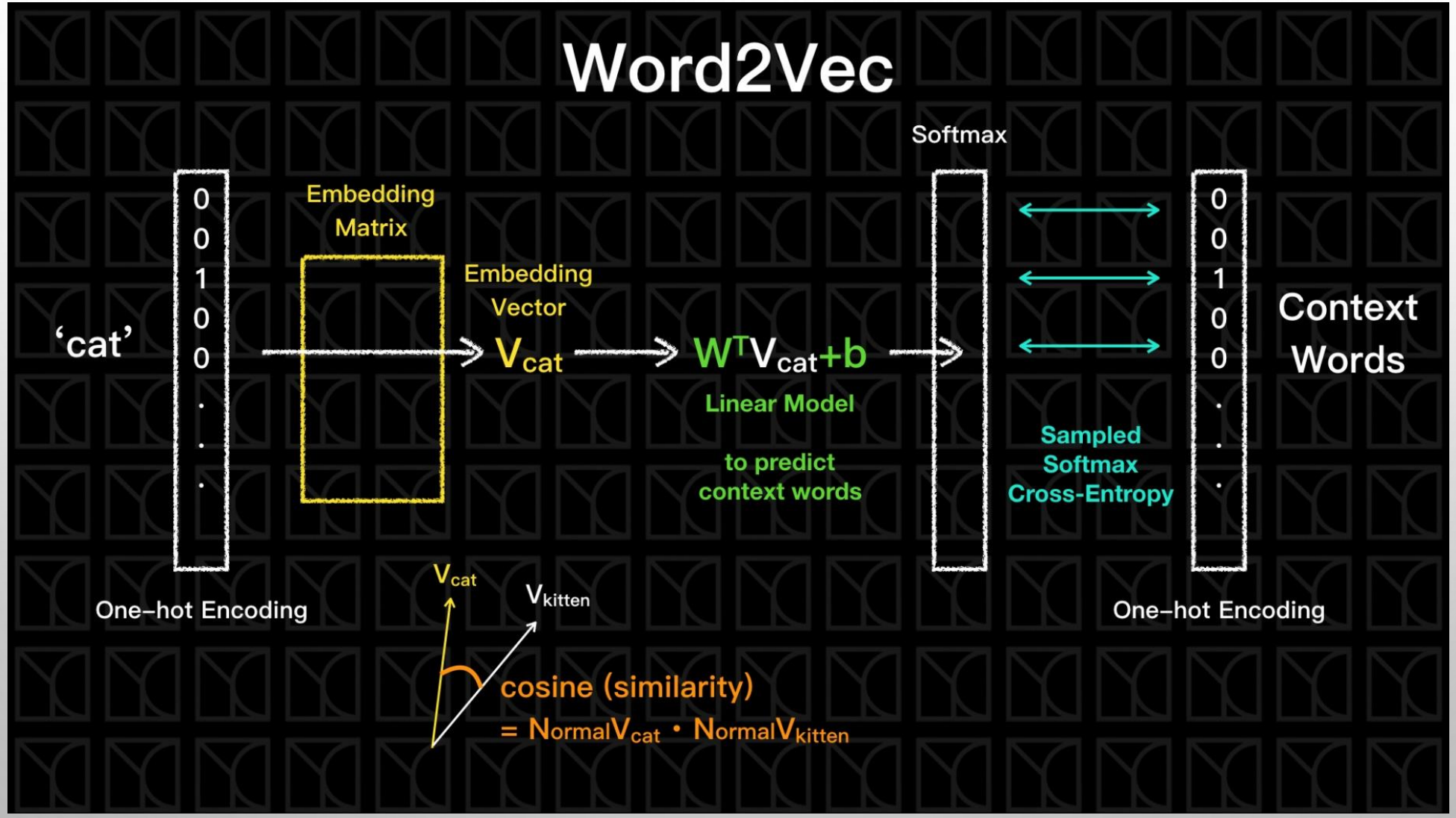
# WORD2VEC

# 基礎概念

- 顧名思義是將文字轉為詞向量
- 屬於UNSUPERVISED LEARNING
- 最核心的概念是用前一個詞預測下一個詞
- 是一種訓練文字語意的類神經網路
- HIDDEN LAYER只有一層
- 可以很好地表達不同詞之間的相似和類比關係
- 使用ONE-HOT ENCODING
- 分為兩種模型
  - 1.SKIP-GRAM
  - 2.CBOW

# 流程圖


Word2Vec
'cat'
0
0
1
0
0
Embedding Matrix
Embedding Vector
$V_{cat}$
$W^T V_{cat}+b$
Linear Model
to predict context words
Softmax
Sampled Softmax Cross-Entropy
0
0
1
0
0
Context Words
One-hot Encoding
One-hot Encoding
$V_{cat}$
$V_{kitten}$
cosine (similarity)
= Normal$V_{cat}$ • Normal$V_{kitten}$

# ONE-HOT ENCODING

- 也稱之為ONE OF N ENCODING
- 例:"THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG"

The = [1, 0, 0, 0, 0, 0, 0, 0, 0]

quick = [0, 1, 0, 0, 0, 0, 0, 0, 0]

brown = [0, 0, 1, 0, 0, 0, 0, 0, 0]

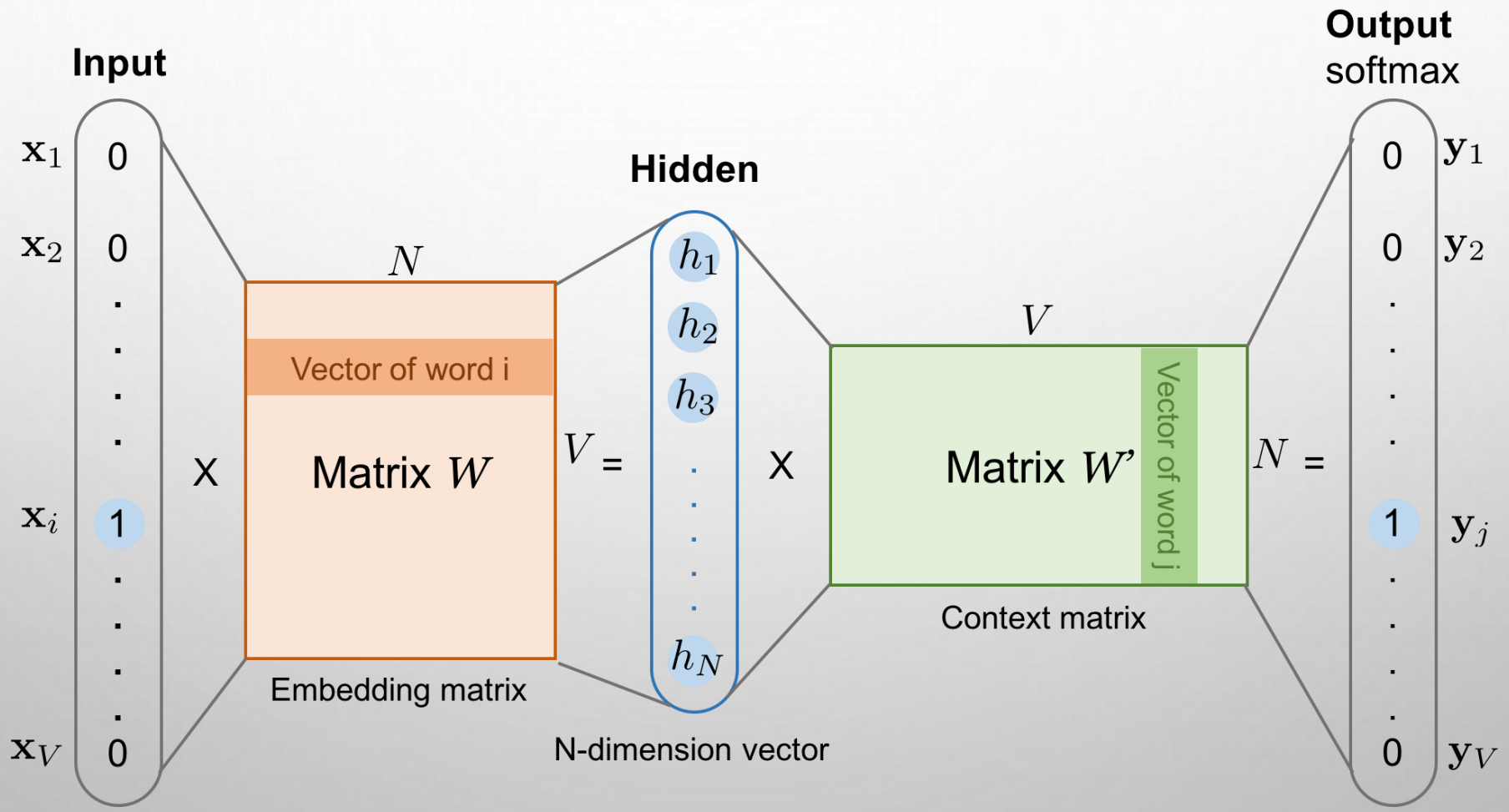
…

….

dog = [0, 0, 0, 0, 0, 0, 0, 0, 1]

# WORD EMBEDDING

- 可達到降維的效果



# skip-gram 和 CBOW(continuous bagging of words)

- SKIP-GRAM

  用中間詞預測上下文

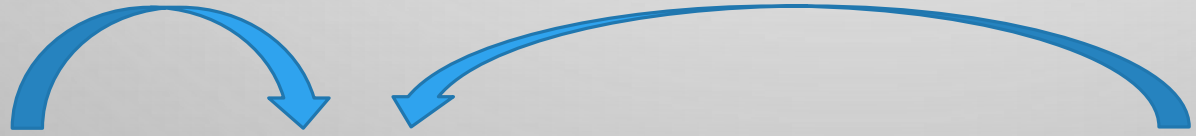
- CBOW

  用上下文預測中間詞

skip-gram:

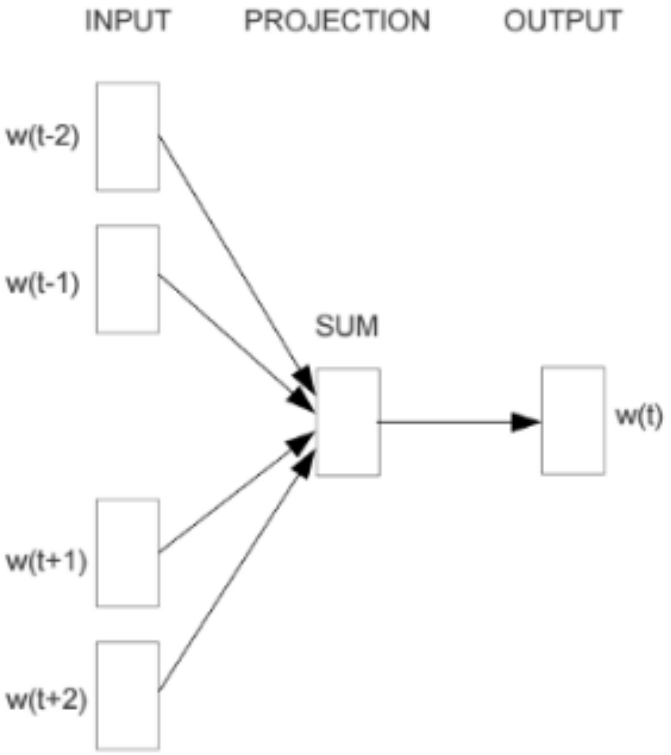
Eruka is Guanting's girlfriend and Wanrong is his ex-girlfriend.

CBOW:

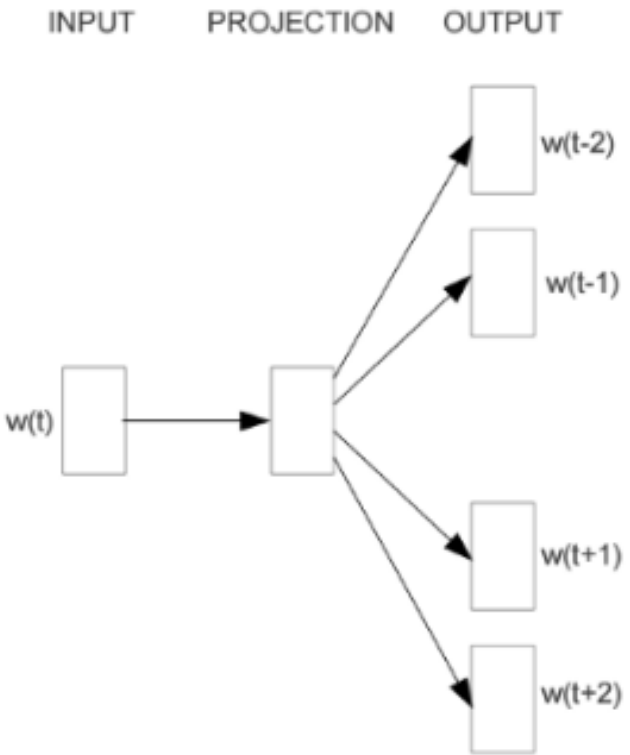


Eruka is Guanting's girlfriend and Wanrong is his ex-girlfriend.

# 兩種模型架構圖

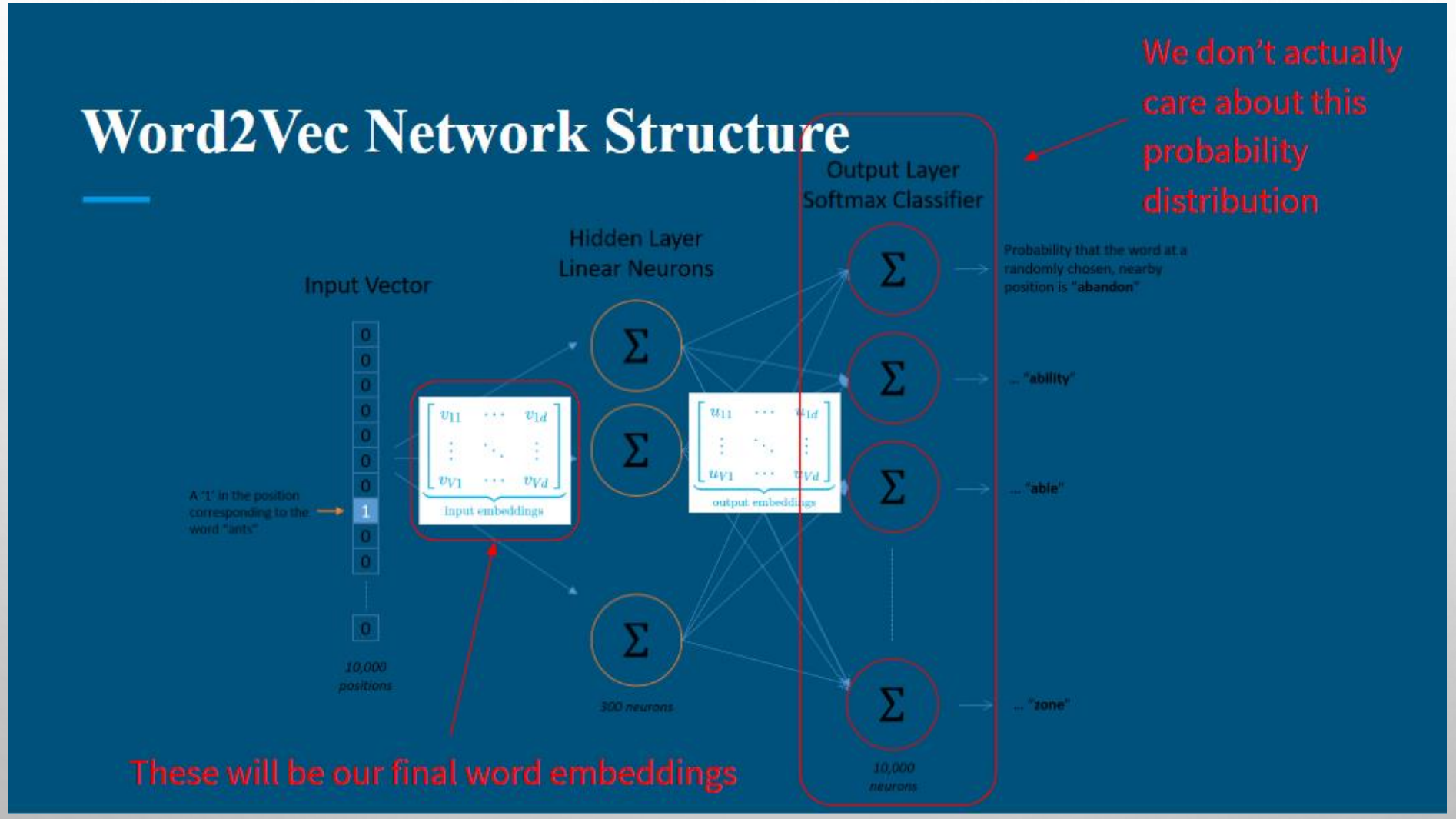



CBOW



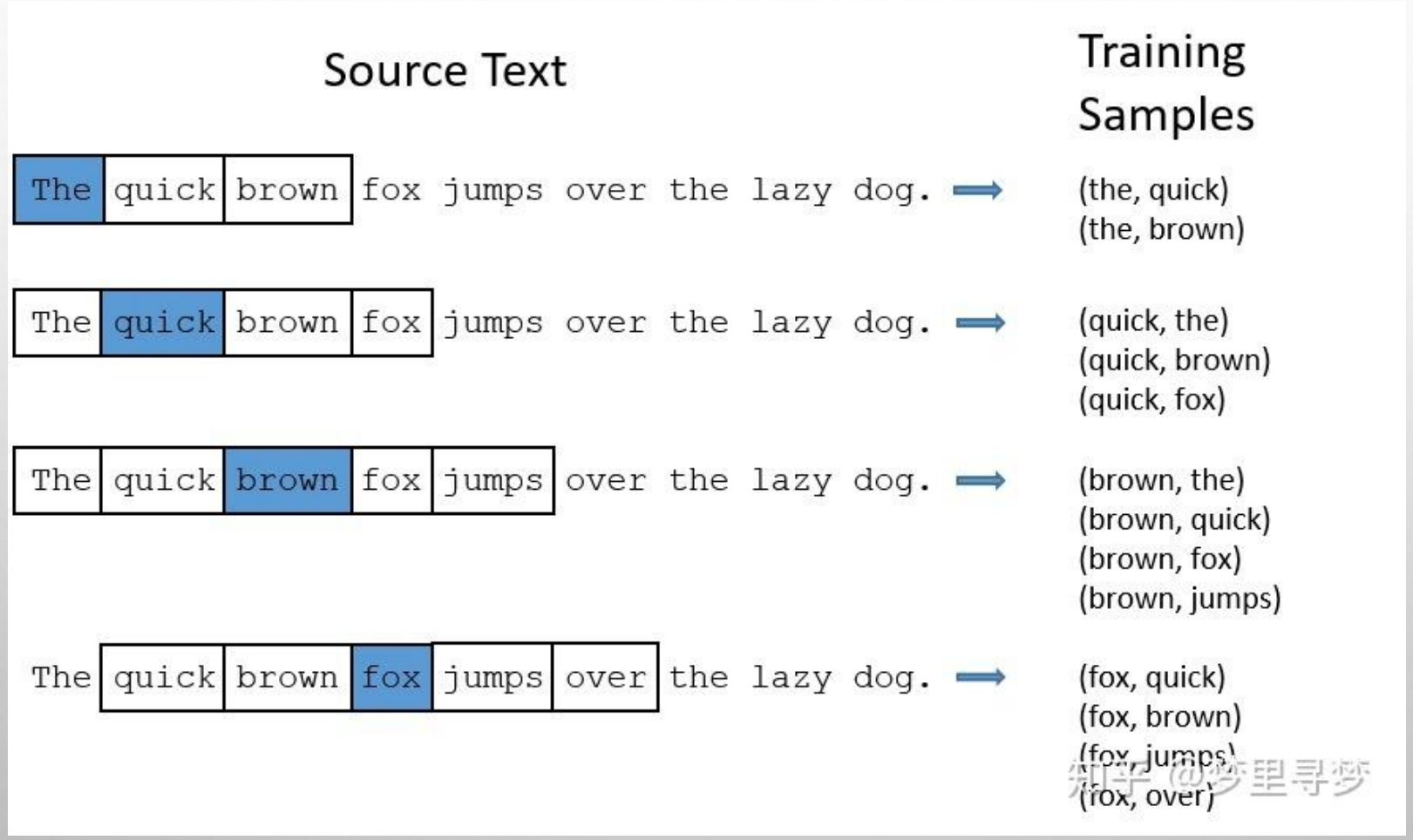

Skip-gram

# 以SKIP-GRAM為例



# INPUT 格式

window size=2



# TRAINING

- MINIMIZE CROSS ENTROPY

# OUTPUT格式

- 每個詞的機率，長度和INPUT的詞向量長度相同。

- 找出機率最大的詞。

# WORD EMBEDDING 結果

Use cosine similarity

V(China)-V(Beijing)+V(Tokyo)=V(Japan)