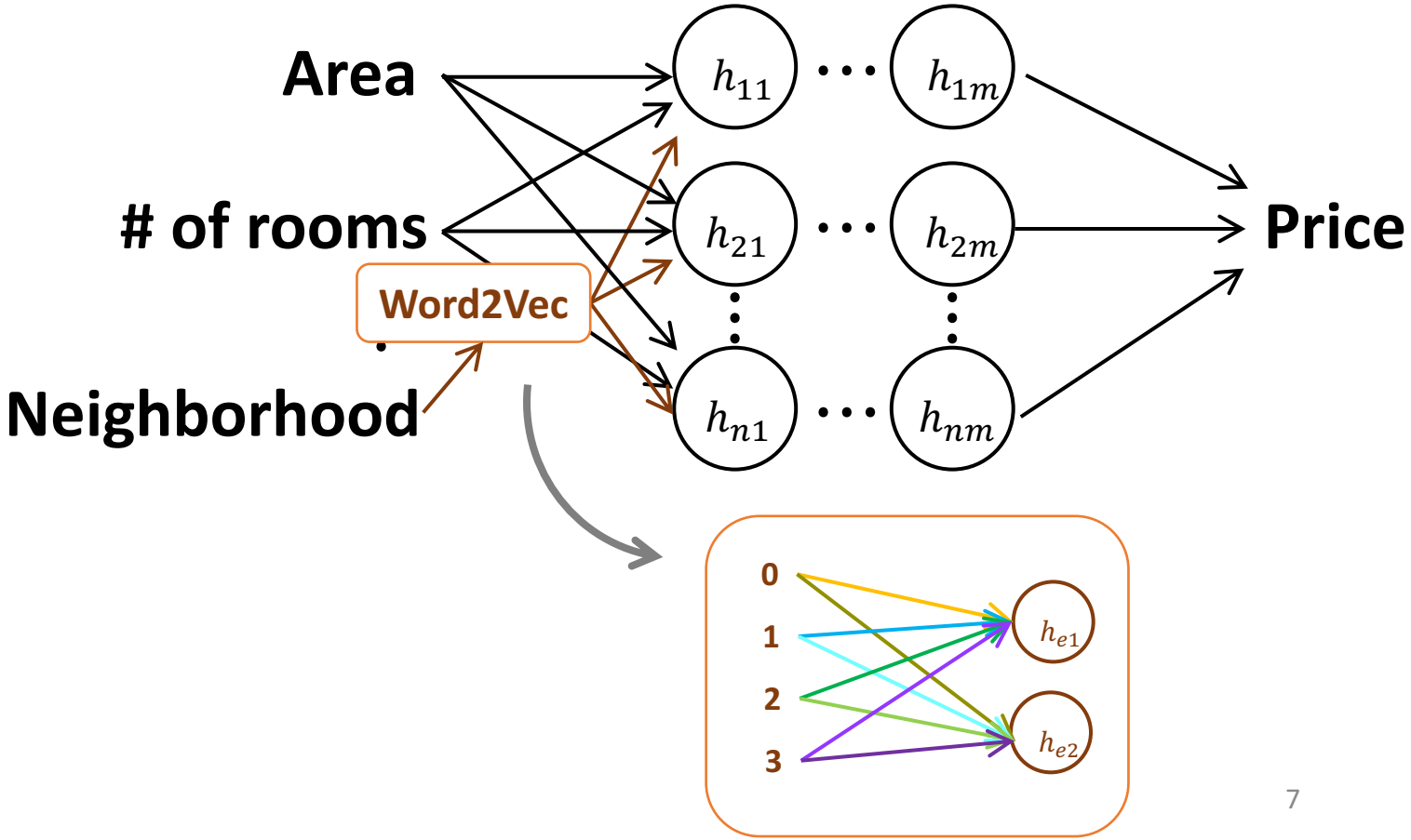
Embeddings

- Recap from last Class



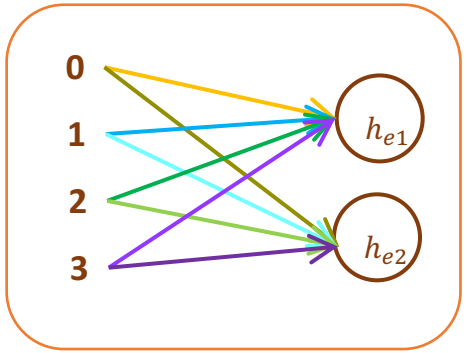
Embeddings

- Recap from last Class



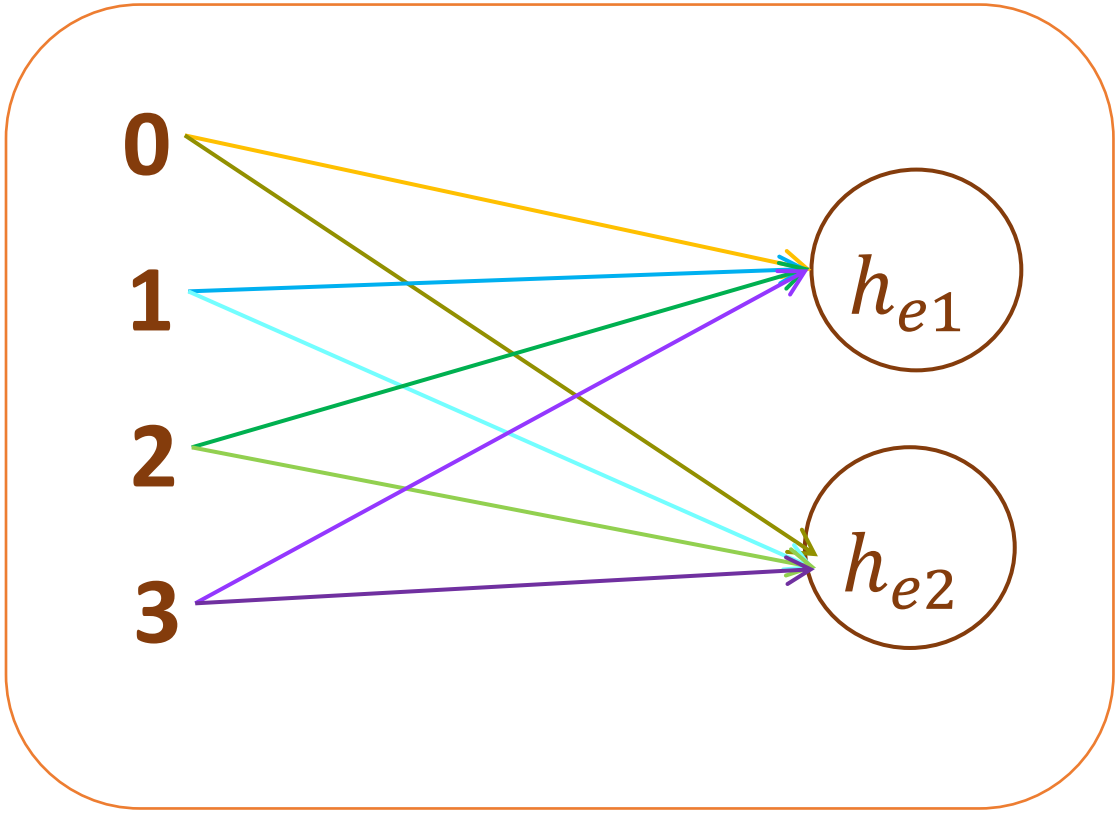
Embeddings

- Recap from last Class



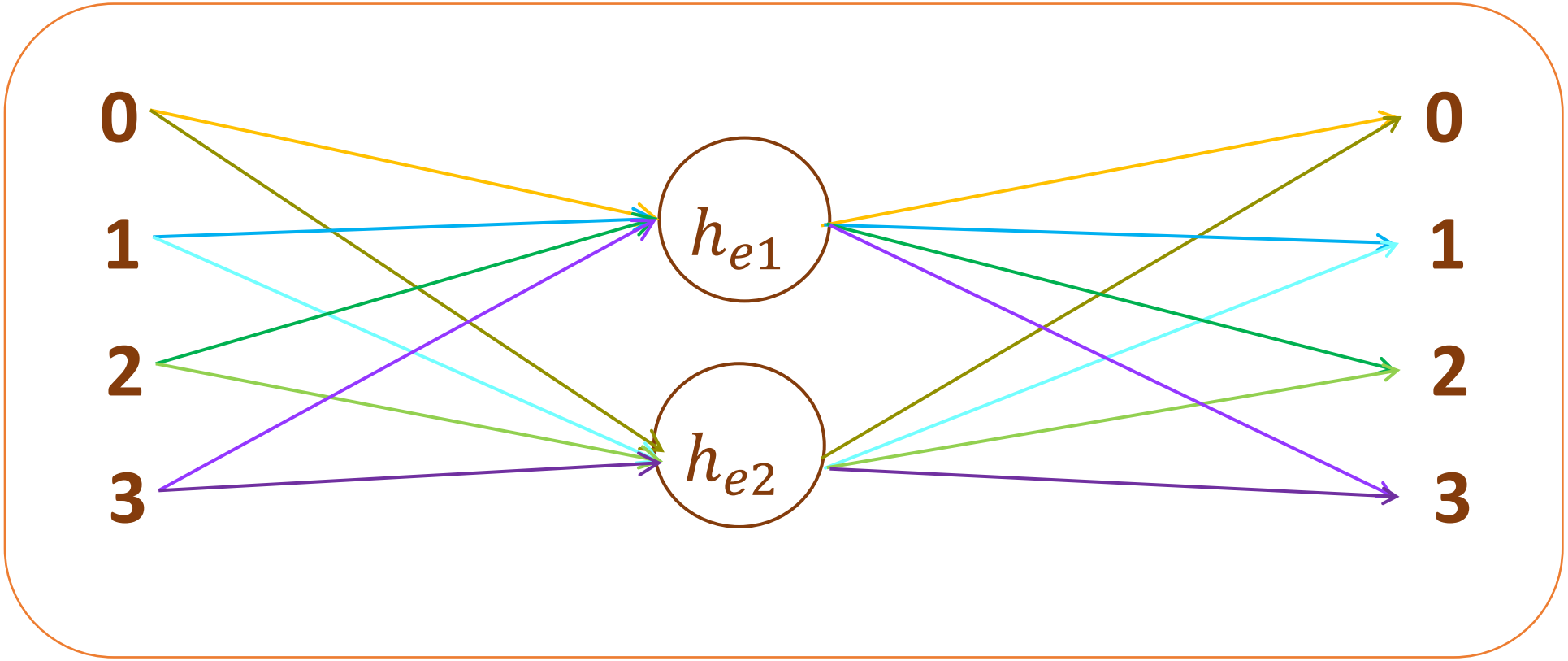
Embeddings

- Recap from last Class



Embeddings

- Recap from last Class

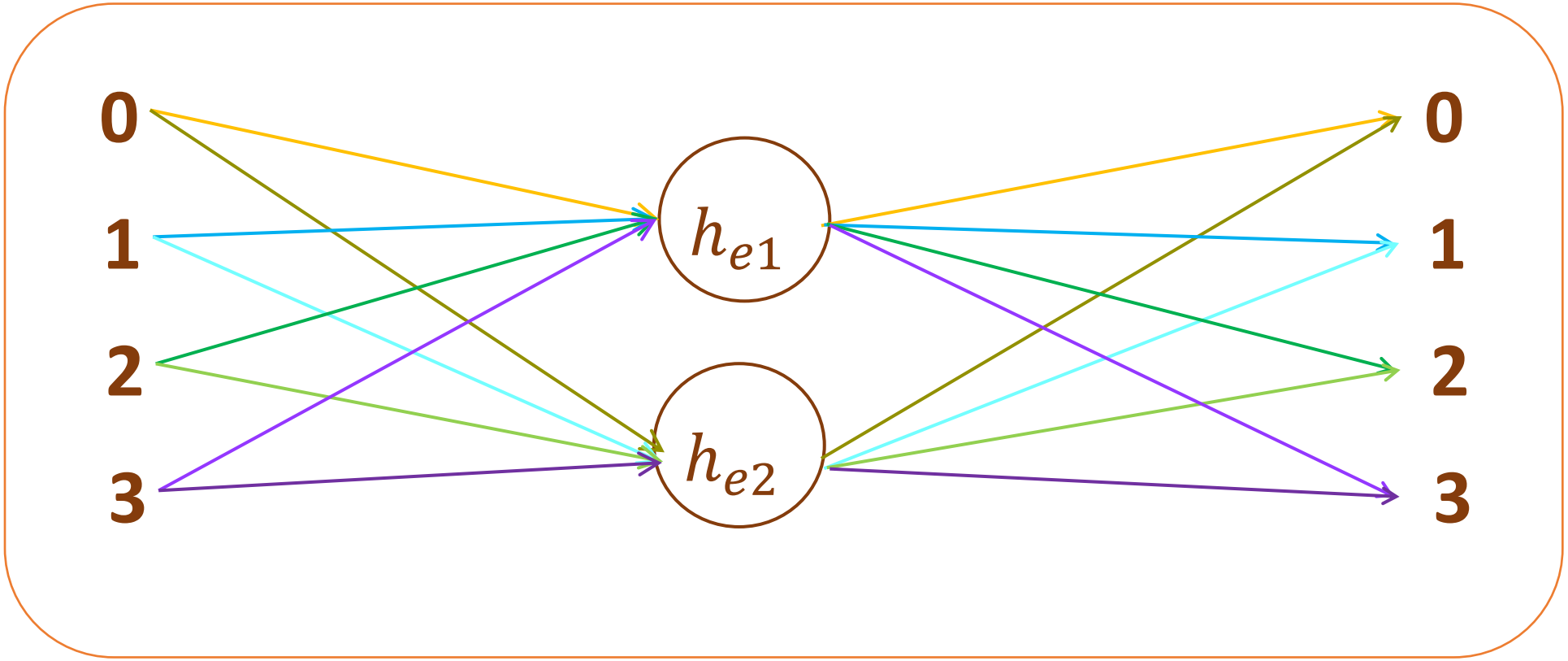


Embeddings

- Recap from last Class

I am the king

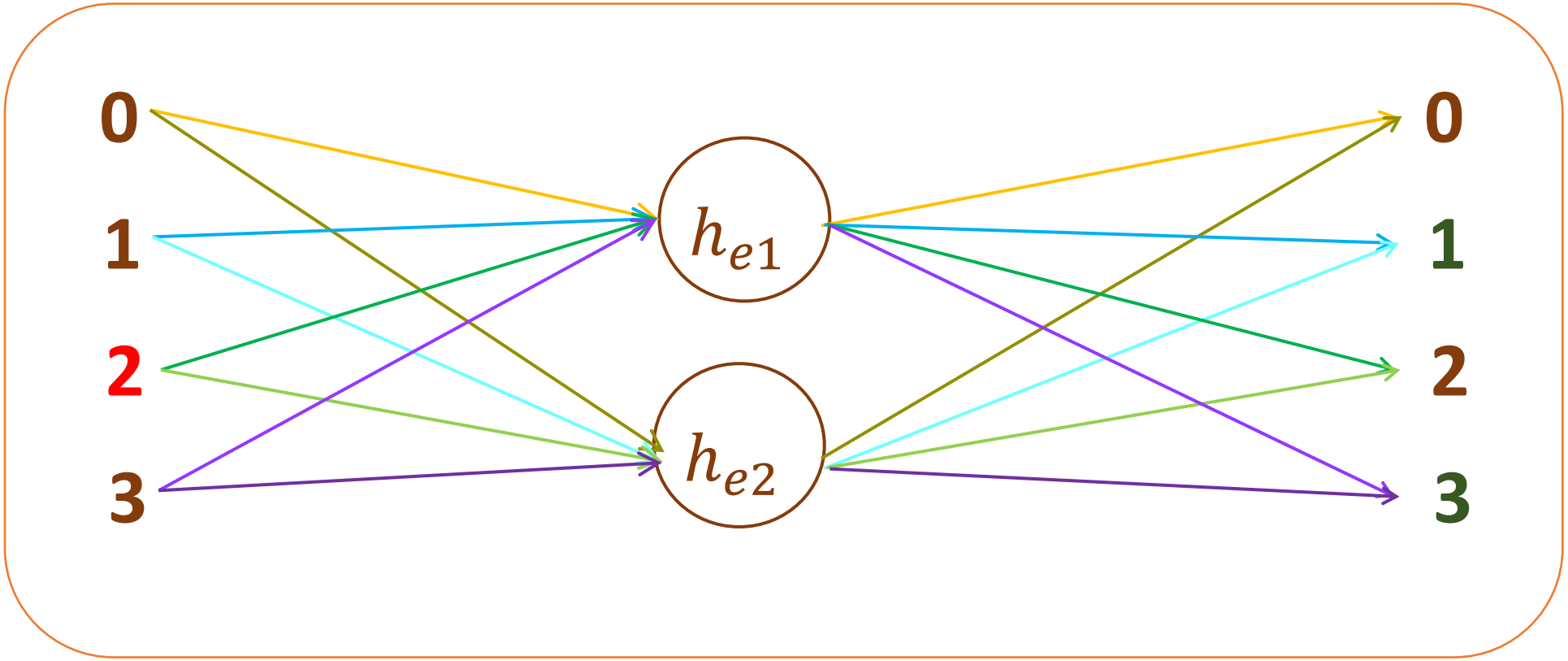
0 1 2 3



Embeddings

- Recap from last Class

I am the king
0 1 2 3



Embeddings

- How to build the dataset?
  - https://www.youtube.com/watch?v=64qSgA66P-8

Word2Vec data generation (skipgram)

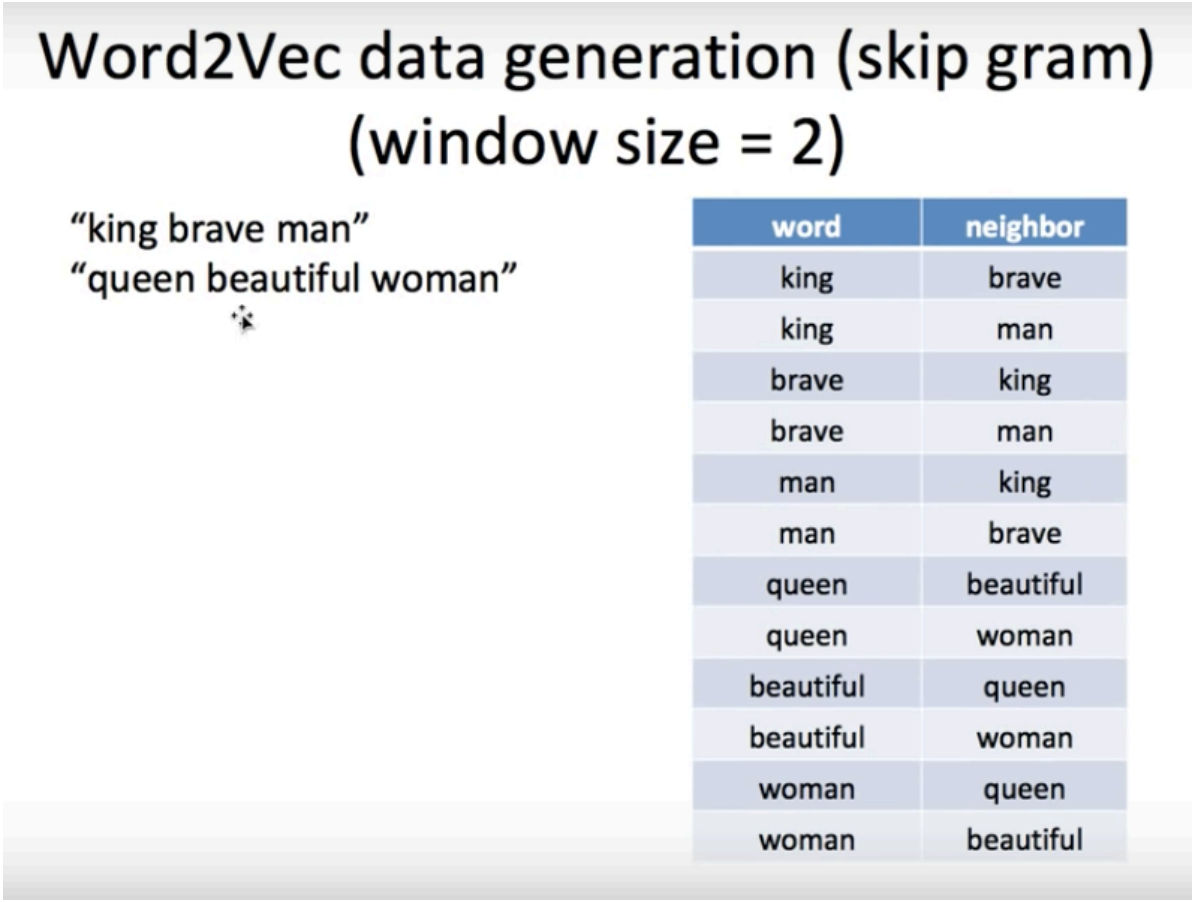
(window size = 1)

"king brave man"
"queen beautiful woman"

word	neighbor
king	brave
brave	king
brave	man
man	brave
queen	beautiful
beautiful	queen
beautiful	woman
woman	beautiful

Embeddings

- How to build the dataset?
  - https://www.youtube.com/watch?v=64qSgA66P-8



word	neighbor
king	brave
king	man
brave	king
brave	man
man	king
man	brave
queen	beautiful
queen	woman
beautiful	queen
beautiful	woman
woman	queen
woman	beautiful

Word2Vec from Scratch